Statistica Sinica **35** (2025), 411-430
doi:https://doi.org/10.5705/ss.202022.0322

IDENTIFYING THE MOST APPROPRIATE ORDER FOR CATEGORICAL RESPONSES

Tianmeng Wang and Jie Yang*

University of Illinois at Chicago

Abstract: Categorical responses arise naturally within various scientific disciplines. In many circumstances, there is no predetermined order for the response categories, and the response has to be modeled as nominal. In this study, we regard the order of response categories as part of the statistical model, and show that the true order, when it exists, can be selected using likelihood-based model selection criteria. For predictive purposes, a statistical model with a chosen order may outperform models based on nominal responses, even if a true order does not exist. For multinomial logistic models, widely used for categorical responses, we show the existence of theoretically equivalent orders that cannot be differentiated based on likelihood criteria, and determine the connections between their maximum likelihood estimators. We use simulation studies and a real-data analysis to confirm the need and benefits of choosing the most appropriate order for categorical responses.

Key words and phrases: AIC, BIC, categorical data analysis, model selection, multinomial logistic model.

1. Introduction

Categorical responses, in which the measurement scale consists of a set of categories, arise naturally in many scientific disciplines. Examples include the social sciences for measuring attitudes and opinions, health sciences for measuring responses to a medical treatment, behavioral sciences for diagnosing mental illness, ecology for determining primary land use in satellite images, education for measuring student responses, and marketing for determining consumer preferences, among many others (Agresti, 2018). When the response is binary, generalized linear models are widely used (McCullagh and Nelder, 1989; Dobson and Barnett, 2018). When responses have three or more categories, multinomial logistic models are popular (Glonek and McCullagh, 1995; Zocchi and Atkinson, 1999; Bu, Majumdar and Yang, 2020), and include four kinds of logit models: *baseline-category*, *cumulative*, *adjacent-categories*, and *continuation-ratio* logit models.

Following the notation of Bu, Majumdar and Yang (2020), we consider summarized data in the form of $\{(\mathbf{x}_i, \mathbf{Y}_i), i = 1, \ldots, m\}$ from an experiment

*Corresponding author. E-mail: jyang06@uic.edu

or observational study with $d \geq 1$ covariates and $J \geq 3$ response categories, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})^T$, for $i = 1, \ldots, m$, are distinct level combinations of the d covariates, and $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^T$, with Y_{ij} indicating the number of original observations associated with the covariates $\mathbf{x}_i$ and the jth response category, for $j = 1, \ldots, J$. A multinomial logistic model assumes $\mathbf{Y}_i \sim$ Multinomial$(n_i; \pi_{i1}, \ldots, \pi_{iJ})$ independently, with $n_i = \sum_{j=1}^{J} Y_{ij} > 0$ and positive categorical probabilities π_{ij} associated with $\mathbf{x}_i$, for $i = 1, \ldots, m$.

For *nominal* responses, that is, the response categories do not have a natural ordering (Agresti, 2013), baseline-category logit models, also known as multiclass logistic regression models, are commonly used. Following Bu, Majumdar and Yang (2020), the baseline-category logit model with partial proportional odds (*ppo*) can be described in general as

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \eta_{ij} = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta}\ , \tag{1.1}$$

where $\mathbf{h}_j^T(\cdot) = (h_{j1}(\cdot), \ldots, h_{jp_j}(\cdot))$ and $\mathbf{h}_c^T(\cdot) = (h_1(\cdot), \ldots, h_{p_c}(\cdot))$ are known predictor functions, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})^T$ and $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_{p_c})^T$ are unknown regression parameters, for $i = 1, \ldots, m$, $j = 1, \ldots, J-1$. As special cases, $\mathbf{h}_j^T(\mathbf{x}_i) \equiv 1$ leads to a proportional odds (*po*) model that assumes the same parameters for all categories, except the intercepts, and $\mathbf{h}_c^T(\mathbf{x}_i) \equiv 0$ leads to a nonproportional odds (*npo*) model that allows all parameters to change across categories. For additional explanations and examples about ppo, po, and npo models, see Bu, Majumdar and Yang (2020).

In model (1.1), the Jth category is treated as the baseline category. It is well known that the choice of baseline category does not matter, because the resulting models are equivalent (Hastie, Tibshirani and Friedman, 2009, Sec. 4.4). However, the equivalence of choices of baseline categories is true only for npo models. As we show in Section 4.4, for po or general ppo models, those with different baseline categories are not equivalent, and thus the baseline category should be chosen carefully.

The other three logit models assume that the response categories have a natural ordering or a hierarchical structure, and are known as *ordinal* or *hierarchical* models, respectively. Following Bu, Majumdar and Yang (2020), these three logit models with ppo can be written as follows:

$$\log\left(\frac{\pi_{i1} + \cdots + \pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{iJ}}\right) = \eta_{ij} = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta}, \text{ cumulative;} \tag{1.2}$$

$$\log\left(\frac{\pi_{ij}}{\pi_{i,j+1}}\right) = \eta_{ij} = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta}, \text{ adjacent categories;} \tag{1.3}$$

$$\log\left(\frac{\pi_{ij}}{\pi_{i,j+1} + \cdots + \pi_{iJ}}\right) = \eta_{ij} = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta}, \text{ continuation ratio.} \tag{1.4}$$

These are special cases of the multivariate generalized linear models (McCullagh, 1980) or multivariate logistic models (Glonek and McCullagh, 1995).

Note that cumulative logit models have been extended to cumulative link models and ordinal regression models (McCullagh, 1980; Agresti, 2013; Yang, Tong and Mandal, 2017). A baseline-category logit model can be modified with a probit link, and is known as a multinomial probit model (Aitchison and Bennett, 1970; Agresti, 2013; Greene, 2018). Furthermore, the continuation-ratio logit model can be changed with a complementary log-log link (O'Connell, 2006) for data analysis. We focus on multinomial logistic models, because the logit link is the most commonly used.

For some applications, the ordering of the response categories is clear. For example, trauma data (Chuang-Stein and Agresti, 1997; Agresti, 2010; Bu, Majumdar and Yang, 2020) includes $J = 5$ ordinal response categories, namely, `death`, `vegetative state`, `major disability`, `minor disability`, and `good recovery`, known as the Glasgow Outcome Scale (Jennett and Bond, 1975). A cumulative logit model with npo has been recommended for modeling such data (Bu, Majumdar and Yang, 2020).

For some other applications, the ordering is either unknown or difficult to determine. As a motivating example, the police data described in Section 5 contain covariates about individuals killed by the police in the United States for the period 2000 to 2016. The responses have four categories, namely, `shot`, `tasered`, `shot and tasered`, and `other`. To model the responses that are relevant to the police's actions on various covariates of the suspects, one strategy is to treat the response as nominal and use the baseline-category logit model (1.1). Another strategy is to determine an appropriate order for the categories, and then to use one of the other three logit models (1.2), (1.3), and (1.4). Our analysis in Section 5 shows that a continuation-ratio npo model with a chosen order performs best, and that the second strategy is significantly better.

