Reliable inference in categorical regression analysis for non-randomly coarsened observations

Julia Plass[1*] Marco E.G.V. Cattaneo[2] Thomas Augustin[1]
Georg Schollmeyer[1] Christian Heumann[1]

[1] *Department of Statistics, LMU Munich, Ludwigstr. 33, 80539 Munich, Germany*

[2] *Department of Clinical Research, University of Basel, Spitalstr. 12, 4031 Basel, Switzerland*

[*] corresponding author: julia.plass@stat.uni-muenchen.de

Summary

In most surveys, one is confronted with missing or, more generally, coarse data. Traditional methods dealing with these data require strong, untestable and often doubtful assumptions, e.g. coarsening at random. But due to the resulting, potentially severe bias, there is a growing interest in approaches that only include tenable knowledge about the coarsening process, leading to imprecise, but reliable results. In this spirit, we study regression analysis with a coarse categorical dependent variable and precisely observed categorical covariates. Our (profile) likelihood-based approach can incorporate weak knowledge about the coarsening process, and thus offers a synthesis of traditional methods and cautious strategies refraining from any coarsening assumptions. This also allows a discussion of the uncertainty about the coarsening process, besides sampling uncertainty and model uncertainty. Our procedure is illustrated with data of the panel study "Labour market and social security" conducted by the Institute for Employment Research, whose questionnaire design produces coarse data.

Key words: coarse data, (cumulative) logit model, missing data, partial identification, PASS data, (profile) likelihood

1 Introduction

In almost all surveys the problem of item-nonresponse occurs (e.g. Jackson et al., 2010; Sikov, 2018). In the case of missing data, the extent of the bias and the opportunities to correct for it strongly depend on the underlying missingness mechanism (as introduced by Rubin, 1976). Therefore, one of the fundamental challenges in the statistical analysis of missing data is the restricted possibility to test the commonly assumed missingness mechanisms without adding strong assumptions: While a test for missing completely at random can already be found in Little (1988), it is fundamentally impossible to distinguish between missing at random and missing not at random (e.g. Jaeger, 2006). Despite the awareness of this problem, untestable assumptions on the missingness process are still frequently included in situations where their validity might actually be doubtful. Examples beyond the missing at random assumption are approaches relying on a specific pattern-mixture or selection model (e.g. developed by Heckman, 1976). In this way, point-identifiability, i.e. uniqueness of parameters of the probability law underlying the observable data, is spuriously forced, because it is an important prerequisite for the applicability of traditional statistical methods for handling missing data, as for instance the EM algorithm or imputation techniques (e.g. Chen and Haziza, 2018; Little and Rubin, 2014; Zhang, 2003).

Especially due to the potentially substantial bias induced by wrongly imposing such point-identifying assumptions and the lack of indicators fully expressing the impact of nonresponse in survey estimates (e.g. Nishimura et al., 2016), a proper reflection of the available information about the underlying missingness assumption is indispensable (e.g. Manski, 2003). To this end, one departs from insisting on point-identifying assumptions by turning to strategies that only include the achievable knowledge, typically resulting in set-valued estimators. In this way, approaches based on the methodology of partial identification start with no missingness assumptions at all, but then add successively assumptions compatible with the obtainable knowledge (e.g. Manski, 2003). Similarly, sensitivity analyses for selection models (e.g. Kenward et al., 2001; Verbeke and Molenberghs, 2009) take several different models of missing data processes into account. Further approaches avoiding strong assumptions about the missingness mechanism are proposed in Cattaneo and Wiencierz (2012), Denœux (2014) and Zhang (2010). Recently, Manski (2015) gave a new impetus to this topic by stressing the advantage

of interval-valued point estimates for official statistics with survey nonresponse.

Against this background, we refer to the approach by Zhang (2010) (cf. also Cattaneo and Wiencierz (2012)), who uses the profile likelihood to describe statistical evidence with missing data without imposing any assumptions on the missingness process. We generalize this approach in three ways:

1. We study coarse data as a generalization of missing data.

2. We offer a way to incorporate auxiliary knowledge about the coarsening process as a generalization of the cautious case refraining from any coarsening assumptions.

3. We consider categorical regression analysis as a generalization of the categorical i.i.d. case.

In this way, we do not restrict ourselves to the issue of nonresponse, but also look at the problem of partially observed values where subsets of the full state space are observed, also referred to as the coarse data problem (e.g. Heitjan and Rubin, 1991). An advantage of our approach is the capability to make use of potentially available partial knowledge about the coarsening process, which now no longer has to be left out of consideration (as in approaches that are based on strict assumptions as well as in cautious approaches that rely on no coarsening assumptions at all). In particular, we have the opportunity to consider "near coarsening at random" instead of the strong "(exact) coarsening at random" models, and thus to improve the credibility of traditional approaches.

The need of such an approach quantifying the underlying uncertainty due to data incompleteness is also apparent from the following recent practical example, intensively discussed in the German media: Results of a survey on the job-seeking refugees in Germany were published by the Federal Employment Agency and provoked a heated debate, mainly due to different ways of dealing with item-nonresponse concerning the possession of a school-leaving qualification. While ignoring the 24.7% nonrespondents leads to the result that 34.3% job-seeking refugees are without school-leaving qualification and assumes the refusals to answer to be made randomly, the newspaper "Bild" disseminated an extreme interpretation of the Federal Institute for Vocational Education and Training's conjecture that job-seeking refugees without school-leaving qualification rather tend to refuse to answer and simply counted all nonrespondents to this group, hence speaking of 59% in this context (cf. e.g. Hoeren, 2017; Brack, 2017). A

clear communication of the underlying uncertainty would have avoided some discussions and should generally be part of every trustworthy data analysis. As a reaction to the incident several statistical agencies pointed to the importance of reflecting on the reasons why the respondents refused their answers (cf. e.g. Brücker and Schupp, 2017). The approach presented in this paper is able to express the underlying uncertainty attributed to nonresponse and could utilize weak, but tenable knowledge about the coarsening obtained from the main reasons for nonresponse.

The main goal of this paper is the estimation of the coefficients of categorical regression models with coarse data, without including strong, untenable assumptions about the coarsening. Motivated by two examples regarding the income questions from the panel study "Labour market and social security" (PASS, Trappmann et al., 2010) conducted by the Institute for Employment Research (i.e. the Research Institute of the (German) Federal Employment Agency), the focus is set on the logit model for binary response data and the cumulative logit model for ordinal response data. Throughout the paper, we restrict ourselves to cases of coarse categorical (including nominal and ordinal scale) response variables and precisely observed categorical covariates.

Besides the uncertainty about the coarsening process, we discuss some aspects of sampling uncertainty and model uncertainty in the context of coarse data. The impact of model assumptions in the context of incomplete data is an important aspect of study in literature (cf. e.g. Hüllermeier, 2014; Sánchez and Couso, 2018; Schollmeyer and Augustin, 2015). We study in particular the effect of the regression model assumptions on the estimated coarsening parameters.

Our paper is structured as follows: In Section 2 we motivate the collection of coarse data and introduce the running example based on the PASS data. Afterwards, in Section 3 we explain the way we look at the problem and elaborate the profile likelihood approach that is used here in order to determine regression coefficients that are basically reliant upon no coarsening assumptions at all. This approach is refined in Section 4 in order to be able to utilize also weak knowledge about the coarsening process. After having addressed the uncertainty due to coarse data, in Section 5 we also consider sampling uncertainty and discuss the uncertainty inherent in the parametric assumptions of the regression model in this coarse data context. Section 6

concludes by giving a summary and some remarks on further research.

2 Coarse categorical data: The running example

In most surveys, respondents can choose their answers between several predetermined options. Nevertheless, providing answers associated to a specific level of accuracy may be considered problematic for different reasons: First, respondents might be able to give a more precise answer, but there is no possibility to express it. Second, the other way round, respondents potentially may at most be able to decide for a *set* of categories, but not for the one category they actually belong to, since they are not acquainted enough with the topic of the question. Third, respondents may deliberately refuse their precise answer for reasons of data privacy. While the consequence in the first situation is (only) loss of information, in the second and third situation non-ignorable nonresponse or measurement errors occur in a classical questionnaire design.

All these problems could be attenuated by asking in different ways allowing the respondent to answer with the desired level of accuracy. For example, an explicit collection of coarse categories is used in the panel study "Labour market and social security" (PASS study, Trappmann et al., 2010) conducted by the Institute for Employment Research. Here the question about income, which is known to be highly sensitive (e.g. Tourangeau and Yan, 2007), is asked by means of the following questioning technique illustrated in Figure 1: Respondents refusing to disclose their precise income (in the following called nonrespondents) are asked to answer additional questions starting from providing rather wide income classes (e.g. < 1000 € or not) that are successively narrowed (e.g. < 500 €, between 500 € and 750 €, or between 750 € and 1000 €). In this way, answers with different levels of coarseness are received by simultaneously ensuring the individual degree of data privacy demanded by the respondents. This strategic questioning technique to increase response rates is sometimes referred to as non-response follow-up (e.g. Olson, 2013, where this is distinguished from "follow-up attempts", i.e. repeated efforts to contact respondents). Depending on the research question, various ways to integrate the answers from the respondents reporting their precise, non-categorical income are conceivable. We first point to some general options before we mention how we proceed here: To include all answers in the most precise level inferable from the data, a mixture model (e.g. McLachlan

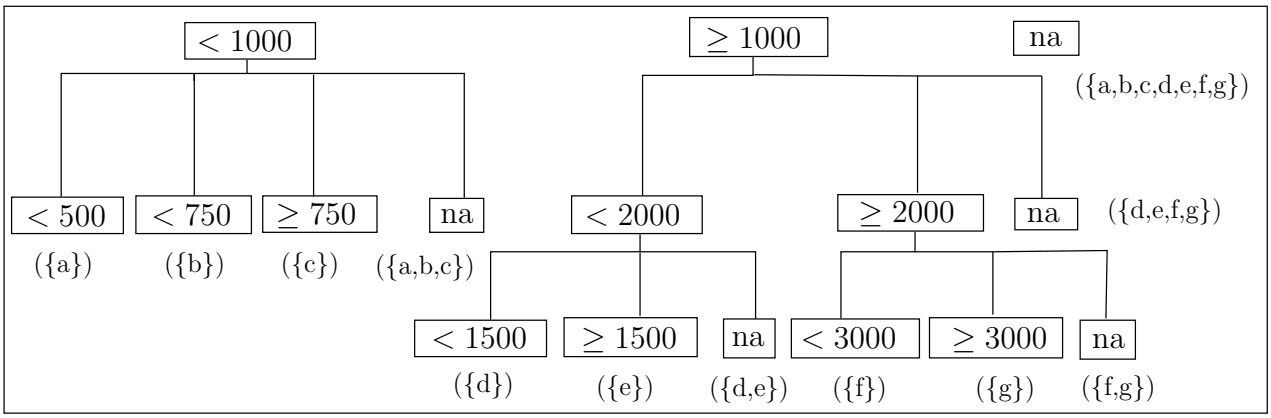

Figure 1: In the PASS study, for nonrespondents the income questions are individually adjusted, providing for instance categories abbreviated by "$<$ 500 €", "$<$ 750 €" (actually meaning $<$ 750 € and $\geq$ 500 €) and "$\geq$ 750 €" ($\geq$ 750 € and $<$ 1000 €) to original nonrespondents who already reported to be in class $<$ 1000 € in an earlier question. The notation in brackets refers to **Example 2**, introduced later on, where the cardinality of the sets gives some indication about the level of accuracy. For ease of presentation, we restrict to the granularity of categories depicted here. In fact, the PASS data partly provide even finer categories.

and Peel, 2004) may be used differentiating between nonrespondents and respondents. In some situations, e.g. in the context of poverty measurement, an answer on a certain ordinal level might be sufficient, hence the precise answers could be classified to the most precise income categories reported by the nonrespondents, allowing a joint analysis. An alternative might be a joint likelihood approach accounting for respondents, nonrespondents and different groups of partial respondents by distinct likelihood contributions (cf. Drechsler et al., 2015, who use an imputation-based technique and illustrate their results by the PASS data as well). Here, we restrict ourselves to the answers of the nonrespondents and treat income as ordinal, by considering as the true categories the ones that would result in case of a response to all follow-up questions. In a second step a mixture model or a comparative analysis of respondents and nonrespondents could follow.

Our question in focus will be the investigation of some covariates' impact on a true categorical response variable partly observed in a coarse way. In the example, the true categorical income is used as a response variable distinguishing the following two settings, referred to as "Example 1" and "Example 2" later on:

Example 1: Binary response variable

Here we restrict the available income data to the answers obtained from the first question.

