LMU

LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR STATISTIK

Coarse categorical data under epistemic and ontologic uncertainty: Comparison and extension of some approaches

Julia Plaß

Master Thesis

Supervision: Prof. Dr. Thomas Augustin, Dipl.-Math. Georg Schollmeyer
Department of Statistics – University of Munich

July 15, 2013

Abstract

There are different reasons for coarse data, namely epistemic and ontologic uncertainty. While epistemic uncertainty can be induced by survey specific measures and problems as preservation of the respondents' anonymity, indecision of respondents can cause coarse data under ontologic uncertainty. Although coarse data are widely present in this way, it is still an open topic how to deal with data of that kind.

Therefore, in this master's thesis some methods that partly have been used in other areas exclusively until now will be investigated and applied in this context, where it will be concentrated on coarse categorical data. The concept of coarsening at random and methods as partial identification and sensitivity analysis can be helpful in connection with the analysis of epistemic uncertainty, where the theory of random sets and the Dempster-Shafer theory will serve as the basis for the introduction of the *-notation that is able to deal with ontologic uncertainty. Moreover, it will be illustrated by means of simulated data how a categorical dependent variable that is either coarse because of epistemic or ontologic uncertainty can be incorporated within a multinomial logit model. Finally, a comparison of these modelling approaches will emphasize the importance of the distinction between these two types of uncertainty.

Acknowledgement

There are many people who accompanied me through the ups and downs of writing this master's thesis, wherefore I am very grateful. I would particularly like to thank Prof. Dr. Thomas Augustin and Dipl.-Math. Georg Schollmeyer for their outstanding supervision and their always available willingness for extremely helpful discussions and valuable comments. Moreover, I am thankful for Prof. Augustin's suggestion of this exciting topic and the encouraging words especially at the beginning of this project. Furthermore I appreciate the fair comments of his working group at the research seminar.

A special thanks goes to Matthias Speidel who always psyched me up in stressful times and who was always ready for long hours of discussions. Thank you, Matthias, Ari, Flo, Basti, Sandra, Good-Craig, Just-Craig and Francesca, my attentive typo-experts and proofreaders. A last but definitely not least "thank you" goes to my family who always support me in my decisions.

Contents

1. Introduction to the problem of coarse data

"Once I make up my mind, I'm full of indecision."

– Oscar Levant (1906-1972) –

Regardless which area of life is considered, it is characteristic for human beings to balance between several options. Nevertheless, the indecisiveness of respondents is not reflected in most surveys and instead it is common to force a precise answer or to provide a ""Don't know" category" for those that have not made their decision yet. It is obvious that information is lost if one proceedes in this way, as even indecision between several possibilities reveals some information by definitely excluding some options. Thus, it could be worth to account for this so called coarse data under *ontologic uncertainty* (Greek: onto - "being", logia - "theory") that can be induced by indecision of respondents. Data under ontologic uncertainty are coarse by nature, but can be observed in a precise way. Thereby, it is important to emphasize that these coarse answers represent the truth, as even the indecisive respondents do not know which of the possiblities that are consistent with their coarse answer shows the one that fits best to their preferences. For instance, if a respondent reports to be indifferent between electing party "A" and "B", he does not know yet if he intends to elect party "A" or party "B", but already knows that he will decide against party "C".

Data can not only be coarse by nature, but also actually show precise true values that potentially can only be observed in a coarse way. Data of that kind are called coarse data under *epistemic uncertainty* (Greek: epistēmē - "knowledge") and there are different reasons for obtaining such data. Here three reasons, namely the guarantee of anonymization, the prevention of refusals and restricted measurement accuracy, will be addressed in more detail.

The capability to record more and more data about all imaginable areas of life has increased permanently during the last years, so that the need for privacy and thus anonymization of data is more important than ever (Cormode and Srivastava 2010, p. 1015). Survey institutes have to follow the instructions of data protection and hence in many cases questionnaires provide anwers of rather coarse character. In this way, for instance it is asked for the data in terms of grouping classes or at least it is ensured to anonymize data after data collection. As essentially true answers are underlying which are coarsened in a second step and thus can only be observed in an imprecise way, it is clearly evident that anonymization generates data under epistemic uncertainty.

If one disregards reasons of data protection for a moment, at the first glance it seems to be contradictory to coarsen actually precise data in a first step and deanonymize those data afterwards in order to obtain precise results again. But especially in case of sensitive questions (e.g. income, drug consumption), one expects being able to prevent refusals and obtaining more honest answers by recording coarse instead of precise data.

Moreover, sometimes only coarsened answers of precise values are available, because some respondents report their answer with restricted accuracy only, so that for instance data can be rounded or heaped. Rounding is present in many cases where respondents are required to report a particular metric value, for instance the time required for commuting between home and work (e.g. “pairfam”, wave 1, Q227, Nauck, B. and Brüderl, J. and Huinink, J. and Walper, S. 2013). Therby rounding can be induced either by convenience (e.g. half an hour instead of 24 min.) or by the inability or limited possibility to report precise results (e.g. 24 min. and 11 sec instead of 24 min.). Against this, one is concerned with heaping if data are collected with various levels of coarseness. Age heaping is a frequently used example (Heitjan and Rubin 1991, p. 2251), as it is reasonable that the age of babies, children and adults is reported in months, half years and years respectively.

Thus, generally the presence of coarse data under epistemic uncertainty can be ascribed to measures and problems that are connected with the study design.

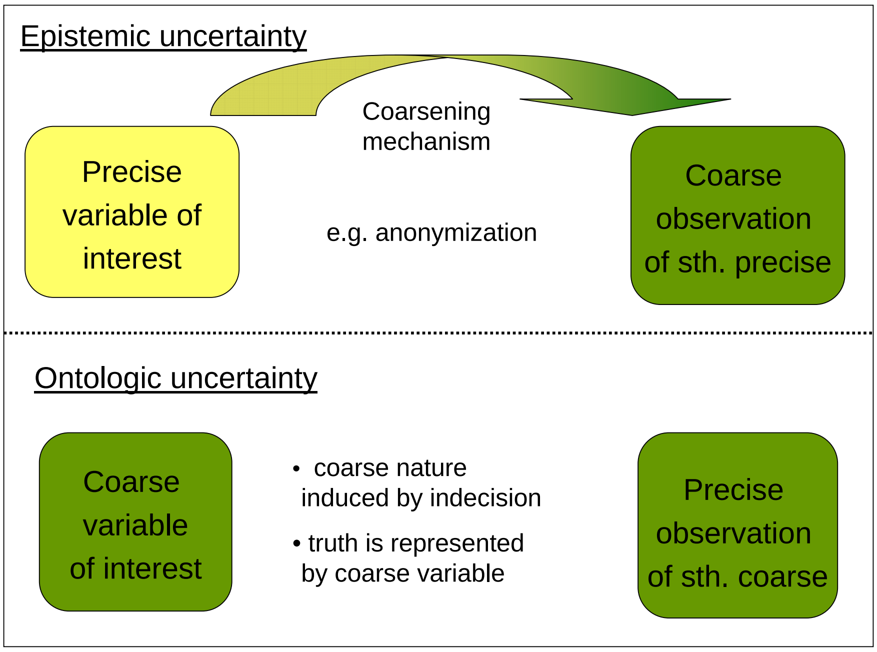

Epistemic uncertainty
Coarsening mechanism
Precise variable of interest
e.g. anonymization
Coarse observation of sth. precise
Ontologic uncertainty
Coarse variable of interest
• coarse nature induced by indecision
• truth is represented by coarse variable
Precise observation of sth. coarse

Figure 1.1.: Difference between epistemic and ontologic uncertainty.

As coarse data under epistemic uncertainty show actually true values that potentially can only be observed in a coarsened way, a coarsening mechanism is underlying in this case. The absence of a coarsening mechanism in the case of ontologic uncertainty, where data are already of coarse nature, represents one of the most important differences between those two situations of coarse data, which are illustrated by means of Figure 1.1. Because of these differences completely different goals have to be persued in the framework of the corresponding analyses of data, so that in the context of data under epistemic uncertainty it is of peculiar interest to consider the underlying coarsening mechanism and to investigate the true underlying answers, where under ontologic uncertainty a framework that is able to involve the actually coarse answers is more important.

