stats

MDPI

Article

Goodness-of-Fit and Generalized Estimating Equation Methods for Ordinal Responses Based on the Stereotype Model

Daniel Fernández [1,2,3,†], Louise McMillan [4], Richard Arnold [4], Martin Spiess [5] and Ivy Liu [4,*]

1 Serra Húnter Fellow, Department of Statistics and Operations Research (DEIO), Universitat Politècnica de Catalunya · BarcelonaTech (UPC), 08028 Barcelona, Spain; daniel.fernandez.martinez@upc.edu
2 Institute of Mathematics of UPC-BarcelonaTech (IMTech), 08028 Barcelona, Spain
3 Centro de Investigación Biomédica en Red de Salud Mental, CIBERSAM, Instituto de Salud Carlos III, Monforte de Lemos 3-5, Pabellón 11, 28029 Madrid, Spain
4 School of Mathematics and Statistics, Victoria University of Wellington, Cotton Building 356, Gate 7, Kelburn Parade, Wellington 6012, New Zealand; louise.mcmillan@vuw.ac.nz (L.M.); richard.arnold@vuw.ac.nz (R.A.)
5 Psychological Methods and Statistics, Hamburg University, 20146 Hamburg, Germany; martin.spiess@uni-hamburg.de
* Correspondence: ivy.liu@vuw.ac.nz; Tel.: +64-44635648
† These authors contributed equally to this work.

Abstract: Background: Data with ordinal categories occur in many diverse areas, but methodologies for modeling ordinal data lag severely behind equivalent methodologies for continuous data. There are advantages to using a model specifically developed for ordinal data, such as making fewer assumptions and having greater power for inference. **Methods**: The ordered stereotype model (OSM) is an ordinal regression model that is more flexible than the popular proportional odds ordinal model. The primary benefit of the OSM is that it uses numeric encoding of the ordinal response categories without assuming the categories are equally-spaced. **Results**: This article summarizes two recent advances in the OSM: (1) three novel tests to assess goodness-of-fit; (2) a new Generalized Estimating Equations approach to estimate the model for longitudinal studies. These methods use the new spacing of the ordinal categories indicated by the estimated score parameters of the OSM. **Conclusions**: The recent advances presented can be applied to several fields. We illustrate their use with the well-known arthritis clinical trial dataset. These advances fill a gap in methodologies available for ordinal responses and may be useful for practitioners in many applied fields.

Keywords: goodness-of-fit; longitudinal data; ordinal data; stereotype model

Citation: Fernández, D.; McMillan, L.; Arnold, R.; Spiess, M.; Liu, I. Goodness-of-Fit and Generalized Estimating Equation Methods for Ordinal Responses Based on the Stereotype Model. *Stats* **2022**, *5*, 507–520. https://doi.org/10.3390/stats5020030

Academic Editor: Wei Zhu

Received: 31 March 2022
Accepted: 30 May 2022
Published: 1 June 2022

1. Introduction

1.1. Ordinal Responses

Many studies use data with ordinal categories (see e.g., [1–5]). For instance, in a questionnaire, Likert scale responses might be "strongly disagree", "disagree", "neutral", "agree", and "strongly agree" [6,7]. It may be easier for participants to provide rankings than absolute scores. In ecological studies, the ordinal Braun–Blanquet scale is used to collect species abundance data as it reduces sampling time compared with obtaining precise numerical estimates of abundance [3,8,9].

An ordinal variable indicates inherent order [10]. It differs from a nominal variable which has categories without any ordering information. Another defining distinction between a nominal and an ordinal variable is the effect of covariates on the outcome. As a covariate changes value in a particular direction, the distribution of the response consistently moves to higher categories, or consistently moves to lower categories, whereas for a nominal response the covariate may have different effects on different categories.

Ordinal responses are often collected and coded as numbers—for example, in the Likert scale above, the levels of agreement might be coded 1, 2, 3, 4, or 5. However,

the degree of dissimilarity (i.e., the spacing) between adjacent categories of the scale might not be the same for all pairs of levels. In the 5-level example, the difference expressed between level 1 and level 2 might be much greater than the difference expressed between level 3 and level 4. Given the possibility of unequal spacing, any analyses based on the assumption of equal spacing may lead to flawed conclusions.

Although the use of ordinal data are common, the methods for analyzing ordinal data often treat them as continuous or nominal. Agresti [10] (Section 1.3) discussed several disadvantages of using ordinary linear models designed for continuous data. Firstly, the results depend on the coding used. Secondly, this approach does not take into account the error created by treating ordinal values as continuous data. Thirdly, the predicted values could be outside the ordinal range. Finally, the application of continuous-data regression methods to ordinal data can produce misleading results due to "floor" and "ceiling" effects on the dependent variable ([10], Section 1.3.1).

Another approach to deal with ordinal responses is to dichotomize them in order to use logistic regression models for binary responses. This approach clearly comes at the cost of loss of information and reduced statistical power. Stromberg [11] empirically demonstrated these effects, showing poor precision and a loss of predictive power.

Unlike methods for numerical data, methods for analyzing ordinal data are not well known to many researchers. Liu and Agresti [12] and Agresti [10] extensively described various ordinal regression models including proportional-odds-type models using cumulative logits or adjacent-categories logits [13] and continuation-ratio logits [14]. The cumulative logits option is popular and is often called the "proportional odds model".

1.2. The Ordered Stereotype Model

This article focuses on the ordered stereotype model (OSM) introduced by Anderson [15], which assumes that the effects of the covariates on the response are proportional but not equal for the different levels of the response. This model is more flexible than the proportional odds model but more parsimonious than the unrestricted models which allow the covariates to have entirely different effects on the different categories. Greenland [16] showed that the stereotype model is a natural option when the progression of the response variable occurs through various stages. One of its main features is that certain parameters of the fitted model can be treated as data-driven numerical values which code the ordinal data as continuous and numerical, but not necessarily with equally spaced levels. This ability to estimate the possibly unequal spacings among ordinal responses is an improvement over other ordinal response models.

The OSM received more attention after it was fully discussed in Agresti [10]. Subsequently, several authors (including the authors of this article) have studied and applied the OSM to a number of fields. Among others, Ananth and Kleinbaum [17] reviewed it for epidemiological studies, Johnson [18] presented a generalization of the OSM for psychology studies, Fullerton [19] reviewed the OSM, among others ordinal regression models, for sociological studies, Liu [20] applied it in the area of education, Fernández and Pledger [21] demonstrated the use of OSM for count data in Ecology, and Williams and Archer [22] combined the OSM with an elastic net penalty as a method capable of modeling an ordinal outcome for high-throughput genomic data sets.

The aim of this article is to introduce and summarize two recent advances of the OSM: the development of three goodness-of-fit tests for cross-sectional ordinal data to assess if an OSM model holds; and the description of an adapted Generalized Estimating Equations method that can be applied to ordinal longitudinal data.