Thus, the categories "$<$ 1000 €", "$\geq$ 1000 €" and "no answer" (i.e. the coarse answer "either $<$ 1000 € or $\geq$ 1000 €") are observed, reducing the coarsening problem to the missing data problem. When we consider ***Example 1****, the categories are abbreviated by "$<$", "$\geq$" and "na" in the following.*

Example 2: Ordinal response variable

Here we account for the whole ordinal structure inherent in the data, and the observed income variable includes different levels of coarseness. In the context of ***Example 2****, the abbreviations given in brackets in Figure 1, i.e. categories "{a}" to "{a,b,c,d,e,f,g}" (abbreviated by {a−g}), are utilized, where the latter one is interpreted as "either a or b or ... or g".*

In this way, we constructed one data situation with a binary and one with an ordinal true response variable (with values "$<$ 1000 €" and "$\geq$ 1000 €" (**Example 1**) and values "{a}" to "{g}" (**Example 2**), respectively). In both examples we use the highest school leaving certificate (first covariate) and age (second covariate) as covariates. Both variables are dichotomized, thus showing values "Abitur no" (0) or "Abitur yes" (1) (the "Abitur" is the general qualification for university entrance in Germany) and "$<$ 40" (0) or "$\geq$ 40" (1), respectively. The contingency tables in Table 1 and Table 2 summarize the considered unweighted data. To avoid very small sample sizes in some subgroup-specific coarse categories, we rely on the data from wave 1 to 4, including information on 877 individuals. Since the PASS study is a panel study, we had to make sure that no individual is included repeatedly, hence excluding duplicates. Questionnaire techniques directly offering different levels of coarse categories (instead of starting solely with a "Don't know" and precise categories like in the PASS study) are expected to reduce the problem of small sample sizes in coarse categories.

Explanation regarding the anonymization of the data: To comply with our data access contract and the non-disclosure regulations of the Federal Employment Agency (cf. Beyer et al., 2014), we have not only to delete in Table 2 all frequencies that are ≤ 3, here marking them by "*", but also to guarantee that any back-calculation of values is impossible. In each line of Table 2 the sums of the frequencies referring to the categories $\{a\}$, $\{b\}$, $\{c\}$, and $\{a,b,c\}$ (group 1) as well as to $\{d\}$, $\{e\}$, $\{d,e\}$, $\{f\}$, $\{g\}$, $\{d-g\}$, and $\{f,g\}$ (group 2) can

Table 1: Contingency table for the data of **Example 1** (Binary response variable).

		Observed income class		
Abitur	age	$<$	$\geq$	na
no (0)	<40 (0)	97	63	102
	≥ 40 (1)	69	115	131
yes (1)	<40 (0)	33	50	41
	≥ 40 (1)	38	79	59

Table 2: Contingency table for the data of **Example 2** (Ordinal response variable).

		Observed income class											
Abitur	age	{a}	{b}	{c}	{a,b,c}	{d}	{e}	{d,e}	{f}	{g}	{d–g}	{f,g}	{a–g}
no (0)	<40 (0)	50	17	18	12	22	11	*	9	*	9	*	102
	≥ 40 (1)	24	18	21	6	23	18	6	16	9	33	10	131
yes (1)	<40 (0)	21	*	*	*	10	7	5	7	8	9	4	41
	≥ 40 (1)	20	9	*	*	*	9	*	14	20	17	10	59

be inferred from Table 1. For this reason, we apply an additional (more heuristic) data-based rule and hide also the next smallest entry in each group containing deleted entries, and then the next smallest entry as well whenever the sum of the frequencies in the deleted entries is smaller than seven. In this way, a considerable variety of possible scenarios remains, making back-calculations uninformative enough.

3 A profile likelihood approach with coarse categorical data

Our main goal is the estimation of the regression coefficients with coarse categorical data when refraining from strong, frequently untenable assumptions about the coarsening. For this purpose, we extend the profile likelihood approach for the latent variable distribution developed by Zhang (2010). But – unlike Zhang – we explicitly consider the dependence of the likelihood on the coarsening process, which offers the opportunity to include different kinds of auxiliary information about the coarsening process in a next step (cf. Section 4) and hence one will no longer be restricted to the case of total ignorance or total knowledge. This section is structured as follows: First, we will present our view of the coarse data problem, including the explicit parametrization forming the basic difference to the approach by Zhang (2010) (cf. Section 3.1). Second, we will elaborate a profile likelihood approach for the latent variable

distribution based on this view (cf. Section 3.2), which will then be refined for the regression context (cf. Section 3.3).

3.1 The general view of the problem

To frame the problem of coarse data technically, we distinguish between an observed and a latent world.

Let $(x_{11}, \ldots, x_{1p}, y_1), \ldots, (x_{n1}, \ldots, x_{np}, y_n)$ be a sample of n independent realizations of the categorical random variables $(X_1, \ldots, X_p, Y)$. Unfortunately, some values y_i are not known precisely, hence the random variable Y refers to the latent world. Instead, we only observe a sample $(x_{11}, \ldots, x_{1p}, \mathfrak{y}_1), \ldots, (x_{n1}, \ldots, x_{np}, \mathfrak{y}_n)$ of n independent realizations of $(X_1, \ldots, X_p, \mathcal{Y})$, where the random set $\mathcal{Y}$ (e.g. Nguyen, 2006) belongs to the observed world and describes the coarsened observation of Y. While the state space of the variable Y is denoted by S_Y, the values of the random set $\mathcal{Y}$ are contained within the state space $S_{\mathcal{Y}} \subset \mathcal{P}(S_Y)$, where we assume the empty set to be generally excluded, but all precise categories $\{y\}$ to be included. Since we aim at a regression analysis here, we are interested in the estimation of the probabilities

$$\pi_{\mathbf{x}y} = P(Y = y | \mathbf{X} = \mathbf{x}), \quad y \in S_Y, \ \mathbf{x} \in S_X \ , \tag{1}$$

given the – assumed to be – precise values $\mathbf{x} = (x_1, \ldots, x_p)^T \in S_X$ of the categorical covariates $X_1, \ldots, X_p$. Throughout the next sections, we speak of the *latent variable distribution* when we refer to the parameters in (1). The associated dependence on the covariates is described by a response function h (here we write $\tilde{h}$ to stress that it is adapted to account also for the reference category)

$$\pi_{\mathbf{x}y} = \tilde{h}(\eta_{\mathbf{x}y}) \ , \tag{2}$$

with linear predictor $\eta_{\mathbf{x}y} = \beta_{0y} + d(\mathbf{x})^T \boldsymbol{\beta}_y$, where d fills the role of transferring the covariates into appropriate dummy-coded ones (cf. e.g. Fahrmeir et al., 2013, p. 31). In this paper, an estimation of the regression coefficients β_{0y} and $\boldsymbol{\beta}_y$ that only includes the available information about the coarsening process will be of special interest.

By applying the theorem of total probability, we can include the *coarsening parameters*

$$q_{\mathfrak{y}|\mathbf{x}y} = P(\mathcal{Y} = \mathfrak{y}|\mathbf{X} = \mathbf{x}, Y = y), \quad \mathfrak{y} \in S_{\mathcal{Y}},\ \mathbf{x} \in S_X,\ y \in S_Y\ , \tag{3}$$

(cf. Section 3.2, in particular in (9)), and hence establish a formal connection between both worlds, i.e. the parameters of the latent variable distribution in (1) and the parameters of the *observed variable distribution*

$$p_{\mathbf{x}\mathfrak{y}} = P(\mathcal{Y} = \mathfrak{y}|\mathbf{X} = \mathbf{x}), \quad \mathfrak{y} \in S_{\mathcal{Y}},\ \mathbf{x} \in S_X\ . \tag{4}$$

Apart from requiring error-freeness in the sense that the true value is contained in the coarse value, $Y \in \mathcal{Y}$, and distinct parameters in the missing data model and the substantive model (cf. Rubin, 1976), we refrain from making assumptions about the coarsening in Section 3.2 and 3.3, but then discuss in Section 4 how frequently available weak knowledge about the coarsening can be included in a powerful way.

3.2 The profile (log-)likelihood for the latent variable distribution

Compared to the global log-likelihood l, which depends on all parameters, the profile log-likelihood is a function of the parameter of interest only and arises from the global log-likelihood by considering all other parameters as nuisance parameters (cf. e.g. Pawitan, 2001, p. 80), taking the values producing the maximum. In our case, a specific parameter $\pi_{\mathbf{x}y}$, $\mathbf{x} \in S_X,\ y \in S_Y$ (cf. (1)) might be of interest. The profile log-likelihood follows as

$$l(\pi_{\mathbf{x}y}) = \max_{\xi} l(\pi_{\mathbf{x}y}, \xi) \tag{5}$$

with nuisance parameters ξ including all coarsening parameters (cf. (3)) and all parameters defining the latent variable distribution (cf. (1)) apart from the one that that is of interest. In our categorical setting, the number of nuisance parameters is finite and does not increase with the number of observations. We typically rely on the relative profile (log-)likelihood, which simply arises from a kind of normalization: The relative profile log-likelihood is obtained by subtracting the maximum value of the log-likelihood function from the profile log-likelihood,

so that the maximum value of the profile log-likelihood function is always 0. To write down the (relative) profile (log-)likelihood for a parameter of the latent variable distribution, we start with the global (log-)likelihood in terms of all parameters of the latent world (cf. (2) and (3))

$$\gamma = (\pi_{\mathbf{x}y}, q_{\mathfrak{y}|\mathbf{x}y})_{\mathbf{x}\in S_X, \mathfrak{y}\in S_\mathcal{Y}, y\in S_Y} \ . \tag{6}$$

For this purpose, we can rely on the parametrization of a selection model (cf. e.g. Heckman, 1976; Diggle and Kenward, 1994; Sikov, 2018) and the basic argumentation developed in Plass et al. (2015). We start in the observed world with the random set $\mathcal{Y}$ and interpret all elements of $S_\mathcal{Y}$ as categories of their own (e.g. category $\{a, b\}$ is regarded as its own precise category). For fixed covariate values $\mathbf{x} \in S_X$, the cell counts $(n_{\mathbf{x}\mathfrak{y}})_{\mathfrak{y}\in S_\mathcal{Y}}$ are multinomially distributed, and thus the log-likelihood in terms of all parameters of the observed world (cf. (4))

$$\vartheta = (p_{\mathbf{x}\mathfrak{y}})_{\mathbf{x}\in S_X, \mathfrak{y}\in S_\mathcal{Y}} \tag{7}$$

can be written as

$$l(\vartheta) = \sum_{\mathbf{x}\in S_X, \mathfrak{y}\in S_\mathcal{Y}} n_{\mathbf{x}\mathfrak{y}} \cdot \ln\left(p_{\mathbf{x}\mathfrak{y}}\right) \ . \tag{8}$$

Next, the information from the observation model relating the latent to the observed world is included via a mapping $\Phi : \gamma \mapsto \vartheta$. This mapping describes the transfer – based on the theorem of total probability – between the parametrization in terms of the components of γ in (6), i.e. the parameters of the latent variable distribution and the coarsening parameters, and the ones of ϑ in (7), i.e. the parameters of the observed variable distribution. Consequently, the prescription of the reparametrization Φ is given by

$$p_{\mathbf{x}\mathfrak{y}} = \sum_{y\in\mathfrak{y}} \pi_{\mathbf{x}y} \cdot q_{\mathfrak{y}|\mathbf{x}y} \ , \tag{9}$$

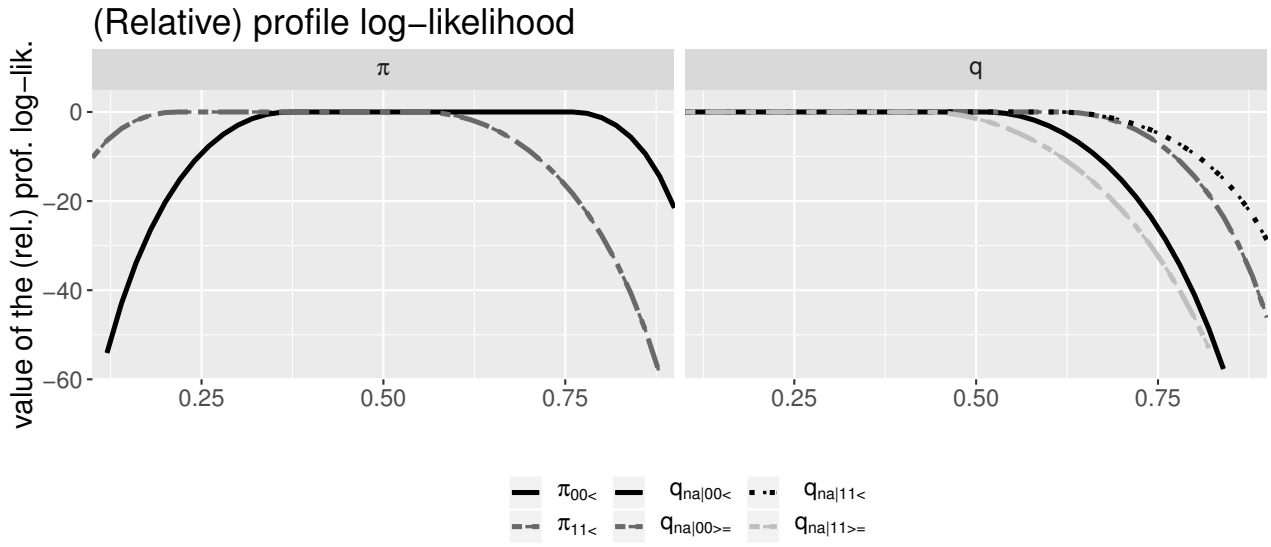

Figure 2: The (relative) profile log-likelihood functions for the parameters in γ, here restricting to subgroups "00" and "11".

