

Uncertain Choices: The Heterogeneous Multinomial Logit Model

Tutz, Gerhard

Veröffentlichungsversion / Published Version

Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:

Tutz, G. (2021). Uncertain Choices: The Heterogeneous Multinomial Logit Model. *Sociological Methodology*, 51(1), 86-111. <https://doi.org/10.1177/0081175020979689>

Nutzungsbedingungen:

Dieser Text wird unter einer CC BY-NC Lizenz (Namensnennung-Nicht-kommerziell) zur Verfügung gestellt. Nähere Auskünfte zu den CC-Lizenzen finden Sie hier: <https://creativecommons.org/licenses/by-nc/4.0/deed.de>

Terms of use:

This document is made available under a CC BY-NC Licence (Attribution-NonCommercial). For more information see: <https://creativecommons.org/licenses/by-nc/4.0>

Uncertain Choices: The Heterogeneous Multinomial Logit Model

Sociological Methodology

2021, Vol. 51(1) 86–111

© The Author(s) 2021



DOI: 10.1177/0081175020979689

<http://sm.sagepub.com>**Gerhard Tutz**¹ 

Abstract

In this article, a modeling strategy is proposed that accounts for heterogeneity in nominal responses that is typically ignored when using common multinomial logit models. Heterogeneity can arise from unobserved variance heterogeneity, but it may also represent uncertainty in choosing from alternatives or, more generally, result from varying coefficients determined by effect modifiers. It is demonstrated that the bias in parameter estimation in multinomial logit models can be substantial if heterogeneity is present but ignored. The modeling strategy avoids biased estimates and allows researchers to investigate which variables determine uncertainty in choice behavior. Several applications demonstrate the usefulness of the model.

Keywords

heterogeneous multinomial logit model, heterogeneity, heterogeneous choice model, location-scale model, location-shift model, multinomial responses

INTRODUCTION

The modeling of heterogeneity in binary and ordinal response models has been a topic of intensive research. In particular, Allison's (1999) demonstration that comparisons of binary model coefficients across groups can be misleading if one has underlying heterogeneity of residual variances has stimulated research in the area. Williams (2009), Mood (2010), Rohwer (2015), Karlson, Holm, and Breen (2012), and Breen, Holm, and Karlson (2014) have all investigated ways to deal with this problem.

One approach is based on the heterogeneous choice model, in which an explicit term is included that accounts for variance heterogeneity (see Williams 2009, 2010). McCullagh (1980) considered an earlier version of the model under the name location-scale model, but the importance of modeling variance heterogeneity was not recognized until much later. Tutz and Berger (2016, 2017) proposed alternative models to account for variance heterogeneity in ordinal models. The considered location-shift models use an additive parameterization unlike the heterogeneous choice model, which uses a multiplicative predictor. Nevertheless, location-scale and location-shift models

¹Ludwig-Maximilians-Universität München, München, Germany

Corresponding Author:

Gerhard Tutz, Ludwig-Maximilians-Universität München, Akademiestraße 1, 80799 München, Germany.

Email: tutz@stat.uni-muenchen.de

typically show similar goodness of fit. However, both models assume the response is ordinal.

CUB-type mixture models (i.e., combination of discrete uniform and binomial distribution) account for heterogeneity in ordinal responses. These models assume the observed response results from a mixture of an ordinal response model and an uncertainty component. The latter is determined by a uniform discrete distribution over categories. It is assumed to represent respondents' uncertainty. Explanatory variables can determine the probabilities of the mixture. D'Elia and Piccolo (2005), Iannario and Piccolo (2010), Iannario (2012), Tutz et al. (2017), and Iannario et al. (2020) all considered models of this type, and Piccolo and Simone (2019) provided an extensive overview.

It is surprising that prior work has considered heterogeneity only for ordinal responses, not for unordered responses. In many surveys (e.g., surveys that cover party preference), respondents are asked to choose from a set of categories that represent distinct but not ordered alternatives. In the special case of binary choices, the heterogeneous model applies but CUB-type models fail because they work only for more than two categories. However, for more than two response categories, both types of model explicitly assume that categories are ordered.

Here, we consider the case of unordered response categories, in which response categories are mere labels without any inherent ordering. Social scientists use choice models for this type of response when modeling party preference, for example, or the choice of brands in economic applications. In such cases, respondents choose from a set of k categories, which simply represent alternatives. The response is measured on a nominal scale. Instead of using the numbers $1, \dots, k$, any k labels or any permutation of the numbers could be used. The multinomial logit model (MLM) is the most widely used model for unordered responses; it models the response probability as a function of explanatory variables.

A model is proposed that accounts for additional heterogeneity not captured by the linear parametric terms in the MLM. Heterogeneity is modeled as a function of explanatory variables that modifies the response probabilities. Heterogeneity can be interpreted as variance heterogeneity but also as respondents' uncertainty. The resulting model is a truly multicategorical model, which, in contrast to the classical MLM, cannot be estimated by considering subsets of two response categories. The main benefits of the modeling strategy are that (1) the approach accounts for potential heterogeneity in unordered choices; (2) we avoid bias in parameter estimates, which can be substantial if heterogeneity is present but ignored; and (3) we obtain information on the dependence of choice uncertainty on explanatory variables.

THE HETEROGENEOUS MLM

Let $Y \in \{1, \dots, k\}$ denote the response in nominal categories $1, \dots, k$, and \mathbf{x} denote a vector of explanatory variables. The widely used MLM has the general form

$$P(Y = r|\mathbf{x}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_r)}{\sum_{s=1}^k \exp(\mathbf{x}^T \boldsymbol{\beta}_s)}, \tag{1}$$

where for simplicity \mathbf{x} is assumed to contain an intercept. Side constraints are needed, for example, $\boldsymbol{\beta}_1^T = (0, \dots, 0)$, which makes category 1 the reference category, or $\sum_{s=1}^k \boldsymbol{\beta}_s^T = (0, \dots, 0)$, which is symmetric and does not refer to any fixed category. The model contains k linear predictors $\eta_r = \mathbf{x}^T \boldsymbol{\beta}_r$, one for each category (although the effective number is $k - 1$). In contrast to typical ordinal models, the parameter vectors are category specific, which means the model has a rather large number of parameters.

Accounting for Heterogeneity

Let \mathbf{z} denote an additional vector of explanatory variables, which can be distinct from \mathbf{x} but may also contain components of \mathbf{x} . The *heterogeneous MLM* (HMLM) proposed here contains an additional heterogeneity term, which is determined by \mathbf{z} and a vector of parameters $\boldsymbol{\gamma}$. It has the form

$$P(Y = r|\mathbf{x}, \mathbf{z}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_r e^{\mathbf{z}^T \boldsymbol{\gamma}})}{\sum_{s=1}^k \exp(\mathbf{x}^T \boldsymbol{\beta}_s e^{\mathbf{z}^T \boldsymbol{\gamma}})}. \tag{2}$$

For any two categories, we thus have

$$\log\left(\frac{P(Y = r|\mathbf{x}, \mathbf{z})}{P(Y = s|\mathbf{x}, \mathbf{z})}\right) = \mathbf{x}^T (\boldsymbol{\beta}_r - \boldsymbol{\beta}_s) e^{\mathbf{z}^T \boldsymbol{\gamma}}. \tag{3}$$

The predictors in the model comprise two components: the location terms, $\mathbf{x}^T \boldsymbol{\beta}_r$, which are category specific, and the scaling term, $e^{\mathbf{z}^T \boldsymbol{\gamma}}$, which is not category specific. The effect of the scaling term is considered in the following, where for simplicity we assume that $z \in \{0, 1\}$ is a group indicator. We obtain the following properties:

For $\boldsymbol{\gamma} \rightarrow -\infty$, we obtain $P(Y = r|\mathbf{x}, z = 1) = 1/k$, which means a person from group 1 chooses categories at random and shows maximal uncertainty, that is, a *noncontingent response style*. This occurs if people have a tendency to respond carelessly, randomly, or nonpurposefully (Baumgartner and Steenkamp 2001; Van Vaerenbergh and Thomas 2013).

For $\boldsymbol{\gamma} \rightarrow \infty$, we obtain $P(Y = r|\mathbf{x}, z = 1) = 1$ for one of the categories $r \in \{1, \dots, k\}$, provided the parameters $\boldsymbol{\beta}_r$ are nonzero and vary across categories. This means that a person from group 1 knows exactly which category to choose, the person has a distinct preference (more details and a proof are given in the Appendix).

With varying parameter $\boldsymbol{\gamma}$, the whole spectrum from a random response to a distinct choice is covered. This interpretation of the scaling component focuses on uncertainty. Before discussing alternative interpretations, let us briefly mention some further properties.

If $k=2$, the model reduces to the heterogeneous choice model for binary responses, which is a special case of the heterogeneous choice model designed for ordinal responses (see, e.g., Mood 2010; Williams 2009).

If we use the first category as a reference category by setting $\boldsymbol{\beta}_1^T = (0, \dots, 0)$, we obtain

$$\log\left(\frac{P(Y=r|\mathbf{x}, \mathbf{z})}{P(Y=1|\mathbf{x}, \mathbf{z})}\right) = \log\left(\frac{P(Y=r|Y \in \{1, r\}, \mathbf{x}, \mathbf{z})}{P(Y=1|Y \in \{1, r\}, \mathbf{x}, \mathbf{z})}\right) = \mathbf{x}^T \boldsymbol{\beta}_r e^{\mathbf{z}^T \boldsymbol{\gamma}} = \mathbf{x}^T \boldsymbol{\beta}_r^*, \quad (4)$$

where $\boldsymbol{\beta}_r^* = \boldsymbol{\beta}_r e^{\mathbf{z}^T \boldsymbol{\gamma}}$ is the effective parameter that determines the strength of the effect of the explanatory variables \mathbf{x} , which varies over values of \mathbf{z} . With $\boldsymbol{\gamma}^T = (0, \dots, 0)$, we obtain the simple MLM, which can be seen as a collection of binary logit models of the form of equation (4). In the multinomial model, we can also investigate the effect of variables by fitting separate binary models, although estimates are less efficient (see, e.g., Agresti 2013). This does not hold for the heterogeneous model. For $k > 2$, the uncertainty model is a truly multinomial model. We cannot investigate the effect of variables separately, because the factor $e^{\mathbf{z}^T \boldsymbol{\gamma}}$ is the same in all the binary models that are contained.