2. Methods

2.1. OSM: Formulation and Basics

Let Y_i be an ordinal response with q levels, i.e., $Y_i \in \{1, 2, \ldots, q\}$, for observation i, where $i = 1, \ldots, n$. The ordered stereotype model [15] has the following form:

$$w_{ik} \stackrel{\text{def}}{=} \log\left(\frac{P[Y_i = k \mid \boldsymbol{x}_i]}{P[Y_i = 1 \mid \boldsymbol{x}_i]}\right) = \alpha_k + \phi_k \boldsymbol{\beta}' \boldsymbol{x}_i, \tag{1}$$

where $i = 1, \ldots, n$, $k = 2, \ldots, q$, and $\alpha_1 = 0$. The model imposes a monotone non-decreasing constraint:

$$0 = \phi_1 \leq \phi_2 \leq \cdots \leq \phi_q = 1 \tag{2}$$

to ensure the ordered nature of Y_i [15].

Model (1) treats the first category ($k = 1$) as the reference category. The covariates $\boldsymbol{x}_i$ can be categorical or numerical. The parameters $\boldsymbol{\beta}$ quantify the effects of $\boldsymbol{x}_i$ on w_{ik}. Note that the monotonic relationship imposed by the constraint (2) is enforced for the overall linear predictor term $\boldsymbol{\beta}' \boldsymbol{x}_i$, rather than for any single covariate $\boldsymbol{x}_{i\ell}$.

For illustration, Figure 1 shows four simulated scenarios in which probability distributions of each category k of a $q = 4$-level ordinal variable in Model (1) are depicted. Within each panel, the covariate-dependent term in the linear predictor, $\eta = \boldsymbol{\beta}' \boldsymbol{x}_i$, is varied, and the effects of changing the parameters $\{\phi_k\}$ and $\{\alpha_k\}$ are shown between the panels. Thus, each graph represents a different scenario and, within each graph, there is one curve for each response category $k \in \{1, \ldots, 4\}$ against varying η.

Without imposing the non-decreasing constraint (2), Model (1) is appropriate for a nominal response variable that lacks an intrinsic ordering. Anderson [15] motivated Model (1) from a baseline-categories logit model with $w_{ik} = \alpha_k + \boldsymbol{\beta}_k' \boldsymbol{x}_i$. To make the model more parsimonious, $\boldsymbol{\beta}_k' \boldsymbol{x}_i$ is then replaced by $\phi_k \boldsymbol{\beta}' \boldsymbol{x}_i$. Thus, the OSM achieves the parsimony of a single parameter to describe a predictor effect by using the same scores for each predictor ([10], Chapter 4.3.1). Figure 1a is an example of the probabilities of the stereotype model without the ordering constraint on the score parameters ϕ, and the sequence of curves is out of order (i.e., 1, 3, 2, 4).

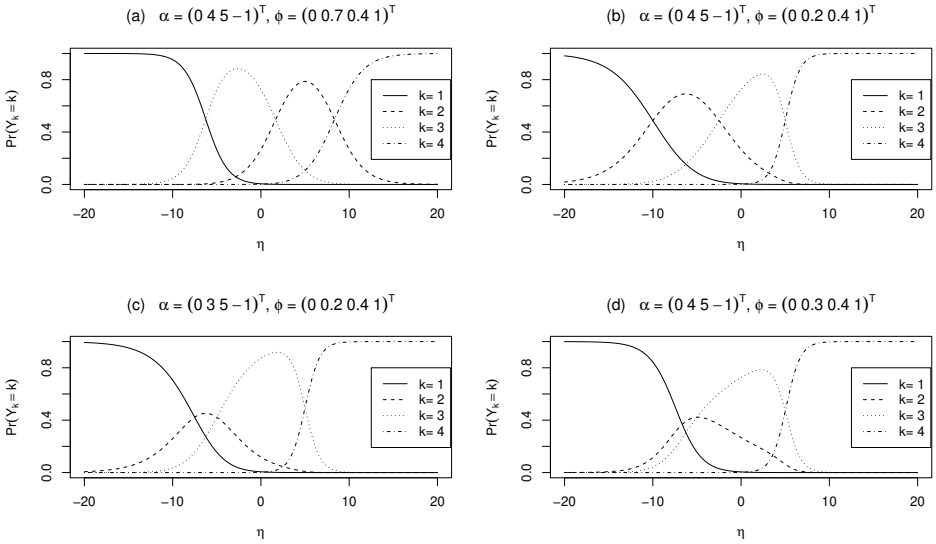

Figure 1. Illustration of the effects of parameters $\{\alpha_k\}$ and $\{\phi_k\}$ as a function of $\eta = \boldsymbol{\beta}' \boldsymbol{x}_i$. This figure is taken under permission of the author from [23]. (©2020 The Authors. *Statistics in Medicine* published by John Wiley & Sons, Ltd., Hoboken, NJ, USA).

The parameters $\{\alpha_k\}$ determine the shape of the distribution, and, if Y were a numerical random variable, would control aspects such as skewness and kurtosis.

We can observe the effect of changing the values of $\{\alpha_k\}$ when the values (α_2, α_3) are changed from $(4, 5)$ in Figure 1b to $(3, 5)$ in Figure 1c. The reduced value of α_2 reduces the probability of category 2 occurring.

We interpret the parameters $\{\phi_1, \phi_2, \ldots, \phi_q\}$ as the scores of the response Y_i. These scores lie between the fixed extremes of $\phi_1 = 0$ and $\phi_q = 1$, but they do not need to be equally spaced. We can observe how the score parameters affect $P[Y_i = k]$ associated with $\boldsymbol{\beta}'\boldsymbol{x}_i$ in Figure 1; the scores for categories $k = 2$ and $k = 3$ are brought closer together from $(\phi_2, \phi_3) = (0.2, 0.4)$ in Figure1b to $(0.3, 0.4)$ in Figure1d. The curves for categories $k = 2$ and $k = 3$ overlap more strongly in Figure1d as a result.

From Model (1), the probabilities $\theta_{ik} = P(Y_i = k \mid \boldsymbol{x}_i)$ can be derived as

$$\theta_{ik} = P[Y_i = k \mid \boldsymbol{x}_i] = \frac{e^{w_{ik}}}{\sum_{\ell=1}^{q} e^{w_{i\ell}}}. \tag{3}$$

When $\phi_j = \phi_{j'}$ with $j \neq j'$,

$$\log\left(\frac{P[Y_i = j \mid \boldsymbol{x}_i]}{P[Y_i = j' \mid \boldsymbol{x}_i]}\right) = \alpha_j - \alpha_{j'},$$

which is independent of $\boldsymbol{x}_i$. Thus, the response levels j and j' can be combined as a single level when trying to find the association between $\boldsymbol{x}_i$ and Y_i. See (Agresti [10], Ch. 4) for further discussion. After the categories are combined, the intercept will change, but the term $\phi_j\boldsymbol{\beta}'\boldsymbol{x}_i$ in Model (1) remains unchanged.

Although the ordinal data themselves do not provide information about the spacings between response categories, if the ordered stereotype model (1) holds, then observed variations in response propensity associated with variations in covariates can reveal information about spacing. If only a small change in a covariate is required to move from, say, level 2 to level 3, but a larger variation is required to move from level 3 to level 4, we can infer that levels 2 and 3 are closer to each other than level 3 is to level 4. Note that, when $\{\phi_k\}$ are equally spaced, Model (1) is equivalent to the adjacent-categories logit model shown by (Agresti [10], Ch. 4). Additionally, it is often sensible to conduct a likelihood-ratio test comparing the OSM with score parameters to the special case with fixed, equally spaced score parameters, which corresponds to the proportional-odds form of the adjacent-categories logit model. Such a test determines if the score parameters depart significantly from being equally spaced, which allows working with more parsimonious models.