for all $\mathbf{x} \in S_X,\ \mathfrak{y} \in S_\mathcal{Y}$. Finally, we can plug (9) into (8), which gives us the log-likelihood in terms of the parameters of the latent world, i.e.

$$l(\gamma) = \sum_{\mathbf{x} \in S_X, \mathfrak{y} \in S_\mathcal{Y}} n_{\mathbf{x}\mathfrak{y}} \cdot \ln\Big(\sum_{y \in \mathfrak{y}} \pi_{\mathbf{x}y} \cdot q_{\mathfrak{y}|\mathbf{x}y}\Big)\ . \tag{10}$$

Having determined the global (log-)likelihood by (10), maximum likelihood estimates can be obtained and profile (log-)likelihoods can be calculated according to (5). Each profile log-likelihood function can be graphically represented by varying the value of the parameter of interest on a grid and each time calculating the log-likelihood at the values of the nuisance parameters that jointly maximize the log-likelihood for this specific value of the parameter of interest. While for the observed world $l(\vartheta)$ can usually be uniquely maximized yielding $\hat{\vartheta}$, due to the non-injectivity of the mapping Φ, this is generally impossible for $l(\gamma)$ – except when very strict assumptions about the coarsening, such as coarsening at random (cf. Section 4), are imposed. The range of the plateau of the profile (log-)likelihood characterizes the maximum likelihood estimator of the parameter of interest. In fact, due to the invariance of the likelihood under parameter transformations, the set-valued maximum likelihood estimator

$$\hat{\Gamma} = \{\gamma \mid \Phi(\gamma) = \hat{\vartheta}\} \tag{11}$$

Table 3: Estimation of the parameters of the latent world (**Example 1**; the counts can be inferred from Table 2).

$\hat{\pi}_{\mathbf{x}<}$	$\hat{q}_{na\|\mathbf{x}<}$	$\hat{q}_{na\|\mathbf{x}\geq}$
$\hat{\pi}_{00<} \in [0.37,\ 0.76]$	$\hat{q}_{\mathrm{na}\|00<} \in [0,\ 0.51]$	$\hat{q}_{\mathrm{na}\|00\geq} \in [0,\ 0.62]$
$\hat{\pi}_{01<} \in [0.22,\ 0.63]$	$\hat{q}_{\mathrm{na}\|01<} \in [0,\ 0.66]$	$\hat{q}_{\mathrm{na}\|01\geq} \in [0,\ 0.53]$
$\hat{\pi}_{10<} \in [0.27,\ 0.60]$	$\hat{q}_{\mathrm{na}\|10<} \in [0,\ 0.55]$	$\hat{q}_{\mathrm{na}\|10\geq} \in [0,\ 0.49]$
$\hat{\pi}_{11<} \in [0.22,\ 0.55]$	$\hat{q}_{\mathrm{na}\|11<} \in [0,\ 0.61]$	$\hat{q}_{\mathrm{na}\|11\geq} \in [0,\ 0.43]$

results. We can represent it for the single components of γ by its one-dimensional projections, obtaining

$$\hat{\pi}_{\mathbf{x}y} \in \left[\frac{n_{\mathbf{x}\{y\}}}{n_{\mathbf{x}}},\ \frac{\sum_{\mathfrak{y}\ni y} n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}}}\right], \quad \hat{q}_{\mathfrak{y}|\mathbf{x}y} \in \left[0,\ \frac{n_{\mathbf{x}\mathfrak{y}}}{n_{\mathbf{x}\{y\}} + n_{\mathbf{x}\mathfrak{y}}}\right], \tag{12}$$

for all $\mathbf{x} \in S_X,\ y \in S_Y$ and all $\mathfrak{y} \in S_{\mathcal{Y}}$ such that $\{y\} \subsetneq \mathfrak{y}$ with $n_{\mathbf{x}} > 0$. It is important to keep in mind that points in these intervals are constrained by the restrictions in (9). The result in (12) can be shown to correspond to the one obtained from so-called cautious data completion, virtually plugging in all potential precise sample outcomes compatible with the observations (cf. Augustin et al., 2014, §7.8).

Furthermore, we can calculate the (relative) profile log-likelihood for all parameters of the latent world in **Example 1**, i.e. $\pi_{\mathbf{x}<}$, $q_{na|\mathbf{x}<}$ and $q_{na|\mathbf{x}<}$ for all four subgroups $\mathbf{x} \in S_X = \{\text{“00”}, \text{“01”}, \text{“10”}, \text{“11”}\}$, which are interpreted as “Abitur=0, age=0”, “Abitur=0, age=1”, “Abitur=1, age=0” and “Abitur=1, age=1”, respectively. For reasons of presentation, Figure 2 restricts to the results from subgroup “00” and “11”. The set-valued maximum likelihood estimate $\hat{\Gamma}$ is not a singleton: There are multiple estimated combinations of coarsening parameters and latent variable distributions that are compatible with the restriction in (9) and thus lead to the estimated observed variable distribution. Different scenarios for the estimation of $\pi_{00<}$ are conceivable, ranging from attributing all coarse categories “na” to “$\geq$” to including them all in category “$<$”, thus obtaining (cf. (12))

$$\hat{\pi}_{00<} \in [\underline{\hat{\pi}}_{00<},\ \overline{\hat{\pi}}_{00<}] \quad \text{with} \quad \underline{\hat{\pi}}_{00<} = \frac{97}{262} \approx 0.37 \quad \text{and} \quad \overline{\hat{\pi}}_{00<} = \frac{97 + 102}{262} \approx 0.76\ .$$

The resulting estimates (i.e. the one-dimensional projections of $\hat{\Gamma}$) for **Example 1** are shown in Table 3.

3.3 The profile (log-)likelihood for the regression coefficients

Starting from the global (log-)likelihood for the parameters of the latent world described by (10), we first aim at determining the global (log-)likelihood in terms of the regression coefficients and the coarsening parameters. Proceeding as in (5), we can then derive the profile (log-)likelihoods for the regression coefficients from this global (log-)likelihood.

In a generalized linear model, the relation between the response variable distribution and the linear predictor is described by the response function. We can thus obtain the global log-likelihood in terms of the regression coefficients and the coarsening parameters by simply replacing all parameters defining the latent variable distribution in (10) by the response function in (2), obtaining

$$l(\beta_{0y}, \boldsymbol{\beta}_y, (q_{\mathfrak{y}|\mathbf{x}y})_{\mathbf{x}\in S_X, \mathfrak{y}\in S_\mathcal{Y}, y\in S_Y}) = \sum_{\mathbf{x}\in S_X, \mathfrak{y}\in S_\mathcal{Y}} n_{\mathbf{x}\mathfrak{y}} \cdot \ln\big(\sum_{y\in\mathfrak{y}} \tilde{h}(\beta_{0y} + d(\mathbf{x})^T\boldsymbol{\beta}_y) \cdot q_{\mathfrak{y}|\mathbf{x}y}\big)\ . \tag{13}$$

For instance, the profile log-likelihood for β_{0y} is then calculated as

$$l(\beta_{0y}) = \max_{\xi} l(\beta_{0y}, \xi) \tag{14}$$

with nuisance parameters ξ including all coarsening parameters and all regression coefficients except β_{0y}. In Section 5.2 we will see that, depending on the data situation and the parametric assumptions of the regression model, the resulting profile (log-)likelihoods for the regression coefficients can either show again a plateau or have a unique maximum. To already give an intuition for this point, one can imagine that rather weak parametric assumptions and a large number of coarse observations still allow for multiple possible underlying precise values, reflected by the plateau of the profile (log-)likelihood, while strong parametric assumptions and a small number of coarse observations tendentially lead to a unique maximum. We now illustrate the calculation of profile log-likelihood functions for regression coefficients using the data situations of Example 1 and Example 2.

Example 1: Due to the binary dependent variable ($S_Y = \{<, \geq\}$), we choose a logit model

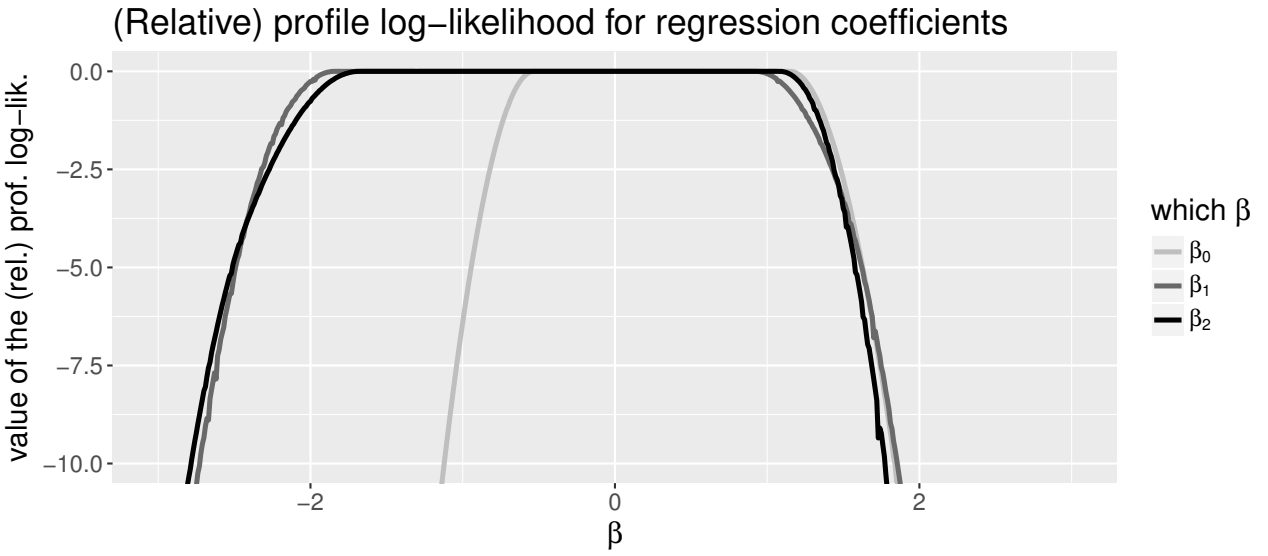

Figure 3: With the data of **Example 1**, the (relative) profile log-likelihood function for each regression coefficient β_0, β_1, and β_2 shows a plateau.

with the response function

$$\pi_{\mathbf{x}<} = P(Y = \text{“}<\text{”} \,|\mathbf{x}) = \frac{\exp(\beta_0 + d(\mathbf{x})^T\boldsymbol{\beta})}{1 + \exp(\beta_0 + d(\mathbf{x})^T\boldsymbol{\beta})} \tag{15}$$

for the category of interest, here “<”, and

$$\pi_{\mathbf{x}\geq} = \frac{1}{1 + \exp(\beta_0 + d(\mathbf{x})^T\boldsymbol{\beta})}\,, \tag{16}$$

for the reference category, here “≥”. Plugging (15) and (16) in (13) gives the global log-likelihood in terms of the regression coefficients $l(\beta_0, \beta_1, \beta_2, q_{na|00<}, q_{na|00\geq}, \ldots, q_{na|11<}, q_{na|11\geq})$, where β_1 and β_2 refer to the covariates Abitur and age, respectively. The profile log-likelihood for each regression coefficient (depicted in Figure 3) can be calculated by evaluating the global log-likelihood on a grid of values for the regression coefficient of interest and for each value choosing the nuisance parameters at the corresponding likelihood-maximizing values. By considering the plateau of the profile log-likelihood for each regression coefficient, we can infer the cautious maximum likelihood estimates for the regression coefficients, which refer to the case without assumptions on the coarsening process. The profile (log-)likelihood cannot be maximized uniquely in this data situation (further details are given in Section 5.2, where this case is referred to as situation 1).