Interpretation of Parameters and Motivations of the Model

The effect of the scaling component has not always been presented clearly in the classical heterogeneous model; in particular, it is often interpreted solely as representing variances. Therefore, we consider in the following several ways to interpret effects.

Variance Heterogeneity and Random Utilities. One way to motivate the multinomial model is to consider it as a random utility model. Let U_r be an unobservable random utility associated with the r th response category. For example, U_r is the (subjective) utility of a brand among a choice of brands $1, \dots, k$ or the ‘‘attractiveness’’ of the r th political party. Let U_r be determined by $U_r = u_r + \varepsilon_r$, where u_r is a fixed value, representing the fixed utility associated with the r th response category, and $\varepsilon_1, \dots, \varepsilon_k$ are iid random variables with distribution function $F(\cdot)$. In addition, let the response Y be determined by the *principle of maximum random utility*, which specifies that the link between the observable Y and the unobservable random utility is given by

$$Y=r \quad \Leftrightarrow \quad U_r = \max_{j=1, \dots, k} U_j.$$

Thus, we choose the alternative that maximizes the random utility. If we assume that $\varepsilon_r, \dots, \varepsilon_k$ are iid variables with distribution function $F(x) = \exp(-\exp(-x))$ (i.e., the Gumbel or maximum extreme value distribution), we obtain

$$P(Y=r) = \frac{\exp(u_r)}{\sum_{j=r}^k \exp(u_j)}$$

(e.g., McFadden 1973; Yellott 1977). By setting $u_r = \mathbf{x}^T \boldsymbol{\beta}_r$ (with restriction $\boldsymbol{\beta}_1 = \mathbf{0}$), we obtain the simple multinomial model.

Let us now assume, more generally, that the random utilities are given by $U_r = u_r + \sigma \varepsilon_r$, where $\sigma = e^{-z^T \boldsymbol{\gamma}}$. Then, the maximum utility approach, $Y = r \Leftrightarrow U_r = \max_{j=1, \dots, k} U_j$ yields the heterogeneous model

$$P(Y = r | \mathbf{x}, \mathbf{z}) = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_r / e^{-z^T \boldsymbol{\gamma}})}{\sum_{s=1}^k \exp(\mathbf{x}^T \boldsymbol{\beta}_s / e^{-z^T \boldsymbol{\gamma}})} = \frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_r e^{z^T \boldsymbol{\gamma}})}{\sum_{s=1}^k \exp(\mathbf{x}^T \boldsymbol{\beta}_s e^{z^T \boldsymbol{\gamma}})},$$

where again u_r is parameterized as $u_r = \mathbf{x}^T \boldsymbol{\beta}_r$.

The derivation suggests the modification of effect strength $\boldsymbol{\beta}_r$ to $\boldsymbol{\beta}_r e^{z^T \boldsymbol{\gamma}}$ is due to varying heterogeneity of the variances in the underlying random utilities. If variances differ, then effect strengths of the explanatory variables \mathbf{x} also differ. If $k = 2$, the derivation is in accordance with the derivation of the ordinal heterogeneous model when applied to binary choices (see, e.g., Mood 2010; Williams 2009). In his frequently cited binary response example, Allison (1999) examined biochemists' careers with the binary response "promotion to associate professor from assistant professor (yes/no)" and demonstrated that effect strengths of covariates (e.g., number of published articles) can be affected by differing variances in the male and female groups of scientists.

Note that a scaling problem might affect the interpretation of parameters even when variances do not explicitly depend on covariates. Parameters in the model are then given by $\boldsymbol{\beta}_r / \sigma$, and parameters are only identified if the scaling parameter σ is fixed. This raises problems if one compares parameters across nested models. With random utilities given by $U_r = \mathbf{x}^T \boldsymbol{\beta}_r + \sigma \varepsilon_r$ we cannot expect the standard deviation σ to be the same in the full model and a reduced model with fewer explanatory variables unless the corresponding parameters are zero. For binary response models, Karlson et al. (2012) proposed a method to compare regression coefficients between nested models.

Uncertainty. In questionnaires in which respondents choose among unordered alternatives, heterogeneity in variances of underlying utilities can be seen as uncertainty. If the variance is large, the preference for specific categories becomes less distinct; if variance is small, the category with the largest value in the location term $\mathbf{x}^T \boldsymbol{\beta}_r$ has a high probability of being chosen. Thus, heterogeneity in variances turns into uncertainty.

However, the scaling term $e^{z^T \boldsymbol{\gamma}}$ can also be seen as measuring uncertainty without reference to underlying variances. Reference to variances always assumes an underlying latent trait model, which is an additional and not necessary assumption. Latent variable models provide a motivation for categorical response models, but they are not needed when interpreting parameters unless one wants to investigate the latent structure itself (for a discussion of binary responses, see Kuha and Mills 2017). As shown earlier, the parameter $\boldsymbol{\gamma}$ determines if (groups of) respondents have a tendency to strongly prefer specific categories or tend to choose categories more or less at random, yielding a categorical uniform response distribution. Thus, heterogeneity can be seen as representing a noncontingent response style. Response styles have been mainly investigated in item response settings with ordered responses (see, e.g., Falk and Cai 2016; Johnson and Bolt 2010; Wetzel and Carstensen 2017). However, one can also

expect to find response styles if responses are measured on a nominal rather than ordinal scale level.

Varying Coefficients and Interactions. More generally, the model can also be seen within the framework of varying coefficients, proposed by Hastie and Tibshirani (1993) and extended by Cai, Fan, and Li (2000), Antoniadis, Gijbels, and Verhasselt (2012), and Park et al. (2015). For simplicity, let the scaling component contain the single binary variable z , that is, one has a scaling effect $z\gamma$. Then, all effects of explanatory variables in the location term are described by $\beta e^{z\gamma}$. This means they vary over values of z . The variable z is an *effect modifier*: it modifies all effects in the location term. It is a specific form of interaction between variables in the location term and variables in the heterogeneity term. To be more concise, if we let z also be present in the location term, then we obtain the following for the predictor:

$$\eta_r = (\beta_{r0} + z\beta_{rz} + \sum_j x_j \beta_{rj}) e^{z\gamma} = \tilde{\beta}_{r0} + z\tilde{\beta}_{rz} + \sum_j x_j \tilde{\beta}_{rj} + \sum_j x_j z \tilde{\beta}_{rj}^{(jz)},$$

where $\tilde{\beta}_{r0} = \beta_{r0}$, $\tilde{\beta}_{rj} = \beta_{rj}$, $\tilde{\beta}_{rz} = \beta_{rz} + (\beta_{r0} + \beta_{rz})(e^\gamma - 1)$, and $\tilde{\beta}_{rj}^{(jz)} = \beta_{rj}(e^\gamma - 1)$. The last form of the predictor shows a model in which interactions between all the variables x_1, \dots, x_p and z are included. The crucial point is that these interactions are not varying freely, they are constrained because they were generated by heterogeneity. The constraints can be given in the form

$$\tilde{\beta}_{r1}^{(1z)} / \tilde{\beta}_{r1} = \dots = \tilde{\beta}_{rp}^{(pz)} / \tilde{\beta}_{rp}, \quad (5)$$

which means the proportion between the interaction effect $\tilde{\beta}_{rj}^{(jz)}$ and the corresponding main effect $\tilde{\beta}_{rj}$ is the same for all variables, namely, $e^\gamma - 1$. This means that the heterogeneous model includes interactions between the covariates and the variable in the heterogeneity term, but interactions have a specific form. The model is not equivalent to the general interaction model that contains all interactions between covariates and z , but it can be estimated using the interaction model with constraints. The importance of interaction terms in ordinal heterogeneous responses has been emphasized by Allison (1999) and investigated more closely by Rohwer (2015) and Tutz (2018, 2020). It is also present in the nominal model.

From a slightly different viewpoint, interactions can be seen as varying coefficients. Varying-coefficient models allow one to model effects that are modified by other variables. An advantage is that no reference to latent motivating variables is needed. These models aim to identify which effects are not stable across the variation of other variables and provide a general concept for interpretation of effects in models that account for the type of heterogeneity considered here. For the binary response case, Tutz (2020) investigated varying coefficients and the simpler form of constraints.

The representation as a model with interactions holds also in the case where \mathbf{x} and z are distinct. Then z is not included in the location terms, that is, $\beta_{rz} = 0$. We obtain the same interaction parameters and main effects of the x variables as above, only the main effect of z simplifies to $\tilde{\beta}_{rz} = \beta_{r0}(e^\gamma - 1)$. We also see that the heterogeneity effect,

although not present in the original location term, generates a location shift in the interaction representation of the model.

Measurement of Variability in Nominal Responses. Let us comment briefly on the general problem of measuring variability in a nominal response. Because the variable Y is measured on a nominal scale, the classical measure for variability, namely the variance of Y , is useless. Measures that makes sense for nominal categories are the (normalized) Gini heterogeneity, $G = (1 - \sum_{r=1}^k \pi_r^2) \frac{k}{k-1}$, and the impurity based on entropy, $I_E = -\sum_{r=1}^k \pi_r \log(\pi_r)$. Both measures are zero if one of the probabilities has value 1 (and the rest 0), and take their maximal value in the case of the uniform distribution $\pi_1 = \dots = \pi_k$. Thus, small values indicate concentration in one of the categories, and large values indicate strong variability with the same probability in all categories. Strong variability corresponds to strong heterogeneity and therefore low concentration.

This is in line with the effect of the heterogeneity term in the HMLM, in which the probabilities are determined by covariates. Because the probabilities depend on covariates, concentration is different for persons with different explanatory variables. Thus, the category-specific location terms, which are present in the simple and the extended nominal logit model, already generate heterogeneous concentration. The heterogeneity term, which is not category specific, modifies this basic structure, yielding stronger concentration if $\mathbf{z}^T \boldsymbol{\gamma} \rightarrow \infty$, and weaker concentration if $\mathbf{z}^T \boldsymbol{\gamma} \rightarrow -\infty$. Thus, this can be seen as the heterogeneity not captured in the category-specific location terms.

In summary, whatever the interpretation of $e^{z\gamma}$, as uncertainty or variance, the presence of the heterogeneity term means the probabilities vary over values of z ; the size of β parameters should not be interpreted without accounting for the scaling component. Heterogeneity effects of the sort considered here mean that probabilities change driven by heterogeneity although there might be no or a weak location effect (see also the application below). The essence of the heterogeneity captured by the heterogeneous model is variability: persons with different values of z show different variability in responses, and therefore have differing response probabilities, because in categorical data variability and location are not separated. The only exception is the extreme case $\mathbf{x}_r^T \boldsymbol{\beta} = 0$ for all r . In this case, persons have response probabilities $1/k$ and $z\gamma$ does not change response probabilities.