Parameter estimation methods for Model (1) include the standard maximum likelihood (ML) method [15] or the generalized least squares (GLS) method [24]. Holtbrugge and Schumacher [25] proposed a method to estimate the parameters in the stereotype model by using an iteratively reweighted least square algorithm and Greenland [16] and Preedalikit et al. [26] separately proposed alternating algorithms based on two iterative steps. Feldmann and König [27] proposed a maximum likelihood parameter estimation based on discriminant analysis. In a Bayesian approach, Ahn et al. [1,28] presented a comprehensive inference method for fitting the OSM in case-control studies. Lunt and Unit [29] described how to estimate the parameters of the OSM via the Stata package *soreg* and the advantages of this method over other Stata commands and Kuss [24] introduced two methods to estimate this model using SAS.

To the best of our knowledge, there are only three packages in the statistical software R [30] for fitting the Stereotype Model (SM): *gnm* [31], *VGAM* [32], and *ordinalgmifs* [33]. However, none of them impose the monotone non-decreasing constraint (2) in the process of estimation and, therefore, none of them can fit the model for ordinal responses (OSM), only for nominal responses (SM). Our **R** package *clustord* [34] contains the method *osm* that fits the ordered stereotype regression model using the reparametrization method in Fernández et al. [35].

2.2. Recent Advances in the OSM

2.2.1. Goodness-of-Fit Tests

Regarding checking the model fit, there are few methods available for ordinal models. Most of the existing methods focus on the proportional odds model (POM). For example, Fagerland and Hosmer [36] introduced a test extending the binary case [37]; Pulkstenis and Robinson [38] developed a modified Pearson χ^2 statistic; Lin and Chen [39] proposed a test that applies a non-parametric local linear smoothing technique; and Liu et al. [40] showed a graphical method using cumulative sums of residuals. Additionally, Lipsitz et al. [41] proposed a test for several regression models for ordinal responses, but it is not always suitable for small samples. Finally, Li and Shepherd [42] and Liu et al. [43] proposed new residuals for ordinal regression models.

This paper discusses three approaches to evaluate Model (1) based on equivalent tests for the POM. These methods construct the tests using the new spacing of the ordinal response categories dictated by the estimated score parameters $\{\phi_k\}$.

Fernández and Liu [44] modified the Hosmer–Lemeshow test for the proportional odds model [36] to propose a new test. The (possibly) uneven spacing of the ordinal response categories, determined by the fitting of the OSM, is used in the test to compute the weighted score for each observation, which is required to set the partition in Hosmer–Lemeshow tests, and also for replacing the default equally-spaced response category labels for $\{Y_i\}$ with their corresponding scores ϕ_k (rescaled from the original $[0, 1]$ to be into the range $[1, q]$). Unlike traditional Hosmer–Lemeshow tests, this test computes two partitions instead of only one. It applies the first partition based on deviances, before applying a second partition based on the weighted scores. This first partition is not applied in the traditional Hosmer–Lemeshow test. It is a novel step proposed for the OSM test, and its inclusion makes it easier to detect the lack of fit than with traditional Hosmer–Lemeshow tests, which have the potential disadvantage of missing an important deviation from fit during the grouping process. Fernández and Liu [44] described the technical details of the steps for this test and evaluated the performance of the reliability of the test (OSM_{HL}) by setting up a comprehensive simulation study, in which the null distribution and the power of the test statistic were assessed. The results of this empirical study showed that the proposed test performed best when the model contained continuous covariates compared to the traditional Pearson's chi-squared test. However, the OSM_{HL} test is sensitive to the choice of the number of groups based on the deviance. When the number of groups in the first partition (based on the deviance) is too large, the grouping scheme separates the observations extremely, which leads to a large value of the test statistic, even if the null model holds. Therefore, the test might incorrectly signal a lack of fit in such cases. For that reason, we recommend always grouping into only two groups during the first partition.

The other two goodness-of-fit tests for the OSM were presented in Fernández et al. [45] with modifications based on the Lipsitz test for the proportional odds model [41]. Firstly, the Lipsitz test involves partitioning subjects into groups based on assigning equally-spaced scores to the response categories, whereas our two new tests incorporate the fitted score parameters of the OSM during the partitioning process, and these fitted score parameters may not be equally spaced. Secondly, the Lipsitz test assesses the goodness of fit of the null model by comparing it with an alternative model. In the alternative model, the group effects are added. The goodness-of-fit test is equivalent to a test of no group effects using the standard statistical tests, including a likelihood-ratio, Wald, or score test statistic. In these new tests, we also construct an alternative model with putative group effects, using two approaches. In the first approach, (OSM_L), we construct the alternative model using the OSM. In the second approach, (OSM_{LML}), we use the fitted score parameters from the OSM to replace the original ordinal responses $\{Y_i\}$ with their corresponding numerical scores ϕ_k, and thus we fit a linear model with the rescaled responses. Fernández et al. [45] described the technical details of the steps for these two tests.

Fernández and Liu [44] ran a comprehensive simulation study to compare the performance of these two tests (OSM_L and OSM_{LML}) with OSM_{HL}, which is a version of the

Fagerland and Hosmer test for the proportional odds model, modified to calculate the fitted probabilities from the OSM. For large data, the OSM_L test performs well. The OSM_{LML} test is better when there are few covariates. The simulation study also showed that, although the OSM_L and OSM_{LML} achieved their correct nominal significance rates, and have a good power against covariate mis-specification, they were not highly specific to the OSM. Therefore, rejection of the null model using both tests provides enough lack of fit evidence of the null model. However, another model may also hold even when one fails to reject the null model. This drawback is also common with other goodness-of-fit tests.

In addition to the three tests described above, Liu and Fernández [46] discussed a modified Hosmer–Lemeshow method for large data sets for the OSM. It is well known that the power of goodness-of-fit tests increases when the sample size increases. The null hypothesis of perfect fit is likely to be rejected for large samples, regardless of how good the model is. Nattino et al. [47] proposed a modified Hosmer–Lemeshow approach that does not depend on the sample size to evaluate the adequacy of a logistic regression model. Their new test statistic is based on a noncentrality parameter that measures the level of lack of fit. Liu and Fernández [46] generalized their modified approach to the test OSM_{HL}, and used a small simulation study to show the performance when the sample size ranges from 25,000 to 1 million. The modified Hosmer–Lemeshow test seems to be more conservative for the OSM than for logistic regression models. Further investigation is needed in order to improve the power.

2.2.2. Generalized Estimating Equations

Thus far, we have discussed the OSM and its recent advances in a cross-sectional context. However, there are also very few estimation methods for this model for longitudinal data. These data differ from cross-sectional data in that for each individual i there are multiple observations taking place at different times t. The data Y_{it} are then grouped within individuals, and are associated with individual level covariates that may also vary over time $\boldsymbol{x}_{it}$. Crucial to the analysis of such data are the proper inclusion of any within-individual correlation among these repeated observations.