Example 2: To account for the ordinal structure of the response variable, we base our analysis on the cumulative logit model (cf. e.g. Fahrmeir et al., 2013, p. 334–337). This model relies on the idea that the ordinal response categories are obtained instead of the exact values of a latent continuous variable $\tilde{Y}$, thus introducing a second layer of latency in our case. For this variable a regression model $\tilde{Y} = -d(\mathbf{x})^T\boldsymbol{\beta} + \epsilon$ with $\epsilon \sim F$ is assumed, where F is the (standard) logistic distribution function. Assuming intercepts increasing with the order of the respective category: $-\infty = \beta_{0y^{(0)}} < \beta_{0y^{(1)}} < \cdots < \beta_{0y^{(m)}} = \infty$, the connection to our categorical variable of interest Y is given by $Y = y^{(l)} \iff \beta_{0y^{(l-1)}} < \tilde{Y} \leq \beta_{0y^{(l)}}$, $l = 1, \ldots, m$, where $y^{(l)}$ is the lth category within the ordered categories $y^{(1)}, \ldots, y^{(l)}, \ldots, y^{(m)}$. While the intercepts are category-specific, the regression coefficients $\boldsymbol{\beta}$ are not: this is also referred to as proportional-odds assumption (cf. e.g. Fahrmeir et al., 2013, p. 336). The ordinal structure is included by basing the analysis on the cumulative probabilities described by the distribution function F, hence considering the response function

$$
\begin{aligned}
&P(Y \leq y^{(l)} \,|\mathbf{x}) = F(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta}), \quad \text{with} \\
&F(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta}) = \frac{\exp(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta})}{1 + \exp(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta})}, \quad \text{and with} \\
&\pi_{\mathbf{x}y^{(l)}} = \; P(Y = y^{(l)} \,|\mathbf{x}) = F(\beta_{0y^{(l)}} + d(\mathbf{x})^T\boldsymbol{\beta}) - F(\beta_{0y^{(l-1)}} + d(\mathbf{x})^T\boldsymbol{\beta}), \quad l = 1, \ldots, m,
\end{aligned}
\tag{17}
$$

(cf. e.g. Fahrmeir et al., 2013, p. 335). Using the relation formalized via the response function in (17) within the global (log-)likelihood representation in (13) again gives us the starting point for the derivation of the (relative) profile (log-)likelihood for each regression coefficient, here for the category-specific intercepts β_{0a}, β_{0b}, β_{0c}, β_{0d}, β_{0e}, β_{0f} and the regression coefficients β_1 and β_2 of **Example 2**. In Figure 4 the (relative) profile log-likelihood functions for all regression coefficients are depicted (the calculated profile log-likelihood functions have been smoothed to avoid the irregularities due to numerical optimization). For reasons of data protection, we refer to one possible, arbitrarily chosen data scenario that is compatible with the data in Table 2. Anyway, this choice of the data situation has no recognizable impact on the results. The maximum likelihood estimates for the regression coefficients are again obtained by considering the maxima/maximum of the respective function. Figure 4 suggests a unique maximum (pointing to situation 2, as specified in Section 5.2). Due to the numerical

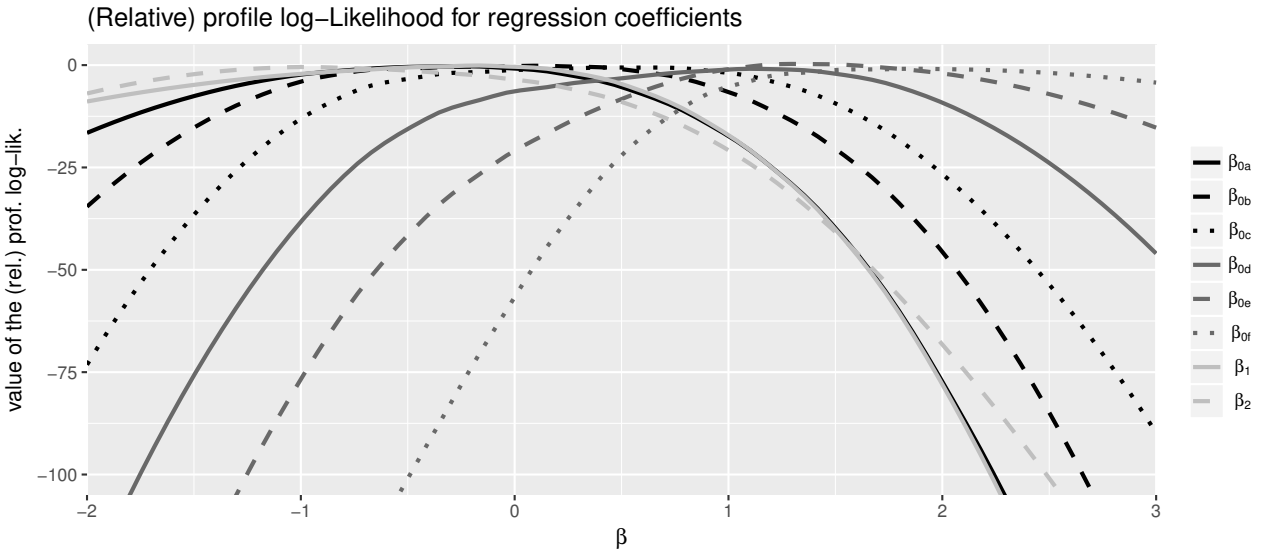

Figure 4: Relying on the data in Table 2, for all regression coefficients the respective profile-likelihood is shown.

difficulties of the cumulative model to deal with the increasing order of the category-specific intercepts, which also appear here (and possibly even become more intense), the result has to be treated with some caution.

4 Exploiting partial knowledge about the coarsening

In Section 3, we showed a cautious modeling approach, basically refraining from any coarsening assumptions. Our results already pointed to the problem that the obtained intervals in the cautious case can be quite wide (cf. e.g. the cautious regression estimators inferable from Figure 3), and thus might be too conservative in some applications. In practice one frequently considers the other extreme situation of spurious precision, where strong coarsening assumptions, mostly coarsening at random (CAR), are imposed. Heitjan and Rubin (1991) introduced the concept of CAR, which for each observable value $\mathfrak{y}$ requires constant coarsening probabilities $q_{\mathfrak{y}|y}$ regardless of the true underlying value $y \in \mathfrak{y}$. Adapting CAR to our contingency table framework, the requirement has to be valid for all subgroups split by the considered covariates: for instance, within each group split by Abitur and age the true income category cannot have any influence on the coarsening probability. Under CAR, parameters are identified and the corresponding profile (log-)likelihood functions can be maximized uniquely (cf. e.g. Jaeger, 2006). However, in many cases one is not sure whether strong, point-identifying assumptions

such as CAR are indeed justified or whether a remarkable bias has to be expected (cf. e.g. Plass et al. (2015, p. 254), where an illustrative simulation study has been conducted showing the bias resulting from estimating under CAR when it is actually not valid).

For these reasons, we suggest a synthesis of both extreme situations. The user is thus no longer trapped between including no knowledge or claiming full knowledge, but is instead able to exploit just the weak information that is available about the coarsening. For the missing-data problem, literature reveals some possibilities to incorporate (partial) knowledge (cf. ideas based on the methodology of partial identification or a systematic sensitivity analysis). Mostly, one restricts either the incompleteness process or the response propensities (e.g. Kenward et al., 2001; Manski, 2015): here we concentrate on the first option in the context of coarse data and formulate constraints on $q_{\mathfrak{y}|\mathbf{x}y}$, which then results in a generalization of CAR.

To prepare our idea, it can help to look at Nordheim (1984), who suggests a way to generalize the missing at random assumption (MAR) by including the ratio between missingness parameters into the analysis of non-randomly missing and misclassified data. In Plass et al. (2015) we extend this idea by making assumptions about the coarsening probability ratios

$$R_{\mathbf{x},y,y',\mathfrak{y}} = \frac{q_{\mathfrak{y}|\mathbf{x}y}}{q_{\mathfrak{y}|\mathbf{x}y'}}, \quad \mathfrak{y} \in S_{\mathcal{Y}},\ y,\ y' \in \mathfrak{y},\ \mathbf{x} \in S_X \ , \tag{18}$$

where the special case of CAR is expressed by setting all these ratios equal to 1.

In most practical cases it is unrealistic to claim knowledge about the exact values of the ratios in (18). Nevertheless, it seems quite realistic that former studies or substance-matter considerations allow rough statements about the magnitude of the ratios in the sense of $R_{\mathbf{x},y,y',\mathfrak{y}} \in [\underline{R},\ \overline{R}]$ with $\underline{R}, \overline{R} \geq 0$. To practically incorporate such weak knowledge about the coarsening and hence to obtain a reliable estimation of the regression coefficients, we can again rely on the profile (log-)likelihood (cf. Section 3.3), including the corresponding linear constraints

$$q_{\mathfrak{y}|\mathbf{x}y} \geq q_{\mathfrak{y}|\mathbf{x}y'} \cdot \underline{R} \ \text{ and } \ q_{\mathfrak{y}|\mathbf{x}y} \leq q_{\mathfrak{y}|\mathbf{x}y'} \cdot \overline{R}\ , \mathfrak{y} \in S_{\mathcal{Y}},\ y,\ y' \in \mathfrak{y},\ \mathbf{x} \in S_X \ , \tag{19}$$

into the original optimization problem.

For example, there are several practical situations where CAR is principally conceivable, but its exact applicability is rather questionable. In such cases, it can be reasonable to replace

CAR by specific neighborhood assumptions (as e.g. addressed in Manski, 2015, for MAR), requiring that the coarsening probabilities lie in the neighborhood of the CAR case. We refer to these assumptions as near CAR and formalize them by choosing $R_{\mathbf{x},y,y',\mathfrak{y}}$ to lie within the interval $[\frac{1}{\tau_1},\ \tau_2]$, where $\tau_1,\ \tau_2 \geq 1$ specify the neighborhood. Rough values of $\tau_1,\ \tau_2 \geq 1$ may be derived from external information dependent on the specific data application and give the basis for a systematic sensitivity analysis.

Another, in some sense dual, kind of strict coarsening that can be relaxed in a next step is given by subgroup independence (SI, cf. Plass et al. (2017) for more details in the i.i.d. case). SI is characterized by the independence of the coarsening probabilities from the corresponding covariate values, thus assuming that for each fixed $\mathfrak{y} \in S_{\mathcal{Y}}$ and $y \in \mathfrak{y}$ the coarsening parameters $q_{\mathfrak{y}|\mathbf{x}y}$ have the same value for all $\mathbf{x} \in S_X$. Like in the context of CAR, we can consider probability ratios

$$R_{\mathbf{x},\mathbf{x}',y,\mathfrak{y}} = \frac{q_{\mathfrak{y}|\mathbf{x}y}}{q_{\mathfrak{y}|\mathbf{x}'y}}, \quad \mathfrak{y} \in S_{\mathcal{Y}},\ y \in \mathfrak{y},\ \mathbf{x}, \mathbf{x}' \in S_X \ , \tag{20}$$

where the special case of SI is expressed by setting all these ratios equal to 1. By including corresponding constraints, analogous to (19), weak assumptions about the values of the ratios in (20) can be imposed. In particular, near SI can be constructed in analogy to near CAR.

New developments, such as paradata, might positively affect the popularity of our idea. Paradata are a by-product of the regular data collection and refer to the data collection process itself (cf. e.g. Durrant and Kreuter, 2013; Kreuter and Olson, 2013). Examples are interviewer observations about the housing situation of (unit-)nonrespondents to account for nonresponse bias, or interviewer assessments of the respondents' embarrassment with regard to certain questions or response times (cf. e.g. Couper and Kreuter, 2013). Information of this kind might enrich knowledge about the coarsening process. In many cases, this knowledge is expected to be of a rather weak nature, such as "respondents with a long response time rather tend to give a coarse answer compared to respondents answering quickly". While traditional approaches have to leave this information unconsidered, our approach is able to incorporate this weak knowledge properly.