The deliberations and examples that have been presented in the context of explaining coarse data that are either induced by epistemic or ontologic uncertainty show that coarse data are widely present. Nevertheless, no consensus has been reached on how to deal with coarse data, wherefore the goal of this thesis will consist of considering and comparing some approaches that address coarse data under epistemic and ontologic uncertainty. In this framework some methods that partly have been used in other areas exclusively until now will be investigated and applied in this context. In order to maintain a clear scope of this thesis, it will be concentrated on coarse categorical data. For illustration of most ideas in case of epistemic uncertainty a simple example will be considered with two true categories "A" and "B", which are observed as "A", "B" or "A XOR B". "XOR" is a commonly used sign representing "either...or". Against this, under ontologic uncertainty the notation will be such that three categories "$\{A\}$", "$\{B\}$" and "$\{A,\ B\}$" can be regarded as true categories, where the latter one expresses indecision between category "A" and "B".

For the purpose of pursuing these intentions, it will be proceeded as follows:

In Chapter 2 the basic problem of coarse data under epistemic uncertainty will be explained, where the simplifying property of "coarsening at random" as well as partial identification and sensitivity analysis, two procedures that rely on implying justified assumptions only, will show possible approaches that try to deal with this question. As the latter two methods are prevalent in the framework of missing data only, they will be applied in the context of coarsened data in this chapter.

Against this, in the context of dealing with ontologic uncertainty, addressed in Chapter 3, the introduction of the *-notation, which relies on an analysis on the power set and allows to represent data that are coarse by nature, is of peculiar interest. The conceptions of that notation are based on some ideas of the random set theory and the Dempster-Shafer theory.

After some general approaches for dealing with coarse data under epistemic and ontologic uncertainty have been investigated, it will be analysed how a coarse dependent variable can be involved within a regression model. As in this thesis categorical data are addressed, the multinomial logit model will represent the basic model in this context.

In Chapter 4 a coarse dependent variable under epistemic uncertainty will be incorporated into an *iid* model as well as a multinomial logit model with two covariates. Thereby, the resulting identification problem and some approaches of Chapter 2 that try to deal with it will be of main interest. As analysis will be based on simulated data, true values are available and thus these methods can be evaluated by looking at the relative empirical. Moreover, an alternative imputation based approach will be sketched.

In Chapter 5 an extended multinomial logit model will be proposed that is able to include a coarse dependent variable under ontologic uncertainty. After having focused on the particularity of this model in general, for reasons of consistency the same models as in Chapter 3 will be regarded again. Moreover, some ideas of Dempster-Shafer theory concerning prediction of results when decisions have been made will be shown, where the implication of some assumptions that further restrict this interval valued result will be discussed.

Chapter 6 will collect the ideas of Chapter 4 and Chapter 5 by comparing the multinomial logit based approach under epistemic and ontologic uncertainty, where the importance of the distinction between these two types of uncertainty will be emphasized.

Finally, in Chapter 7 the main results will be summarized and some open questions will be focused that could not be investigated in the framework of this thesis.

2. Approaches for dealing with coarse categorical data under epistemic uncertainty

After differences between epistemic and ontologic uncertainty have been worked out in Chapter 1, the importance of distinguishing between those concepts in the development of appropriate approaches should be obvious. Therefore, the already existing approaches dealing with data under epistemic uncertainty will be the main focus in this chapter, while the case of ontologic uncertainty will be covered in Chapter 3.

It is reasonable to show the basic formal situation of working with data under epistemic uncertainty first, which can be regarded as the starting point for several approaches that will be presented in this chapter. Moreover I want to show the notation, which is used here and is partly adopted from Heitjan and Rubin [1991], in this context.

The essential problem of the presence of epistemic uncertainty as already described in Chapter 1 consists of the fact that the true values of the characteristic of interest, denoted here by random variable Y, can not be observed in an exact way, but only in a coarsened form instead, denoted here by random variable $\mathcal{Y}$, written in a calligraphic way. At a first glance the labeling seems to be somewhat unusual (typically X and Y are devoted respectively), but in respect of asuming only the dependent variable being coarsened in this thesis, this decision can make sense indeed.

Furthermore the characteristic of interest Y is categorical as this thesis will concentrate on this case. Its true, but potentially unobserved, values lie in the sample space Ω with the result that $\mathcal{Y}$ is able to take values in the power set of Ω, containing 2^{Ω} elements. To keep things simple the sample space of $\mathcal{Y}$ can be reduced by regarding only those sets in the power set to which a positive probability is assigned. For instance, if the true values of a characteristic take the values “A”, “B” or “C”, the power set contains 2^3 sets, namely “A”, “B”, “C”, “A XOR B”, “A XOR C”, “B XOR C”, “A, B XOR C” and the empty set. Because of the mentioned restriction only the possible sets of the underlying situation are included in the sample space of the observed quantity $\mathcal{Y}$.

Having this notation in mind, the following basic equation (2.1) can be explained which can be derived by the “Theorem of total probability”:

$$P(\mathcal{Y} = \mathfrak{y}) = \sum_y P(\mathcal{Y} = \mathfrak{y}|Y = y) \cdot P(Y = y). \tag{2.1}$$

As the emphasis of this thesis will be on the categorical case as well as for reasons of simplicity, it has been decided to show this equation for the discrete case. Nevertheless, in Section 2.1 this equation will be extended for general data.

The probability $P(Y = y)$ of the true but potentially unobserved variable is the quantity of interest. If the nature of the coarsening, which is expressed by $P(\mathcal{Y} = \mathfrak{y}|Y = y)$, was known, this probability directly could be calculated, because the probability of the observed variable $P(\mathcal{Y} = \mathfrak{y})$ is known or at least can be estimated. Nevertheless in most situations the coarsening process is unknown and therefore one can have big problems to derive the requested quantity $P(Y = y)$.

The approaches which I will describe in this chapter focus on this problem and try to find solutions in different ways. Because of the fact that the problem will be solved if the coarsening process is known, these approaches focus on the quantity which describes the coarsening, namely $P(\mathcal{Y} = \mathfrak{y}|Y = y)$. Moreover probability $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ will be important in this chapter, because it is able to express the underlying epistemic nature of uncertainty by emphasizing that there is a true value of the characteristic which will be coarsened in a second step. Hereafter this probability $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ is sometimes denoted by $q(y|\mathfrak{y})$.

As already mentioned in Chapter 1 for illustration the case of observing categories “A”, “B” and “A XOR B” will be considered at many points of this thesis. In this connection coarsening mechanisms $P(\mathcal{Y} = (A\ XOR\ B)|Y = A)$

and $P(\mathcal{Y} = (A\ XOR\ B)|Y = B)$ will be denoted by q_1 and q_2 respectively. In case of q_1 and q_2 being equal, which will form a case of special interest, the notation of $q_1 = q_2 = q$ will be applied.

While ignorability constitutes rather a property than a real approach for dealing with this initial problem (see Section 2.1), sensitivity analysis and partial identification can be regarded as two attempts which both try to solve the mentioned problem by getting an idea about $P(\mathcal{Y} = \mathfrak{y}|Y = y)$, but by proceeding from a different angle (see Section 2.2 and 2.3). Therefore, it will be worth to contrast these two approaches (see Section 2.4).

Although these approaches can be applied in the general case, this thesis mainly will concentrate on the presence of categorical data. But because of the fact that the concept of ignorability has been considered for the general case by Heitjan and Rubin [1991], I will summarize their findings in a general way first, before illustrations by several examples will concern the categorical case.

Partial identification and sensitivity analysis based approaches have primarily been developed in the context of the missing data problem. Thus, after summarizing the already existing background in conjunction with missing data, the objective will consist of applying these definitions in the framework of coarsened data. For reasons of simplicity I will focus on the categorical case in the course of this. Moreover it could be interesting to establish a connection to related areas like the problem of missing (see Subsection 2.1.6) or misclassified data (see Subsection 2.2.4) in this chapter.