Kuss [48] and Johnson [18] (via ML and GLS approaches) proposed the use of existing statistical software to estimate the OSM for longitudinal data. However, the GLS approach requires knowing the correct structure of the mean and associations among the ordered response variables, in order to obtain consistent estimation of the parameters of interest. On the other hand, the ML approach is computationally complex, due to the complicated structure of the joint distribution of the response variables. Thus, the complexity cost of the ML estimation approach is high even with modern computing capacities.

Liang and Zeger [49] proposed a generalized estimating equations (GEE) approach for longitudinal data. The method only needs to specify the mean structure of the response variables marginally—and thus the marginal forms of the model (1) can be used without needing to construct a full joint model. Dependencies over time are then modeled via a "working" correlation matrix. This correlation matrix contains a simple estimate of the correlation between the repeated measures for a single individual. It is diagonal if the observations are independent, but may encode, for example, an autoregressive time dependence. The functional form is chosen by the modeler, and this choice can be checked with goodness of fit tests. The parameter estimators of the mean model are consistent and asymptotically normally distributed under weak regularity conditions, even if the working correlation matrix is incorrect. Because the GEE approach does not need to specify the joint distribution of the response variables, it has the advantage of avoiding the calculation of high-dimensional integrals, and is thus computationally efficient.

Spiess et al. [23] recently developed a GEE approach to estimate the OSM for longitudinal data. The approach uses a finite sample correction, and is based on working covariance matrices, which are not required to be correctly specified. To the best of our knowledge, that was the first time this approach has been developed. In this work, simulation studies confirmed the properties of GEE estimators as described in the literature for other models.

Additionally, the authors observed that, if the true correlations are high, then adopting a working correlation matrix that explicitly models these associations may lead to substantial efficiency improvements for the estimators, in comparison with those when the identity matrix is adopted. In the final part of this paper, the authors evaluated the performance of the estimators by replacing the working correlation structure with the local odds ratio structure [50].

3. Case Study

This section uses a real life example of the arthritis clinical trial to show the application of the proposed goodness-of-fit tests to cross-sectional ordinal data and of the GEE estimator to longitudinal ordinal data.

3.1. Arthritis Clinical Trial

A randomized clinical trial was designed to evaluate the effectiveness of the drug Auranofin relative to a placebo for the treatment of rheumatoid arthritis [51,52]. Following Lipsitz et al. [51] and Touloumis et al. [50], we consider the completely observed cases in our analysis by assuming that the missing values were missing completely at random.

This study includes 302 observed individuals. The data set is available from the original arthritis data set in the R package *multgee* [53] by calling *data(arthritis)*. We treat the self-assessment of rheumatoid arthritis as an ordinal response variable with a 5-level Likert scale from "very poor" (1) to "very good" (5). The dataset includes the self-assessment responses before the trial (as a baseline) and at $T = 3$ follow-up time points at 1, 3, and 5 months after the treatment. The available covariates of interest were sex, age, and type of treatment (placebo or Auranofin drug). Due to its ordered nature, the covariate Baseline could be treated as categorical or numerical.

We made a few modifications to the original data set. Firstly, we removed nine individuals who had missing values for the response at at least one of the time points. Thus, the final sample size used was $n = 293$, which is the same as in Lipsitz et al. [41]. Secondly, we noted that very few of the observations were in the category "very good" (5), and this led to convergence problems in the case of the GEE estimators. Thus, we merged the two categories "good" (4) and "very good" (5), and thenceforth worked with $q = 4$ ordinal response variables, for both the goodness-of-fit tests and the GEE estimators. Table 1 lists the variables.

Table 1. Variables and their possible values in the rheumatoid arthritis dataset.

Variable	Description	Values
Baseline	Self-assessment before the trial	1 = very poor 2 = poor
t1, t3, t5	and 1, 3, and 5 months follow-up, respectively	3 = fair 4 = good or very good
Sex	Gender of the individual	0 = female 1 = male
Age	Years. Recorded at the baseline.	Range 21–66
Trt	Treatment	0 = placebo group 1 = drug group

3.2. Goodness-of-Fit Test for the Arthritis Data Set

We start with the illustration of the goodness-of-fit tests described in this summary article (details of the formulation and technical steps are detailed in [44,45]). We initially fitted the OSM to the arthritis dataset. We used our own R code, which includes the constraint (2) following Fernández et al. [35] because methods in other R packages do not enforce the ordinal nature of the responses.

In the model fitting, we treated sex, age, type of treatment, and self-assessment score of arthritis at the baseline as predictors. Additionally, as the described goodness-of-fit tests were developed for cross-sectional studies, we only used the self-assessment of arthritis at 5 months (t5) as the ordinal response for these tests, and removed the variables related to the other follow-up time points (i.e., t1 and t3). Table 2 shows the fitted OSM parameter values.

Table 2. Parameter estimates for the OSM (1). The ordinal response is self-assessment of arthritis at 5 months (t5). The significant effects are shown in bold.

Coefficient	Estimate	S.E.	95% C.I.
$\widehat{\alpha}_2$	**−0.145**	**0.060**	**(−0.263, −0.027)**
$\widehat{\alpha}_3$	**−0.782**	**0.096**	**(−0.969, −0.594)**
$\widehat{\alpha}_4$	**−2.508**	**0.117**	**(−2.737, −2.279)**
$\widehat{\beta}_1$ (Sex)	0.225	0.246	(−0.257, 0.707)
$\widehat{\beta}_2$ (Age)	−0.020	0.012	(−0.044, 0.004)
$\widehat{\beta}_3$ **(Trt)**	**1.333**	**0.237**	**(0.869, 1.797)**
$\widehat{\beta}_4$ **(Baseline)**	**2.304**	**0.260**	**(1.795, 2.812)**
$\widehat{\phi}_2$	**0.402**	**0.167**	**(0.075, 0.729)**
$\widehat{\phi}_3$	**0.672**	**0.142**	**(0.394, 0.950)**

We observe that the drug treatment significantly improves arthritis, but neither sex nor age are significant. Furthermore, patients who had a good self-assessment at the baseline (i.e., before the trial) tend to have a better self-assessment at 5 months. The score parameter estimates ϕ_2 and ϕ_3 imply that they differ significantly from $\phi_1 = 0$ and $\phi_4 = 1$. Figure 2 visually compares the default equal spacing of categories with the fitted spacing from $\{\widehat{\phi}_k\}$ (rescaled to range from 1 to q). The bottom axis shows the equally spaced scale and the top axis shows the fitted score scale as indicated by the data. Although there is not a big difference, we can distinguish the non-equally spaced categories in the top axis.

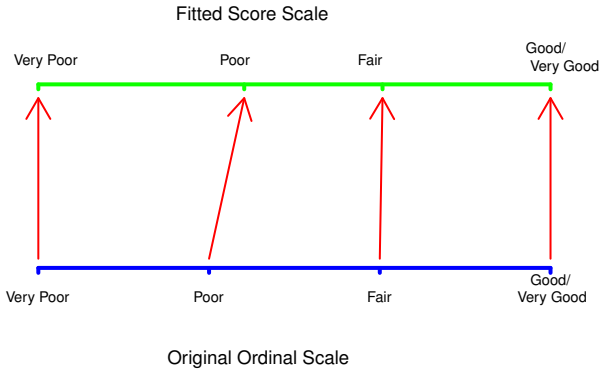

Figure 2. Reassigned ordinal scale: Scale comparison between the equally spaced scale and the score scale indicated by the fitted score parameters $\{\widehat{\phi}_k\}$ for the arthritis dataset.