To illustrate the inclusion of auxiliary information about the coarsening process, we again

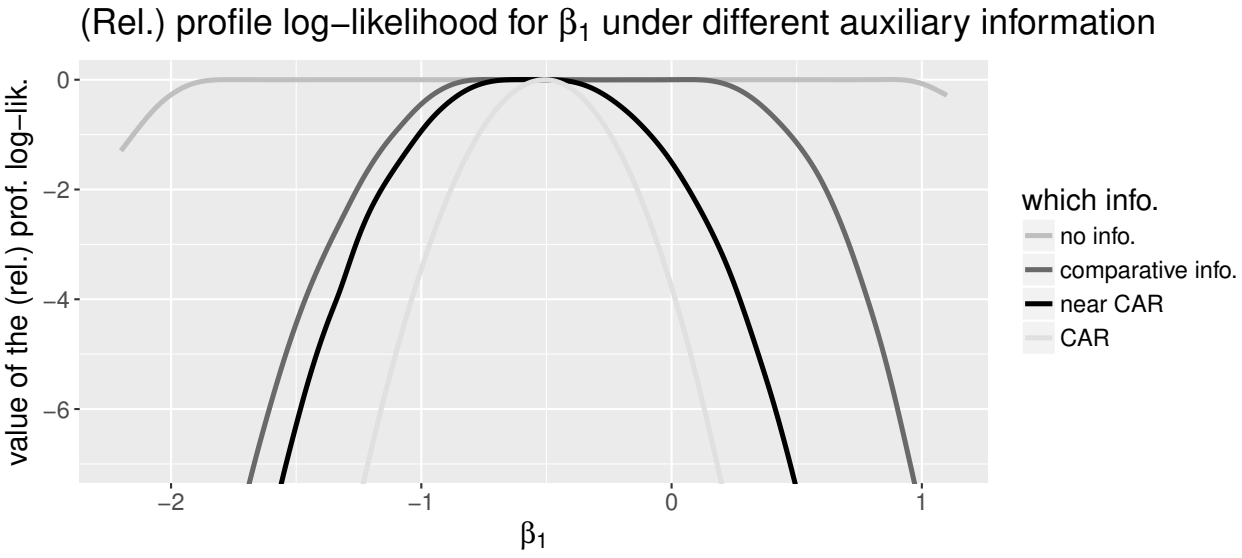

Figure 5: The relative profile log-likelihood function for β_1 based on different types of auxiliary information and the data of **Example 1**.

Table 4: Regression estimates under different information (**Example 1**).

no information	comparative information	near CAR	CAR
$\hat{\beta}_1 \in [-1.84,\ 0.92]$	$\hat{\beta}_1 \in [-0.77,\ 0.05]$	$\hat{\beta}_1 \in [-0.67,\ -0.47]$	$\hat{\beta}_1 = -0.51$

refer to the setting of the examples. There are frequently situations where assumptions such as "respondents with a high income rather tend to give no answer compared to the ones with a low income" might be justified from an application standpoint. This weak knowledge about the missingness probabilities can be formalized by including constraints such as $q_{na|\mathbf{x}<} \leq q_{na|\mathbf{x}\geq}$, or equivalently $R_{\mathbf{x},<,\geq,na} \in [0,1]$ (**Example 1**), and $q_{\{a,b,c\}|\mathbf{x}a} \leq q_{\{a,b,c\}|\mathbf{x}b} \leq q_{\{a,b,c\}|\mathbf{x}c}$ (**Example 2**). With the data of **Example 1**, the profile log-likelihood functions for β_1 (Abitur effect) under this comparative information and other (partial) assumptions (namely CAR and near CAR with τ_1=1.25, τ_2=1.3) are depicted in Figure 5, while the corresponding reliable regression estimates are given in Table 4. While the regression estimate is uncommitted about the presence and direction of an Abitur effect if no knowledge about the coarsening is included, under CAR a precise value is obtained. The results under partial knowledge show a remarkable refinement of the one obtained with no knowledge, solving the trade-off between information and credibility in situations where some weak knowledge is available.

5 Further aspects of the involved uncertainties

In Section 4 we presented a way to make use of frequently available weak knowledge about the coarsening and hence focused on the uncertainty about the coarsening process. One has to deal with (at least) two further kinds of uncertainties: First, uncertainty occurs due to the availability of a finite sample size only. In Section 5.1 we will construct likelihood-based confidence intervals, which not only capture this sampling uncertainty, but simultaneously still account for the uncertainty about the coarsening process. Second, considering a regression problem, model assumptions are made and model uncertainty arises. In Section 5.2 we will discuss the impact of the parametric assumptions of the regression model in the coarse data problem.

5.1 Uncertainty due to the finite sample size

There are already several proposals that consider the sampling error induced by the availability of a finite sample only, here called sampling uncertainty, and the uncertainty about the data incompleteness at the same time. In this way, confidence intervals for partially identified parameters have been constructed (cf. Imbens and Manski, 2004; Horowitz and Manski, 2000; Stoye, 2009a; Vansteelandt et al., 2006). One of the advantages of the (profile) likelihood approach is the possibility to infer confidence intervals in a direct way. These likelihood-based confidence intervals are appealing due to their (compared to Wald intervals) better performance in case of a small sample size (cf. e.g. Neale and Miller, 1997).

Usually, likelihood-based confidence intervals are constructed by cutting the relative profile log-likelihood function at level $\delta = (-0.5\chi^2_{1,1-\alpha})$ (cf. e.g. Venzon and Moolgavkar, 1988), where $1-\alpha$ denotes the (approximate) confidence level. The confidence interval consists then of all values of the parameter of interest for which the relative profile log-likelihood is larger than or equal to δ. Likelihood-based confidence intervals in the presence of missing data have already been studied for the probabilities $\pi_{\mathbf{x}y}$, $\mathbf{x} \in S_X,\ y \in S_Y$, relying on the profile likelihood presented in Section 3.2 (cf. Cattaneo and Wiencierz, 2012; Zhang, 2010). We can proceed analogously and obtain asymptotic $(1-\alpha)$ confidence intervals for the regression coefficients by cutting the corresponding relative profile log-likelihood at level δ.

In Figure 6, we exemplify the construction of likelihood-based confidence intervals, using the

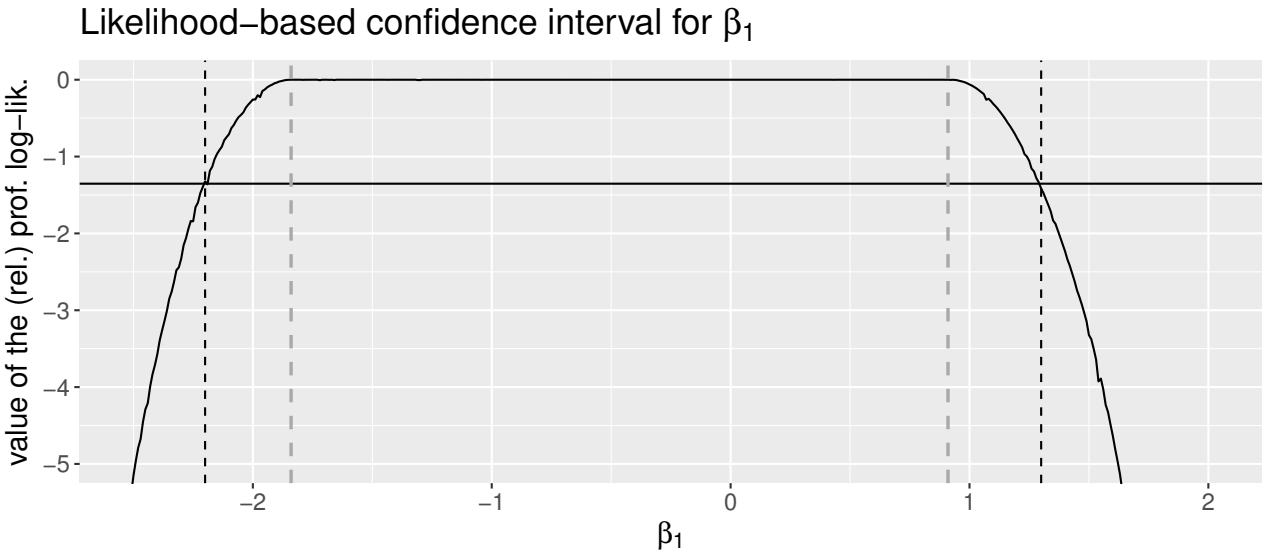

Figure 6: The δ-cut is symbolized by the solid line, while the black dashed lines mark the bounds of the confidence interval, here with confidence level $1 - \alpha = 0.9$. The extent of the sampling uncertainty is visible by comparing these bounds with the bounds of the maximum-likelihood estimate characterized by the gray lines.

Abitur effect β_1 and the data of **Example 1**. In this illustration, we refrain from any assumption about the coarsening, but confidence intervals could also be constructed based on a profile (log-)likelihood including some (partial) knowledge about the coarsening process along the lines of Section 4. By considering for instance the case without information and comparing the confidence interval for β_1 (i.e. [-2.20, 1.29]) with the corresponding point estimation yielding the interval [-1.84, 0.92] in Table 4, one can infer the magnitude of the sampling uncertainty.

5.2 Model uncertainty

In general, there are different interpretations of model uncertainty. Two main sources of model uncertainty are related to assumptions about

(i) the model family (e.g. a generalized linear model based on a probit or logit link) and

(ii) the variables and interactions included (cf. variable selection, e.g. Kuo and Mallick (1998)).

Model assumptions and the inherent model uncertainty in case of partially identified models have already been studied in literature (cf. e.g. Ponomareva and Tamer, 2011; Schollmeyer and Augustin, 2015). In particular, in this case – unlike in the identified case – the interpretation of a linear model (in the sense of Freedman (1987) who distinguishes between a structural or a

descriptive view) is crucial, and can lead to different results. Referring to categorical regression analysis, we will characterize different kinds of results that can also fundamentally differ from the ones obtained in the precise/identified case. However, we assume that the right model family has been specified and concentrate on (ii). In our categorical setting, the saturated model corresponds to not making any assumptions and is in this sense fully nonparametric. For this reason, when we speak about model uncertainty we mean the uncertainty due to the parametric assumptions of the regression model (i.e. at least one effect or interaction of the saturated model is set equal to zero).

In this section, we restrict ourselves to the setting of **Example 1**. Although results are expected to be transferable to cases with non-binary response variables, further research should tackle this topic.

As noted above, the **saturated model** is fully nonparametric. For this reason, and due to the bijectivity and continuity of the link function $g(\pi_{\mathbf{x}y})$ (cf. e.g. Fahrmeir et al., 2013, p. 304), the bounds of the estimated regression coefficients (cf. Section 3.3) can also be calculated as a direct transformation of the bounds of the estimated parameters of the latent variable distribution (cf. Section 3.2). In Appendix A, this result is illustrated for the setting of **Example 1**, where the logit model is appropriate.

Substantially different conclusions can be derived in the presence of parametric assumptions in the regression model, i.e. if a **non-saturated model** is specified. Since several effects or interactions are set equal to zero, the number of parameters that have to be estimated is reduced. As a consequence, in the classical case, the regression coefficients and the nonparametric response variable distribution are no longer generally compatible with each other. However, in the coarse data problem, due to the multiple conceivable scenarios this incompatibility is no longer generally valid, and we discover systematically differing situations. By relying on (9), for these situations it is directly inferable whether and how the intervals for the coarsening estimators are sharpened due to the parametric assumptions of the regression model (cf. also the three pictures in Figure 8 of Appendix D and the derivation in Appendix E):

1. Compatibility:

   (a) The obtained regression estimators produce the estimated bounds of the latent variable distribution calculated without parametric assumptions (i.e. the bounds in

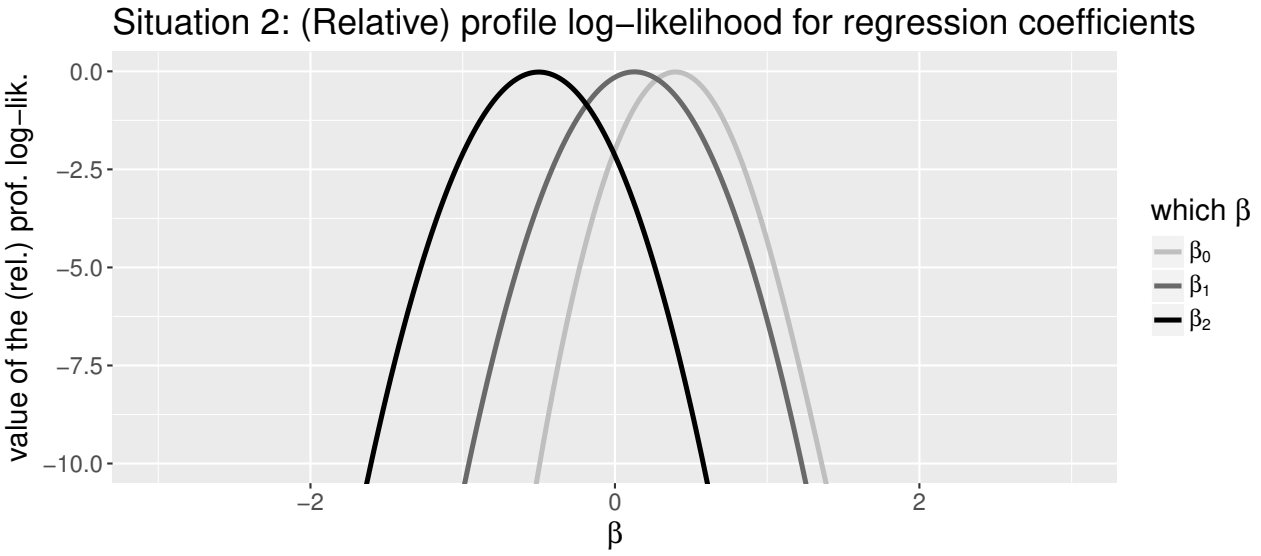

Figure 7: Profile log-likelihood functions based on an (arbitrary) data constellation classified to situation 2.