2.1. Ignorability of coarsening

Concerning this initial problem, the easiest way to start consists of thinking about situations in which the coarsening can be ignored. There is a concept called “coarsened at random” which can be regarded as a property that simplifies a lot. In this section I will explain and illustrate this fundamental concept. Moreover it is interesting to describe some already existing extensions of this concept as well as to investigate its relation to the concept of “missing at random”.

In order to structure this procedure I decided to divide this section into several subsections, namely 2.1.1 Basic Situation, 2.1.2 Likelihood under nonstochas-

tic coarsening, 2.1.3 Likelihood under stochastic coarsening, 2.1.4 The concept of “coarsened at random”, 2.1.5 Further extensions of the concept of “coarsened at random” and 2.1.6 Relation to the missing data problem.
The following introduction to the concept of “coarsening at random” (CAR) (Subsection 2.1.1 to 2.1.4) is strictly guided by Heitjan and Rubin [1991] and equations are adapted from there. Nevertheless here the problem is motivated in a different way and there are some own explanations and illustrations. As distinguishing between nonstochastic and stochastic coarsening and their consequences in my opinion is a central aspect for the contentual understanding of CAR and the underlying equations, I have decided to work out these forms of coarsening in a detailed way and emphasized their different role in the development of the likelihood. Moreover a notation is applied, which differs slightly from the one used by them, in order to refer to the problem described above.

2.1.1. Basic Situation

Extending equation (2.1) to its general continious form, one yields the following likelihood (similar to equation (2.12) of Heitjan and Rubin [1991])

$$L(\theta, \gamma, \mathfrak{y}) = \int_{\mathfrak{y}} q(\mathfrak{y}|y, \gamma)\ f(y, \theta) dy, \tag{2.2}$$

where both components $q(\mathfrak{y}|y, \gamma)$ and $f(y, \theta)$ are analogues to $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ and $P(Y = y)$ from equation (2.1), respectively. Therefore, component $f(y, \theta)$ denotes the distribution of the true value of the characteristic which is potentially unobserved, where θ is the parameter of interest, and $q(\mathfrak{y}|y, \gamma)$ denotes the conditional distribution of the observed value given the true value, where γ is the parameter of coarsening which will be described in more detail in Subsection 2.1.3. The generality of equation (2.2) results from the flexibility in using different dominating measures, like the counting measure for the discrete case obtaining equation (2.1) from above, the lebesgue measure for the continous case, or a mixture of the lebesgue and the dirac measure for cases with continous parts and jump discontinuities.
This likelihood I want to regard as the general likelihood here and use it as a starting point for deriving the simplified grouped Likelihood L_G in Subsection 2.1.2 and the correct Likelihood L_C in Subsection 2.1.3. These forms of

likelihoods basically correspond to this general likelihood, but the underlying difference is resulting by the different coarsening mechanisms $q(\mathfrak{y}|y,\gamma)$ that can be used in the different situations of Subsection 2.1.2 and Subsection 2.1.3. In Subsection 2.1.2 I will address the situation that a nonstochastic coarsening mechanism is underlying, which will lead to a simplified $q(\mathfrak{y}|y,\gamma)$, while in Subsection 2.1.3 the precence of stochastic coarsening will be viewed.

In order to understand why the form of $q(\mathfrak{y}|y,\gamma)$ in Subsection 2.1.2 is easier compared to $q(\mathfrak{y}|y,\gamma)$ of Subsection 2.1.3, it might be useful to realize the differnece between a nonstochastic and a stochastic coarsening mechanism.

If one is faced with a nonstochastic coarsening mechanism this means that the underlying degree of coarsening is predetermined and known. This I want to illustrate by the following example: One is concerned with a questionnaire which focuses the question “How many hours do you watch TV per week?”. If only answers in intervals are available, like “I don’t watch TV at all”, “1-5 hours”, “5-10 hours”, “10-20 hours” and “more than 20 hours” and it is assumed that all respondents answer in a correct way, then the coarsening is predefined in the sense that there is a unique coarsened form for every true answer. So if there is a respondent who answers that he watches TV four hours per day, it is obvious that “1-5 hours” will be the observed answer. So one does not have to reflect on which mapping $Y \longrightarrow \mathcal{Y}$ to choose, because it is known. Other examples for nonstochastic coarsening mechanisms are censored data if fixed and known censoring times are used or rounded data if one uses a fixed rounding rule (Heitjan 1993).

Otherwise in the case of a stochastic coarsening mechanism one does not know before which degree of coarsening will be needed and the decision for a special degree is rather at random. To give an example again one could imagine a situation with heaped data. An example which is often used in the context of heaped data is the answer to the question “How many cigarettes do you smoke per day?” (Heitjan 1994). There might be people who state their exact number of cigarettes but others who report their number in complete packs. In this case it is more difficult to decide which degree of coarsening is present and therefore which mapping $Y \longrightarrow \mathcal{Y}$ to choose, because it is neither known nor predetermined. Please note that even if this kind of coarsening is called “stochastic”, the choice for a special degree of coarsening does not have to be at random, as there could be other factors which can influence the decision for a particular level of coarsening. For instance, it could be plausible that respondents who smoke large amounts of cigarettes per day rather report their answer in packages or that old people want to state their answer more exactly and thus report the real number of cigarettes. Therefore, it could be possible that the level of coarsening depends either from the variable of interest itself or from other variables. For this, the term "stochastic" coarsening in a certain manner can be misleading.

For understanding the difference between the grouped likelihood L_G (see Subsection 2.1.2) as well as the correct likelihood L_C (see Subsection 2.1.3) and the fact that there results a different $q(\mathfrak{y}|y,\gamma)$ (see Equation 2.2) for both forms of likelihoods, it is important to keep these types of coarsenings in mind.

2.1.2. Likelihood under nonstochastic coarsening

If one is concerned with a nonstochastic coarsening mechanism, things can be simplified, because the conditional probability $q(\mathfrak{y}|y,\gamma)$ does not depend on the level of coarsening and thus is independent from the parameter γ. Hence, it can be expressed in the following way (basically adopted from equation (2.1) of Heitjan and Rubin [1991]):

$$q(\mathfrak{y}|y,\gamma) = r(\mathfrak{y}|y,\theta) = \begin{cases} 1, & \text{if } \mathfrak{y} = \mathcal{Y}(y) \\ 0, & \text{if } \mathfrak{y} \neq \mathcal{Y}(y). \end{cases} \tag{2.3}$$

Therefore, in the case of grouping the conditional distribution which describes the coarsening process is only determined by the fact that the observed random variable $\mathcal{Y}$ has to be a function of the true variable Y, i.e. $\mathcal{Y}$=$\mathcal{Y}$(Y). This means that $\mathfrak{y}$ has to be interpreted as the subspace of Ω in which the true value of Y lies. Because of the fact that the degree of coarsening is predetermined, one does not have to think about which degree of coarsening is chosen and how the coarsening can be modelled as it has to be done in the presence of stochastic coarsening (see Subsection 2.1.3). That forms the essential difference between nonstochastic and stochastic coarsening concerning the development of the likelihood functions.

If one inserts $q(\mathfrak{y}|y,\gamma)$ of equation (2.3) into equation (2.2), the grouped likelihood L_G will result which can be used if a nonstochastic coarsening is underlying. Although this grouped likelihood L_G is quite desirable, things could be further facilitated by evaluating this Likelihood at the center of the grouping classes and yielding the approximated Likelihood L_A. As this procedure does not incorporate that these centers are not equal to the observed values, one can not assume generally that L_G and L_A are proportional to each other and therefore inferences concerning to L_A can be incorrect.

2.1.3. Likelihood under stochastic coarsening

Until now I have regarded the case of nonstochastic coarsening only. But additionaly it is important to focus on cases whose underlying coarsening mechanism is stochastic and the degree of coarseness is not predetermined. Therefore, the following notation is introduced which is mainly adopted from Heitjan and Rubin [1991].