Note that heterogeneity in the model is linked to covariates. It is not modeled on the individual's level in the form of random effects, which might be interesting but hard to obtain without repeated measurements. It is a population-averaged approach, in contrast to conditional modeling approaches that consider responses given covariates and subject-specific parameters (for the distinction between conditional and population-averaged approaches, see, e.g., Neuhaus, Kalbfleisch, and Hauck 1991).

Identifiability

Identifiability of the parameters in a model is critical, because only then is reliable inference on parameters possible. Parameters of the heterogeneous multinomial model are identifiable if \mathbf{x} and \mathbf{z} are distinct (see the Appendix). The general case $\mathbf{x} = \mathbf{z}$ is

more difficult. In one particular case, parameters are certainly not identifiable, namely if there is just one binary predictor variable. Then the predictors in the logit model have the form $x\beta_r e^{x\gamma}$. The model is equivalent to a model with predictors $x\beta_r$, as one can always set $\beta_r = \beta_r e^{x\gamma}$. In other words, one can always set $\gamma = 1$, because γ is not identifiable.

However, parameters are identified if the number of predictors in \mathbf{x} is larger than one ($p > 1$) and there is just one variable in the heterogeneity term. This is useful because in many applications the researcher wants to investigate if a specific variable modifies the effect strengths. This was the case in Allison's biochemists example, which focused on gender as an effect modifier to investigate equal opportunity issues. Identifiability in the case of just one heterogeneity variable can also be used to investigate if variables are needed in the heterogeneity term by including one variable at a time.

Nevertheless, one also wants to allow for a vector of explanatory variables in the heterogeneity terms, although it will be typically shorter than the vector of variables in the location term. In the Appendix, it is shown that, in general, the parameters of the heterogeneous model are identifiable if $p > 1$ and at least one variable contained in the location term has no effect in the heterogeneity term. The Appendix also includes the case with only one effect modifier. Note that the general result does not exclude that also the model that includes all variables in the heterogeneity term might be identifiable in specific settings.

IGNORING HETEROGENEITY

To demonstrate that parameter estimates can be severely biased if heterogeneity is ignored, we show some results of a simulation study. We include two explanatory variables, one binary, following a Bernoulli variable $B(1, 0.5)$, and one continuous, following a normal distribution, $N(0, 1)$. In the case $k = 3$, the parameters are $\beta_2^T = (0.3, \beta_{21}, \beta_{22})$, $\beta_3^T = (0.5, 0, 0)$. Only estimates of β_{21}, β_{22} are shown, which are chosen by $\beta_{21} = 0.3$, $\beta_{22} = 0.25$. In the header of plots, estimates of β_{21} are referred to as binary variables, and estimates of β_{22} as continuous variables. In the plots, the true values are given as gray lines. Estimates of the heterogeneity parameter and the difference in deviances are also given. Because the models are nested, we can compute how much better the heterogeneity model fits the data, which also indicates if the model can be simplified to the simple logit model.

Figure 1 shows estimates of γ , β_{21} , and β_{22} if $\gamma = 0.6$, and the heterogeneity is in the continuous variable. We see that the estimate of the continuous variable is strongly biased if heterogeneity is ignored. Estimates of the heterogeneous model are much closer to the true values than are estimates of the simple model. The heterogeneity parameter is also estimated rather well. The differences in deviances show the heterogeneous model fits data much better, and it indicates that in a data analysis one would find the improvement significant in most cases. Figure 2 shows estimates if the heterogeneity ($\gamma = 0.6$) is in the categorical variable. The results are similar, but also estimates of the binary variable are strongly biased if heterogeneity is ignored.

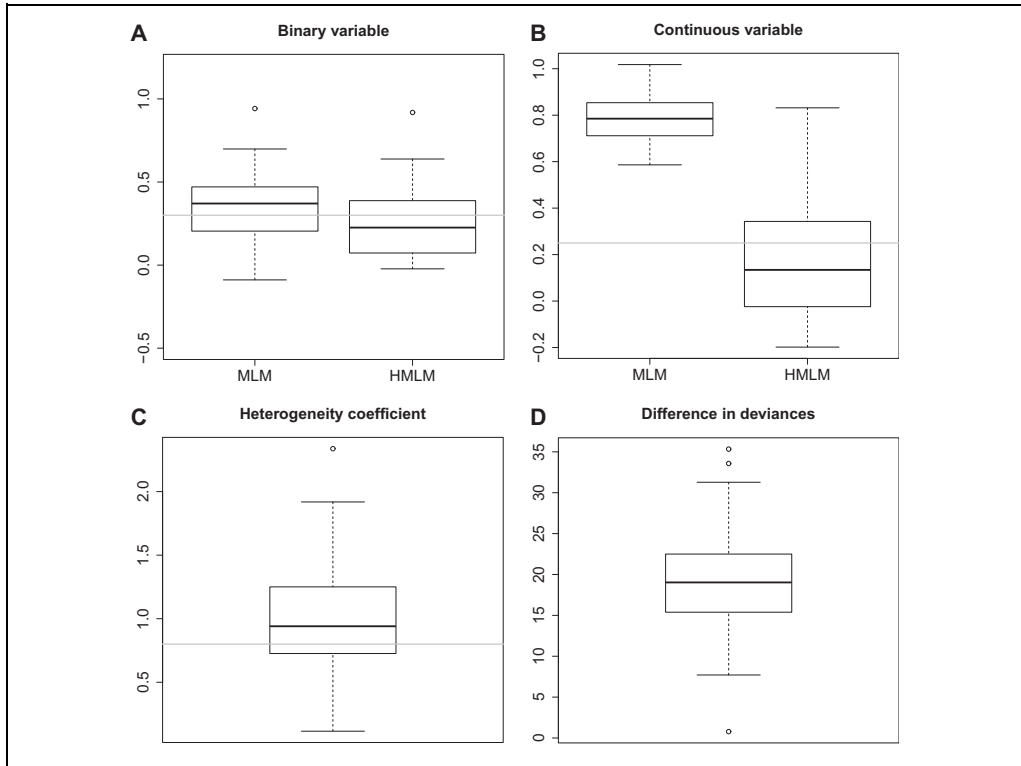


Figure 1. Estimates of parameters in simulation study with heterogeneity in continuous variable.

Note: $n = 500$, $k = 3$, $\gamma = 0.6$.

The effect of varying values of γ is depicted in Figure 3. The figure shows estimates of the coefficients of the binary and the continuous variable if heterogeneity is present in the continuous variable. The first row shows estimates of the binary variable coefficient, which are quite similar for the MLM and the HMLM. The more interesting coefficient is the continuous variable coefficient in the second row: the bias increases strongly with increasing values of γ if it is ignored. In contrast, estimates obtained by the HMLM do not show severe bias. It is seen that coefficients of variables that are linked to heterogeneity might be strongly biased if one uses the simple MLM. Figure 4 shows the corresponding estimates of γ . We see that the HMLM provides reasonable estimates of the heterogeneity parameter. Similar pictures are obtained if heterogeneity is in the categorical variable.

MODELING WITH HETEROGENEITY

The following sections demonstrate the usefulness of the modeling approach in several applications. The Appendix provides details on how to obtain the estimates.

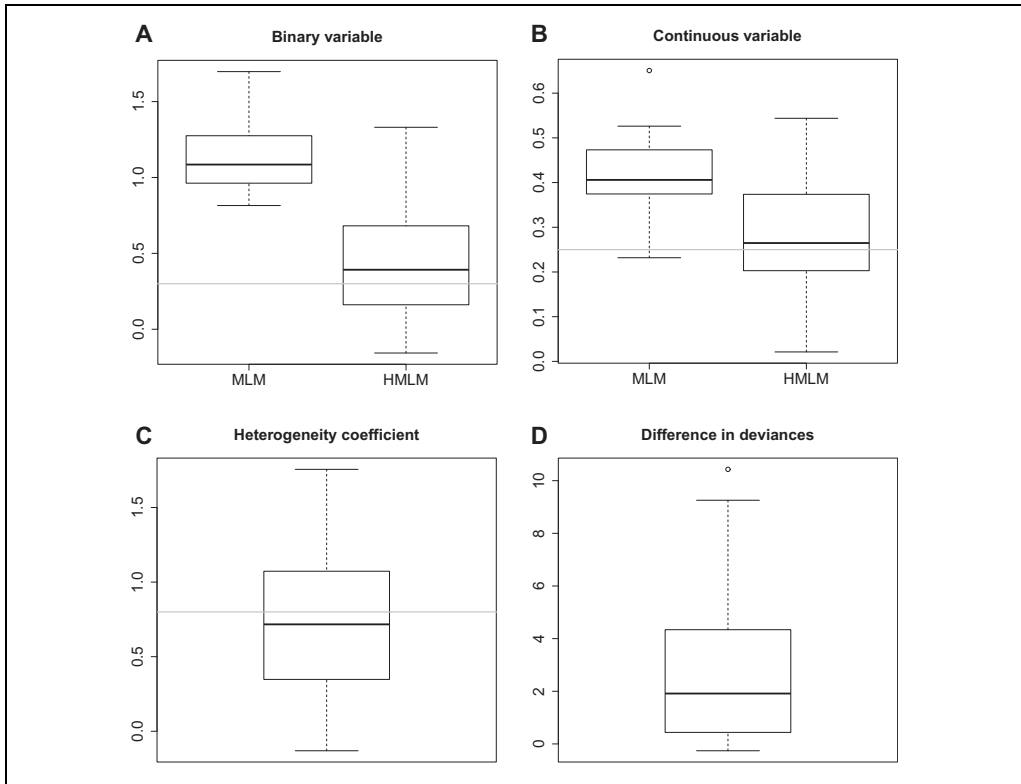


Figure 2. Estimates of parameters in simulation study with heterogeneity in categorical variable.

Note: $n = 500$, $k = 3$, $\gamma = 0.6$.

Party Choice

We consider modeling of party choice with data from the German Longitudinal Election Study. The data are included in the R package *EffectStars* (Schauberger 2019). The response categories refer to the dominant parties in Germany: the Christian Democratic Union (CDU; category 1), the Social Democratic Party (SPD; category 2), the Liberal Party (FDP; category 3), the Green Party (category 4), and the Left Party (Die Linke; category 5). The explanatory variables are age (standardized), gender (1 = male, 0 = female), and regional provenance (west; 1 = former West Germany, 0 = otherwise). The sample size is $n = 816$.