A critical question that needs answering is whether we can identify the true order of the response categories, when it exists. Our answer is summarized as follows. First, we will show in Section 3 that if there is a true order with a true model, it will attain the maximum likelihood asymptotically, so that it can be identified using a likelihood-based model selection technique, such as the AIC or BIC. Second, depending on the type of logit model, some orders are indistinguishable or equivalent, because they attain the same maximum likelihood (see Tbl. 1 for a summary of the equivalence among the orders identified in Sec. 2). Third, depending on the range of covariates or predictors, some order that is not equivalent to the true one may approximate the maximum likelihood so well that it is not numerically distinguishable from the true order (see Sec. 4.3).

In practice, nevertheless, even there is no true order among the response categories, we can still use likelihood-based model selection techniques to choose a working order supported by the data. As such, an ordinal model based on the

Table 1. Equivalence Among Orders of Response Categories.

Logit model	ppo or po	npo
Baseline-category	Same if the baseline is unchanged (Theorem 2)	All orders are the same (Theorem 5)
Cumulative	Same as its reversed order (Theorem 3)	Same as its reversed order (Theorem 3)
Adjacent-categories	Same as its reversed order (Theorem 4)	All orders are the same (Theorem 6)
Continuation-ratio	All orders are distinguishable (Section 4.1)	Same if switching last two (Theorem 7)

working order will outperform nominal models in terms of prediction accuracy (Sec. 4.4 and 4.5). We provide a real-data example in Section 5 that shows how to reduce the prediction errors significantly based on the working order. Overall, we suggest that practitioners view identifying the most appropriate order of response categories as part of the model selection procedure.

2. Equivalence Among Orders of Response Categories

In this section, before investigating which order of the response categories is best, we first answer a more fundamental question that occurs when two different orders lead to the same maximum likelihood. In this case, if one uses the AIC or BIC to select the best model (see Hastie, Tibshirani and Friedman, 2009 for a good review), these two orders are indistinguishable, or equivalent. Such a phenomenon has been observed for some baseline-category models (Hastie, Tibshirani and Friedman, 2009), and here we show that it exists fairly generally in other multinomial logistic models as well (see Tbl. 1 for a summary).

Given the original data $\{(\mathbf{x}_i, \mathbf{Y}_i), i = 1, \ldots, m\}$, $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^T$ consists of the counts of observations falling into the response categories in the original order or labels $\{1, \ldots, J\}$. If we consider a regression model with a different order $\{\sigma(1), \ldots, \sigma(J)\}$ of the response categories, where $\sigma : \{1, \ldots, J\} \to \{1, \ldots, J\}$ is a permutation, this is equivalent to fitting the model using the permuted data $\{(\mathbf{x}_i, \mathbf{Y}_i^\sigma), i = 1, \ldots, m\}$, where $\mathbf{Y}_i^\sigma = (Y_{i\sigma(1)}, \ldots, Y_{i\sigma(J)})^T$. We denote $\mathcal{P}$ as the collection of all permutations on $\{1, \ldots, J\}$. Then, each permutation $\sigma \in \mathcal{P}$ represents an order of the response categories, also denoted by σ.

Now, we consider two orders or permutations $\sigma_1, \sigma_2 \in \mathcal{P}$. For $i = 1, 2$, we denote $l_i(\boldsymbol{\theta})$ as the likelihood function with order σ_i or $\{\sigma_i(1), \ldots, \sigma_i(J)\}$. We say that σ_1 and σ_2 are *equivalent*, denoted as $\sigma_1 \sim \sigma_2$, if $\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l_1(\boldsymbol{\theta}) = \max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} l_2(\boldsymbol{\theta})$, where $\boldsymbol{\Theta}$ is the parameter space, that is, the set of all feasible $\boldsymbol{\theta}$. It is straightforward that "$\sim$" is an *equivalence relation* among the permutations in $\mathcal{P}$. That is, $\sigma \sim \sigma$ for all σ; $\sigma_1 \sim \sigma_2$ if and only if $\sigma_2 \sim \sigma_1$; and $\sigma_1 \sim \sigma_2$ and

$\sigma_2 \sim \sigma_3$ imply $\sigma_1 \sim \sigma_3$ (Wallace, 1998, Sec. 1.4).

2.1. Partial proportional odds (ppo) models

According to Bu, Majumdar and Yang (2020) (see also Glonek and McCullagh, 1995 and Zocchi and Atkinson, 1999), all four logit models (1.1) to (1.4) with ppo (that is, the most general case) can be rewritten in a unified form

$$\mathbf{C}^T \log(\mathbf{L}\boldsymbol{\pi}_i) = \boldsymbol{\eta}_i = \mathbf{X}_i\boldsymbol{\theta}, \quad i = 1, \ldots, m, \tag{2.1}$$

where $\mathbf{C}$ is a $J \times (2J-1)$ constant matrix, $\mathbf{L}$ is a $(2J-1) \times J$ constant matrix, depending on the type of logit model, $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iJ})^T$ are category probabilities at $\mathbf{x}_i$ satisfying $\sum_{j=1}^{J} \pi_{ij} = 1$, $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{iJ})^T$ are the linear predictors, $\mathbf{X}_i$ is the model matrix consisting of $\mathbf{h}_j(\mathbf{x}_i)$ and $\mathbf{h}_c(\mathbf{x}_i)$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T, \boldsymbol{\zeta}^T)^T$ consists of $p = p_1 + \cdots + p_{J-1} + p_c$ regression parameters. See Bu, Majumdar and Yang (2020) for more details and examples.

According to Theorem 5.1 in Bu, Majumdar and Yang (2020), for cumulative logit models, the parameter space $\boldsymbol{\Theta} = \{\boldsymbol{\theta} \in R^p \mid \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j < \mathbf{h}_{j+1}^T(\mathbf{x}_i)\boldsymbol{\beta}_{j+1}, \text{ for } j = 1, \ldots, J-2, i = 1, \ldots, m\}$ depends on the range of covariates. For the other three logit models, $\boldsymbol{\Theta}$ is typically R^p itself. Apparently, neither the parameter space $\boldsymbol{\Theta}$ nor the model (2.1) is affected by a permutation of the data $\mathbf{Y}_i$.

By reorganizing the formulae in Section S.11 of the Supplementary Material of Bu, Majumdar and Yang (2020), we write the category probabilities π_{ij} as explicit functions of η_{ij} (and thus of $\boldsymbol{\theta}$) in Lemma 1. To simplify the notation, we denote $\rho_{ij} = \text{logit}^{-1}(\eta_{ij}) = e^{\eta_{ij}}/(1+e^{\eta_{ij}})$, and thus $\rho_{ij}/(1-\rho_{ij}) = e^{\eta_{ij}}$, for $j = 1, \ldots, J-1$, and $\rho_{i0} \equiv 0$, for $i = 1, \ldots, m$.