For the general case of stochastic coarsening the precision of reporting has to be included to specify the prescription of the coarsening, which is a-priori undefined for stochastic coarsening by definition. This is done by establishing the precision of reporting as a new random variable G with sample space Γ, whose value g affects the coarsening in that way that it decides which of the mappings $Y \longrightarrow \mathcal{Y}$ should be applied for determining the coarsening process. Because of introducing random variable G, random variable $\mathcal{Y}$ can be expressed now as a function not only of Y, but also as a function of G. Therefore, r from equation (2.3), which determines how the true variable of interest Y turns into $\mathcal{Y}$, changes to (see equation (2.8) from Heitjan and Rubin [1991])

$$r(\mathfrak{y}|y, g, \theta, \gamma) = \begin{cases} 1, & \text{if } \mathfrak{y} = \mathcal{Y}(y, g) \\ 0, & \text{if } \mathfrak{y} \neq \mathcal{Y}(y, g). \end{cases} \tag{2.4}$$

So value $\mathfrak{y}$ will be observed if on the one hand the true value lies in $\mathfrak{y}$ and on the other hand it is determined by a corresponding g, i.e. $\mathfrak{y} = \mathcal{Y}(y, g)$.

But there is a kind of problem, because even though the observed value y can give an idea about the underlying value of G, the precision of reporting cannot be observable in general. Due to this, densitiy $h(g|y, \gamma)$ that models the precision of reporting has to be involved, where γ describes the corresponding parameter. Thus, the resulting $q(\mathfrak{y}|y,\gamma)$ can be calculated by the following equation (in the main equation (2.9) of Heitjan and Rubin [1991]):

$$q(\mathfrak{y}|y,\gamma) = \int_{\Gamma} r(\mathfrak{y}|y,g)\ h(g|y,\gamma)dg. \tag{2.5}$$

If one inserts $q(\mathfrak{y}|y,\gamma)$ of equation (2.5) into equation (2.2) the correct likelihood for the general case of stochastic coarsening results, which will be denoted by L_C. This likelihood L_C is preferable, because it incorporates not only the coarsening of Y (by term r), but is also able to handle types of coarsening whose degree of coarsening is stochastic (by extended term r and term h). Nevertheless, it is obvious that the calculation of this correct likelihood L_C is much more complex than the one of the grouped likelihood L_G in which $q(\mathfrak{y}|y,\gamma)$ reduces to r (see Equation (2.3)). This fact forms the motivation for the development of the property "coarsened at random".

2.1.4. The concept of "coarsened at random"

Even if L_G might be desirable because of its easier calculation, it can sometimes be incorrect, because it does not account for the stochastic nature of the coarsening process. Therefore, focusing on the following question could be interesting:

> ***Question:*** *Under which properties is it possible to use the grouped likelihood instead of the correct likelihood, i.e. in which cases is $L_G \propto L_C$?*

Parameter distinctness forms an important property in this context. This property is well known from the "missing at random" assumption in the missing data context and means here that the parameter γ which governs the coarsening mechanism and the parameter θ of the true distribution of the variable of interest belong to disjoint parameter spaces (Rubin 1976). "Coarsened at random" forms a second property in this framework, formulated by Heitjan and Rubin [1991, p. 2248] (adjusted notation):

> *"The data y are coarsened at random (CAR) if, for the fixed observed value of $\mathfrak{y}$ and for each value of γ, $q(\mathfrak{y}|y,\gamma)$ takes the same value for all*

> *$y \in \mathbf{y}$, that is, for all values of y that are consistent with the observed coarse data $\mathfrak{y}$."*

Simplified and in own words this means that under CAR probability $q(\mathfrak{y}|y, \gamma)$ which determines the coarsening process is the same no matter which true value y is underlying. This has to be fulfilled for each coarsening mechanism that could be imaginable.
Imagine for illustration the problem that the answers to an item in a questionnaire ("a", "b", "c" or "d") cannot be observed exactly, but only in a coarsened variant instead. That could be useful if the respondents' anonymity is wished to be preserved. In this case "coarsened at random" means for example that for the fixed observed value "c or d" and for each feasable γ, probability $q((c\ XOR\ d)|y, \gamma)$ takes the same value for all true values that correspond with the observed data, namely true value "$y = c$" and "$y = d$". Thus the probability that determines the coarsening mechanism is constant ($q(\mathcal{Y} = \mathfrak{y}|Y = y)$=const) no matter which true value y is underlying as long as it fits to the observed value $\mathfrak{y}$.
From this little example one can notice directly that CAR entails quite much information about the coarsening process. Therefore, one has to pay attention if in the regarded situation parameter distinction and CAR are actually valid, because Heitjan and Rubin [1991, p. 2249] have shown that using L_G can be misleading if these properties are not satisfied.
Although Heitjan and Rubin [1991] state their conclusion in a more detailed way (equal likelihood ratio and posterior distribution of L_G and L_C), the following answer to the question above should be sufficient here:

> ***Answer:*** *If data are CAR and parameter distinctness of θ and γ is satisfied then L_G and L_C are proportional to each other and it is permitted to calculate L_G instead of L_C.*

A coarsening process for which CAR as well as parameter distinctness are valid can be called "ignorable" (notation transfered from missing data problem, e.g. Little and Rubin 2002). Because the observed variable does not depend on the true value as long as it corresponds to the observed value, the likelihood does not depend on the true, but potentially unobserved, values and therefore inferences that ignore the coarsening process can be made. This explaines the intuition of calling such a process “ignorable”.

To get an overview of some examples and applications concerning the concept of CAR, Heitjan [1993] might be helpful. For instance, different types of coarsening are regarded in this context and it is shown that rounding, type I censoring (see Kalbfleisch and Prentice [2011, p. 41]), which is present if the censoring times are fixed, and type II censoring (see Kalbfleisch and Prentice [2011, p. 41]), which investigates censoring after the fixed d-th failure, are always CAR. Moreover the problem of “Competing Risks” is handeled there, which is of interest if there are other reasons than the treated one, that lead to censored failure times. Therefore, it could sometimes be suggestive to face the question “time until death” instead of “time until death from a special illnesss which is of interest”.

To get a clearer understanding of the nature of CAR, it might be useful to focus on some depictions of CAR that Gill et al. [1997, p. 8] show for the discrete case, namely $CAR(\mathcal{Y}|Y)$, $CAR(Y|\mathcal{Y})$ and $FACTOR(\mathcal{Y})$. $CAR(\mathcal{Y}|Y)$ and $CAR(Y|\mathcal{Y})$ are closely linked to each other as they differ by a weighting factor only as will be shown more detailed in Equation 2.18. $CAR(\mathcal{Y}|Y)$ equals the definition which has already been given on page 14 and one advantage of this kind of representation is that the two stage procedure is expressed in the sense that first the random variable of interest Y is generated and after that it is observed in a coarsened form $\mathcal{Y}$ by a distinct process. The chronology of considering $CAR(Y|\mathcal{Y})$ is different as the assumption is made for given observed data. If for instance one has observed $\mathcal{Y}$ = “(a XOR b)”, under this CAR presentation “coarsening at random” only implies that the true value is an element of “a XOR b”, so it is either “a” or “b” (Gill et al. 1997, p. 6), which represents an obvious fact. The third CAR display concernes the factorization of the marginal distribution of $\mathcal{Y}$ for the discrete case ($P(\mathcal{Y} = A) = P(Y \in A)\ P(\mathcal{Y} = A|Y \in A)$) (Gill et al. 1997, p. 6). Gill et al. [1997, p. 7-8] have shown that given the observed data one can create the variable of interest Y, such that the observed variable represents a coarsening of Y and CAR holds. Moreover the corresponding factorization is unique. For that reason Gill et al. [1997] conclude their statement “CAR is everything”.

In my opinion one has to treat this statement with caution, because it only means that for every observed data there is a way to construct a variable of interest such that CAR is satisfied , but it does not express that CAR probabilities are justified in every case and thus interpreting this statement as "Everything is CAR" could be misleading. Therefore, I summarize the property of CAR as an assumption that may simplify the corresponding likelihood (see page 15), but whose plausibility has to be checked first. Two methods that are able to embed CAR without relying on this assumption only, namely partial identification and sensitivity analysis, will be presented in Subsection 2.2 and 2.3.

After having described the general concept of "coarsened at random", I will explain some already developed further extensions from this basic concept.

2.1.5. Further extensions of the concept of "coarsened at random"

Heitjan [1994] has developed an interesting extension to the already described basic concept of "coarsened at random" by introducing a new random variable as well as the concept of "coarsened completely at random" (CCAR).