(12)). ⇒ The parametric assumption has no impact on the estimated coarsening parameters.

(b) The regression estimators represent tighter bounds than the nonparametric estimate of the latent variable distribution in (12). The intuition behind this results is given by the restriction on (some) interactions, which are set equal to zero (hence limiting the possible precise data scenarios). ⇒ The parametric assumption has a slight impact on the estimated coarsening parameters, sharpening the corresponding intervals.

2. Incompatibility: The parametric assumptions prevent that the nonparametric estimate of the latent variable distribution in (12) is reproduced by means of the regression estimators (this is what usually happens in the classical case without coarse observations). ⇒ The parametric assumption has a strong impact on the estimated coarsening parameters, making them precise (cf. the unique maximum of the profile (log-)likelihood in Figure 7, which refers to an illustrative data constellation classified to this situation).

Technically, we obtained this distinction by writing down a (linear) optimization problem for the setting of **Example 1** (cf. Appendix B). This optimization problem aims at finding the bounds of the regression coefficients under the condition that the nonparametrically estimated latent variable distribution can be achieved. Whenever a solution exists, we are in situation 1,

where the specific situation 1a is obtained when the inequalities of the optimization problem can be satisfied with equality.

By rearranging the system of inequalities in the optimization problem, we can derive a condition to distinguish situation 1 from situation 2 (cf. Appendix C). This condition indicates that with a growing amount of coarse observations we end up in situation 1, while a small amount rather implies situation 2 (with the precise data situation as a special case). It turns out that **Example 1** is classified into situation 1, and more specifically 1a.

6 Concluding remarks

In this paper, we developed an approach for regression analysis with coarse categorical response variables and precisely observed categorical covariates. Our results are reliable because we are no longer forced to make a particular, often unjustified coarsening assumption in order to achieve point-identified parameters. Instead, we could reveal a practical possibility for the user to include frequently available rough statements from his domain knowledge about the coarsening (such as "respondents with a low income rather tend more often to give a coarse answer than respondents with an average income") to refine the results obtained from an analysis based on no assumptions about the coarsening at all. To determine regression estimators that only rely on tenable coarsening assumptions, we have chosen an estimation procedure that extends the profile likelihood approach for the latent variable distribution by Zhang (2010) in several ways. In particular, we represent the likelihood in terms of the coarsening parameters to be able to include (weak) available knowledge about the coarsening process. We then further reparametrize this likelihood by exploiting the relation between the latent variable distribution and the regression parameters formalized by the response function of the respective categorical regression model. Maximizing the corresponding profile likelihood for each regression coefficient gives us regression estimators that only rely on tenable coarsening assumptions. The approach developed has been examined for the cumulative logit model and in more detail for the logit model. While corresponding results for other categorical regression model (e.g. analogous probit models) should be achieved mutatis mutandis when addressing coarse response

variables, the inclusion of coarse covariates is expected to be less straightforward, also due to different conceivable interpretations of conditioning on coarse observations.

We applied all findings to the PASS data. A comparison of the regression estimates based on weak assumptions with the ones of a traditional method including the coarsening at random assumption and a cautious method incorporating no assumptions at all confirms that a synthesis has been found. Depending on the research question, our results might (still) be assessed as too little informative. However, a possibly small content of information should not be regarded as a weakness of an approach based on the methodology of partial identification, but associated to sparse additional knowledge.

Apart from the uncertainty about the coarsening process, we considered two further kinds of uncertainty:

I. The sampling uncertainty. We additionally addressed the uncertainty due to the availability of a finite sample size only and constructed corresponding likelihood-based confidence intervals.

II. The model uncertainty. We discovered the interplay between the parametric assumptions of the regression model and the uncertainty about the coarsening assumption for binary data modelled by logistic regression and figured out that the impact of the parametric assumptions of the regression model on the estimated coarsening parameters can go from no effect, via a slight effect through to a very strong effect. Considering this impact also for more general categorical regression models beyond the logit model should be part of further research.

Since the incorporation of the uncertainty about the coarsening process represented the main focus of this paper, we concentrated on a separate discussion of I and II in the presence of the uncertainty about the coarsening (cf. Section 5.1 and Section 5.2). Further research should include a joint study of all three kinds of uncertainty.

The major limitation of our work is given by the restricted setting: Throughout this paper, we confined ourselves to coarse categorical data from small state spaces consisting of few categories only. While in categorical cases showing large state spaces one can focus on the most important coarse categories to avoid the explosion of the number of coarse categories, the problem has to be approached totally differently for continuous variables (cf. e.g. Schollmeyer and Augustin,

2015). Since most questionnaires involve questions showing a manageable number of categorical answers (cf. e.g. the German General Social Survey, or the European Social Survey) and many applications from official statistics only include a small number of covariates, there is a large variety of situations where our proposals can be powerfully employed.

The likelihood approach for the latent variable distribution (cf. Section 3.2) turns out to be a fruitful field of study for further research: The connection between the latent and the observed world gives the opportunity to transfer already existing likelihood-based methods for precise categorical data to the setting of coarse data (e.g. statistical tests, cf. Plass et al., 2017). Another promising topic is the inclusion of weak auxiliary knowledge (cf. Section 4): First, the collection of paradata might enrich (weak) knowledge about the coarsening process, which not only shows the relevance of the presented approach, but also gives rise to practical examples for further research. Second, our way to include weak knowledge can also be applied to other problems relying on strong assumptions, such as misclassification, propensity score matching, and statistical matching, where starting points are already provided in Molinari (2008), Stoye (2009b) (who studied an approach based on partial identification to estimate treatment effects without considering propensity scores), and D'Orazio et al. (2006), respectively. In particular, propensity score matching and statistical matching traditionally rely on strict assumptions, namely the strongly ignorable treatment assignment and the conditional independence assumption, respectively, where our strategy would allow for a practically highly relevant relaxation of these prerequisites.

Acknowledgements

We are very thankful for the helpful remarks of the reviewers. Furthermore, we are grateful to the Institute for Employment Research, Nuremberg, especially Mark Trappmann and Anja Wurdack, for the access to the PASS data and their support in practical matters. Finally, we thank Paul Fink, who discussed data disclosure issues with us. The first author also thanks the LMUMentoring program, providing financial support for young female researchers.

References

Augustin, T., Walter, G., and Coolen, F. (2014). Statistical inference. In Augustin, T., Coolen, F., de Cooman, G., and Troffaes, M., editors, *Introduction to Imprecise Probabilities*, pages 135–189. Wiley.

Beyer, O., Hüser, A., Rudloff, K., and Rüst, M. (2014). Statistische Geheimhaltung: Rechtliche Grundlagen und fachliche Regelungen der Statistik der Bundesagentur für Arbeit (Statistical nondisclosure: Legal fundamentals and professional regulations of the Statistics Department of the Federal Employment Agency [translated by the authors]). Available online at `https://statistik.arbeitsagentur.de/Statischer-Content/Grundlagen/Statistische-Geheimhaltung/Generische-Publikationen/Statistische-Geheimhaltung.pdf`, Accessed: 2019-01-23.

Brack, G. (2017). Wie viele Flüchtlinge sind ohne Schulabschluss? (How many refugees are without school leaving certificate? [translated by the authors]). Available online at `http://www.br.de/nachrichten/faktencheck/fluechtlinge-ohne-schulabschluss-faktencheck-faktenfuchs-100.html`, Accessed: 2019-01-23.

Brücker, H. and Schupp, J. (2017). Annähernd zwei Drittel der Geflüchteten haben einen Schulabschluss (Approximately two third of the refugees graduated [translated by the authors]). Available online at `https://www.iab-forum.de/annaehernd-zwei-drittel-der-gefluechteten-haben-einen-schulabschluss/`, Accessed: 2019-01-23.

Cattaneo, M. and Wiencierz, A. (2012). Likelihood-based imprecise regression. *Int. J. Approx. Reason.*, 53:1137–1154.

Chen, S. and Haziza, D. (2018). Recent developments in dealing with item non-response in surveys: a critical review. *Int. Stat. Rev.* Early View, `https://doi.org/10.1111/insr.12305`.

Couper, M. and Kreuter, F. (2013). Using paradata to explore item level response times in surveys. *J. R. Stat. Soc. A Stat.*, 176:271–286.

Denœux, T. (2014). Likelihood-based belief function: justification and some extensions to low-quality data. *Int. J. Approx. Reason.*, 55:1535–1547.

Diggle, P. and Kenward, M. (1994). Informative drop-out in longitudinal data analysis. *J. R. Stat. Soc. C Stat.*, 43:49–93.

D'Orazio, M., Di Zio, M., and Scanu, M. (2006). Statistical matching for categorical data: Displaying uncertainty and using logical constraints. *J. Off. Stat.*, 22:137–157.

Drechsler, J., Kiesl, H., and Speidel, M. (2015). MI double feature: Multiple imputation to address nonresponse and rounding errors in income questions. *Austrian Journal of Statistics*, 44:59–71.

Durrant, G. and Kreuter, F. (2013). The use of paradata in social survey research. *J. R. Stat. Soc. A Stat.*, 176:1–3.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). *Regression: Models, Methods and Applications*. Springer.

Freedman, D. (1987). A rejoinder on models, metaphors, and fables. *J. Educ. Stat.*, 12:206–223.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. *Ann. Econ. Soc. Meas.*, 5:475–492.

Heitjan, D. and Rubin, D. (1991). Ignorability and coarse data. *Ann. Stat.*, 19:2244–2253.

Hoeren, D. (2017). 59 Prozent der Flüchtlinge haben keinen Schulabschluss (59 percent of the refugees do not have a school leaving certificate [translated by the authors]). Available online at `http://www.bild.de/politik/inland/fluechtlinge/59-prozent-haben-keinen-schulabschluss-52943448.bild.html`, Accessed: 2019-01-23.

Horowitz, J. and Manski, C. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *J. Am. Stat. Assoc.*, 95:77–84.

Hüllermeier, E. (2014). Learning from imprecise and fuzzy observations: Data disambiguation through generalized loss minimization. *Int. J. Approx. Reason.*, 55:1519–1534.

Imbens, G. and Manski, C. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72:1845–1857.

Jackson, D., White, I., and Leese, M. (2010). How much can we learn about missing data?: an exploration of a clinical trial in psychiatry. *J. R. Stat. Soc. A Stat.*, 173:593–612.

Jaeger, M. (2006). On testing the missing at random assumption. In Fürnkranz, J., Scheffer, T., and Spiliopoulou, M., editors, *ECML '06, Proceedings of the 17th European Conference on Machine Learning*, Lecture Notes in Artificial Intelligence, pages 671–678. Springer.

Kenward, M., Goetghebeur, E., and Molenberghs, G. (2001). Sensitivity analysis for incomplete categorical data. *Stat. Model.*, 1:31–48.

Kreuter, F. and Olson, K. (2013). Paradata for nonresponse error investigation. In Kreuter, F., editor, *Improving Surveys with Paradata*, pages 11–42. Wiley.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhya B*, 60:65–81.

Little, R. (1988). A test of missing completely at random for multivariate data with missing values. *JASA*, 83:1198–1202.

Little, R. and Rubin, D. (2014). *Statistical Analysis with Missing Data*. 2nd edition, Wiley.

Manski, C. (2003). *Partial Identification of Probability Distributions*. Springer.

Manski, C. (2015). Credible interval estimates for official statistics with survey nonresponse. *J. Econometrics*, 191:293–301.

McLachlan, G. and Peel, D. (2004). *Finite Mixture Models*. Wiley.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *J. Econometrics*, 144:81–117.

Neale, M. and Miller, M. (1997). The use of likelihood-based confidence intervals in genetic models. *Behav. Genet.*, 27:113–120.