Lemma 1. *For the four logit models* (1.1) *to* (1.4),

$$\pi_{ij} = \begin{cases} \dfrac{\rho_{ij}/(1-\rho_{ij})}{1+\sum_{l=1}^{J-1}\rho_{il}/(1-\rho_{il})} & \text{, baseline category} \\ \rho_{ij} - \rho_{i,j-1} & \text{, cumulative} \\ \dfrac{\prod_{l=j}^{J-1}\rho_{il}/(1-\rho_{il})}{1+\sum_{s=1}^{J-1}\prod_{l=s}^{J-1}\rho_{il}/(1-\rho_{il})} & \text{, adjacent categories} \\ \prod_{l=0}^{j-1}(1-\rho_{il})\cdot\rho_{ij} & \text{, continuation ratio,} \end{cases} \tag{2.2}$$

for $i = 1, \ldots, m$ *and* $j = 1, \ldots, J-1$. *In addition,*

$$\pi_{iJ} = \begin{cases} \left\{1 + \sum_{l=1}^{J-1} \frac{\rho_{il}}{1-\rho_{il}}\right\}^{-1} & \text{, baseline category} \\ 1 - \rho_{i,J-1} & \text{, cumulative} \\ \left\{1 + \sum_{s=1}^{J-1} \prod_{l=s}^{J-1} \frac{\rho_{il}}{1-\rho_{il}}\right\}^{-1} & \text{, adjacent categories} \\ \prod_{l=1}^{J-1} (1-\rho_{il}) & \text{, continuation ratio,} \end{cases} \tag{2.3}$$

for $i = 1, \ldots, m$.

Because $\eta_{ij} = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j + \mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta}$, Lemma 1 indicates that π_{ij} are functions of $\mathbf{x}_i$ and $\boldsymbol{\theta}$, and do not depend on $\mathbf{Y}_i$ or $\mathbf{Y}_i^\sigma$, which is true for general multinomial logit models (2.1). The following theorem provides a sufficient condition for $\sigma_1 \sim \sigma_2$.

Theorem 1. *Consider the multinomial logit model* (2.1) *with independent observations and two permutations,* $\sigma_1, \sigma_2 \in \mathcal{P}$. *Suppose for any* $\boldsymbol{\theta}_1 \in \boldsymbol{\Theta}$, *there exists a* $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$, *and vice versa, such that,*

$$\pi_{i\sigma_1^{-1}(j)}(\boldsymbol{\theta}_1) = \pi_{i\sigma_2^{-1}(j)}(\boldsymbol{\theta}_2), \tag{2.4}$$

for all $i = 1, \ldots, m$ *and* $j = 1, \ldots, J$. *Then,* $\sigma_1 \sim \sigma_2$. *Furthermore,* $\sigma\sigma_1 \sim \sigma\sigma_2$, *for any* $\sigma \in \mathcal{P}$.

Here, $\sigma\sigma_1$ in Theorem 1 represents the composition of σ and σ_1. That is, $\sigma\sigma_1(j) = \sigma(\sigma_1(j))$, for all j. Using this notation, $(\sigma\sigma_1)^{-1} = \sigma_1^{-1}\sigma^{-1}$.

The proof of Theorem 1 is relegated to the Supplementary Material.

Theorem 2. *Consider the baseline-category logit model* (1.1) *with ppo. Suppose* $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i)$, *for all* $i = 1, \ldots, m$. *Then, all orders of response categories that keep* J *invariant are equivalent.*

Theorem 2 includes baseline-category logit models with po, because $\mathbf{h}_j(\mathbf{x}_i) \equiv 1$ for po models. It actually includes many npo or ppo models used in practice, where $\mathbf{h}_1 = \cdots = \mathbf{h}_{J-1}$. For example, main-effects models that assume $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i) = (1, \mathbf{x}_i^T)^T$ are widely used.

Theorem 2 also implies that if $\sigma(J) \neq J$, then σ may not be equivalent to the original order id, or the identity permutation. We provide such a numerical example in Section 4.4.

Theorem 3. *Consider the cumulative logit model* (1.2) *with ppo. Suppose* $\mathbf{h}_j(\mathbf{x}_i) = \mathbf{h}_{J-j}(\mathbf{x}_i)$, *for all* $i = 1, \ldots, m$ *and* $j = 1, \ldots, J-1$. *Then, any order* σ_1 *is equivalent to its reverse order* σ_2, *which satisfies* $\sigma_2(j) = \sigma_1(J+1-j)$, *for* $j = 1, \ldots, J$.

Theorem 3 includes cumulative logit models with po, because $\mathbf{h}_j(\mathbf{x}_i) \equiv 1$. It also includes many npo or ppo models used in practice that satisfy $\mathbf{h}_1 = \cdots = \mathbf{h}_{J-1}$.

Theorem 4. *Consider the adjacent-categories logit model* (1.3) *with ppo. Suppose* $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i)$*, for all* $i = 1, \ldots, m$*. Then, any order* σ_1 *is equivalent to its reverse order* σ_2 *, which satisfies* $\sigma_2(j) = \sigma_1(J + 1 - j)$*, for all* $j = 1, \ldots, J$*.*

2.2. Nonproportional odds (npo) models

By removing the item $\mathbf{h}_c^T(\mathbf{x}_i)\boldsymbol{\zeta}$ from (1.1) to (1.4), we obtain explicit forms of the four logit models with npo. For npo models, $\boldsymbol{\theta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{J-1}^T)^T$, $p = p_1 + \cdots + p_{J-1}$, and $\eta_{ij} = \mathbf{h}_j^T(\mathbf{x}_i)\boldsymbol{\beta}_j$, for $i = 1, \ldots, m$ and $j = 1, \ldots, J - 1$. Compared with po models, npo models involve more regression parameters, and thus are more flexible. For more details about multinomial logistic models with npo, please see, for example, Section S.8 in the Supplementary Material of Bu, Majumdar and Yang (2020).

Theorem 5. *Consider the baseline-category logit model with npo. Suppose* $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i)$*, for all* $i = 1, \ldots, m$*. Then, all orders of response categories are equivalent.*

Theorem 5 confirms that the choice of baseline category does not matter for multiclass logistic regression models (Hastie, Tibshirani and Friedman, 2009). What is new here is the explicit correspondence between $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ provided in the proof of Theorem 5. Based on the correspondence, if we obtain the maximum likelihood estimate (MLE) for $\boldsymbol{\theta}_1$, we can easily derive the MLE for $\boldsymbol{\theta}_2$ explicitly, without running another optimization.

Theorem 6. *Consider the adjacent-categories logit model with npo. Suppose* $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i)$*, for all* $i = 1, \ldots, m$*. Then, all orders of response categories are equivalent.*

The result of Theorem 6 is truly surprising. The order of the response categories in an ordinal model does not matter! The transformation (S.1) and its inverse (S.3) in the proof of Theorem 6 in the Supplementary Material are not trivial either.

Theorem 7. *For the continuation-ratio logit model with npo,* $\sigma_1 \sim \sigma_2$ *if* $\sigma_2 = \sigma_1(J - 1, J)$*, where* $(J - 1, J)$ *is a transposition that switches* $J - 1$ *and* J*.*

3. Asymptotic Optimality of True Order

In this section, we discuss the optimality of the true order of the response categories, when it exists. In short, the model with the true order is asymptotically optimal in terms of an AIC or BIC likelihood-based model selection criterion.