As the degree of coarseness, namely G, cannot always be observable in a precise way, he establishes an extra random variable H which is able to give some information about G. In Subsection 2.1.3 density $h(g|y, \gamma)$ has been used to model the coarsening in case of value g being unknown, such that denoting this random variable as H is appropriate. In this way, Heitjan [1994] faces a situation in which the true, but potentially unobserved, variables of interest are Y and G, while variables $\mathcal{Y}$ and H are observed.

Instead of giving some formal definitions, I decided to illustrate the concept of "coarsening completely at random" and the application of this new random variable H by an example such that these new definitions become comprehensable. This example is similar to the running example of Heitjan and Basu [1996]. To understand the difference between situations under CAR and CCAR and needing and not needing a random variable H respectively, I will distinguish four cases within this example. If one is interested in the formal background, it might be helpful to read through Heitjan [1994], where the property of "coarsened completely at random" is shown in a frequentist framework and the one of "coarsened at random" in a Bayes framework. Thus the difference between CAR and CCAR can be motivated by the distinction between these two modes (Heitjan and Basu 1996, p. 208).

> **Example 1** I will focus on the question "How many bottles of a special beverage did you consume during the last month?". Then data could be heaped (see Chapter 1) and there might be respondents who will answer in bottles ($G = 0$), while others report their answer in terms of beverage crates ($G = 1$), probably those who drink this beverage very often. Assuming that there are eight bottles in one of those crates, coarsening function
>
> $$\mathcal{Y} = \begin{cases} \{Y_i\}, & \text{if } G_i = 0 \\ \{8\lfloor \frac{Y_i}{8} \rfloor, ..., 8\lfloor \frac{Y_i}{8} \rfloor + 7\}, & \text{if } G_i = 1 \end{cases}$$
>
> could result, where $\lfloor \ \rfloor$ represents the floor function.

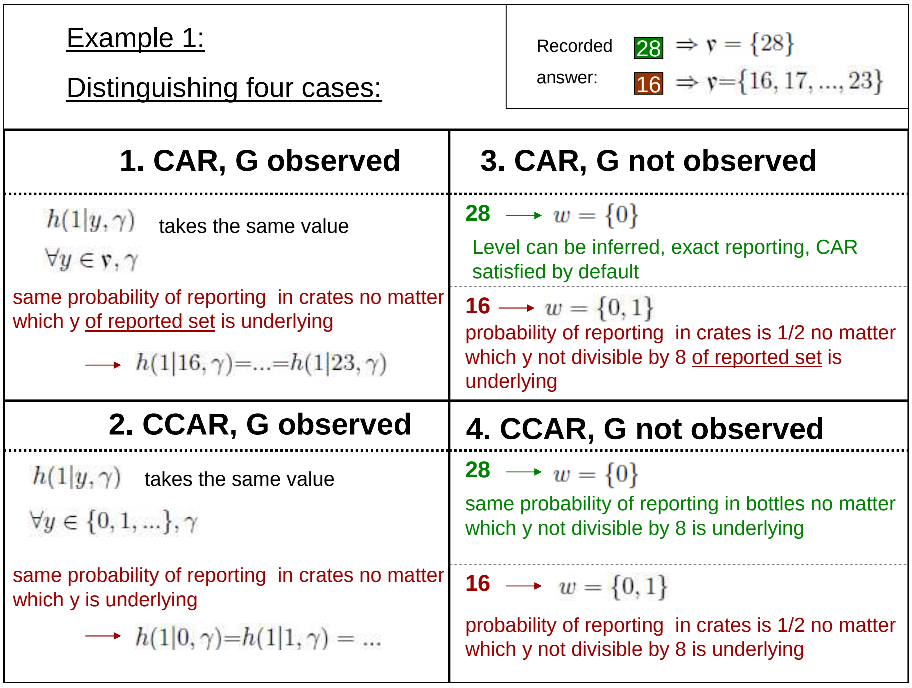
1.) First of all, I regard the case of *CAR under the assumption that G can be fully observed* for every respondent, this means that $H = \{G\}$. This assumption accounts for further simplification of the CAR property in the sense that instead of $q(\mathfrak{y}|y, \gamma)$ from equation (2.5) only h has to be the same for a given observed value of $\mathcal{Y}$ no matter which true value of Y that corresponds with the observed data (i.e. $y \in \mathfrak{y}$) is underlying. That statement equals Corrollary 1 of Heitjan and Rubin [1991, p. 2249]. In the following I try to explain this finding rather in a contentual than a formal way.
The reason for this simplification consists of the fact that under the assumption that the value of G is known and fixed, the task of r is no longer to determine the right degree of coarsening, but only to conduct the grouping. Being no longer a function of g, one can exclude r from the integral in equation (2.5) by

$$q(\mathfrak{y}|y, \gamma) = \int_\Gamma r(\mathfrak{y}|y, g)\ h(g|y, \gamma)dg = r(\mathfrak{y}|y) \int_\Gamma h(g|y, \gamma)dg$$

and therefore only h has to be included into the definition of CAR.
Illustrated by means of this example the property of CAR can be expressed in the following way: If every respondent reports his answer in terms of bottles

($g = 0$), CAR is satisfied by default, being only one element in the corresponding reported set (see coarsening function). Otherwise, namely in the case that there are some respondents who answer in terms of crates ($g = 1$), one has to check if $h(1|y, \gamma)$, i.e. the conditional probability of reporting crates, is the same no matter which true value y that is consistent with the observed data is underlying. For instance, if a crate reporter (known $g = 1$) consumed 16 bottles, a reported set of $\mathfrak{y}$=$\{16, 17, ..., 23\}$ would result (see coarsening function) and therefore it would have to be satisfied that $h(1|16, \gamma)$=$h(1|17, \gamma)$=...=$h(1|23, \gamma)$ for all γ.
2.) Secondly, I focus on a situation of *CAR* again, but now the value of G *is unknown,* i.e. at a first glance one does not have any evidence whether the respondents originally reported their answer in bottles or in crates. But if an answer of a respondent was $\mathfrak{y} = 28$, it would be obvious that this answer is in terms of bottles (because 28 is not divisible by eight), while an answer of $\mathfrak{y} = 16$ (divisible by eight) wouldn't give some information on the form of reporting (bottles or crates) and therefore $w = \{0, 1\}$.
CAR is valid if for every respondent who belongs to the latter case ($w = \{0, 1\}$), for all γ, probability $h(1|y, \gamma)$ is equal to $\frac{1}{2}$ for all true values y that are not divisible by eight and are consistent with the observed data (Heitjan 1994, p. 705), e.g. $h(1|17, \gamma) = ... = h(1|23, \gamma) = \frac{1}{2}$ if one faces the reported set from above. The additional condition in the sense that the value of these conditional densities not only have to be equal, but also have to be $\frac{1}{2}$ results from the fact that compared to case one the value of g is unknown and therefore one has to decide which degree of coarsening is underlying. As in this example only two levels of coarsenings are possible, namely reporting the number in terms of bottels ($w = 0$) or in terms of crates ($w = 1$), assigning both levels the same probability given a special possible true value (e.g. $h(1|17) = h(0|17)$), leads to probability $\frac{1}{2}$. This additional condition forms the difference between G being fully observed (case 1) and G being unobserved (case 2).
It could be a little bit surprising that even if G is not directly observed a condition on h, and not on $q(\mathfrak{y}|y, \gamma)$ like described in the previous subsection, is imposed. The reason for this consists of the fact that r, which usually governs the choice of the underlying degree of coarseness, is not needed here. This can be said, because in this situation either the underlying degree of coarseness is