For illustration, let us first investigate if age is an effect-varying variable. Table 1 shows parameter estimates of the MLM and HMLM models with age in the scaling component. We see that the heterogeneity effect of age should not be neglected (value = .481, s.e. = .172). Older respondents show more distinct preferences for political parties than do younger respondents. The parameters obtained for the heterogeneity model differ from the parameters obtained without accounting for heterogeneity. In particular, the age parameters are much closer to zero for the heterogeneity model.

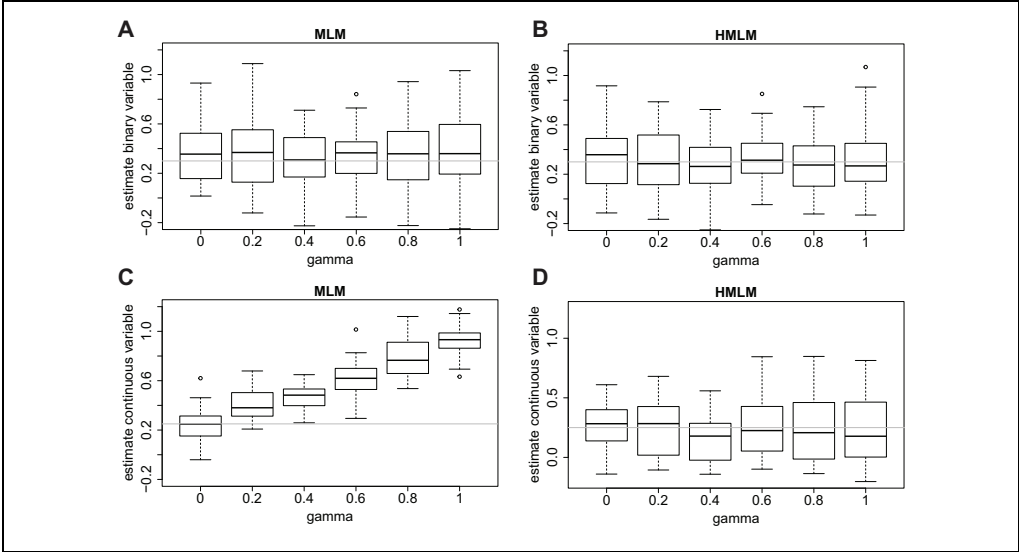


Figure 3. Estimates of parameters for varying γ in continuous variable.
Note: $n = 500$, $k = 3$. First row: estimates of binary variable parameter; second row: estimates of continuous variable parameter.

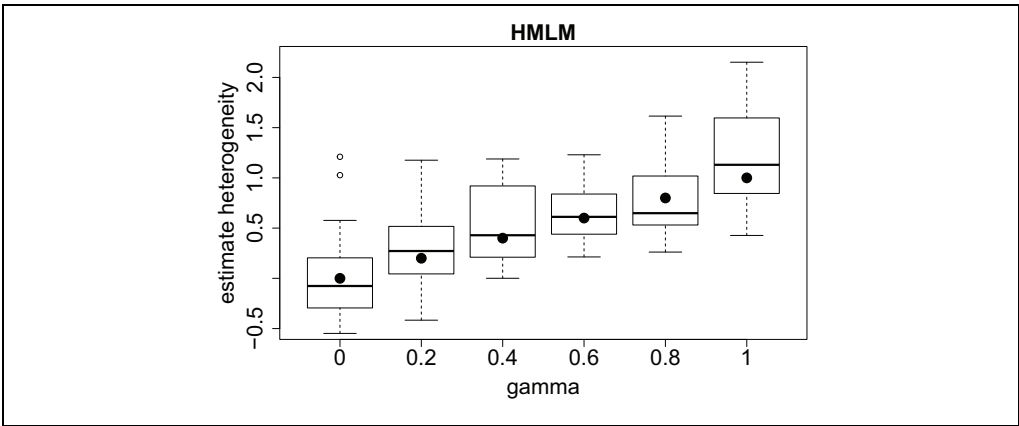


Figure 4. Estimates of heterogeneity parameter for varying γ in continuous variable.
Note: Filled circles represent the true values, $n = 500$, $k = 3$.

However, that does not mean age has a weaker effect on the response. Figure 5 shows the effect of age for males living in the western part of the country. The dotted lines show the effects on probabilities in the multinomial model without heterogeneity; the gray lines represent the effects in the heterogeneous model. The change of probabilities across age is modified, but the effect of age has a very similar tendency. The high significance of the heterogeneity effect suggests part of this effect might be due to heterogeneity.

Table 1. Data Fit for Party Choice Data with and without Heterogeneity in Age

	Estimates without Heterogeneity				Estimates with Heterogeneity in Age			
	Intercept	Gender	West	Age	Intercept	Gender	West	Age
Parameters								
2	-.7161	.1585	.5670	-.2001	-.6098	.1617	.4641	-.0610
3	-1.2655	.6416	-.1055	-.4179	-1.2751	.7091	-.0773	.0568
4	-1.5812	.3795	.4687	-.7962	-1.2741	.4433	.2897	-.3227
5	-.7417	.7914	-.6794	-.2730	-.6192	.6942	-.7563	.0377
Heterogeneity								.4813
Standard error								
2	.2061	.1855	.2145	.09424	.1918	.1446	.1900	.1034
3	.2473	.2424	.2532	.12249	.2375	.2226	.2160	.1970
4	.2696	.2372	.2760	.12637	.2876	.2361	.2818	.2534
5	.2133	.2245	.2226	.11275	.1951	.2111	.1883	.1501
Heterogeneity								.1725
Log-likelihood			-1,201.429				-1,194.916	

Age is not the only variable that might modify effect strengths. Including one variable at a time shows the variable gender is significant, but not the variable west. Table 2 (right-hand columns) shows parameter estimates of the HMLM model with heterogeneity effects of gender and age. We see that they should not be neglected and included simultaneously. We also fitted a model that includes all variables in the scaling component (left-hand columns). The fit also shows the heterogeneity effect of gender and age should not be neglected. Modification of the age effect is rather similar to that seen in Figure 5 and is not shown.

Contraceptive Prevalence Survey

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey, available from the UCI Machine Learning Repository (Contraceptive Method Choice Data Set). The samples are married women who were either not pregnant or did not know if they were pregnant at the time of interview. The response is the contraceptive method used (1 = no use, 2 = long-term use, 3 = short-term use). The explanatory variables are wife's age in years (agew), wife's education (eduw; 1 = low, 2, 3, 4 = high), husband's education (eduh; 1 = low, 2, 3, 4 = high), number of children ever born (children), and wife's religion (relw; 0 = non-Islam, 1 = Islam). The sample size is $n = 1,473$. Fitting models shows that for most variables, there is no heterogeneity effect. The exception is wife's education. Table 3 shows the models without and with heterogeneity in that variable. The effect strengths of variables is much weaker if we account for heterogeneity in wife's education. For example, the effects of children are .33 and .34 if the MLM is fitted but .24 and .25 if the HMLM is fitted. The effects of variables might be overestimated if heterogeneity is ignored.

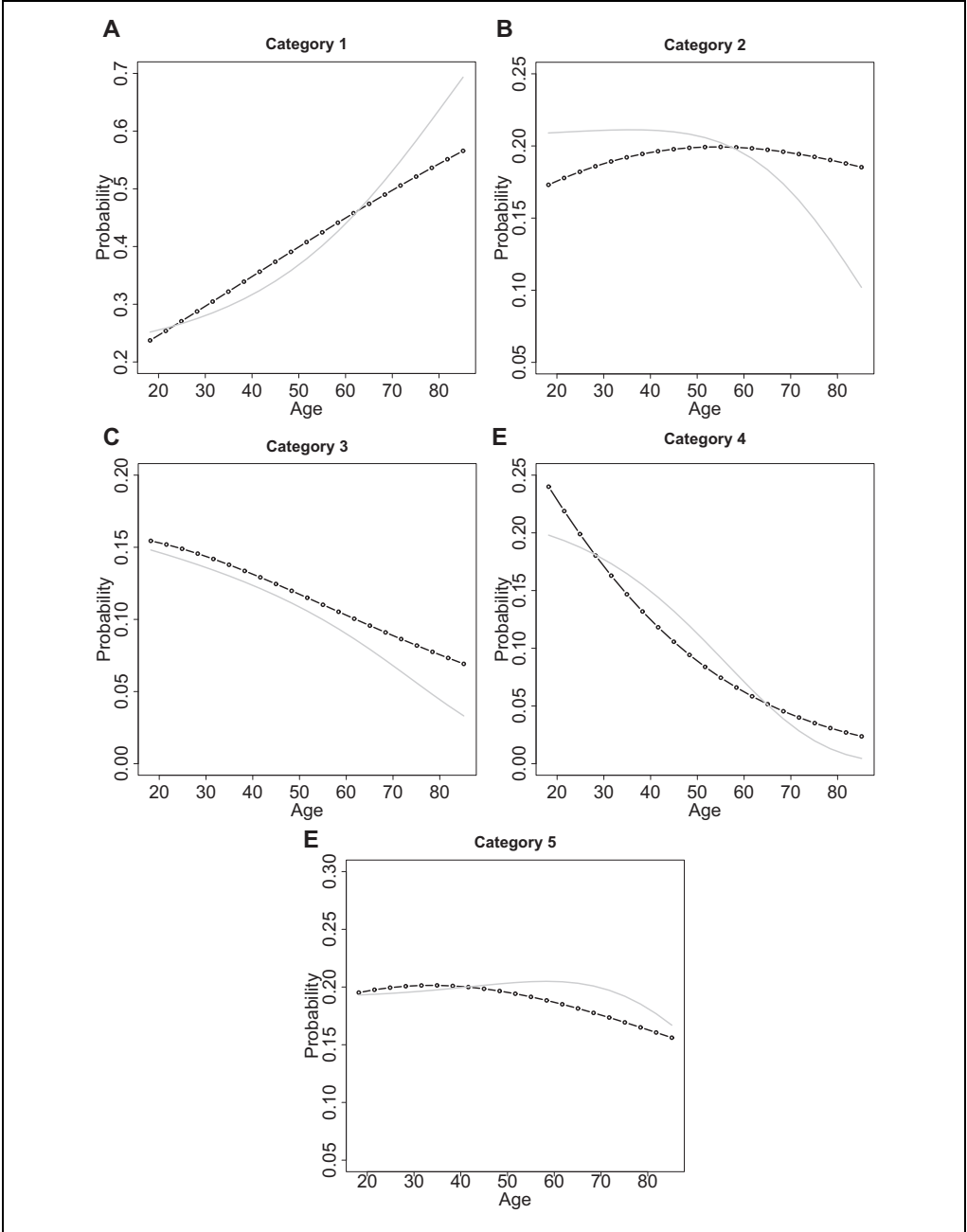


Figure 5. Effects of age on probabilities with age as heterogeneity variable for party choice data.