Suppose an experiment is performed under the multinomial logit model (2.1), with predetermined design points $\mathbf{x}_1, \ldots, \mathbf{x}_m$, the true parameter values $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta} \subseteq R^p$, and the true order $\sigma_0 \in \mathcal{P}$ of the response categories. Recall that the

original experiment assigns n_i subjects to $\mathbf{x}_i$, with the total number of subjects $n = \sum_{i=1}^m n_i$. To avoid trivial cases, we assume $n_i > 0$ for each i (otherwise, we may delete any $\mathbf{x}_i$ with $n_i = 0$).

In order to consider the asymptotic properties of the parameter and order estimators, we consider independent and identically distributed (i.i.d.) observations (X_l, Y_l), for $l = 1, \ldots, N$, generated as follows: (i) $X_1, \ldots, X_N$ are i.i.d. from a discrete distribution taking values in $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ with probabilities $n_1/n, \ldots, n_m/n$, respectively; (ii) given $X_l = \mathbf{x}_i$, Y_l follows Multinomial$(1; \pi_{i\sigma_0^{-1}(1)}(\boldsymbol{\theta}_0), \ldots, \pi_{i\sigma_0^{-1}(J)}(\boldsymbol{\theta}_0))$, that is, Y_l takes values in $\{1, \ldots, J\}$ with probabilities $\pi_{i\sigma_0^{-1}(1)}(\boldsymbol{\theta}_0), \ldots, \pi_{i\sigma_0^{-1}(J)}(\boldsymbol{\theta}_0)$, respectively. The summarized data can still be denoted as $\{(\mathbf{x}_i, \mathbf{Y}_i), i = 1, \ldots, m\}$, where

$$\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iJ})^T \sim \text{Multinomial}\left(N_i; \pi_{i\sigma_0^{-1}(1)}(\boldsymbol{\theta}_0), \ldots, \pi_{i\sigma_0^{-1}(J)}(\boldsymbol{\theta}_0)\right),$$

and $N_i = \sum_{j=1}^J Y_{ij}$ is the total number of subjects assigned to $\mathbf{x}_i$.

Given the count data Y_{ij}, the log-likelihood function under the multinomial logit model (2.1) with parameters $\boldsymbol{\theta}$ and a permutation $\sigma \in \mathcal{P}$ applied to Y_{ij} to determine the true order is

$$\begin{aligned} l(\boldsymbol{\theta}, \sigma) &= \sum_{i=1}^m \sum_{j=1}^J Y_{i\sigma(j)} \log \pi_{ij}(\boldsymbol{\theta}) + \sum_{i=1}^m \log(N_i!) - \sum_{i=1}^m \sum_{j=1}^J \log(Y_{i\sigma(j)}!) \\ &= \sum_{i=1}^m \sum_{j=1}^J Y_{ij} \log \pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta}) + \sum_{i=1}^m \log(N_i!) - \sum_{i=1}^m \sum_{j=1}^J \log(Y_{ij}!) \\ &= \sum_{i=1}^m \sum_{j=1}^J Y_{ij} \log \pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta}) + \text{constant}, \end{aligned}$$

Then, the MLE $(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ that maximizes $l(\boldsymbol{\theta}, \sigma)$ maximizes

$$l_N(\boldsymbol{\theta}, \sigma) = \sum_{i=1}^m \sum_{j=1}^J Y_{ij} \log \pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta})$$

as well.

Lemma 2. *If $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ and $\hat{\sigma} \in \mathcal{P}$ satisfy $\pi_{i\hat{\sigma}^{-1}(j)}(\hat{\boldsymbol{\theta}}) = Y_{ij}/N_i$ for all i and j, then $(\hat{\boldsymbol{\theta}}, \hat{\sigma})$ must be an MLE.*

To explore the asymptotic properties of log-likelihood and MLEs, we denote $l_0 = \sum_{i=1}^m \sum_{j=1}^J n_i \pi_{ij}(\boldsymbol{\theta}_0) \log \pi_{ij}(\boldsymbol{\theta}_0)/n \in (-\infty, 0)$, which is a finite constant. We further denote $n_0 = \min\{n_1, \ldots, n_m\} \geq 1$ and $\pi_0 = \min\{\pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0), i = 1, \ldots, m; j = 1, \ldots, J\} \in (0, 1)$.

Lemma 3. *As $N \to \infty$, $N^{-1} l_N(\boldsymbol{\theta}_0, \sigma_0) \to l_0 < 0$ almost surely.*

Let $(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N)$ denote an MLE that maximizes $l_N(\boldsymbol{\theta}, \sigma)$ or, equivalently, $l(\boldsymbol{\theta}, \sigma)$. The following theorem indicates that the true values $(\boldsymbol{\theta}_0, \sigma_0)$ asymptotically maximize $l_N(\boldsymbol{\theta}, \sigma)$ as well. That is, the true parameter values $\boldsymbol{\theta}_0$ and the true order σ_0 are asymptotically optimal in terms of the likelihood principle.

Theorem 8. *As* $N \to \infty$, $N^{-1}|l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) - l_N(\boldsymbol{\theta}_0, \sigma_0)| \to 0$ *almost surely. Furthermore,* $N^{-1} l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) \to l_0$ *and* $l_N(\boldsymbol{\theta}_0, \sigma_0)/l_N(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) \to 1$ *almost surely as well.*

Following Burnham and Anderson (2004), we define

$$\begin{aligned} \mathrm{AIC}(\boldsymbol{\theta}, \sigma) &= -2l(\boldsymbol{\theta}, \sigma) + 2p, \\ \mathrm{BIC}(\boldsymbol{\theta}, \sigma) &= -2l(\boldsymbol{\theta}, \sigma) + (\log N)p. \end{aligned}$$

Then, the usual $\mathrm{AIC} = \mathrm{AIC}(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) \leq \mathrm{AIC}(\boldsymbol{\theta}_0, \sigma_0)$, and the usual $\mathrm{BIC} = \mathrm{BIC}(\hat{\boldsymbol{\theta}}_N, \hat{\sigma}_N) \leq \mathrm{BIC}(\boldsymbol{\theta}_0, \sigma_0)$. As a direct conclusion of Theorem 8, we have the following corollary.

Corollary 1. $N^{-1}|\mathrm{AIC} - \mathrm{AIC}(\boldsymbol{\theta}_0, \sigma_0)| = N^{-1}|\mathrm{BIC} - \mathrm{BIC}(\boldsymbol{\theta}_0, \sigma_0)| \to 0$ *almost surely, as* $N \to \infty$.

Theorem 8 and Corollary 1 confirm that the true parameter and order $(\boldsymbol{\theta}_0, \sigma_0)$ are among the best options asymptotically under likelihood-based model selection criteria. Nevertheless, note that there are two cases where one may not be able to identify the true order σ_0 easily.

Case one: If there exists another order $\sigma \sim \sigma_0$, as discussed in Section 2, then there exists another $\boldsymbol{\theta}$ such that $\pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta}) = \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)$, for all i and j. In this case, $l(\boldsymbol{\theta}, \sigma) = l(\boldsymbol{\theta}_0, \sigma_0)$, and the true order σ_0 with $\boldsymbol{\theta}_0$ is not distinguishable from the order σ with $\boldsymbol{\theta}$.