Table 3 depicts the results of the goodness-of-fit tests described in this article, i.e., OSM_{HL}, OSM_{L}, and OSM_{LML}, and also, for comparison purposes, our version of the Fagerland and Hosmer goodness-of-fit test that uses probabilities calculated from the fitted OSM. All four tests use the same partition into 10 groups.

Table 3. Goodness-of-fit test results for the OSM fitted to the arthritis clinical trial data set. All tests use the same partition into 10 groups.

Test	Statistical Value	p-Value
Fagerland and Hosmer (OSM version)	21.58	0.198
OSM_{HL}	22.36	0.398
OSM_{L}	11.64	0.227
OSM_{LML}	14.03	0.112

As we can observe in Table 3, there is no evidence of lack of fit at the 5% significance level, regardless of the test applied. These goodness-of-fit results using the OSM (Table 2) are similar to the results in Lipsitz et al. [41] for the proportional odds model.

3.3. GEE Estimator for the Arthritis Data Set

In the previous section, we only used a single ordinal response (the self-assessment of arthritis at 5 months) to fit the OSM and test its goodness-of-fit. However, as Lipsitz et al. [51] and Touloumis et al. [50] did, we used all three follow-up assessments (t1, t3, and t5) in order to compute the GEE estimator approach for longitudinal data. In that manner, the t3 and t5 variables describing the self-assessment at 3-months and 5-months after treatment are entered into the model as two dummy responses, and the t1 variable describing the 1-month follow-up is treated as the reference category in the model.

We used the GEE approach with a comprehensive range of working correlation matrices (see details of the correlation matrices in [23]) to fit the arthritis data set considering the longitudinal ordinal responses. Before showing the results, we need to establish some nomenclature: the different working correlation matrices assuming independence, equicorrelation, unstructured, Toeplitz, or AR(1) are indicated by subscripts I, E, U, T, A, respectively. In general, GEE estimators are indicated as GEE_{ℓ}, where $\ell \in \{I, E, U, T, A\}$. Alternatively, to work with different correlation structures to specify the dependence, Spiess et al. [23] adopt a local odds-ratio approach [50] to model the dependence. The common structure based on local odds ratios includes the uniform, the category exchangeable, the time exchangeable, and the row-and-column effect (RC) [54] structures, which are denoted by GEE_{UN}, GEE_{CE}, GEE_{TE}, GEE_{RC}, respectively.

The results of fitting the OSM using a diverse range of working correlation matrices are shown in Table 4 and the working correlation matrices over time for the arthritis clinical trial data set and for the GEE_{E} (lower triangular matrix) and the GEE_{U} (upper triangular matrix) estimators are shown in Table A1 in Appendix A.

Inspecting the results in Table 4, we conclude that the estimates of the set of parameters α_3, α_4, ϕ_3, $\boldsymbol{\beta}_{t2}$, $\boldsymbol{\beta}_{Trt}$, $\boldsymbol{\beta}_{b3}$, and $\boldsymbol{\beta}_{b4}$ do not strongly depend on the choice of working correlation matrix. Moreover, in all cases α_3, α_4, and $\boldsymbol{\beta}_{t2}$ are not statistically different from zero, as the 95% confidence intervals cover zero. Additionally, we observed inconsistencies in the estimates of the remaining parameters α_2, ϕ_2, $\boldsymbol{\beta}_{t3}$, and $\boldsymbol{\beta}_{b2}$. However, given that all the estimated correlations in the correlation matrices are small, the choice of the working correlation matrix does not change the conclusion of the analysis.

The GEE estimators all lead to similar conclusions. For the $\{\alpha_k\}$ parameters, the 95% confidence intervals cover zero for all GEE estimators, apart from α_2 under GEE_{UN} and GEE_{CE}, although, for these cases, one side of the 95% confidence interval is very near zero. Regarding the score parameters $\{\phi_k\}$, all GEE estimators give similar values, enforcing the monotonic constraint (2), consistently implying that ϕ_2 and ϕ_3 differ from $\phi_1 = 0$ and $\phi_4 = 1$. The 95% confidence intervals overlap for the second and third categories, suggesting that it is possible we could collapse those categories into a single category. However, the description of collapsing categories in Agresti [10] implies that the point estimates should be closer and the final decision of merging categories should be taken by the practitioner who knows the data context, so in this case we will not collapse them.

Regarding the follow-up time point effect parameters $\{\boldsymbol{\beta}_{t1}, \boldsymbol{\beta}_{t3}, \textit{and } \boldsymbol{\beta}_{t5}\}$, we can observe there is no significant difference at the 3-month time point relative to the 1-month

time point under any of the GEE estimators. However, there is a positive effect of the 5-month time point relative to the 1-month time point in most of the GEE estimators. This result implies that a difference in self-assessment may be observed at the earliest after 3 months, but not before 3 months.

The effect of the Auranofin treatment (β_{Trt}) has a positive effect on the responses. Its estimated increase in the log odds relative to the placebo group is around 1.2 in the majority of GEE estimators. Finally, there seems to be no difference between the effects of the baseline self-assessments 1 = "very poor" (β_{b1}) and 2 = "poor" (β_{b2}) on self-assessment at the follow-up time points. However, baseline self-assessments of 3 = "fair" (β_{b3}) and 4 = "good or very good" (β_{b4}) have an effect on self-assessments at the follow-up time points. It implies that patients who were less affected by arthritis before the trial are still doing well after the treatment. Furthermore, our conclusion may be sensitive to the presence of unobserved patient/individual heterogeneity, which was not considered in our models.

Table 4. Estimate of parameters (E), standard errors (S.E.), lower (L), and upper (U) 95% confidence bounds, assuming normality, for the arthritis clinical trial data set. GEE estimators are indicated correspondingly as GEE_ℓ, where $\ell \in \{I, E, U, T, A\}$. The GEE_I, GEE_{UN}, GEE_{CE}, GEE_{TE}, and GEE_{RC} estimators assume independence, the uniform, the category exchangeable, the time exchangeable, and the RC structure, respectively. The parameters β_{b2}, β_{b3}, and β_{b4} represent the baseline assessment variables, β_{t3} and β_{t5} are the month follow-up (reference category is the first month follow-up), and β_{Trt} is the effect of the treatment (placebo group being the reference group). These results are also reported in [23].