Note: Dotted lines show effects in the multinomial model without heterogeneity; gray lines represent effects in the heterogeneous model.

Table 2. Data Fit for Party Choice Data with and without Heterogeneity

	Estimates with Full Heterogeneity				Estimates with Reduced Heterogeneity			
	Intercept	Gender	West	Age	Intercept	Gender	West	Age
Parameters								
2	-.8495	-.0510	.5463	.0209	-.7777	.01241	.6569	-.0234
3	-1.2483	-.5066	-1.4617	.3610	-1.2070	.01351	-.1880	.2139
4	-1.1776	-1.5490	-.9876	.0408	-1.1977	-.50054	.2303	-.1779
5	-.6078	.3647	-2.7186	.2810	-.4775	.39998	-1.1826	.1469
Heterogeneity		-1.1324	-.6914	.7140		-.8514		.6542
Standard error								
2	.2800	.4315	.3598	.2113	.2513	.2305	.3001	.1711
3	.3401	.8377	2.0124	.3246	.2880	.4723	.2878	.4036
4	.3390	1.3548	1.8483	.4972	.3344	.6645	.4072	.5923
5	.2875	.4125	2.3691	.2116	.2359	.3070	.3076	.2350
Heterogeneity		.3231	.7541	.1764		.3680		.3341
Log-likelihood		-1,191.405				-1,191.634		

Table 3. Model Fits with and without Heterogeneity in Wife's Education for Contraception Data

	Estimates without Heterogeneity					Estimates with Heterogeneity in Wife's Education				
	agew	Children	relw	eduw	eduh	agew	Children	relw	eduw	eduh
Parameters										
2	-.0349	.3321	-.5247	.9918	.0485	-.0266	.2429	-.3244	.7698	.0452
3	-.1054	.3476	-.3803	.3800	.1175	-.0724	.2570	-.2461	.3111	.0807
Heterogeneity									.1304	
Standard error										
2	.0114	.0420	.1982	.1090	.1290	.00835	.04824	.14346	.12483	.08643
3	.0107	.0380	.1967	.0819	.0970	.01573	.04846	.12994	.06658	.06947
Heterogeneity									.0704	
Log likelihood		-1,407.99				-1,406.361				

Satisfaction Data

The data contain students' satisfaction with the faculty at Università degli Studi di Napoli Federico II, available in the R package CUB (data set CUBevaluation2008). Response categories are level of global satisfaction: 1 = neutral, 2 = not satisfied, 3 = satisfied, and 4 = very satisfied, which is only partially ordered. Explanatory variables are gender (0 = male, 1 = female), age in years, and change of faculty (1 = changed faculty, 0 = did not change faculty). Sample size is $n = 4,042$. Change was the only variable to show significant heterogeneity, which seems sensible as students who change faculty should be less certain about the new faculty. Table 4 shows the fits of the model without and with heterogeneity in the variable change. We again see that the effect of variables might be overestimated if heterogeneity is ignored. For

Table 4. Model Fits with and without Heterogeneity in Change of Faculty for Satisfaction Data

	Estimates without Heterogeneity			Estimates with Heterogeneity in Change		
	Gender	Age	Change	Gender	Age	Change
Parameters						
2	-.5343	-.0159	-.0324	-.18014	-.01490	-.6220
3	-.4218	-.0246	.1951	-.15867	-.01403	-1.1198
4	-.6420	.0320	.6118	-.24159	.00885	-1.2253
Heterogeneity	—	—				1.0593
Standard error						
2	.1953	.0283	.2532	.1338	.0117	.2939
3	.1771	.0254	.2301	.1138	.0109	.4582
4	.1723	.0242	.2250	.1542	.0117	.6082
Heterogeneity	—	—				.5797
Log likelihood		-4,144.275			-4,141.711	

example, the effects of gender are $-.53$, $-.42$, and $-.64$ if the MLM is fitted but merely $-.18$, $-.15$, and $-.24$ if the HTML is fitted.

FURTHER ISSUES

We now briefly consider marginal modeling and how models behave when applied to subsets of response categories. It is also investigated how the heterogeneous model can be used to deal with the problem of irrelevant alternatives. In a change of perspective, I emphasize that subsets of response categories refer to differing populations.

Marginal Effects and Modeling Strategies

One approach to circumvent identifiability problems uses the predicted probability metric to investigate marginal effects of predictors. Long and Mustillo (2018) show how this approach can be used to compare groups in binary regression models. With $g \in \{0, 1\}$ denoting the group the basic concept is to fit the interaction model, $P(Y = 1 | \mathbf{x}, g) = F(g \times \mathbf{x}^T \boldsymbol{\beta}^{(1)} + (1 - g) \times \mathbf{x}^T \boldsymbol{\beta}^{(0)})$, where $F(\cdot)$ is a response function (e.g., the logistic distribution function). The parameters $\boldsymbol{\beta}^{(1)}$, $\boldsymbol{\beta}^{(0)}$ represent the coefficients in the two groups. The predicted probabilities can be used to test if the groups differ in response probabilities or to investigate the marginal effects of covariates, that is, if the change in the probability of the outcome is the same in both groups when a covariate changes. Long and Mustillo gave an example that demonstrates the usefulness of the concept. The basic approach can also be used for multinomial responses by fitting MLMs separately in groups or by using a single equation model with interactions.

Using probabilities instead of focusing on coefficients is helpful if one wants to compare specific groups, but it is less appropriate as a general modeling strategy that includes potential heterogeneity. Avoiding investigating coefficients is an advantage,

but this approach has some limitations. In particular, the focus is on comparison of groups, not investigating general marginal effects. In addition, the comparisons work under the assumption that the variables that are not currently investigated are held fixed at specific values. This means conditional effects are investigated, not marginal effects in the sense of collapsing over the other variables. As Agresti and Tarantola (2018) noted, the “marginal effect” terminology is a bit misleading but seems to be in common use. Because the effects are conditional, the conclusions depend on the specific chosen values of the other variables, and it is harder to do when some of these variables are continuous. One also must select what sort of change, discrete or average, one wants to investigate.

A crucial difference with models that explicitly specify heterogeneity components is that marginal approaches use classical regression models with interactions. The model $P(Y = 1 | \mathbf{x}, g) = F(g \times \mathbf{x}^T \boldsymbol{\beta}^{(1)} + (1 - g) \times \mathbf{x}^T \boldsymbol{\beta}^{(0)})$ has linear predictors and group-specific coefficients. It allows for all interactions between the covariates and the specific group under consideration but does not use a multiplicative term. Thus, it avoids the potential instability of models with multiplicative terms (Keele and Park 2006). Heterogeneity models assume specific interactions are generated by interactions, but further interactions can always be included in the location term. The models could also be used to evaluate predicted marginal effects, although that seems not to have been done yet. As the party choice example shows, predicted probabilities for the MLM and HMLM need not be very different; both models indicate that age is relevant when modeling party choice (Figure 5).

The curves in Figure 5 show the effect of explanatory variables on the probabilities, which is a main objective of marginal approaches. This allows researchers to test if groups differ in terms of predicted probabilities. Within the heterogeneity model approach, the effect on probabilities is an effect of parameters. A variable such as age has an effect if it cannot be neglected in the location or the heterogeneity term, which can be tested. Then, the effect on probabilities is tested indirectly, but not directly as in marginal approaches. The advantage is that one does not have to condition on specific values of the other variables; the downside is that one does not use the natural marginal effect metric, which uses the probabilities. But as in binary response cases, one can compare groups regarding their marginal effects by using the predicted probabilities, as proposed by Long and Mustillo (2018) (see the Appendix for details). Simple ways to interpret effects of explanatory variables using the marginal metric may also be obtained by deriving generalizations of descriptive measures (see Agresti and Tarantola 2018).

One advantage of pure location models as used in the marginal approach is that estimates are easier to obtain than in multiplicative models. Keele and Park (2006) demonstrated this for the binary heteroskedastic model and showed that larger sample sizes are needed to obtain reliable estimates in heterogeneity models. Also, misspecification might have a stronger effect in heterogeneity models than in simple location models. This means care is needed when selecting variables in the location and the heterogeneity term. In some applications, one might suspect heterogeneity in specific variables for substantive reasons, and investigate if this suspicion is warranted. If there

are no clear candidates but one wants to account for possible heterogeneity, one must select the variables that actually contribute to improve the fit. It seems sensible to include variables in the heterogeneity term only if they have strong effects that should not be neglected. If effects are weak, models that ignore heterogeneity might be preferable. Keele and Park even argued that in some cases it might be better to estimate standard models. However, the choice certainly depends on the strengths of the effects and therefore on the concrete application.

The location term is less critical, but it can be useful to include interaction effects, which might affect the relevance of variables in the heterogeneity term. In general, model choice and therefore variable selection is harder than in classical models, because one has two terms in which variables can be present, and because of the multiplicative structure. Even in classical regression models, stepwise selection procedures have some disadvantages and have been widely replaced by selection tools that are based on penalization as the lasso and its various extensions (Tibshirani 1996; Yuan and Lin 2006; Zhao, Rocha, and Yu 2009). In future research, similar methods could be used to address selection problems in heterogeneity models using differing penalties for inclusion in the location and the heterogeneity term, but methods for this advanced form of variable selection are not yet available even for the simpler binary heteroscedastic model.

Effect Modifiers and Independence from Irrelevant Alternatives

The MLM has a property typically referred to as independence from irrelevant alternatives, which has been called a blessing and a curse (McFadden 1986). It may be seen as a blessing because if it holds, it makes it possible to infer choice behavior with multiple alternatives using data from simple experiments like paired comparisons. Yet it is a rather strict assumption that may not hold for heterogeneous patterns of similarities among alternatives. In the following, we consider how this property can be addressed by allowing for the presence of effect modifiers.