Case two: In practice, given the set of experimental settings $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$, it is not unlikely, especially when the range of experimental settings is narrow, that there exists a $(\boldsymbol{\theta}, \sigma)$ such that $\pi_{i\sigma^{-1}(j)}(\boldsymbol{\theta}) \approx \pi_{i\sigma_0^{-1}(j)}(\boldsymbol{\theta}_0)$, for all i and j. Then, the order σ with $\boldsymbol{\theta}$ achieves roughly the same likelihood. In this case, the difference between σ_0 and σ could be insignificant, and may not improve unless one increases the sample size N and the range of $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ (e.g., see Sec. 4.1 and 4.3). Compared with po models, npo models have more parameters and are more flexible. As a result, Case two may occur more often for npo models.

In both Case one and Case two, the true order σ_0 is not significantly better than all other orders, even with an increased sample size N. Nevertheless, the results of our simulation studies (Sec. 4) show that there are two situations in which the true order σ_0 can be identified easily. One is with larger absolute

Table 2. Trauma Simulation Study with $N = 802$.

Model	True Order		Best Order	
	AIC_0	Rank	AIC_*	$AIC_0 - AIC_*$
Baseline-category po	96.77	73	96.40	0.37
Cumulative po	94.54	7	93.84	0.70
Cumulative npo	102.15	51	101.70	0.45
Adjacent-categories po	92.66	3	92.08	0.58
Continuation-ratio po	96.25	20	94.38	1.87
Continuation-ratio npo	102.43	21	102.03	0.40

values of the regression coefficients (Sec. 4.2). The other is with a larger range of experimental settings (Sec. 4.3). Both situations reduce the possibility of Case two, but neither can fix Case one.

4. Simulation Studies

4.1. Simulation study based on trauma data

In the example mentioned in Section 1, the trauma data (Chuang-Stein and Agresti, 1997, Tbl. 1; Agresti, 2010, Tbl. 7.6) are the responses of $N = 802$ trauma patients from four treatment groups under different dose levels. The response categories (GOS) have a clear order, namely: (1) `death`, (2) `vegetative state`, (3) `major disability`, (4) `minor disability`, and (5) `good recovery`. That is, the original order $\sigma_0 = \text{id}$ is the true order. In this case, $J = 5$. For illustration purposes, the only predictor x is chosen as the dose level, taking values in $\{1, 2, 3, 4\}$.

In this section, we present simulation studies to explore whether we can identify the true order of the response categories under different multinomial logit models. For each model, for example, the baseline-category logit model with po (see *baseline-category po* in Tbl. 2), (i) we fit the model using the original data $(Y_{ij})_{ij}$ against the dose level $x_i = i$ to obtain estimated parameter values $\hat{\boldsymbol{\theta}}_o$; (ii) we simulate a new data set $(Y'_{ij})_{ij}$ with total sample size N using the model with regression coefficients $\hat{\boldsymbol{\theta}}_o$; that is, for the simulated data, we have a true order $\sigma_0 = \text{id}$ and true parameter values $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}_o$; (iii) for each possible order $\sigma \in \mathcal{P}$, we fit the model using permuted simulated data $(Y'_{i\sigma(j)})_{ij}$ and calculate the corresponding AIC value; and (iv) we check the difference between the AIC at σ_0 (denoted as AIC_0) and the AIC at $\hat{\sigma}$ (denoted as AIC_*) that minimizes the AIC value, as well as the rank of AIC_0 among all orders. Ideally, we have $AIC_0 - AIC_* = 0$ with rank 1 out of $5! = 120$; that is, the original order achieves the smallest AIC value. Nevertheless, in practice, the result depends on the sample size N and the set of experimental settings.

Table 3. Trauma Simulation Study with $N = 8020$.

Model	True Order		Best Order	
	AIC_0	Rank	AIC_*	$AIC_0 - AIC_*$
Baseline-category po	133.19	1	133.19	0
Cumulative po	130.52	1	130.52	0
Cumulative npo	135.55	1	135.55	0
Adjacent-categories po	130.31	1	130.31	0
Continuation-ratio po	130.54	1	130.54	0
Continuation-ratio npo	141.51	21	139.96	1.55

In Table 2, we list the simulation results with $N = 802$, the original sample size. According to Theorems 5 and 6, all orders under a baseline-category npo model or an adjacent-categories npo model are indistinguishable in terms of the AIC. Thus, we omit these two models, and list the other six commonly used multinomial logit models. In Table 2, the true order is not evident under any of the multinomial logit models; that is, the rank is not one, or the AIC value is not the smallest. To check whether the true order can be regarded approximately as the best one, or whether the difference between the true order and the best order is significant, we denote $\Delta = AIC_0 - AIC_*$, the difference in terms of their AIC values. According to Burnham and Anderson (2004), $\Delta \leq 2$ suggests that the true order is considered substantially the best; $4 \leq \Delta \leq 7$ indicates that the true order is considerably less than optimal; and $\Delta > 10$ suggests the true order is essentially worse than the best one. Because no AIC difference in the last column of Table 2 is greater than two, we conclude that the AIC differences between the true order and the best order are not significant. Note that because the number of parameters does not change across orders, using the BIC is equivalent to using the AIC here.

The simulation results (not listed in Tbl. 2) also show that all 120 orders under the continuation-ratio po model lead to distinct AIC values. That is, all orders are distinguishable, or no two orders are equivalent, supporting the corresponding statement in Table 1.

In Table 3, we increase the sample size N to 8020, 10 times as large as the original one, to numerically check the asymptotic optimality of the true order. That is, the new data $(Y'_{i1}, \ldots, Y'_{iJ}) \sim \text{Multinomial}(10n_i; \pi_{i1}(\hat{\boldsymbol{\theta}}_o), \ldots, \pi_{iJ}(\hat{\boldsymbol{\theta}}_o))$, where $\hat{\boldsymbol{\theta}}_o$ is fitted from the original data $(Y_{ij})_{ij}$ and $n_i = \sum_{j=1}^{J} Y_{ij}$. Clearly, all models except the continuation-ratio npo model perform best in terms of the true order of the response categories. This confirms our conclusion in Section 3 that the true order is asymptotically optimal. In this case, the continuation-ratio npo model behaves differently. With 10 times the original sample size, the true order still ranks 21st, with an even bigger AIC difference 1.55 (not statistically significant either). If we further increase the sample size to $N = 40 \times 802$, the true

Table 4. Best Order and Treatment Effects

Treatment effect ζ	AIC (true)	Rank	AIC (best)	AIC (3rd best when true=best)
-0.1755	94.54	7	93.84	-
-0.3510	98.93	3	98.23	-
0.3510	90.84	1	90.84	91.61
-0.7020	85.17	1	85.17	86.58

order ranks third with an AIC difference 0.05. Actually, even for a very large N, there are still six orders among the top tier, the true order, transpositions $(3, 4)$ and $(3, 5)$, and their equivalent orders (Thm. 7). This simulation study provides a numerical example for Case two described in Section 3.

4.2. True order and treatment effects

In Section 3, we noted that larger absolute values of the regression coefficients may make the true order easier to identify. In this section, we illustrate such a scenario using a cumulative po model, which has fewer coefficients and is easier to modify.