Nguyen, H. (2006). *An Introduction to Random Sets*. CRC.

Nishimura, R., Wagner, J., and Elliott, M. (2016). Alternative indicators for the risk of non-response bias: A simulation study. *Int. Stat. Rev.*, 84:43–62.

Nordheim, E. (1984). Inference from nonrandomly missing categorical data: An example from a genetic study on Turner's syndrome. *J. Am. Stat. Assoc.*, 79:772–780.

Olson, K. (2013). Do non-response follow-ups improve or reduce data quality?: a review of the existing literature. *J. R. Stat. Soc. A Stat.*, 176:129–145.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference Using Likelihood*. Oxford University Press.

Plass, J., Augustin, T., Cattaneo, M., and Schollmeyer, G. (2015). Statistical modelling under epistemic data imprecision: Some results on estimating multinomial distributions and logistic regression for coarse categorical data. In Augustin, T., Doria, S., Miranda, E., and Quaeghebeur, E., editors, *ISIPTA '15, Proceedings of the Ninth International Symposium on Imprecise Probability: Theories and Applications*, pages 247–256. SIPTA.

Plass, J., Cattaneo, M., Schollmeyer, G., and Augustin, T. (2017). On the testability of coarsening assumptions: A hypothesis test for subgroup independence. *Int. J. Approx. Reason.*, 90:292–306.

Ponomareva, M. and Tamer, E. (2011). Misspecification in moment inequality models: Back to moment equalities? *Econom. J.*, 14:186–203.

Rubin, D. (1976). Inference and missing data. *Biometrika*, 63:581–592.

Sánchez, L. and Couso, I. (2018). A framework for learning fuzzy rule-based models with epistemic set-valued data and generalized loss functions. *Int. J. Approx. Reason.*, 92:321–339.

Schollmeyer, G. and Augustin, T. (2015). Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data. *Int. J. Approx. Reason.*, 56:224–248.

Sikov, A. (2018). A brief review of approaches to non-ignorable non-response. *Int. Stat. Rev.*, 86:415–441.

Stoye, J. (2009a). More on confidence intervals for partially identified parameters. *Econometrica*, 77:1299–1315.

Stoye, J. (2009b). Partial identification and robust treatment choice: an application to young offenders. *J. Stat. Theory Pract.*, 3:239–254.

Tourangeau, R. and Yan, T. (2007). Sensitive questions in surveys. *Psychol. Bull.*, 133:859–883.

Trappmann, M., Gundert, S., Wenzig, C., and Gebhardt, D. (2010). PASS: A household panel survey for research on unemployment and poverty. *Schmollers Jahrbuch*, 130:609–623.

Vansteelandt, S., Goetghebeur, E., Kenward, M., and Molenberghs, G. (2006). Ignorance and uncertainty regions as inferential tools in a sensitivity analysis. *Stat. Sinica*, 16:953–979.

Venzon, D. and Moolgavkar, S. (1988). A method for computing profile-likelihood-based confidence intervals. *J. R. Stat. Soc. C Appl.*, 37:87–94.

Verbeke, G. and Molenberghs, G. (2009). *Linear Mixed Models for Longitudinal Data*. Springer.

Zhang, P. (2003). Multiple imputation: Theory and method. *Int. Stat. Rev.*, 71:581–592.

Zhang, Z. (2010). Profile likelihood and incomplete data. *Int. Stat. Rev.*, 78:102–116.

A Regression estimators in a saturated model

In a saturated model the bounds of the estimated regression coefficients can also be calculated as a direct transformation of the bounds of the estimated parameters of the latent variable distribution. For sake of illustration, we refer to the setting of **Example 1**, considering a logit model with the response function in (15) and (16). Equivalently, the model can be represented by the link function

$$g(\pi_{\mathbf{x}<}) = \ln\left(\frac{\pi_{\mathbf{x}<}}{1-\pi_{\mathbf{x}<}}\right) = \beta_0 + d(\mathbf{x})^T\boldsymbol{\beta}\,. \tag{21}$$

Considering a saturated model, we specify the linear predictor as $\beta_0 + \beta_1 \cdot \text{Abitur} + \beta_2 \cdot \text{age} + \beta_{12} \cdot \text{age·Abitur}$, where age·Abitur denotes the interaction between both covariates. As argued below, the bounds of the four estimated regression coefficients are then determined by transforming the bounds of the four estimators $\hat{\pi}_{00<}$, $\hat{\pi}_{01<}$, $\hat{\pi}_{10<}$ and $\hat{\pi}_{11<}$, hence obtaining

$$\begin{aligned}
\hat{\beta}_0 &\in \left[\ln\left(\tfrac{\underline{\hat{\pi}}_{00<}}{1-\underline{\hat{\pi}}_{00<}}\right),\ \ln\left(\tfrac{\overline{\hat{\pi}}_{00<}}{1-\overline{\hat{\pi}}_{00<}}\right)\right], \qquad \hat{\beta}_1 \in \left[\ln\left(\frac{\underline{\hat{\pi}}_{10<}}{1-\underline{\hat{\pi}}_{10<}}\right) - \overline{\hat{\beta}}_0,\ \ln\left(\frac{\overline{\hat{\pi}}_{10<}}{1-\overline{\hat{\pi}}_{10<}}\right) - \underline{\hat{\beta}}_0\right] \\
\hat{\beta}_2 &\in \left[\ln\left(\tfrac{\underline{\hat{\pi}}_{01<}}{1-\underline{\hat{\pi}}_{01<}}\right) - \overline{\hat{\beta}}_0,\right. \qquad \hat{\beta}_{12} \in \left[\ln\left(\frac{\underline{\hat{\pi}}_{11<}}{1-\underline{\hat{\pi}}_{11<}}\right) - \overline{\hat{\beta}}_1 - \overline{\hat{\beta}}_2 - \underline{\hat{\beta}}_0,\right. \\
&\qquad \left.\ln\left(\tfrac{\overline{\hat{\pi}}_{01<}}{1-\overline{\hat{\pi}}_{01<}}\right) - \underline{\hat{\beta}}_0\right], \qquad\qquad \left.\ln\left(\frac{\overline{\hat{\pi}}_{11<}}{1-\overline{\hat{\pi}}_{11<}}\right) - \underline{\hat{\beta}}_1 - \underline{\hat{\beta}}_2 - \overline{\hat{\beta}}_0\right],
\end{aligned} \tag{22}$$

where $\underline{\hat{\beta}}_0$, $\underline{\hat{\beta}}_1$, and $\underline{\hat{\beta}}_2$ ($\overline{\hat{\beta}}_0$, $\overline{\hat{\beta}}_1$, and $\overline{\hat{\beta}}_2$) represent the estimated lower (upper) bounds of the regression coefficients. Since the saturated model is fully nonparametric, it does not induce any restrictions on the coarsening parameters.

Derivation of the estimators in (22):

- Define $\lambda_0 = \ln(\frac{\pi_{00<}}{1-\pi_{00<}})$, $\lambda_1 = \ln(\frac{\pi_{10<}}{1-\pi_{10<}})$, $\lambda_2 = \ln(\frac{\pi_{01<}}{1-\pi_{01<}})$, $\lambda_3 = \ln(\frac{\pi_{11<}}{1-\pi_{11<}})$
- λ_0, λ_1, λ_2, λ_3 can be arbitrarily chosen in the intervals formed by the corresponding bounds:

$$
\begin{aligned}
&\underline{\lambda}_0 \le \lambda_0 = \beta_0 \le \overline{\lambda}_0 && \to \underline{\beta}_0 = \underline{\lambda}_0;\ \ \overline{\beta}_0 = \overline{\lambda}_0 \\
&\underline{\lambda}_1 \le \lambda_1 = \beta_0 + \beta_1 \le \overline{\lambda}_1 && \to \underline{\beta}_1 = \underline{\lambda}_1 - \overline{\lambda}_0 = \underline{\lambda}_1 - \overline{\beta}_0;\ \overline{\beta}_1 = \overline{\lambda}_1 - \underline{\beta}_0 \\
&\underline{\lambda}_2 \le \lambda_2 = \beta_0 + \beta_2 \le \overline{\lambda}_2 && \to \beta_2 \text{ analogously} \\
&\underline{\lambda}_3 \le \lambda_3 = \beta_0 + \beta_1 + \beta_2 + \beta_{12} \le \overline{\lambda}_3
\end{aligned}
$$

$$
\begin{aligned}
\beta_{12} &= \lambda_3 - \lambda_1 - \lambda_2 + \lambda_0 \\
\underline{\beta}_{12} &= \underline{\lambda}_3 - \overline{\lambda_1} - \overline{\lambda}_2 + \underline{\lambda}_0 \\
&= \underline{\lambda}_3 - (\overline{\beta}_1 + \underline{\beta}_0) - (\overline{\beta}_2 + \underline{\beta}_0) + \underline{\beta}_0 \\
&= \underline{\lambda}_3 - \overline{\beta}_1 - \underline{\beta}_0 - \overline{\beta}_2 \ \ , \ \text{analogously for } \ \overline{\beta}_{12}\ .
\end{aligned}
$$

B Optimization problem in a non-saturated model (Example 1)

Here we refer to $\pi_{\mathbf{x}<} = h(\beta_0 + \beta_1 \cdot \text{Abitur} + \beta_2 \cdot \text{age})$ of **Example 1** with the response function in (15) and (16) and present the optimization problem for the effect of Abitur:

$$
\begin{aligned}
&\beta_1 \to \min/\max \quad \text{subject to} \\
&\underline{\hat{\pi}}_{00<} \le \frac{\exp(\beta_0)}{1+\exp(\beta_0)} \le \overline{\hat{\pi}}_{00<}, && \underline{\hat{\pi}}_{10<} \le \frac{\exp(\beta_0+\beta_1)}{1+\exp(\beta_0+\beta_1)} \le \overline{\hat{\pi}}_{10<}, \\
&\underline{\hat{\pi}}_{01<} \le \frac{\exp(\beta_0+\beta_2)}{1+\exp(\beta_0+\beta_2)} \le \overline{\hat{\pi}}_{01<}, && \underline{\hat{\pi}}_{11<} \le \frac{\exp(\beta_0+\beta_1+\beta_2)}{1+\exp(\beta_0+\beta_1+\beta_2)} \le \overline{\hat{\pi}}_{11<}\ .
\end{aligned}
\tag{23}
$$

By using the link function in (21), this optimization problem can be transformed into one with linear constraints:

$$
\begin{aligned}
&\beta_1 \to \min/\max \quad \text{subject to} \\
&\ln\Big(\frac{\underline{\hat{\pi}}_{00<}}{1-\underline{\hat{\pi}}_{00<}}\Big) \le \beta_0 \le \ln\Big(\frac{\overline{\hat{\pi}}_{00<}}{1-\overline{\hat{\pi}}_{00<}}\Big), && \ln\Big(\frac{\underline{\hat{\pi}}_{10<}}{1-\underline{\hat{\pi}}_{10<}}\Big) \le \beta_0+\beta_1 \le \ln\Big(\frac{\overline{\hat{\pi}}_{10<}}{1-\overline{\hat{\pi}}_{10<}}\Big), \\
&\ln\Big(\frac{\underline{\hat{\pi}}_{01<}}{1-\underline{\hat{\pi}}_{01<}}\Big) \le \beta_0+\beta_2 \le \ln\Big(\frac{\overline{\hat{\pi}}_{01<}}{1-\overline{\hat{\pi}}_{01<}}\Big), && \ln\Big(\frac{\underline{\hat{\pi}}_{11<}}{1-\underline{\hat{\pi}}_{11<}}\Big) \le \beta_0+\beta_1+\beta_2 \le \ln\Big(\frac{\overline{\hat{\pi}}_{11<}}{1-\overline{\hat{\pi}}_{11<}}\Big)\ .
\end{aligned}
\tag{24}
$$

C Condition to distinguish situation 1 from situation 2

We take the opitmization problem in Appendix B as a starting point and aim at deriving a condition that tells us whether the optimization problem is solvable (situation 1) or not (situation 2).