Pars.	GEE_I				GEE_E				GEE_U				GEE_T			
	E	S.E.	L	U	E	S.E.	L	U	E	S.E.	L	U	E	S.E.	L	U
α_2	0.800	0.552	−0.283	1.883	0.746	0.453	−0.141	1.633	0.788	0.451	−0.095	1.672	0.742	0.451	−0.143	1.626
α_3	0.824	0.674	−0.496	2.145	0.755	0.543	−0.310	1.820	0.804	0.542	−0.258	1.865	0.748	0.544	−0.318	1.814
α_4	−0.485	0.852	−2.156	1.185	−0.654	0.723	−2.070	0.762	0.615	0.730	−2.045	0.815	−0.668	0.727	−2.092	0.757
ϕ_2	0.349	0.167	0.021	0.677	0.349	0.169	0.018	0.680	0.339	0.170	0.005	0.672	0.349	0.167	0.021	0.677
ϕ_3	0.623	0.102	0.422	0.823	0.612	0.122	0.373	0.851	0.605	0.123	0.365	0.846	0.612	0.122	0.373	0.850
β_{t3}	−0.130	0.269	−0.656	0.397	−0.114	0.260	−0.624	0.397	0.113	0.258	−0.619	0.393	−0.119	0.261	−0.629	0.392
β_{t5}	0.505	0.254	0.007	1.004	0.538	0.266	0.017	1.059	0.526	0.263	0.012	1.041	0.530	0.264	0.012	1.047
β_{Trt}	1.191	0.471	0.267	2.115	1.240	0.410	0.437	2.043	1.216	0.409	0.415	2.017	1.239	0.412	0.430	2.047
β_{b2}	1.271	0.900	−0.494	3.036	1.458	0.765	−0.042	2.957	1.467	0.768	−0.037	2.972	1.494	0.767	−0.009	2.998
β_{b3}	2.449	0.937	0.613	4.285	2.602	0.751	1.130	4.075	2.552	0.755	1.072	4.033	2.616	0.755	1.136	4.095
β_{b4}	5.331	1.637	2.123	8.538	5.356	1.482	2.451	8.262	5.312	1.458	2.454	8.171	5.375	1.473	2.487	8.262

Pars.	GEE_A				GEE_{UN}				GEE_{CE}				GEE_{TE}			
	E	S.E.	L	U	E	S.E.	L	U	E	S.E.	L	U	E	S.E.	L	U
α_2	0.791	0.460	−0.111	1.693	1.063	0.493	0.096	2.030	1.067	0.498	0.092	2.042	0.833	0.458	−0.065	1.731
α_3	0.773	0.557	−0.319	1.865	1.142	0.593	−0.020	2.305	1.153	0.591	−0.004	2.311	0.831	0.548	−0.244	1.906
α_4	−0.566	0.731	−1.999	0.867	−0.172	0.793	−1.726	1.382	−0.191	0.797	−1.754	1.372	−0.623	0.754	−2.102	0.855
ϕ_2	0.342	0.178	−0.007	0.692	0.269	0.208	−0.139	0.677	0.269	0.208	−0.138	0.676	0.320	0.174	−0.020	0.660
ϕ_3	0.624	0.126	0.377	0.870	0.563	0.139	0.290	0.835	0.557	0.139	0.285	0.829	0.590	0.124	0.346	0.834
β_{t3}	−0.125	0.262	−0.638	0.388	−0.118	0.245	−0.597	0.362	−0.111	0.243	−0.587	0.365	−0.101	0.253	−0.597	0.394
β_{t5}	0.514	0.265	−0.005	1.033	0.458	0.242	−0.016	0.932	0.458	0.242	−0.016	0.932	0.529	0.255	0.030	1.029
β_{Trt}	1.231	0.412	0.423	2.038	0.986	0.399	0.203	1.768	0.976	0.396	0.199	1.752	1.148	0.397	0.371	1.926
β_{b2}	1.348	0.767	−0.156	2.852	1.072	0.801	−0.497	2.641	1.144	0.809	−0.442	2.729	1.471	0.792	−0.081	3.024
β_{b3}	2.526	0.762	1.032	4.019	2.082	0.820	0.474	3.689	2.090	0.822	0.479	3.702	2.537	0.775	1.018	4.057
β_{b4}	5.341	1.579	2.247	8.436	4.841	1.542	1.819	7.863	4.863	1.521	1.883	7.844	5.321	1.450	2.479	8.163

Pars.	GEE_{RC}			
	E	S.E.	L	U
α_2	0.811	0.456	−0.084	1.705
α_3	0.806	0.538	−0.249	1.861
α_4	−0.686	0.749	−2.155	0.783
ϕ_2	0.322	0.169	−0.010	0.654
ϕ_3	0.589	0.120	0.353	0.824
β_{t3}	−0.064	0.252	−0.559	0.430
β_{t5}	0.533	0.253	0.036	1.029
β_{Trt}	1.165	0.395	0.391	1.939
β_{b2}	1.565	0.792	0.013	3.117
β_{b3}	2.594	0.771	1.083	4.105
β_{b4}	5.351	1.400	2.607	8.096

4. Discussion

Ordinal data occur often in applied and social sciences. Because the normal distribution assumption does not hold for ordinal variables, using an ordinary linear model to analyze ordinal data might result in an incorrect conclusion. One possible source of error is "floor" and "ceiling" effects that occur when substantial proportions of subjects give either maximum or minimum scores, so that the true extent of the measures cannot be determined accurately. (For more details of these effects, see [10]).

The use of ordinal regression models has several benefits over ordinary regression models, such as making as few assumptions as possible. It also gives greater power for detecting relevant trends compared to the baseline-categories logit model that ignores the ordering information ([10], Section 1.2). In this article, we focused on the ordered stereotype model (OSM) because we consider that it has advantages over other ordinal models. In particular, one of the main benefits of this model is that it allows us to determine response scores by using the score parameter estimates.

This article is a summary of recent advances of the OSM. We presented three new goodness-of-fit tests, of which one is based on the traditional Hosmer–Lemeshow test (OSM_{HL}), and the other two are based on the method from Lipsitz et al. [41] (OSM_{L} and OSM_{LML}). All of these tests incorporate the new spacing information dictated by the data to group observations. The OSM_{HL} test is easy to use for applied researchers because the derivation of the proposed test is similar to the traditional Hosmer–Lemeshow test. In the OSM_{L} and OSM_{LML} tests, we compare the null model with the alternative model that has grouping effects. Rejection of the null hypothesis implies that the null model does not fit the data well.

Another recent development of the OSM described in this article is a GEE approach for longitudinal data. The estimators have the same properties as the ones for logistic regression models. The estimators are consistent even if the dependence structure is misspecified. It should be noted that the properties hold when missing values are treated as missing completely at random. If that is not the case, then one could implement a missing data strategy, e.g., imputation. To allow a more flexible assumption of missingness, such as missing at random, one should adopt the ML approach using a random effect model. In addition, if the pattern of missing values is non-monotone, then the methods mentioned above will likely give invalid inferences. Finally, the proposed GEE method is restricted to cases where repeated responses are observed at the same time points. In future work, we will generalize the method to consider responses obtained at varying time intervals.

In practice, the OSM has been used less often in applied research compared to other ordinal response models (see [24] for some exceptions), despite its advantages. It might be due to the lack of standard software for model fitting that requires special consideration of constraints on the parameters. There have however been considerable recent developments of macros and functions in standard software to estimate the stereotype model. Kuss [24] modified several standard procedures in SAS to obtain the maximum likelihood estimates. Lunt and Unit [29] developed a Stats module called soreg that implements the OSM. Yee and Hastie [55] used reduced-rank multinomial logistic models to fit the OSM, available in the R package VGAM (Vector Generalized Additive Model) [32]. Finally, the R package ordinalgmifs [33] provides the function *ordinalgmifs* that can be used to fit the OSM. The code of all tests described in this article and the OSM fitting was written in statistical software R [30]. The authors are developing an R package implementing the methods described, and, meanwhile, the code is available from the authors upon request.