The cumulative po model for the trauma data consists of $p = 5$ parameters $\boldsymbol{\theta} = (\beta_1, \beta_2, \beta_3, \beta_4, \zeta)^T$, where ζ represents the treatment effect of the dose level x. By fitting the model using the original data, we obtain the estimated parameters $\hat{\boldsymbol{\theta}} = (-0.7192, -0.3186, 0.6916, 2.057, -0.1755)^T$. Similarly to Section 4.1, we treat $\sigma_0 = \text{id}$ and $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$ as the true values, and simulate a new data set $(Y'_{ij})_{ij}$ with the original sample size $N = 802$. We then check for each possible order σ using the simulated data set. In contrast to Section 4.1, we run four simulation studies, with $\zeta = -0.1755, -0.3510, 0.3510, -0.7020$, respectively, to observe how the magnitude of the treatment effects affects the differences between the orders.

From Table 4, the true order tends to be the best order as $|\zeta|$ increases. For $\zeta = 0.3510$ and -0.7020, the true order attains the best or minimum AIC value. According to Theorem 3 and its proof, the reversed order with parameters $\boldsymbol{\theta}_2 = (-\beta_4, -\beta_3, -\beta_2, -\beta_1, -\zeta)^T$ achieves the same AIC value, which can be viewed as the second best result. From $\zeta = 0.3510$ to $\zeta = -0.7020$, the AIC difference between the true order and the third best order increases from $91.61 - 90.84 = 0.77$ to $86.58 - 85.17 = 1.41$, indicating that a larger treatment effect might make the true order easier to identify.

4.3. True order and experimental settings

In this section, we discuss how the experimental setting $\mathbf{x}_i$ affects the identification of the true order. We use the trauma data under the continuation-ratio npo model as an example, because this is a difficult case in which to identify the true order, according to the results shown in Table 3.

Table 5. Data from Baseline-category po Model with Baseline Category $J = 4$.

$x_i = i$	Y_{i1}	Y_{i2}	Y_{i3}	Y_{i4}	n_i
1	22	33	10	35	100
2	31	40	14	15	100
3	23	43	22	12	100
4	27	49	18	6	100

We explore two ways of changing the set $\{x_i, i = 1, \ldots, m\} = \{1, 2, 3, 4\}$ in the trauma data. In the first, we increase the range to $X_A = \{1, 2, \ldots, 16\}$. In the second, we make finer changes to the experiment settings and obtain $X_B = \{1, 1.25, 1.50, \ldots, 3.75, 4\}$, the range of which is still $[1, 4]$. In both methods, the number m of experimental settings increases significantly.

Similarly to Section 4.1, we treat $\sigma_0 = \mathrm{id}$ and $\boldsymbol{\theta}_0 = \hat{\boldsymbol{\theta}}$, estimated for the continuation-ratio npo model from the original data, as the true values. In this section, we simulate 100 new data sets $(Y_{ij}^{(b)})_{ij}$ independently, with $N = 100 \times 210m$ and $b = 1, \ldots, 100$, using the continuation-ratio npo model with σ_0 and $\boldsymbol{\theta}_0$. For each $b = 1, \ldots, 100$ and each $\sigma \in \mathcal{P}$, we fit the continuation-ratio npo model using the permuted data $(Y_{i\sigma(j)}^{(b)})_{ij}$ against the dose level x_i. The corresponding AIC values are denoted as $\mathrm{AIC}_\sigma^{(b)}$, for $b = 1, \ldots, 100$. To compare the true order σ_0 with each σ of the other 119 orders, we run a one-sided paired t-test on $(\mathrm{AIC}_{\sigma_0}^{(b)})_b$ and $(\mathrm{AIC}_\sigma^{(b)})_b$. A significant p-value indicates that the AIC value associated with σ_0 is significantly smaller than the AIC value of σ.

Under the first scenario X_A, 118 out of 119 p-values are statistically significant, indicating that the true order is significantly better than all other orders, except the equivalent one listed by Theorem 7. That is, an increased range may make the true order easier to identify.

Under the second scenario X_B, there are still three p-values greater than 0.05. Further tests indicate that these four orders, including the true order, are indistinguishable. That is, increasing m while maintaining the range of x_i may not be an efficient way to improve the identifiability of the true order.

4.4. Choice of baseline category

In this section, we use cross-validation to show that the choice of baseline category makes a difference for baseline-category po models.

Table 5 provides simulated data from a baseline-category po model, with the fourth category as the true baseline category. The parameters used for simulating the data are $\boldsymbol{\theta} = (\beta_1, \beta_2, \beta_3, \zeta)^T = (-0.8, -0.3, -1.0, 0.5)^T$. There are $n = 400$ observations in total. We randomly split the observations into two parts, 267 as training data and 133 as testing data. We repeat the random partition 100 times. For each random partition and each order $\sigma \in \mathcal{P}$, (1) we use the training data to fit the baseline-category po model with order σ (similarly to Sec. 4.1), and

denote the fitted model as Model σ; (2) we predict the labels of the responses of the testing data using Model σ, and use the cross-entropy loss (Hastie, Tibshirani and Friedman, 2009) to measure the prediction errors. In this way, we have 100 prediction errors (cross-entropy loss) for each σ. For any two orders, we can run a one-sided paired t-test to check whether one order's prediction error is significantly lower than that of the other (similarly to Sec. 4.3).

We conclude the following from this cross-validation study: (i) all orders that share the same baseline category have the same cross-entropy loss, indicating that they are indistinguishable in terms of their prediction errors, and supporting the results of Theorem 2; (ii) supported by the pairwise t-tests, the orders with the true baseline ($J = 4$) have significantly smaller cross-entropy losses than the other orders, with p-values $5.69 \times 10^{-42}, 9.54 \times 10^{-47}$, and 8.99×10^{-51}, respectively, showing that the correct choice of baseline category matters in practice.

4.5. When true order does not exist

In this section, we investigate an order misspecification issue when the true order does not exist. More specifically, for a baseline-category npo model, all orders are equivalent, according to Theorem 5. In other words, there is no true order. Nevertheless, given such a data set, we can still find the best model with the best order, called the *working* order. The simulation study below shows that when the true order does not exist, we have the following results: (1) with moderate sample sizes, a working order with a different model can be selected, but may not be significantly better than the true model in terms of the AIC; and (2) asymptotically, the true model without a true order may be significantly better than any other model with any working order.

In this simulation study, we use the baseline-category npo model fitted from the original trauma data as the true model, and simulate a data set with the set of covariate levels $X_A = \{1, 2, \ldots, 16\}$, which is a more informative experimental setting (Sec. 4.3). For each level $x_i \in X_A$, we simulate $n_A = 200$ observations, with the total sample size $N = 16n_A = 3200$. For the simulated data with $N = 3200$, the best model according to the AIC is a continuation-ratio npo model, with AIC = 321.31, at its best (working) order {`death`, `major disability`, `vegetative state`, `minor disability`, `good recovery`}, whereas the AIC value of the true model is 322.48. In other words, the best model with the working order is not significantly better than the true model. If we increase the sample size N to $3{,}200 \times 100$, the continuation-ratio npo model is still the best model, with AIC = 604.93 at the best order {`vegetative state`, `death`, `minor disability`, `major disability`, `good recovery`}, which again is not significantly better than the true model, with AIC = 606.75. If we further increase N to $3{,}200 \times 10{,}000$, the true model becomes the best model, with AIC = 914.80, and is significantly better than the continuation-ratio npo model, with AIC = 944.92 at its best order.