known (for values not divisible by eight) or one is concerned with this additional assumption (for values not divisible by eight) which forms a fixed rule for the determination of the degree of coarseness.
3.) Thirdly, the case of *CCAR with G is fully observed* will be addressed. For showing the assumption of “coarsend completely at random” the conditional density $h(1|y, \gamma)$ not only has to be the same for all possible true values that correspond with the reported set (see case 1), but all possible true values $y \in \{0, 1, 2...\}$ (Heitjan 1994, p. 703). In respect of this example “coarsened completely at random” means that the conditional density for respondents who report their answer in terms of crates is the same no matter which true value is underlying, i.e. $h(1|y, \gamma)$ has to take the same value for each γ and for all $y \in \{0, 1, 2...\}$. As in case 1 assumptions do not have to be imposed on probability $h(0|y, \gamma)$, because CCAR is satisfied by default if respondents answer in bottles. The assumption of CCAR is stronger compared to CAR, because this assumption is not only based on the true values of the reported set, but all possible true values. At this point one can note a parallel to the “missing data problem”, because the assumption of “missing completely at random” imposes conditions on both the observed and unobserved values as well. More considerations concerning the relation to the missing data problem can be found in Subsection 2.2.6, 2.3.4 and 2.1.6.
4.) Fourthly, the case of *CCAR with unknown value of G* is left. Heitjan [1994, p. 705] realized that two conditions have to be fulfilled in this case. Illustrated by the example CCAR is valid if: a) For all respondents that report a number which is not divisible by eight and thus $w = \{0\}$, the conditional density of reporting in terms of bottles, namely $h(0|y, \gamma)$, has to be the same for all γ and all true values y not divisible by eight, b) For all all respondents that report a number which is divisible by eight and thus $w = \{0,\ 1\}$, the conditional density of reporting in terms of bottles has to be one half, i.e. $h(0|y, \gamma)=\frac{1}{2}$ for all γ and all y not divisible by eight (Heitjan 1994, p. 705). In case b) a stronger condition results because one does not know which degree of coarsening is underlying (see case 3).
To give a summary of the regarded cases, Figure 2.1 can be helpful. By comparing first and second row of this Figure, one can notice that the basic difference between CAR and CCAR consists of the fact that within the CAR assump-



Example 1:

Distinguishing four cases:

Recorded answer: 28 $\Rightarrow \mathfrak{y} = \{28\}$

16 $\Rightarrow \mathfrak{y}=\{16, 17, ..., 23\}$

1. CAR, G observed	3. CAR, G not observed
$h(1\|y,\gamma)$ takes the same value $\forall y \in \mathfrak{y}, \gamma$ same probability of reporting in crates no matter which y of reported set is underlying $\longrightarrow h(1\|16,\gamma)=...=h(1\|23,\gamma)$	28 $\longrightarrow w = \{0\}$ Level can be inferred, exact reporting, CAR satisfied by default 16 $\longrightarrow w = \{0, 1\}$ probability of reporting in crates is 1/2 no matter which y not divisible by 8 of reported set is underlying
2. CCAR, G observed	**4. CCAR, G not observed**
$h(1\|y,\gamma)$ takes the same value $\forall y \in \{0, 1, ...\}, \gamma$ same probability of reporting in crates no matter which y is underlying $\longrightarrow h(1\|0,\gamma)=h(1\|1,\gamma) = ...$	28 $\longrightarrow w = \{0\}$ same probability of reporting in bottles no matter which y not divisible by 8 is underlying 16 $\longrightarrow w = \{0, 1\}$ probability of reporting in crates is 1/2 no matter which y not divisible by 8 is underlying

Figure 2.1.: Illustrating the four cases that have been distinguished in Example 1: 1. CAR + G observed, 2. CAR + G unobserved, 3. CCAR + G observed, 4. CCAR + G unobserved.

tion it is required that the underlying true values y of $h(1|y, \gamma)$ come from the reported set, where there is no such postulation in the context of the CCAR assumption. Moreover, by contrasting left and right column one can recognise that if G cannot be observed, it has to be distinguished between two cases, namely whether the reported answer is divisible by eight or not. While in the framework of CAR the level of coarseness could be inferred for answers that are not divisible by eight, simplifications of that kind were not possible in the case of CCAR. Moreover, if G cannot be observed an additional assumption has to be imposed in the sense that the corresponding probabilities h not only have to be equal, but also have to be one half.

The concept of "coarsened completely at random" will be addressed in Subsection 2.1.6 again, in order to give some indication of the relation to the missing data problem.

The extension by Heitjan [1994] is not the only one that has been made. For instance, Jaeger [2005] distinguished two forms of CAR in the categorical case, namely weakly coarsened at random (w-car) and strongly coarsened at random (s-car). During s-car definition focuses on the conditional distribution of the coarse data, w-car is based on the joint distribution of complete and coarsened data. By differentiating between these two forms, he can partly make propositions concerning ignorability without the assumption of distinct parameters. He worked out that s-car can serve as the only assumption for obtaining ignorability, whereas in case of w-car in general further assumptions, like an underlying saturated model, are necessary.

After having a clearer understanding about the concept of "coarsened at random" and some possible extensions, it could be interesting to analyse their difference as well as their similarity to the missing data problem.

2.1.6. Relation to the missing data problem

For working out the relation to the missing data problem, I want to use the following example.

> **Example 2** Imagine that the members of a workshop are expected to evaluate this workshop by a scale from 1 to 5, with one being the worst rating and five being the best one.
> *Case 1:* Only a few of those ratings can be observed exactly. Otherwise it is only observed whether positive feedback (e.g. 4 or 5) has been given, or not (e.g. 1, 2 or 3 respectively).
> *Case 2:* Some respondents didn't give an answer to this question at all and therefore some observations are missing.

First, I want to demonstrate the differences and the similarity of the nature of a coarsening and a missing mechanism by means of Example 2.

An obvious difference between case one and case two and therefore between the coarsened data and the missing data problem consists of the fact that in case one more information is available. In the situation of coarsened data the observer knows for some cases only that the true value lies in a special set (e.g. "4 or 5" in case of a positive statement), which is an element of the power set. By contrast in the situation of missing data one has even less information, because the observer does not know anything about the missing observations and every element of the sample space Ω could be possible (e.g. “1,2,3,4 or 5”). Because Ω (example: set “1,2,3,4,5”) is also an element of the power set $\mathcal{P}(\Omega)$, the missing data problem can be regarded as a special case of the coarsened data problem with G_i=1 if there is given maximal precision of coarsening and thus data are fully observed and G_i=0 if there is given minimal precision of coarsening and thus data are missing. Thus in the special case of missing data the coarsening function can be described as

$$\mathcal{Y} = \begin{cases} \{\omega\}, & \text{if } G = 1 \\ \Omega, & \text{if } G = 0 \end{cases}, \tag{2.6}$$

where ω is a single element of Ω.

Furthermore in Subsection 2.1.5 an example has shown that the degree of coarseness G sometimes cannot be observed in the framework of coarsened data and therefore a random variable H has been introduced. As in the context of missing data it is clearly evident whether data are missing ($G = 0$) or precisely observed ($G = 1$), an introduction of random variable H is not necessary, because G is always observed.

Moreover I want to illustrate the equivalence of the problem specific underlying properties. Therefore, it might be useful to recall different types of missingness first and then refer those definitions to the properties in the context of coarsened data in a second step. Little and Rubin [2002] distinguish between three types of missingness, nameley “missing completely at random” (MCAR), “missing at random” (MAR) and “not missing at random” (NMAR).

The missing data mechanism is called MCAR if the missingness neither depends on the missing nor on the observed data. So data are not MCAR in the described example either if the missingness of a feedback is influenced by the rating itself (i.e. in the sense that respondents who didn’t like the workshop, do not give a feedback at all) or if the missingness of a feedback is affected by other answers (like age) which have been reported (i.e. in the sense that younger respondents rather do not give a feedback). In a second step one can try to apply this definition in the context of coarsened data by remembering that missing data can be viewed as coarsened data with $G_i = 1$ if the data are observed and $G_i = 0$ if the data are unobserved (and therefore G equals the missing data indicator of Little and Rubin [2002] (but expressed in the opposite way here)) and G is fully observed. Under CCAR, $h(g|y, \gamma)$ has to take the same value for all $y \in \Omega$, i.e. that the probability that a special coarsening is underlying (and refered to this case: probability of missingness) must be equal no matter which values within the sample space are underlying. This means that under CCAR missingness (e.g. of a feedback) must not depend on the observed data nor on the missing data (both because y is an element of the full sample space). Therefore, one can conclude that the MCAR can be viewed as a special case of CCAR.