We believe that the use of the OSM described in this article may be advantageous for researchers in statistics and practitioners in the applied fields. The estimation of the spacing among response categories is an improvement over other ordinal data models. For example, one might use the estimated scores to calculate numerical summary statistics [56] for other practitioners to understand easily.

The development of the OSM for multi-level ordinal data (clustered and longitudinal data) might be a field to explore for future research. Another possible future research direc-

tion would be to generalize alternative grouping methods used for the binary case [57–59] instead of using the one based on Hosmer–Lemeshow and compare them. Finally, our next research goal is developing a more flexible OSM to incorporate different scores to provide insight on the spacing information that depends on different sets of predictors.

Author Contributions: Conceptualization, D.F., I.L., L.M., R.A. and M.S.; methodology, D.F., I.L., L.M., R.A. and M.S.; software, D.F., I.L., L.M., R.A. and M.S.; formal analysis, D.F., I.L., L.M., R.A. and M.S.; investigation, D.F., I.L., L.M., R.A. and M.S.; writing—original draft preparation, D.F., I.L., L.M., R.A. and M.S.; writing—review and editing, D.F., I.L., L.M., R.A. and M.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research has been supported by Marsden grant E2987-3648 administrated by the Royal Society of New Zealand, by grant 2017 SGR 622 (GRBIO) administrated by the Departament d'Economia i Coneixement de la Generalitat de Catalunya (Spain) and by the Ministerio de Ciencia e Innovación (Spain) [PID2019-104830RB-I00/DOI (AEI): 10.13039/501100011033].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data used in this research are publicly available as detailed in the article.

Acknowledgments: Daniel Fernández is a Serra Húnter Fellow.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Arthritis Clinical Trial

Table A1 shows the working correlation matrices GEE_E (lower triangular) and GEE_U (upper triangular) for the arthritis clinical trial data set (Section 3).

Table A1. Working correlation matrices over time and for the GEE_E (lower triangular) and GEE_U (upper triangular) estimators, after fitting OSM in the arthritis clinical trial data set.

Follow-Up (1-Month) (t1)			Follow-Up (3-Month) (t3)			Follow-Up (5-Month) (t5)		
1	0	0	0.116	−0.107	−0.107	0.221	−0.047	−0.047
0	1	0	−0.107	0.116	−0.107	−0.047	0.221	−0.047
0	0	1	−0.107	−0.107	0.116	−0.047	−0.047	0.221
0.191	−0.078	−0.078	1	0	0	0.236	−0.078	−0.078
−0.078	0.191	−0.078	0	1	0	−0.078	0.236	−0.078
−0.078	−0.078	0.191	0	0	1	−0.078	−0.078	0.236
0.191	−0.078	−0.078	0.191	−0.078	−0.078	1	0	0
−0.078	0.191	−0.078	−0.078	0.191	−0.078	0	1	0
−0.078	−0.078	0.191	−0.078	−0.078	0.191	0	1	0

References

1. Ahn, J.; Mukherjee, B.; Banerjee, M.; Cooney, K.A. Bayesian inference for the stereotype regression model: Application to a case–control study of prostate cancer. *Stat. Med.* **2009**, *28*, 3139–3157. [CrossRef] [PubMed]
2. Cupp, M.A.; Owugha, J.; Florschutz, A.; Beckingham, A.; Kisan, V.; Manikam, L.; Lakhanpaul, M. Birthing a better future: A mixed-methods evaluation of multimedia exposition conveying the importance of the first 1001 days of life. *Lancet* **2018**, *392*, S27. [CrossRef]
3. Furman, B.T.; Leone, E.H.; Bell, S.S.; Durako, M.J.; Hall, M.O. Braun-Blanquet data in ANOVA designs: Comparisons with percent cover and transformations using simulated data. *Mar. Ecol. Prog. Ser.* **2018**, *597*, 13–22. [CrossRef]
4. McNellie, M.J.; Dorrough, J.; Oliver, I. Species abundance distributions should underpin ordinal cover-abundance transformations. *Appl. Veg. Sci.* **2019**, *22*, 361–372. [CrossRef]
5. Loda, T.; Löffler, T.; Erschens, R.; Zipfel, S.; Herrmann-Werner, A. Medical education in times of COVID-19: German students' expectations–A cross-sectional study. *PLoS ONE* **2020**, *15*, e0241660. [CrossRef]
6. Likert, R. A technique for the measurement of attitudes. *Arch. Psychol.* **1932**, *22*, 5–55.
7. Göb, R.; McCollin, C.; Ramalhoto, M. Ordinal Methodology in the Analysis of Likert Scales. *Qual. Quant.* **2007**, *41*, 601–626. [CrossRef]
8. Braun-Blanquet, J. *Plant Sociology: The Study of Plant Communities*; McGraw Hill: New York, NY, USA, 1932.
9. Wikum, D.A.; Shanholtzer, G.F. Application of the Braun-Blanquet cover-abundance scale for vegetation analysis in land development studies. *Environ. Manag.* **1978**, *2*, 323–329. [CrossRef]