Subsets of Response Categories and the Red Bus–Blue Bus Problem. If the logit model holds, we obtain for a subset of response categories $S \subset \{1, \dots, k\}$

$$P_S(Y=r|\mathbf{x})=P(Y=r|Y \in S, \mathbf{x})=\frac{\exp(\mathbf{x}^T \boldsymbol{\beta}_r)}{\sum_{s \in S} \exp(\mathbf{x}^T \boldsymbol{\beta}_s)}, \quad r \in S,$$

which is a logit model with response categories S . Let us look at a simple problem, in which the linear term reduces to a category specific intercept $\mathbf{x}^T \boldsymbol{\beta}_r = \beta_{0r}$, which represents the utility of alternative r . Consider the “red bus–blue bus” problem (see Hausman and Wise 1978). Suppose a commuter has the initial choices of driving or taking a red bus, with the odds given by

$$\frac{P_{\{\text{driving, red bus}\}}(\text{driving})}{P_{\{\text{driving, red bus}\}}(\text{red bus})} = 1,$$

which means $\beta_{01} = \beta_{02}$ in the binary choice problem. Then, an additional choice becomes available: a blue bus that is identical in all respects, except color, to the red

bus. Let the logit model hold for the choice among the three alternatives: driving (category 1, parameter β_{01}), red bus (category 2, parameter β_{02}), and blue bus (category 3, parameter β_{03}). Choosing among subsets, the same parameters apply, and the odds of choosing between driving and the red bus, and between the red bus and the blue bus, should be the same; thus, $\beta_{01} = \beta_{02} = \beta_{03}$ and therefore

$$P_{\{1,2,3\}}(\text{driving}) = P_{\{1,2,3\}}(\text{red bus}) = P_{\{1,2,3\}}(\text{blue bus}) = 1/3.$$

This is a counterintuitive result because the additional “irrelevant” blue bus substantially decreases the choice probability of driving. Similar problems hold for all choice systems that share a property called simple scalability (see Hausman and Wise 1978; Tversky 1972).

The Presence of Effect Modifiers. The independence of irrelevant alternatives raises problems if one wants to combine results from different choice sets and one assumes the multinomial model holds. These problems can be avoided when using the heterogeneous logit model if we assume the choice is an effect modifier. We now demonstrate this for the red bus–blue bus problem.

Let us assume the heterogeneous logit model holds with predictor $\beta_{0r}e^{z\gamma}$ for the r th alternative instead of β_{0r} , as in the MLM. Let z denote an indicator for the choice set; $z = 0$ if we have three alternatives, and $z = 1$ if we have a binary choice:

$$P(Y=r|Y \in \{1, 2, 3\}) = \frac{\exp(\beta_{0r})}{\sum_{s=1}^3 \exp(\beta_{0s})},$$

and

$$P(Y=r|Y \in \{r, 1\}) = \frac{\exp((\beta_{0r} - \beta_{01})e^{z\gamma})}{(1 + \exp((\beta_{0r} - \beta_{01})e^{z\gamma}))}, r=2, 3,$$

$$P(Y=2|Y \in \{2, 3\}) = \frac{\exp((\beta_{02} - \beta_{03})e^{z\gamma})}{(1 + \exp((\beta_{02} - \beta_{03})e^{z\gamma}))}.$$

With $\beta_{01} = 0, \beta_{02} = \beta_{03} = -0.693$, we obtain for the full choice set

$$P(Y=1|Y \in \{1, 2, 3\}) = 0.5,$$

$$P(Y=2|Y \in \{1, 2, 3\}) = P(Y=3|Y \in \{1, 2, 3\}) = 0.25,$$

which are sensible values if one chooses from the three alternatives. For binary choices, we have $z = 1$ and therefore obtain

$$P(Y=2|Y \in \{2, 3\}) = 0.5,$$

because $\beta_{02} = \beta_{03}$, and with $\gamma = -5$

$$P(Y=r|Y \in \{r, 1\}) = 0.499, r=2, 3.$$

Thus, if we allow z to indicate the setting of alternatives presented, we obtain much more sensible probabilities in the red bus–blue bus problem. The variable z acts as an

effect modifier, which determines the choice probabilities as a function of the setting. It is an explanatory variable that must be included. Of course, if the effects of an explanatory variable are to be investigated, one needs some variation of the explanatory variable when collecting data. In the present case, that means it is not sufficient to collect data from the setting with three alternatives; one must also have data from settings with two alternatives. In more general cases, z can be seen as a factor that represents the alternatives that are presented with the values of the γ parameter varying across the factor levels. In sociology, the objective is often to analyze the response behavior in questionnaires without assuming that relationships would be identical if response options differed. Then no alternative settings are needed, and the problem of irrelevant alternatives is of no relevance.

Prior approaches to address the problem of similar alternatives in the choice set use more general distributions in the underlying latent trait model. In particular, the nested logit model and more general models based on the generalized extreme-value distribution have been developed (McFadden 1978, 1981). These models are derived from underlying random utilities but have the disadvantage that one must specify beforehand which alternatives are to be considered similar; they are used mainly in transportation research (see, e.g., Cai et al. 2000; Wen and Koppelman 2001) and less to analyze questionnaire data in the social sciences. Olsen (1982) proposed an alternative approach to address the problem.

The Heterogeneous Logit Model in Subpopulations

The red bus–blue bus problem arises if people must choose from different subsets of alternatives. The question is what can be inferred from the choice of categories if a different set of categories has been presented earlier. In the previous section, it was shown that it might be sensible to include the choice set in the predictors as heterogeneity components.

Subsets of categories can also be seen from a different view: the problem is not the transfer to other presented categories, but if models and parameters are the same given that one *fits* models to varying subsets. If the heterogeneous logit model holds for k categories, it should also hold if one considers the conditional response $Y|Y \in S$ for a subset of categories S . However, that means one fits a model to a subpopulation, namely the subpopulation with chosen categories from S . Although the basic preference for categories captured by the location term should not change too much, the heterogeneity component, which is not category specific, might differ in subpopulations.

Let us consider the party choice data, in which there were five parties—the CDU (category 1), the SPD (category 2), the FDP (category 3), the Green Party (category 4), and the Left Party (category 5)—and we found heterogeneity for the variables west and age. If we fit the model in a reduced set, say the first three parties, we have a different population, because we exclude everyone who tends to strongly favor left-wing parties or is strongly interested in ecological issues. Thus heterogeneity might differ.

We briefly consider the variation in estimates for the party data set. Table 5 shows estimates of the heterogeneity parameters for the variables west and age for varying

Table 5. Heterogeneity in Variables West and Age for Varying Response Categories in Party Choice Data

Parties in Response	$e^{\gamma_{west}}$	$e^{\gamma_{age}}$
CDU, SPD, FDP	1.93	1.99
CDU, SPD, FDP, Greens	4.34	1.36
CDU, SPD, FDP, Greens, Left Party	3.26	1.63

sets of response categories. They are given in the exponential form $e^{\gamma_{variable}}$, which represents the multiplicative modification of the parameters in the location term if the heterogeneity variable changes by one unit.

Although there is some variation in estimates, the tendency is the same in all subpopulations: people living in western Germany have a stronger tendency to specific categories than do people from the East, and the same holds for older versus younger respondents. Heterogeneity for the variable age is comparatively stable across subsets, but there is some variation in heterogeneity linked to the variable west.

CONCLUDING REMARKS

The proposed heterogeneous logit model is able to account for heterogeneity that is typically ignored in MLMs. This heterogeneity can be seen as unobserved variance heterogeneity but also as representing uncertainty without reference to latent variables. The model contains multiplicative terms, which are typically harder to estimate than simple linear terms and show greater variability. Models of this type have been criticized in the binary case because they are less stable than simple binary logit models (Kuha and Mills 2017). However, this is to be expected. Estimation of variance is usually harder to do than estimation of location. The alternative, ignoring heterogeneity, yields stable estimates, but they can be severely biased. Therefore, it seems worthwhile to account for potential heterogeneity.

Nevertheless, stable estimation of variance components typically calls for larger data sets. In small data sets, they are hard to identify and estimate. Fortunately, they often turn out to be negligible and can be ignored. For example, in a data set that contains high school students' choices among general, vocational, or academic programs from the UCLA Statistical Consulting site ($n = 200$; <https://stats.idre.ucla.edu/stat/data/hsbdemo.dta>) and a data set on absenteeism from school in rural New South Wales from R package MASS ($n = 146$), heterogeneity can be ignored. However, in some larger data sets, heterogeneity is also not needed, for example, Agresti's (2013:330) political party identification data set, even though the sample size was $n = 1,001$. These applications have not been included, but are mentioned because they illustrate that the modeling of heterogeneity is not always needed.

In principle, one could allow for category-specific heterogeneity terms letting uncertainty depend on alternatives. One disadvantage is that one loses the derivation from the random utilities. More seriously, the number of parameters would be much higher

and stability of estimates would suffer. Therefore, we abstained from considering the more general model with alternative dependent heterogeneity terms.

In the applications, an R program was used. The code for fitting the models will be made available on GitHub (GerhardTutz/GHMNL).

APPENDIX

Extreme Heterogeneity

Let us consider the heterogeneous logit model in the form

$$\log\left(\frac{P(Y=r|\mathbf{x}, \mathbf{z})}{P(Y=1|\mathbf{x}, \mathbf{z})}\right) = \mathbf{x}^T \boldsymbol{\beta}_r e^{z^T \boldsymbol{\gamma}},$$

which implies that $\boldsymbol{\beta}_1^T = (0, \dots, 0)$. We examine which probabilities are obtained if one of the heterogeneity parameters tends to infinity. In the following proposition, we distinguish between several cases. If no alternative is preferred ($\boldsymbol{\beta}_r^T = (0, \dots, 0)$ for all r), the heterogeneity has no effect. However, if parameters differ, it depends on the values $\mathbf{x}^T \boldsymbol{\beta}_r$, if one category tends to probability one, or several of them tend to the same positive probabilities, and the rest get probability zero.