Data are called MAR, if the missingness depends on the observed data only. For instance, under MAR the probability that a feedback of a particular respondent is missing must not depend on the feedback itself, but may depend on another reported variable like age in the sense that younger respondents rather do not give a feedback. When younger people may give in general worse feedback, because the target audience of this workshop was planned to be older, this could be problematic. Again it is possible to apply this situation in the framework of coarsened data. Under CAR, $h(g|y, \gamma)$ has to be the same no matter which $y \in \mathfrak{y}$ is underlying, where $\mathfrak{y}$ is the observed data. Therefore, the coarsening is only allowed to be dependent on the observed data, where this means in the example that under CAR, the missing depends on the observed values only. Thus, MAR represents a special case of CAR.

Until now, only two definitions have been recalled in the context of coarsened data, namely CCAR and CAR, but in the missing data context NMAR forms a third one. Under NMAR, the missingness depends on the missing as well as the observed values. In practice this can lead to big problems, because if one imagines that, for instance, worse feedbacks are missing more probable, a serious bias can result. Taking NMAR as a starting point, I use explicitly the term “not coarsening at random” (NCAR) for the case that assumptions CCAR and CAR are not satisfied, but the coarsening process is known and can be modelled. This kind of NCAR-definition would be valid, if for instance

h, that models the coarsening, is described like in equation (2.10) of Heitjan and Rubin [1991, p. 2247] (adjusted notation)

$$h(g, y, \alpha) = \prod_{i=1}^{n} \{\Phi[\alpha_1 - \alpha_2 y]\}^{1-g_i} \{1 - \Phi[\alpha_1 - \alpha_2 y]\}^{g_i},$$

where α denotes an additional parameter and Φ is the standard normal integral. Thus the level of coarsening is dependent on the true value itself (as long as $y \neq 0$). Illustrated by the example this could mean that respondents who are not satisfied by the workshop rather gave their feedback in a coarsened form like "1 or 2" or "bad feedback", while the others rather reported their feedback exactly.
So even if assumptions like CAR or CCAR are not valid, it could be possible that the coarsening is known and can be modelled. Even if most methods in the framework of censored data are based on censoring that is uninformative (Lagakos 1979, p. 152), there exist a few approaches that are able to adjust for dependent censoring. Dependent censoring is present if the reasons of censoring affect the resulting survival time. For istance, one could be concerned with dependent censoring if reasons that are connected with the therapy are responsible for the removement of some patients from the study or if the problem of competing risk is present, i.e. if failure times are recorded that are not induced by the reason of interest (e.g. death of a patient because of other reasons). A substantial bias can result if methods that deal with uninformative censoring are used in these kinds of situations. This emphasizes the necessity of methods that are able to deal with informative censoring. For instance, it is achievable to deal with dependent right-censored data in the presence of many covariables by condensing the information that can be revealed by those covariables and using two models, namely one for the lifetime given covariates and one for the censoring time given covariates. Further information concerning the estimation of the marginal survival function of the failure time of interest and some resulting robust properties can be found in Zeng [2004]. In case of competing risk one could apply the approach proposed by Moeschberger and David [1971] instead, who extended the independent censoring model

$$U = min(T, Y) \text{ and } d = 1 \text{ if } T \leq Y \text{ or } d = 0 \text{ if } T > Y,$$

with T being the true survival time and d an indicator whether actual survival time has been observed ($d = 1$) or a censored one ($d = 0$), by focusing on the bivariate distribution for (T,Y) and regarding independence of T and Y as a special case. He showed how maximum likelihood estimators for the parameters of the joint distribution can be derived that are able to adjust for general censoring (Lagakos 1979).

All in all the concept of missing can be seen as a special case of coarsening. This I have illustrated by investigating the general nature as well as different types of mechanisms in the context of coarsened and missing data, which is depicted in Figure 2.2. The aspect of the general nature is showed in the box below, where coarsening can be regarded as a mapping from the sample space into the power set while missing represents a mapping from the sample space either into a singelton (namely $\{\omega\}$, where ω is an element of Ω) in case of complete observation or into the full power set and thus the sample space in case of missing observations. The green ellipse shows that for each type of missing, an underlying general form of coarsening exists, for instance MCAR is a special type of CCAR.

Because of the fact that the missing data problem has been investigated quite well, it could be reasonable to refer some of those methods which are used in the area of missing data to the problem of coarsened data. More considerations concerning this can be found in Subsection 4.6

To sum up this section, I can say that ignorability, which is satisfied if CCAR or both CAR and distinct parameters are applicable, is a quite helpful assumption, because under this property a simplified likelihood results. Nevertheless, one could be concerned with a situation in which CCAR and CAR are not valid, but the underlying coarsening mechanism could be known such that one is able to model it. This has been indicated in the context of dependent censoring. Thus by additional assumptions (CAR or CCAR) or modelling the coarsening mechanism, a solution to the initial problem described in the beginning of this chapter (see equation (2.1)) can be concluded. But there are a lot of situations in which the coarsening is unknown and wrongly imposing assumptions as CCAR or CAR as well as characterizing the coarsening process

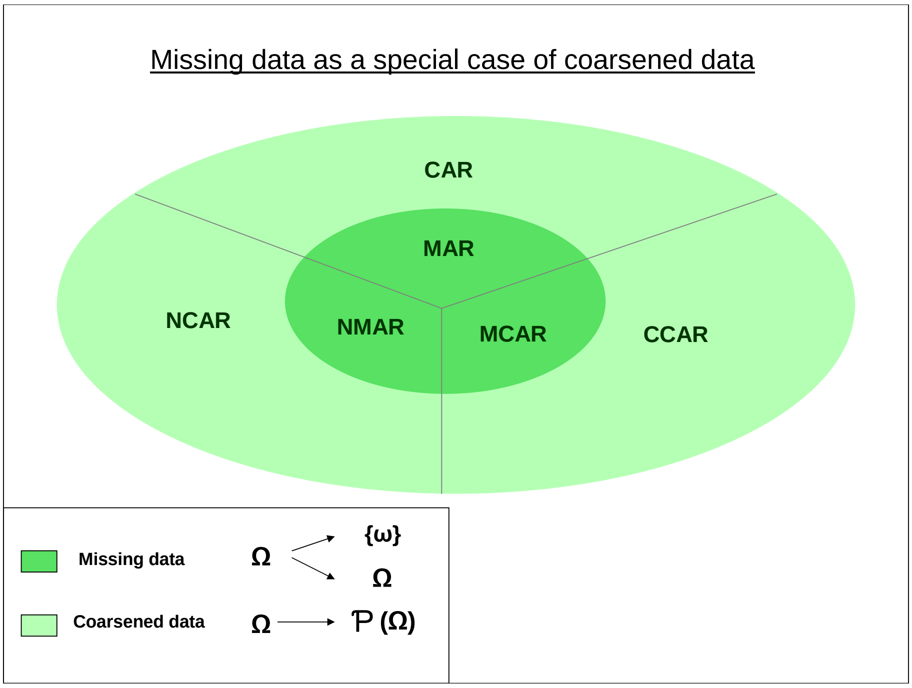


Figure 2.2.: Illustrating missing data as a special case of coarsened data.

in a wrong way can lead to substantial error. Therefore, one has to take care if other methods have to be applied instead which do not rely on these strong assumptions.

With this in view partial identification seems to be a quite preferable approach that starts without any assumption and continues by increasing those gradually. Hence, I want to focus on this approach next by discussing the importance of assumptions within the frame of statistical analysis first.