10. Agresti, A. *Analysis of Ordinal Categorical Data*, 2nd ed.; Wiley Series in Probability and Statistics; Wiley: Hoboken, NJ, USA, 2010.
11. Stromberg, U. Collapsing ordered outcome categories: A note of concern. *Am. J. Epidemiol.* **1996**, *144*, 421–424. [CrossRef]
12. Liu, I.; Agresti, A. The analysis of ordered categorical data: An overview and a survey of recent developments. *Test* **2005**, *14*, 1–73. [CrossRef]
13. McCullagh, P. Regression models for ordinal data. *J. R. Stat. Soc.* **1980**, *42*, 109–142. [CrossRef]
14. McCullagh, P.; Nelder, J.A. *Generalized Linear Models*, 2nd ed.; Chapman & Hall: London, UK, 1989.
15. Anderson, J.A. Regression and Ordered Categorical Variables. *J. R. Stat. Soc. Ser. B* **1984**, *46*, 1–30. [CrossRef]
16. Greenland, S. Alternative models for ordinal logistic regression. *Stat. Med.* **1994**, *13*, 1665–1677. [CrossRef] [PubMed]
17. Ananth, C.V.; Kleinbaum, D.G. Regression models for ordinal responses: A review of methods and applications. *Int. J. Epidemiol.* **1997**, *26*, 1323–1333. [CrossRef]
18. Johnson, T.R. Discrete choice models for ordinal response variables: A generalization of the stereotype model. *Psychometrika* **2007**, *72*, 489–504. [CrossRef]
19. Fullerton, A.S. A conceptual framework for ordered logistic regression models. *Sociol. Methods Res.* **2009**, *38*, 306–347. [CrossRef]
20. Liu, X. Fitting stereotype logistic regression models for ordinal response variables in educational research (Stata). *J. Mod. Appl. Stat. Methods* **2014**, *13*, 31. [CrossRef]
21. Fernández, D.; Pledger, S. Categorising count data into ordinal responses with application to ecological communities. *J. Agric. Biol. Environ. Stat.* **2016**, *21*, 348–362. [CrossRef]
22. Williams, A.A.; Archer, K.J. Elastic Net Constrained Stereotype Logit Model for Ordered Categorical Data. *Biom. Biostat. Int. J.* **2015**, *2*, 00049. [CrossRef]
23. Spiess, M.; Fernández, D.; Nguyen, T.; Liu, I. Generalized estimating equations to estimate the ordered stereotype logit model for panel data. *Stat. Med.* **2020**, *39*, 1919–1940. [CrossRef]
24. Kuss, O. On the estimation of the stereotype regression model. *Comput. Stat. Data Anal.* **2006**, *50*, 1877–1890. [CrossRef]
25. Holtbrugge, W.; Schumacher, M. A comparison of regression models for the analysis of ordered categorical data. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **1991**, *40*, 249–259. [CrossRef]
26. Preedalikit, K.; Liu, I.; Hirose, Y.; Sibanda, N.; Fernández, D. Joint modeling of survival and longitudinal ordered data using a semiparametric approach. *Aust. N. Z. J. Stat.* **2016**, *58*, 153–172. [CrossRef]
27. Feldmann, U.; König, J. Ordinal classification in medical prognosis. *Methods Inf. Med.* **2002**, *41*, 154–163. [PubMed]
28. Ahn, J.; Mukherjee, B.; Gruber, S.B.; Sinha, S. Missing exposure data in stereotype regression model: Application to matched case–control study with disease subclassification. *Biometrics* **2011**, *67*, 546–558. [CrossRef]
29. Lunt, M.; Unit, A. Stereotype ordinal regression. *Stata Tech. Bull.* **2001**, *61*, 1–28.
30. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020.
31. Turner, H.; Firth, D. *Generalized Nonlinear Models in R: An Overview of the gnm Package*; Technical Report; ESRC National Centre for Research Methods: Southampton, UK, 2007.
32. Yee, T.W. The VGAM Package. *R News* **2008**, *8*, 28–39.
33. Archer, K.J.; Hou, J.; Zhou, Q.; Ferber, K.; Layne, J.G.; Gentry, A.E. ordinalgmifs: An R package for ordinal regression in high-dimensional data settings. *Cancer Inform.* **2014**, *13*, 187. [CrossRef]
34. McMillan, L.; Fernandez, D.; Cui, Y.; Matechou, E. Clustord R Package. Available online: https://github.com/vuw-clustering/clustord (accessed on 29 March 2022).
35. Fernández, D.; Arnold, R.; Pledger, S. Mixture-based clustering for the ordered stereotype model. *Comput. Stat. Data Anal.* **2016**, *93*, 46–75. [CrossRef]
36. Fagerland, M.W.; Hosmer, D.W. A goodness-of-fit test for the proportional odds regression model. *Stat. Med.* **2013**, *32*, 2235–2249. [CrossRef]
37. Hosmer, D.W.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; John Wiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
38. Pulkstenis, E.; Robinson, T.J. Goodness-of-fit tests for ordinal response regression models. *Stat. Med.* **2004**, *23*, 999–1014. [CrossRef] [PubMed]
39. Lin, K.C.; Chen, Y.J. Assessing ordinal logistic regression models via nonparametric smoothing. *Commun. Stat.-Methods* **2008**, *37*, 917–930. [CrossRef]
40. Liu, I.; Mukherjee, B.; Suesse, T.; Sparrow, D.; Park, S.K. Graphical diagnostics to check model misspecification for the proportional odds regression model. *Stat. Med.* **2009**, *28*, 412–429. [CrossRef] [PubMed]
41. Lipsitz, S.R.; Fitzmaurice, G.M.; Molenberghs, G. Goodness-of-fit tests for ordinal response regression models. *Appl. Stat.* **1996**, *45*, 175–190. [CrossRef]
42. Li, C.; Shepherd, B.E. A new residual for ordinal outcomes. *Biometrika* **2012**, *99*, 473–480. [CrossRef] [PubMed]
43. Liu, D.; Li, S.; Yu, Y.; Moustaki, I. Assessing partial association between ordinal variables: Quantification, visualization, and hypothesis testing. *J. Am. Stat. Assoc.* **2020**, *116*, 955–968. [CrossRef]
44. Fernández, D.; Liu, I. A goodness-of-fit test for the ordered stereotype model. *Stat. Med.* **2016**, *35*, 4660–4696. [CrossRef]
45. Fernández, D.; Liu, I.; Arnold, R.; Nguyen, T.; Spiess, M. Model-based goodness-of-fit tests for the ordered stereotype model. *Stat. Methods Med Res.* **2020**, *29*, 1527–1541. [CrossRef]

46. Liu, I.; Fernández, D. Discussion on "Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer–Lemeshow test" by Giovanni Nattino, Michael L. Pennell, and Stanley Lemeshow. *Biometrics* **2020**, *76*, 564–568. [CrossRef]
47. Nattino, G.; Pennell, M.L.; Lemeshow, S. Assessing the goodness of fit of logistic regression models in large samples: A modification of the Hosmer–Lemeshow test. *Biometrics* **2020**, *76*, 549–560. [CrossRef]
48. Kuss, O. Modelling physicians' recommendations for optimal medical care by random effects stereotype regression. In Proceedings of the 18th International Workshop on Statistical Modelling, Leuven, Belgium, 7–11 July 2003; Citeseer: University Park, PA, USA, 2003; p. 245.
49. Liang, K.Y.; Zeger, S.L. Longitudinal data analysis using generalized linear models. *Biometrika* **1986**, *73*, 13–22. [CrossRef]
50. Touloumis, A.; Agresti, A.; Kateri, M. GEE for multinomial responses using a local odds ratios parameterization. *Biometrics* **2013**, *69*, 633–640. [CrossRef] [PubMed]
51. Lipsitz, S.R.; Kim, K.; Zhao, L. Analysis of repeated categorical data using generalized estimating equations. *Stat. Med.* **1994**, *13*, 1149–1163. [CrossRef] [PubMed]
52. Bombardier, C.; Ware, J.; Russell, I.J.; Larson, M.; Chalmers, A.; Read, J.L.; Arnold, W.; Bennett, R.; Caldwell, J.; Hench, P.K.; et al. Auranofin therapy and quality of life in patients with rheumatoid arthritis. Results of a multicenter trial. *Am. J. Med.* **1986**, *81*, 565–578. [CrossRef]
53. Touloumis, A. R package multgee: A generalized estimating equations solver for multinomial responses. *J. Stat. Softw.* **2015**, *64*, 1–14. [CrossRef]
54. Goodman, L.A. The analysis of cross-classified data having ordered and/or unordered categories: Association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *Ann. Stat.* **1985**, *13*, 10–69. [CrossRef]
55. Yee, T.W.; Hastie, T.J. Reduced-rank vector generalized linear models. *Stat. Model.* **2003**, *3*, 15–41. [CrossRef]
56. Fernández, D.; Liu, I.; Costilla, R.; Gu, P.Y. Assigning scores for ordered categorical responses. *J. Appl. Stat.* **2020**, *47*, 1261–1281. [CrossRef]
57. Tsiatis, A.A. A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **1980**, *67*, 250–251. [CrossRef]
58. Pulkstenis, E.; Robinson, T.J. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Stat. Med.* **2002**, *21*, 79–93. [CrossRef]
59. Archer, K.J.; Lemeshow, S.; Hosmer, D.W. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Comput. Stat. Data Anal.* **2007**, *51*, 4450–4464. [CrossRef]