$$\ln(\frac{\hat{\underline{\pi}}_{11<}}{1-\hat{\underline{\pi}}_{11<}}) \leq \beta_0+\beta_1+\beta_2 \quad \rightarrow \beta_0+\beta_1+\beta_2 \in [\ln(\frac{\hat{\underline{\pi}}_{11<}}{1-\hat{\underline{\pi}}_{11<}}), \infty] =: I_1$$
$$\ln(\frac{\hat{\overline{\pi}}_{11<}}{1-\hat{\overline{\pi}}_{11<}}) \geq \beta_0+\beta_1+\beta_2 \quad \rightarrow \beta_0+\beta_1+\beta_2 \in [-\infty, \ln(\frac{\hat{\overline{\pi}}_{11<}}{1-\hat{\overline{\pi}}_{11<}})] =: I_2$$
$$\ln(\frac{\hat{\underline{\pi}}_{10<}}{1-\hat{\underline{\pi}}_{10<}}) + \ln(\frac{\hat{\underline{\pi}}_{01<}}{1-\hat{\underline{\pi}}_{01<}}) - \ln(\frac{\hat{\overline{\pi}}_{00<}}{1-\hat{\overline{\pi}}_{00<}}) \leq \beta_0+\beta_1+\beta_2$$
$$\ln(\frac{\hat{\overline{\pi}}_{10<}}{1-\hat{\overline{\pi}}_{10<}}) + \ln(\frac{\hat{\overline{\pi}}_{01<}}{1-\hat{\overline{\pi}}_{01<}}) - \ln(\frac{\hat{\underline{\pi}}_{00<}}{1-\hat{\underline{\pi}}_{00<}}) \geq \beta_0+\beta_1+\beta_2$$

$$\rightarrow \beta_0+\beta_1+\beta_2 \in [\ln(\frac{\hat{\underline{\pi}}_{10<}}{1-\hat{\underline{\pi}}_{10<}}) + \ln(\frac{\hat{\underline{\pi}}_{01<}}{1-\hat{\underline{\pi}}_{01<}}) - \ln(\frac{\hat{\overline{\pi}}_{00<}}{1-\hat{\overline{\pi}}_{00<}}),$$
$$\ln(\frac{\hat{\overline{\pi}}_{10<}}{1-\hat{\overline{\pi}}_{10<}}) + \ln(\frac{\hat{\overline{\pi}}_{01<}}{1-\hat{\overline{\pi}}_{01<}}) - \ln(\frac{\hat{\underline{\pi}}_{00<}}{1-\hat{\underline{\pi}}_{00<}})] =: I_3$$

To have at least one solution, it has to be valid that $I_1 \cap I_3 \neq \emptyset$ and $I_2 \cap I_3 \neq \emptyset$. Hence we obtain the following condition:

$$\ln\left(\frac{\hat{\underline{\pi}}_{11<}}{1-\hat{\underline{\pi}}_{11<}}\right) + \ln\left(\frac{\hat{\underline{\pi}}_{00<}}{1-\hat{\underline{\pi}}_{00<}}\right) \leq \ln\left(\frac{\hat{\overline{\pi}}_{10<}}{1-\hat{\overline{\pi}}_{10<}}\right) + \ln\left(\frac{\hat{\overline{\pi}}_{01<}}{1-\hat{\overline{\pi}}_{01<}}\right)$$
$$\ln\left(\frac{\hat{\underline{\pi}}_{10<}}{1-\hat{\underline{\pi}}_{10<}}\right) + \ln\left(\frac{\hat{\underline{\pi}}_{01<}}{1-\hat{\underline{\pi}}_{01<}}\right) \leq \ln\left(\frac{\hat{\overline{\pi}}_{11<}}{1-\hat{\overline{\pi}}_{11<}}\right) + \ln\left(\frac{\hat{\overline{\pi}}_{00<}}{1-\hat{\overline{\pi}}_{00<}}\right)$$

D Illustration of situations 1a, 1b, and 2

A schematic illustration of the mappings considered in this paper is given in Figure 8, where the squares symbolize the parameter spaces of the respective parameters: The first mapping refers to the relation between the regression coefficients and the parameters of the latent world (formalized via the response function) studied in Section 5.2, the second to the relation between the parameters of the latent and the observed worlds expressed via Φ (cf. Section 3.2).

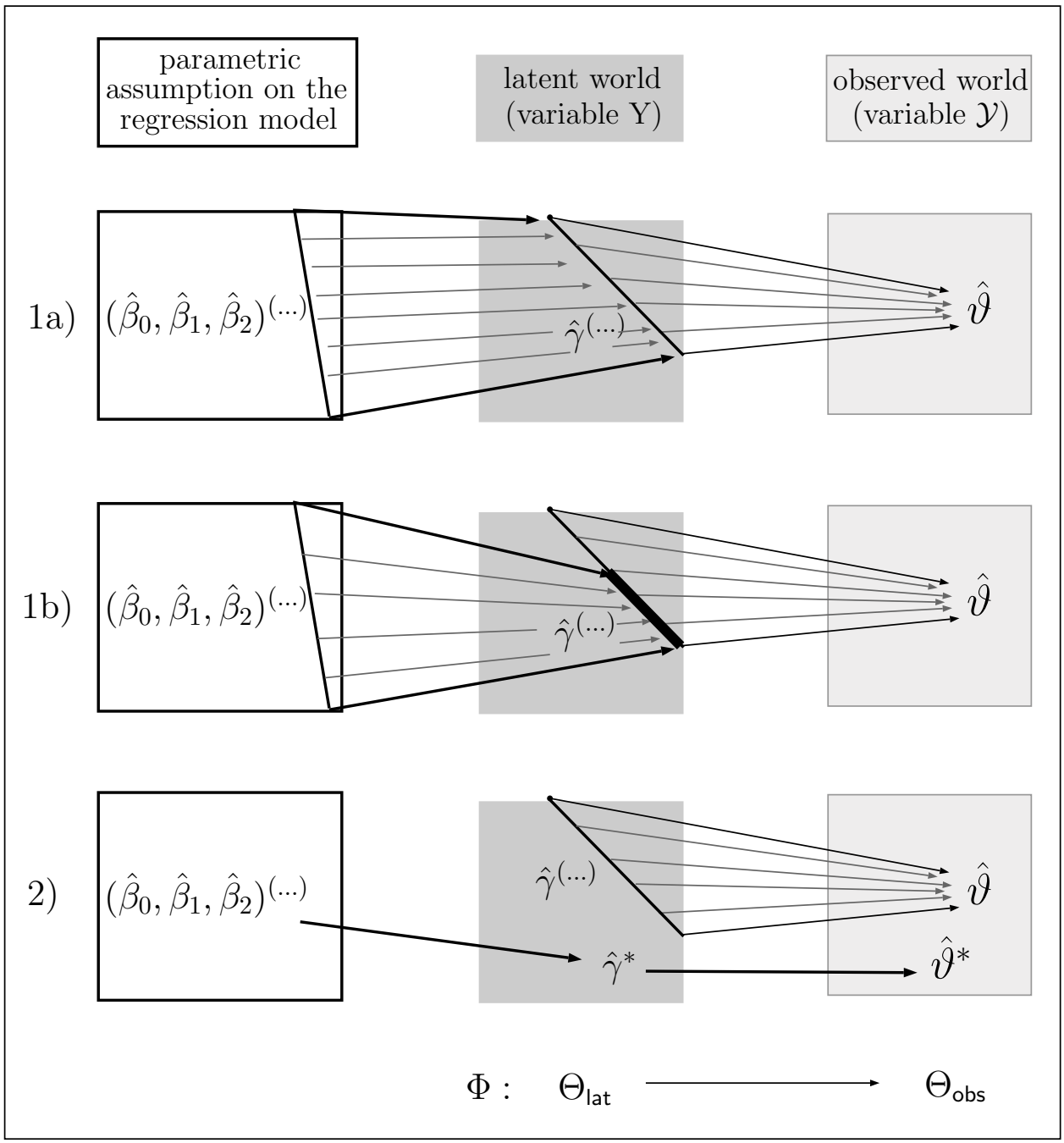

Figure 8: The principally differing situations 1a, 1b, and 2.

E Studying compatible coarsening processes in the different situations

To stress the estimation of the latent variable distribution under the parametric assumptions, we include a star into the notation. In this way, $\pi^*_{\mathbf{x}y} = h'(\eta_{\mathbf{x}y})$ refers to a linear predictor $\eta_{\mathbf{x}y}$ setting at least one interaction to zero, while $\pi_{\mathbf{x}y}$ is calculated without parametric assumptions.

From each situation (situation 1a, 1b, or 2; cf. Section 5.2 and Appendix D) we can directly infer whether and how the intervals for the coarsening estimators sharpen due to the parametric assumptions of the regression model: In situation 1, the estimates of the latent variable distribution fit to the data in the sense that the estimates for the parameters of the observed world, i.e. $\hat{p}_{\mathbf{x}\mathfrak{y}}$, $\mathbf{x} \in S_X$, $\mathfrak{y} \in S_\mathcal{Y}$, are unaffected by the parametric assumptions of the regression model. In this way, one still achieves the same maximal value of the likelihood as in the nonparametric case. Since in situation 1a also the estimated bounds of the latent variable distribution $\pi^*_{\mathbf{x}y}$ coincide with the ones obtained without parametric assumptions $\pi_{\mathbf{x}y}$ (cf. the first picture in Figure 8 in Appendix D), the estimated bounds of the coarsening parameters remain unchanged by the parametric assumptions as well. By contrast, in situation 1b not both estimated bounds of the coarsening parameters are obtained (cf. the second picture in Figure 8 in Appendix D): By applying the relation in (9) for the binary case and solving for the coarsening parameters, we obtain

$$\begin{aligned} p_{\mathbf{x}<} &= \pi^*_{\mathbf{x}<} \cdot (1 - q_{na|\mathbf{x}<}) \\ p_{\mathbf{x}<} &= \pi^*_{\mathbf{x}<} - \pi^*_{\mathbf{x}<} \cdot q_{na|\mathbf{x}<} \\ q_{na|\mathbf{x}<} &= \frac{\pi^*_{\mathbf{x}<} - p_{\mathbf{x}<}}{\pi^*_{\mathbf{x}<}} = 1 - \frac{p_{\mathbf{x}<}}{\pi^*_{\mathbf{x}<}} \end{aligned}$$

and similarly for $\hat{q}_{na|\mathbf{x}\geq}$. Hence, the following estimated bounds for the coarsening parameters are achievable:

$$\hat{q}_{na|\mathbf{x}<} \in \left[1 - \frac{\hat{p}_{\mathbf{x}<}}{\underline{\hat{\pi}}^*_{\mathbf{x}<}},\ \frac{\overline{\hat{\pi}}^*_{\mathbf{x}<} - \hat{p}_{\mathbf{x}<}}{\overline{\hat{\pi}}^*_{\mathbf{x}<}}\right] \quad \text{and} \quad \hat{q}_{na|\mathbf{x}\geq} \in \left[1 - \frac{\hat{p}_{\mathbf{x}\geq}}{\underline{\hat{\pi}}^*_{\mathbf{x}\geq}},\ \frac{\overline{\hat{\pi}}^*_{\mathbf{x}\geq} - \hat{p}_{\mathbf{x}\geq}}{\overline{\hat{\pi}}^*_{\mathbf{x}\geq}}\right], \tag{25}$$

assuming all quantities in the denominator to be greater than zero. Whenever $\underline{\hat{\pi}}^*_{\mathbf{x}<} = \underline{\hat{\pi}}_{\mathbf{x}<}$, then $\underline{\hat{\pi}}^*_{\mathbf{x}<} = \hat{p}_{\mathbf{x}<}$ is valid, so that the lower bound of $\hat{q}_{na|\mathbf{x}<}$ stays zero and is thus not refined (while analogous conclusions can be made for the lower bound of $\hat{q}_{na|\mathbf{x}\geq}$). Due to $\underline{\hat{\pi}}^*_{\mathbf{x}y} \geq \underline{\hat{\pi}}_{\mathbf{x}y}$ and/or $\overline{\hat{\pi}}^*_{\mathbf{x}y} \leq \overline{\hat{\pi}}_{\mathbf{x}y}$, the bounds in (25) are generally not wider than those obtained without parametric assumptions.

In situation 2, the estimated observed variable distribution is no longer the empirical distribution (cf. the third picture in Figure 8 in Appendix D). In general, there are no longer multiple virtual values that are compatible with the observed variable distribution and the

likelihood in terms of γ is maximized uniquely. This is also confirmed by the profile log-likelihood in Figure 7, which refers to an illustrative data constellation classified to situation 2 ($n_{00<} = 60,\ n_{00\geq} = 10,\ n_{00na} = 10,\ n_{10<} = 30,\ n_{10\geq} = 40,\ n_{10na} = 5,\ n_{01<} = 20,\ n_{01\geq} = 50,\ n_{01na} = 2, n_{11<} = 40,\ n_{11\geq} = 10,\ n_{11na} = 5$, which does indeed not satisfy the condition in Appendix C for belonging to situation 1).

To sum up, the parametric assumptions of the regression model may have a different impact on the estimated coarsening parameters, from no impact (situation 1a), via a slight impact (situation 1b) through to a very strong impact making the estimates precise (situation 2).