More concretely, if $\gamma_j \rightarrow \infty$ for one of the components in $\boldsymbol{\gamma}^T = (\gamma_1, \dots, \gamma_m)$, and $z_j > 0$ from $\mathbf{z}^T = (z_1, \dots, z_m)$, we obtain the following:

- (a) If $\boldsymbol{\beta}_1 = \dots = \boldsymbol{\beta}_k$ holds no category is preferred and $P(Y=r|\mathbf{x}, \mathbf{z}) = 1/k$.
- (b) If there is one category r_0 , for which $\mathbf{x}^T \boldsymbol{\beta}_{r_0} > \mathbf{x}^T \boldsymbol{\beta}_r$ for all r , we obtain $P(Y=r_0|\mathbf{x}, \mathbf{z}) = 1$.
- (c) If there is a group of categories S , such that $\mathbf{x}^T \boldsymbol{\beta}_{r_0} > \mathbf{x}^T \boldsymbol{\beta}_r$ for $r_0 \in S, r \notin S$, and $\mathbf{x}^T \boldsymbol{\beta}_{r_0} = \mathbf{x}^T \boldsymbol{\beta}_{r_1}$ for $r_0, r_1 \in S$, we obtain $P(Y=r_0|\mathbf{x}, \mathbf{z}) = 1/|S|$, where $|S|$ is the cardinality of S .

Proof: (i) If all β parameters are zero, $P(Y=1|\mathbf{x}, \mathbf{z}) = \dots = P(Y=k|\mathbf{x}, \mathbf{z})$ follows directly from the model.

(ii) Let S be a subset of categories such that $\mathbf{x}^T \boldsymbol{\beta}_{r_0} > \mathbf{x}^T \boldsymbol{\beta}_r$ for $r_0 \in S, r \notin S$.

Then we have

$$\log\left(\frac{P(Y=r_0|\mathbf{x}, \mathbf{z})}{P(Y=r|\mathbf{x}, \mathbf{z})}\right) = (\mathbf{x}^T \boldsymbol{\beta}_{r_0} - \mathbf{x}^T \boldsymbol{\beta}_r) e^{\mathbf{z}^T \tilde{\boldsymbol{\gamma}}} e^{z_j^T \gamma_j},$$

where $\tilde{\mathbf{z}}$ is the \mathbf{z} -vector without the j th component and $\tilde{\boldsymbol{\gamma}}$ is the corresponding parameter vector.

Because $\mathbf{x}^T \boldsymbol{\beta}_{r_0} - \mathbf{x}^T \boldsymbol{\beta}_r$ is positive, we obtain for $\gamma_j \rightarrow \infty$ that $\log(P(Y=r_0|\mathbf{x}, \mathbf{z})/P(Y=r|\mathbf{x}, \mathbf{z})) \rightarrow \infty$ and therefore $P(Y=r|\mathbf{x}, \mathbf{z}) \rightarrow 0$.

For two categories from S , $r_0, r_1 \in S$, we have $\mathbf{x}^T \boldsymbol{\beta}_{r_0} - \mathbf{x}^T \boldsymbol{\beta}_{r_1} = 0$, and therefore $\log(P(Y=r_0|\mathbf{x}, \mathbf{z})/P(Y=r_1|\mathbf{x}, \mathbf{z})) = 0$ yielding $P(Y=r_0|\mathbf{x}, \mathbf{z}) = P(Y=r_1|\mathbf{x}, \mathbf{z})$. That means for all categories from S the probability is the same. In the special case where S contains just one category, we obtain the result in (b).

Identifiability of Parameters

We now consider the heterogeneous MLM given in equation (2).

- (1) First, we look at the more difficult case $\mathbf{x} = \mathbf{z}$. Let the predictors in the model be given by two sets of parameters, $\beta_1, \dots, \beta_k, \boldsymbol{\gamma}$ and $\tilde{\beta}_1, \dots, \tilde{\beta}_k, \tilde{\boldsymbol{\gamma}}$. The logits, defined by $\text{logit}_r(\mathbf{x}) = P(Y=r|\mathbf{x})/P(Y=1|\mathbf{x})$, are given by $\text{logit}_r(\mathbf{x}) = \{\boldsymbol{\beta}_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r\} e^{\mathbf{x}^T \boldsymbol{\gamma}} = \{\tilde{\boldsymbol{\beta}}_{r0} + \mathbf{x}^T \tilde{\boldsymbol{\beta}}_r\} e^{\mathbf{x}^T \tilde{\boldsymbol{\gamma}}}$, where the intercepts are explicitly included, and $\beta_{10} = \tilde{\beta}_{10} = 0, \boldsymbol{\beta}_1 = \tilde{\boldsymbol{\beta}}_1 = \mathbf{0}$. Because linear transformation of explanatory variables does change the parameters but not the validity of the model, we can, without restriction of generality, assume that each covariate contains the values 0 and 1, which is natural for binary predictors but can also be assumed for continuous variables after centering and scaling. Let the number of predictors p be larger than 1, and at least one variable have no dispersion effect, $\gamma_j = 0$, and $\mathbf{x}^T = (x_1, \dots, x_p)$ denote a vector of explanatory variables.

- (a) We obtain for the differences

$$d_r(\mathbf{x}) = \text{logit}_r(x_1, \dots, x_p) - \text{logit}_r(x_1, \dots, x_j + 1, \dots, x_p) = \beta_{rj} e^{\mathbf{x}^T \boldsymbol{\gamma}}.$$

Let for $s \neq j$ $\mathbf{x}_s = (0, \dots, 1, \dots, 0)$ denote the s th unit vector, which has a 1 in the s th component:

$$\frac{d_r(\mathbf{x}_s)}{d_r(\mathbf{0})} = \frac{\beta_{rj} e^{\gamma_s}}{\beta_{rj}} = e^{\gamma_s},$$

where $\mathbf{0}^T = (0, \dots, 0)$. Thus, $\gamma_s, s = 1, \dots, p$ is determined by the probabilities, which holds for any parameterization yielding $\gamma_s = \tilde{\gamma}_s$.

- (b) In

$$\text{logit}_r(\mathbf{x}) = \{\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r\} e^{\mathbf{x}^T \boldsymbol{\gamma}} = \{\tilde{\beta}_{r0} + \mathbf{x}^T \tilde{\boldsymbol{\beta}}_r\} e^{\mathbf{x}^T \tilde{\boldsymbol{\gamma}}}, r = 1, \dots, k$$

we set $\mathbf{x} = \mathbf{0}$ so that $\boldsymbol{\beta}_{r0} = \tilde{\beta}_{r0}$ holds for $r = 1, \dots, k$.

- (c) For the i th unit vector $\mathbf{x}_i = (0, \dots, 1, \dots, 0)$, we have

$$\text{logit}_r(\mathbf{x}_i) - \text{logit}_r(\mathbf{0}) = (\beta_{r0} + \beta_{ri}) e^{\gamma_i} - \beta_{r0} = \beta_{r0} (e^{\gamma_i} - 1) + \beta_{ri} e^{\gamma_i}.$$

Thus, for the two parameterizations

$$\beta_{r0} (e^{\gamma_i} - 1) + \beta_{ri} e^{\gamma_i} = \tilde{\beta}_{r0} (e^{\tilde{\gamma}_i} - 1) + \tilde{\beta}_{ri} e^{\tilde{\gamma}_i}.$$

Because $\boldsymbol{\gamma} = \tilde{\boldsymbol{\gamma}}$ and $\beta_{r0} = \tilde{\beta}_{r0}$ has been shown to hold, we have $\beta_{ri} = \tilde{\beta}_{ri}$, which concludes the proof.

- (2) In the case where \mathbf{x} and \mathbf{z} are distinct, we have for the differences considered in (a)

$$d_r(\mathbf{x}, \mathbf{z}) = \text{logit}_r(x_1, \dots, x_p, \mathbf{z}) - \text{logit}_r(x_1, \dots, x_j + 1, \dots, x_p, \mathbf{z}) = \beta_{rj} e^{\mathbf{z}^T \boldsymbol{\gamma}}.$$

With $\mathbf{z}^T = (z_1, \dots, z_m)$, one obtains

$$\text{logit}_r(\mathbf{x}, z_1, \dots, z_l + 1, \dots, z_m) = \{\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r\} e^{\mathbf{z}^T \boldsymbol{\gamma} + \gamma_l}$$

and

$$\text{logit}_r(\mathbf{x}, \mathbf{z}) = \{\beta_{r0} + \mathbf{x}^T \boldsymbol{\beta}_r\} e^{\mathbf{z}^T \boldsymbol{\gamma}},$$

yielding

$$\frac{\text{logit}_r(\mathbf{x}, z_1, \dots, z_l + 1, \dots, z_m)}{\text{logit}_r(\mathbf{x}, \mathbf{z})} = e^{\gamma_l}.$$

Thus $\boldsymbol{\gamma}$, and therefore also β_{rj} for all r, j are identifiable.

Obtaining Estimates

The HMLM for observations $i = 1, \dots, n$ has the form

$$\pi_{ir} = P(Y_i = r | \mathbf{x}_i, \mathbf{z}_i) = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_r e^{\mathbf{z}_i^T \boldsymbol{\gamma}})}{\sum_{s=1}^k \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s e^{\mathbf{z}_i^T \boldsymbol{\gamma}})}, \quad j = 1, \dots, k. \tag{6}$$

The response given covariates follows a multinomial distribution given by the vector $\mathbf{y}_i^T = (y_{i1}, \dots, y_{ik})$, in which a response in category r is represented by $y_{ir} = 1$ and $y_{ij} = 0$ for $j \neq r$. With parameters $\boldsymbol{\beta}_r$ collected in $\boldsymbol{\beta}^T = (\boldsymbol{\beta}_2, \dots, \boldsymbol{\beta}_k)$, p -dimensional vector \mathbf{x}_i , and m -dimensional vector \mathbf{z}_i , we obtain for the kernel of the log-likelihood

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\gamma}) &= \sum_{i=1}^n \left\{ \sum_{r=2}^k y_{ir} \log \left(\frac{\pi_{ir}}{1 - \pi_{i2} - \dots - \pi_{ik}} \right) + \log(1 - \pi_{i2} - \dots - \pi_{ik}) \right\} \\ &= \sum_{i=1}^n \left\{ \sum_{r=2}^k y_{ir} (\mathbf{x}_i^T \boldsymbol{\beta}_r) e^{\mathbf{z}_i^T \boldsymbol{\gamma}} - \log \left(1 + \sum_{s=2}^k \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s e^{\mathbf{z}_i^T \boldsymbol{\gamma}}) \right) \right\}. \end{aligned}$$