Table 6. Frequencies of $\text{AIC}_{\text{true}} - \text{AIC}_{\text{other}}$ Categories out of 100 Simulated Trauma Data from Baseline-category npo Model.

N	< 0	$[0, 2)$	$[2, 4)$	$[4, 7)$	$[7, 10)$	≥ 10
$1 \times 3{,}200$	0	57	27	11	2	3
$10 \times 3{,}200$	13	55	25	7	0	0
$100 \times 3{,}200$	36	44	16	3	0	1
$10{,}000 \times 3{,}200$	100	0	0	0	0	0

We repeat the procedure 100 times with various N. For each simulated trauma data set from the baseline-catetory npo model, we calculate the difference between AIC_{true} (the AIC value with the true model) and $\text{AIC}_{\text{other}}$ (the smallest AIC value among all other models and all orders). In Table 6, we list the frequencies of the AIC differences falling into different ranges. For example, the first column "< 0" provides the numbers of simulations out of 100 with $\text{AIC}_{\text{true}} - \text{AIC}_{\text{other}} < 0$. As N increases from $1 \times 3{,}200$ to $10{,}000 \times 3{,}200$, the number of cases increases from 0 to 100, showing that the true model is increasingly likely to outperform other models with any order in terms of AIC values. The other columns in Table 6 show similar patterns, confirming the conclusions described at the beginning of this section.

5. Real-Data Analysis

The US Police Involved Fatalities Data (hereafter, Police data) were downloaded from data.world (`https://data.world/awram/us-police-involved-fatalities`, version June 21, 2020), which was collected by Chris Awarm from three data resources, namely, `https://fatalencounters.org/`, `https://www.gunviolencearchive.org/`, and `Fatal Police Shootings`, from data.world. The original data lists individuals killed by the police in the United States from 2000 to 2016, including information on 12,483 suspects' age, race, mental health status, weapons they were armed with, and whether or not they were fleeing. By way of example, we focus on whether the police's action can be predicted by the aforementioned information related to a suspect.

As summarized in Table 7, there are four categories of (police) responses, namely, `other`, `shot`, `shot and tasered`, and `tasered`. In our notation, $J = 4$. In the original data, there are 60 different types of armed status. Here, we simplify these into three categories: `gun`, if the original input is "gun;" `unarmed`, if the original input is "unarmed" or missing; and `other`, if otherwise. As such, we have 24 possible level combinations of `armed status` ($x_{i1} = 1$ (gun), 2 (other), or 3 (unarmed)), `gender` ($x_{i2} = 0$ (female) or 1 (male)), `flee` ($x_{i3} = 0$ (false) or 1 (true)), `mental illness` ($x_{i4} = 0$ (false) or 1 (true)). Because there is no observation associated with $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^T = (3, 0, 1, 1)^T$, $m = 23$ in this case (Tbl. 7).

Table 7. US Police Involved Fatalities Data (2000 to 2016).

Armed	Gender	Flee	Mental Illness	Other	Shot	Shot and Tasered	Tasered	Total
Gun	Female	False	False	0	134	0	1	135
Gun	Female	False	True	0	57	0	1	58
Gun	Female	True	False	0	8	0	0	8
Gun	Female	True	True	0	4	0	0	4
Gun	Male	False	False	2	3,314	6	35	3,357
Gun	Male	False	True	0	810	4	15	829
Gun	Male	True	False	0	271	5	0	276
Gun	Male	True	True	0	33	1	0	36
Other	Female	False	False	0	53	1	0	54
Other	Female	False	True	1	42	1	0	44
Other	Female	True	False	0	4	1	0	5
Other	Female	True	True	0	2	0	0	2
Other	Male	False	False	1	910	38	10	959
Other	Male	False	True	2	478	21	5	506
Other	Male	True	False	0	114	10	0	124
Other	Male	True	True	0	14	2	0	16
Unarmed	Female	False	False	1	231	0	3	235
Unarmed	Female	False	True	0	61	0	5	66
Unarmed	Female	True	False	0	2	0	0	2
Unarmed	Male	False	False	10	4,338	16	253	4,617
Unarmed	Male	False	True	12	832	5	214	1,063
Unarmed	Male	True	False	0	75	8	0	83
Unarmed	Male	True	True	0	5	1	0	6

Note: The group (Unarmed, Female, Fleed, Has Mental Illness) contains no observation and is omitted.

As an example, we consider the main-effects baseline-category, cumulative, adjacent-categories, and continuation-ratio logit models with po or npo. In our notation, $\mathbf{h}_1(\mathbf{x}_i) = \cdots = \mathbf{h}_{J-1}(\mathbf{x}_i) = (1, \mathbf{1}_{\{x_{i1}=2\}}, \mathbf{1}_{\{x_{i1}=3\}}, x_{i2}, x_{i3}, x_{i4})^T$ for all eight models under consideration. For each model, we choose the best order out of $4! = 24$ of the four response categories, based on the AIC. Of the eight logit models, each with 24 orders, the continuation-ratio npo model with the chosen order (t, s, o, st) or (t, s, st, o) (Tbl. 8) performs best, and can be written as

$$\log\left(\frac{\pi_{ij}}{\pi_{i,j+1}+\cdots+\pi_{iJ}}\right) = \beta_{j1}+\beta_{j2}\mathbf{1}_{\{x_{i1}=2\}}+\beta_{j3}\mathbf{1}_{\{x_{i1}=3\}}+\beta_{j4}x_{i2}+\beta_{j5}x_{i3}+\beta_{j6}x_{i4}\ ,$$

where $j = 1, 2, 3$ and $i = 1, \ldots, 23$. The corresponding BIC values, not shown here, provide a consistent selection result. According to the AIC values, if we choose the continuation-ratio npo model with the best order {`tasered`, `shot`, `other`, `shot and tasered`} against the baseline-category models (po or npo), which are commonly used for categorical responses without an order, the improvement in the prediction accuracy is significant (AIC differences > 5). Note that in this case, it is not trivial to determine the baseline category for po models,

Table 8. Model and Order Selection for Police Data.

Model	AIC with Best Order	Best Order
Baseline-category po	401.33	t is the baseline
Baseline-category npo	197.81	All are the same
Cumulative po	318.14	(st, s, o, t) or (t, o, s, st)
Cumulative npo	194.48	(o, st, s, t) or (t, s, st, o)
Adjacent-categories po	290.17	(st, s, o, t) or (t, o, s, st)
Adjacent-categories npo	197.81	All are the same
Continuation-ratio po	320.22	(t, o, s, st)
Continuation-ratio npo	192.01	(t, s, o, st) or (t, s, st, o)

Note: s = `shot`, t = `tasered`, o = `other`, st = `shot and tasered`.

owing to the existence of the category `other`.

Table 8 also shows large gaps of AIC values between the npo models and the corresponding po models, indicating that npo models are significantly better than the corresponding po models in this case. The differences within the AIC values of the npo models are also much smaller than those within the po models. This provides strong evidence that for the Police data, the parameters for the categories are very different, and thus the proportional odds (po) assumptions are not appropriate (see the Introduction).