When maximizing the log-likelihood it is helpful to use the first derivatives, also known as score functions. They are given by


$$\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial \beta_{rj} = \sum_{i=1}^n \left\{ y_{ij} x_{ij} e^{\mathbf{z}_i^T \boldsymbol{\gamma}} - \frac{x_{ij} e^{\mathbf{z}_i^T \boldsymbol{\gamma}} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_r e^{\mathbf{z}_i^T \boldsymbol{\gamma}})}{1 + \sum_{s=2}^k \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s e^{\mathbf{z}_i^T \boldsymbol{\gamma}})} \right\},$$

for $r = 2, \dots, k, j = 1, \dots, p$, and

$$\partial l(\boldsymbol{\beta}, \boldsymbol{\gamma}) / \partial \gamma_j = \sum_{i=1}^n \left\{ \sum_{r=2}^k y_{ir} (\mathbf{x}_i^T \boldsymbol{\beta}_r) e^{\mathbf{z}_i^T \boldsymbol{\gamma}} z_{ij} - \frac{\sum_{s=2}^k z_{ij} \mathbf{x}_i^T \boldsymbol{\beta}_s e^{\mathbf{z}_i^T \boldsymbol{\gamma}} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s e^{\mathbf{z}_i^T \boldsymbol{\gamma}})}{1 + \sum_{s=2}^k \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s e^{\mathbf{z}_i^T \boldsymbol{\gamma}})} \right\},$$

for $j = 1, \dots, m$. As approximation of the covariance $\text{cov}(\hat{\boldsymbol{\delta}})$ of the total vector $\boldsymbol{\delta}^T = (\boldsymbol{\beta}^T, \boldsymbol{\gamma}^T)$, we use the observed information $-\partial^2 l(\hat{\boldsymbol{\delta}}) / \partial \boldsymbol{\delta} \partial \boldsymbol{\delta}^T$.

ORCID iD

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

References

- Agresti, Alan. 2013. *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley.
- Agresti, Alan, and Cladia Tarantola. 2018. "Simple Ways to Interpret Effects in Modeling Ordinal Categorical Data." *Statistica Neerlandica* 72(3):210–23.
- Allison, Paul D. 1999. "Comparing Logit and Probit Coefficients across Groups." *Sociological Methods & Research* 28(2):186–208.
- Antoniadis, Anestis, Irène Gijbels, and Anneleen Verhasselt. 2012. "Variable Selection in Varying-Coefficient Models Using p-Splines." *Journal of Computational and Graphical Statistics* 21(3): 638–61.
- Baumgartner, Hans, and Jan-Benedict E. M. Steenkamp. 2001. "Response Styles in Marketing Research: A Cross-National Investigation." *Journal of Marketing Research* 38(2):143–56.
- Breen, Richard, Anders Holm, and Kristian Bernt Karlson. 2014. "Correlations and Nonlinear Probability Models." *Sociological Methods & Research* 43(4):571–605.
- Cai, Zongwu, Jianqing Fan, and Runze Li. 2000. "Efficient Estimation and Inferences for Varying-Coefficient Models." *Journal of the American Statistical Association* 95(451):888–902.
- D'Elia, Angela, and Domenico Piccolo. 2005. "A Mixture Model for Preference Data Analysis." *Computational Statistics & Data Analysis* 49:917–34.
- Falk, Carl F., and Li Cai. 2016. "A Flexible Full-Information Approach to the Modeling of Response Styles." *Psychological Methods* 21(3):328.
- Hastie, Trevor, and Robert Tibshirani. 1993. "Varying-Coefficient Models." *Journal of the Royal Statistical Society B* 55(4):757–96.
- Hausman, Jerry A., and David A. Wise. 1978. "A Conditional Probit Model for Qualitative Choice: Discrete Decisions Recognizing Interdependence and Heterogeneous Preference." *Econometrica* 46(2):403–26.
- Iannario, Maria. 2012. "Hierarchical CUB Models for Ordinal Variables." *Communications in Statistics—Theory and Methods* 41(16–17):3110–25.
- Iannario, Maria, Marica Manisera, Domenico Piccolo, and Paola Zuccolotto. 2020. "Ordinal Data Models for No-Opinion Responses in Attitude Survey." *Sociological Methods & Research* 49(1): 250–76.
- Iannario, Maria, and Domenico Piccolo. 2010. "Statistical Modelling of Subjective Survival Probabilities." *Genus* 66(2):17–42.
- Johnson, Timothy R., and Daniel M. Bolt. 2010. "On the Use of Factor-Analytic Multinomial Logit Item Response Models to Account for Individual Differences in Response Style." *Journal of Educational and Behavioral Statistics* 35(1):92–114.
- Karlson, Kristian Bernt, Anders Holm, and Richard Breen. 2012. "Comparing Regression Coefficients between Same-Sample Nested Models Using Logit and Probit: A New Method." *Sociological Methodology* 42(1):286–313.
- Keele, Luke, and David K. Park. 2006. "Difficult Choices: An Evaluation of Heterogeneous Choice Models." Presented at the 2004 Annual Meeting of the American Political Science Association.
- Kuha, Jouni, and Colin Mills. 2017. "On Group Comparisons with Logistic Regression Models." *Sociological Methods & Research* 49(2):498–525.
- Long, J. Scott, and Sarah A. Mustillo. 2018. "Using Predictions and Marginal Effects to Compare Groups in Regression Models for Binary Outcomes." *Sociological Methods & Research*. Retrieved December 7, 2020. <https://journals.sagepub.com/doi/10.1177/0049124118799374>.
- McCullagh, Peter. 1980. "Regression Model for Ordinal Data (with Discussion)." *Journal of the Royal Statistical Society B* 42:109–27.
- McFadden, Daniel. 1973. "Conditional Logit Analysis of Qualitative Choice Behaviour." In *Frontiers in Econometrics*, edited by P. Zarembka. New York: Academic Press.

- McFadden, Daniel. 1978. "Modelling the Choice of Residential Location." In *Spatial Interaction Theory and Residential Location*, edited by A. Karlquist et al. Amsterdam, the Netherlands: North-Holland.
- McFadden, Daniel. 1981. "Econometric Models of Probabilistic Choice." Pp. 198–272 in *Structural Analysis of Discrete Data with Econometric Applications*, edited by C. F. Manski and D. McFadden. Cambridge, MA: MIT Press.
- McFadden, Daniel. 1986. "The Choice Theory Approach to Market Research." *Marketing Science* 5(4): 275–97.
- Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26(1):67–82.
- Neuhaus, John M., J. D. Kalbfleisch, and W. W. Hauck. 1991. "A Comparison of Cluster-Specific and Population-Averaged Approaches for Analyzing Correlated Binary Data." *International Statistical Review* 59:25–35.
- Olsen, Randall J. 1982. "Independence from Irrelevant Alternatives and Attrition Bias: Their Relation to One Another in the Evaluation of Experimental Programs." *Southern Economic Journal* 49(2):521–35.
- Park, Beyeong U., Enno Mammen, Young K. Lee, and Eun Ryung Lee. 2015. "Varying Coefficient Regression Models: A Review and New Developments." *International Statistical Review* 83(1):36–64.
- Piccolo, Domenico, and Rosaria Simone. 2019. "The Class of CUB Models: Statistical Foundations, Inferential Issues and Empirical Evidence." *Statistical Methods and Applications* 28:389–435.
- Rohwer, Goetz. 2015. "A Note on the Heterogeneous Choice Model." *Sociological Methods & Research* 44(1):145–48.
- Schauberger, Gunther. 2019. "EffectStars: Visualization of Categorical Response Models." Retrieved December 7, 2020. <https://cran.r-project.org/web/packages/EffectStars/index.html>.
- Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society B* 58(1):267–88.
- Tutz, Gerhard. 2018. "Binary Response Models with Underlying Heterogeneity: Identification and Interpretation of Effects." *European Sociological Review* 34(2):211–21.
- Tutz, Gerhard. 2020. "Modelling Heterogeneity: On the Problem of Group Comparisons with Logistic Regression and the Potential of the Heterogeneous Choice Model." *Advances in Data Analysis and Classification* 14:517–42.
- Tutz, Gerhard, and Moritz Berger. 2016. "Response Styles in Rating Scales: Simultaneous Modelling of Content-Related Effects and the Tendency to Middle or Extreme Categories." *Journal of Educational and Behavioral Statistics* 41:239–68.
- Tutz, Gerhard, and Moritz Berger. 2017. "Separating Location and Dispersion in Ordinal Regression Models." *Econometrics and Statistics* 2:131–48.
- Tutz, Gerhard, Micha Schneider, Maria Iannario, and Domenico Piccolo. 2017. "Mixture Models for Ordinal Responses to Account for Uncertainty of Choice." *Advances in Data Analysis and Classification* 11(2):281–305.
- Tversky, Amos. 1972. "Elimination by Aspects: A Theory of Choice." *Psychological Review* 79(4): 281–99.
- Van Vaerenbergh, Yves, and Troy D. Thomas. 2013. "Response Styles in Survey Research: A Literature Review of Antecedents, Consequences, and Remedies." *International Journal of Public Opinion Research* 25(2):195–217.
- Wen, Chieh-Hua, and Frank S. Koppelman. 2001. "The Generalized Nested Logit Model." *Transportation Research Part B: Methodological* 35(7):627–41.
- Wetzel, Eunike, and Claus H. Carstensen. 2017. "Multidimensional Modeling of Traits and Response Styles." *European Journal of Psychological Assessment* 33(5):352–64.
- Williams, Richard. 2009. "Using Heterogeneous Choice Models to Compare Logit and Probit Coefficients across Groups." *Sociological Methods & Research* 37(4):531–59.
- Williams, Richard. 2010. "Fitting Heterogeneous Choice Models with *oglm*." *Stata Journal* 10(4):540–67.
- Yellott, John I. 1977. "The Relationship between Luce's Choice Axiom, Thurstone's Theory of Comparative Judgement, and the Double Exponential Distribution." *Journal of Mathematical Psychology* 15:109–44.

- Yuan, Ming, and Yi Lin. 2006. "Model Selection and Estimation in Regression with Grouped Variables." *Journal of the Royal Statistical Society B* 68(1):49–67.
- Zhao, Peng, Guiherme Rocha, and Bin Yu. 2009. "The Composite Absolute Penalties Family for Grouped and Hierarchical Variable Selection." *Annals of Statistics* 37(6A):3468–97.

Author Biography

Gerhard Tutz is a professor of statistics at Ludwig-Maximilians-University, Munich. Current research interests include the modeling of categorical data, latent trait models, and discrete survival analysis. He has published various books in methodological and applied journals.