To validate the selected model, we conduct five-fold cross-validation for the data, with cross-entropy loss as the criterion (Hastie, Tibshirani and Friedman, 2009). We compare our selected continuation-ratio npo model with the baseline-category npo model, which is commonly used for nominal responses. In terms of cross-entropy loss, the continuation-ratio npo model achieves 550.50, which is less than the value of 555.91 for the baseline-category npo model. This is consistent with our conclusion based on AIC values.

The estimated parameters for the chosen continuation-ratio npo model are provided in Table 9, which can be used to interpret the roles and effects of different factors. For example, $\hat{\beta}_{13} = 2.03$ indicates that the estimated odds ratio of `tasered` and "unarmed" is $e^{2.03} = 7.61$, which implies that "unarmed" leads to a much smaller chance of being shot. In contrast, $\hat{\beta}_{15} = -18.02$, with an estimated odds ratio $e^{-18.02} = 1.49 \times 10^{-8}$, implies that suspects who flee have a much greater chance of being shot. Because `shot` is usually regarded as more severe than `tasered`, the estimated parameters imply that if suspects show a greater threat such as being armed (gun or other), or try to flee, the police tend to take more extreme actions, such as shooting a gun.

To investigate whether the best order is chosen because of randomness, we conduct a simulation study similar to that in Section 4.5. We regenerate the Police data simulated from the baseline-category npo model fitted from the original data (see Tbl. 10 for the parameter values). For the simulated Police data, the best model is the cumulative npo model with the order {`shot and tasered`,

Table 9. Estimated Parameters for Police Data under Continuation-ratio npo Model.

j	Intercept $\hat{\beta}_{j1}$	Armed Status $\hat{\beta}_{j2}$ (Other)	Armed Status $\hat{\beta}_{j3}$ (Unarmed)	Gender $\hat{\beta}_{j4}$	Flee $\hat{\beta}_{j5}$	Mental Illness $\hat{\beta}_{j6}$
1	-6.00	-0.44	2.03	1.17	-18.02	1.34
2	6.47	-2.43	-1.09	-0.58	-1.55	-0.59
3	0.26	-1.49	1.69	-2.29	-28.35	1.02

Note: $j = 1, 2, 3$ correspond to `tasered`, `shot`, and `other`, respectively.

Table 10. Estimated Parameters for Police Data under Baseline-category npo Model.

j	Intercept $\hat{\beta}_{j1}$	Armed Status $\hat{\beta}_{j2}$ (Other)	Armed Status $\hat{\beta}_{j3}$ (Unarmed)	Gender $\hat{\beta}_{j4}$	Flee $\hat{\beta}_{j5}$	Mental Illness $\hat{\beta}_{j6}$
1	-2.07	-1.34	6.00	-1.18	-0.82	-0.31
2	0.24	-8.83	-2.04	11.09	-1.39	12.74
3	1.89	0.19	0.40	-1.35	2.93	-1.02

Note: $j = 1, 2, 3$ correspond to `other`, `shot`, `shot and tasered`, respectively.

`shot`, `other`, `tasered`} with AIC 149.77, which is not significantly better than the true model (baseline-category npo), with AIC 151.04. If we increase the sample size by a factor of 10, the true model with AIC 271.90 becomes better than the cumulative npo model, with AIC 274.53. If we further increase it by a factor of 100, the true model with AIC 401.92 is significantly better than the cumulative npo model, with AIC 440.45. In other words, with the original sample size of the Police data, if there is no true order, then the selected model with a working order may not be significantly better than the true model. As the amount of data increases, the true model becomes significantly better than the other models with any working order, supporting the simulation results in Section 4.5.

6. Discussion

In our analysis of real data, we consider both the type of model and the order of the categories as parts of the model selection procedure. Our examples show that the improvement by choosing a more appropriate order can be highly significant. For example, if we focus on adjacent-categories po models for the Police data, the smallest AIC value from the best order is 290.17, whereas the largest one from the worst order is 674.44, implying a significant difference in terms of prediction accuracy.

For the Police data, there is no natural order, owing to the existence of the category `other`. According to the AIC values, the continuation-ratio npo model with the working order (t, s, o, st), or equivalently (t, s, st, o), is significantly better than the baseline-category npo models (Tbl. 8). On the one hand, if there is a true order, it will appear among the best orders consistently (Sec. 3). On the other hand, if there is no true order for the Police data, then there is no

ordinal model with a working order that is significantly better than the baseline-category npo model, with the current or a larger sample size (see the simulation study at the end of Sec. 5). In other words, we can report the working order with confidence if it is significantly better than other orders.

Finally, in the proofs of Theorems 2 to 7, we provide explicit transformation formulae from $\boldsymbol{\theta}_1$ with σ_1 to $\boldsymbol{\theta}_2$ with σ_2 when $\sigma_2 \sim \sigma_1$. This significantly reduces the computational cost of finding MLEs with different orders.

Supplementary Material

The online Supplementary Material contains proofs of Theorems 1 to 8, Lemmas 2 and 3, and Corollary 1.

Acknowkedgments

This work was supported, in part, by the US NSF grant DMS-1924859.

References

Agresti, A. (2010). *Analysis of Ordinal Categorical Data*. 2nd Edition. John Wiley & Sons.

Agresti, A. (2013). *Categorical Data Analysis*. 3rd Edition. John Wiley & Sons.

Agresti, A. (2018). *An Introduction to Categorical Data Analysis*. 3rd Edition. John Wiley & Sons.

Aitchison, J. and Bennett, J. (1970). Polychotomous quantal response by maximum indicant. *Biometrika* **57**, 253–262.

Bu, X., Majumdar, D. and Yang, J. (2020). D-optimal designs for multinomial logistic models. *Annals of Statistics* **48**, 983–1000.

Burnham, K. P. and Anderson, D. R. (2004). Understanding AIC and BIC in model selection. *Sociological Methods & Research* **33**, 261–304.

Chuang-Stein, C. and Agresti, A. (1997). A review of test for detecting a monotone dose-response relationship with ordinal response data. *Statistics in Medicine* **16**, 2599–2618.

Dobson, A. and Barnett, A. (2018). *An Introduction to Generalized Linear Models*. 4th Edition. Chapman & Hall/CRC.

Glonek, G. and McCullagh, P. (1995). Multivariate logistic models. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 533–546.

Greene, W. (2018). *Econometric Analysis*. Pearson Education.

Hastie, T., Tibshirani, R. and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. Springer.

Jennett, B. and Bond, M. (1975). Assessment of outcome after severe brain damage. *Lancet* **305**, 480–484.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society. Series B (Methodological)* **42**, 109–142.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*. 2nd Edition. Chapman & Hall/CRC.

O'Connell, A. (2006). *Logistic Regression Models for Ordinal Response Variables*. Sage.

Wallace, D. A. (1998). *Groups, Rings and Fields*. Springer.

Yang, J., Tong, L. and Mandal, A. (2017). D-optimal designs with ordered categorical data. *Statistica Sinica* **27**, 1879–1902.

Zocchi, S. and Atkinson, A. (1999). Optimum experimental designs for multinomial logistic models. *Biometrics* **55**, 437–444.

(Received September 2022; accepted May 2023)