2.2. Partial identification

"The credibility of inference decreases with the strength of the assumptions maintained"

– Charles Manski (2003) –

This statement is called "law of decreasing credibility" and expresses a frequently ignored problem that in my opinion is worth to assign high relevance. In order to understand why in practice it is rarely taken note of this therein addressed issue, it could be reasonable to recall the concept of identifiability first as defined for example in Casella and Berger [1990, p. 511]:

Definition 1. *A parameter θ for a family of distributions $\{f(x|\theta) : \theta \in \Theta\}$ is identifiable if distinct values of θ correspond to distinct probability distribution functions (pdfs) or probability mass functions (pmfs). That is, if $\theta \neq \theta'$, then $f(x|\theta)$ is not the same function of x as $f(x|\theta')$.*

This means that statistical models are identified whether different values of parameters θ produce different pdfs or pmfs of the observed variables. Otherwise one is concerned with difficulties in doing inferences. One can distinguish between a "just identified" model and two different types of non-identifiable models, namely "overidentified" and "underidentified" models. A model is said to be overidentified if the number of known parameters (n_k) exceeds the number of parameters that have to be estimated (n_e). Thus, these kinds of models exhibit a positive degree of freedom ($n_k - n_e > 0$) and are characterized by systems of equations that can not be solved in an analytic way (Fahrmeir et al. 1996, p. 742). For a better understanding of the mentioned problem underidentified models are more important, which show a negative degree of freedom as the number of known parameters is smaller than the number of parameters that have to be estimated. In this case one usually restricts the model for example by setting single parameters constant or equal. To give an example one could recall the basic equation (2.1) of the beginning of this chapter, where the CAR assumption (see 2.1) represents an approach that tries to deal with the underlying non-identifiability by setting parameters equal, e.g. setting $q(\mathcal{Y} = (A\ XOR\ B)|Y = A) = q(\mathcal{Y} = (A\ XOR\ B)|Y = B)$ as under CAR $q(\mathfrak{y}|y, \gamma)$ is the same no matter which true value that corresponds to the observed data is underlying. Consequently, it is actually common to make several assumptions in order to yield a model that is "just identified" and to be able to report results of statistical analysis in a precise way. Thus, in many cases, despite the dilemma expressed in the "law of decreasing credibility", models with strong underlying assumptions which point identify parameters are chosen.

At the first glance, this procedure seems to be reasonable, because researchers, who interpret statistician's analysis, often ask for unique results in order to be able to make clear conclusions and possibly adapt arrangements. As long as assumptions are plausible and do not yield misleading answers, I think it is useful to involve them. Especially if these assumptions lead to point identified results this is quite preferable. Unfortunately, frequently assumptions are used whose contentual justification is not and often even can not be checked and instead of a correct analysis being the objective, point identified, clear results are required (Koopmans and Reiersol 1950, p. 169). But what benefit does a model entail which yields precise results that easily can be interpreted, but which actually provides wrong conclusions?

In my opinion partial identification represents a suggestive answer to this trade-off between requesting a precise point identified parameter on the one hand and using a model that includes justified assumptions only and hence accounts for uncertainty on the other hand. Therefore, I want to recall the approach of partial identification in this section.

Because of the fact that one can distinguish between different kinds of uncertainty, I want to introduce the terminology which will be used here for these different types first in order to avoid confusion later on. After having addressed the basic idea of partial identification and some fields of application, I will summarize some formal findings of Manski [2003], who investigated the topic of partial identification in the context of missing data. Subsequently, I want to apply those developments in the framework of coarsened data and recall two possible points of view analysis could be based on.

2.2.1. Distinguishing between two kinds of uncertainty

Because of the general disagreement in respect of the notation of different kinds of uncertainty, I explicitly want to establish the notation that will be used here.

In the framework of a statistical analysis it is common to account for the kind of uncertainty that can be attributed to finite sampling. As infinite sampling cannot be realized in practice, quantities cannot be treated as certain. Confidence intervals represent generally used instruments that are able to account for uncertainty due to finite sampling by showing a region of values that cover the true parameter with given probability $1 - \alpha$, where α is the significance level. While Tamer [2010] calls this kind of uncertainty "statistical uncertainty", Vansteelandt et al. [2006] uses the designation "sampling imprecision". In order to avoid confusion, I will use a notation that completely differs from the existing ones and hence in this thesis this kind of uncertainty that is an implication of finite sampling will be called *first kind of uncertainty*. Even if the first kind of uncertainty is usually involved into statistical analyses, it is frequently ignored in the case in which the *second kind of uncertainty* is present. The in this thesis so-called second kind of uncertainty describes uncertainty that is induced by the lack of information, for instance in the presence of missing or coarse data. In literature the second type of uncertainty is sometimes called "ignorance" (Vansteelandt et al. 2006, Molenberghs et al. 1999).

As this thesis is concerned with coarsened data, it is mainly concentrated on how to deal with the second type of uncertainty. Nevertheless, one should not forget - like it is often done - to incorportate the first type of uncertainty. Therefore, in the context of partial identification the inclusion of both kinds of uncertainty will be addressed on page 38, where on page 51 it will be explained how the first kind of uncertainty can be integrated additionaly in the framework of sensitivity analysis. Even if these approaches show how the first kind of uncertainty can be incorporated as well, in this chapter estimation will not be covered and the notation will be in terms of probabilities and not in terms of frequencies, that will be used in Chapter 4 for the first time.

2.2.2. The main idea of partial identification

As explained in the context of identifiability in the beginning of this chapter, researchers who are concerned with non-identifiable models often impose assumptions in order to obtain point identified parameters. A basic difference of partial identification to this common procedure consists of the conception that identification no longer has to be regarded as a binary event by either a parameter being identified or non-identified (Manski 2003). Instead it is admitted to impose some justified assumptions that do not have to induce point identified parameters, but at least identify the parameter of interest in parts (i.e. these assumptions “partial identify” the parameter of interest) compared to the set of parameters that seemed to be possible in the beginning of the analysis. Thus, estimators for parameters of interest which follow from partial identification can either be without any information, partial identified or point identified depending on the underlying plausible assumptions being available in this situation.

The main idea of partial identification consists of the point of view that even if only a few plausible assumptions are made and consequently some second kind of uncertainty remains, results can contain valuable information. Therefore, the basic procedure of partial identification first uses the empirical evidence only by using information implied by the data without involving additional assumptions. In a second step the researcher faces possible assumptions and includes those which seem to be justified and there exists a common consesus about their validity (Tamer 2010, p. 169). Hence, in the frame of partial identification mostly nonparametric models with minimal assumptions and partial identified parameters result.

2.2.3. Fields of application

In order to gain an inside of the preference of partial identification for dealing with the second kind of uncertainty, it could be interesting to present some areas in which partial identification could be imaginable or already carries weight. Afterwards I want to show an illustration of the two step procedure of partial identification.

Partial identification has mostly been neglected in econometrics as well as statistics before the 1990s. Reasons for the rare application of this method could be the difficulty to evaluate the plausibility of serveral assumptions as well as the fact that by means of various solutions it is essential harder to improve policies which is an usual task within the field of econometrics (Tamer 2010, p. 174–175). Nevertheless partial identification has started to be more popular in econometrics. Some examples like the approach of Marschak & Andrews, who deduced bounds for the parameters of a production function by partial identification, as well as the development of Fréchet bounds, that restrict the resulting possible joint distributions of random variables X and Y by an upper and a lower limit for the case that the corresponding marginal distributions of X and Y are given, can be found in Tamer [2010].

Another field of application of partial identification consists of the analysis of missing and misclassified data. In order to illustrate partial identification in those situations of data, I want to present two examples now. While the first example shows partial identification as a useful tool for dealing with missing data, the second illustration offers an partial identification based approach for the situation of missclassified data. Both examples demonstrate the two step procedure which first uses the empirical evidence only and then involves further assumptions.

As a first example I regard the situation of Stoye [2009*b*]. Here the problem is faced whether offenders should be assigned to residential or nonresidential treatment with regard to preventing recidivism. In this situation the probability that offenders who are assigned to a special treatment (residential (r), nonresidential (n)) will again turn to crime (c) ($P(Y_t = c)$, t=r,n) are the probabilities of interest, which are unknown. Moreover counterfactual probabilities, like $P(Y_n = c|T = r)$ and $P(Y_r = c|T = n)$ are uninformative, because it is unknown whether offenders would have been criminal if they had been assigned to the other treatment. Partial identification is used now by first using some information that can be revealed from the data generating process and thus an identification region for the probability of interest yields. Afterwards some contentual considerations are included like the assumption of “Outcome Optimization”, which sticks to the idea that judges always assign the right punishment in the sense that the residential (nonresidential) treatment is only chosen if the probability of recidivism is smaller for the underlying offender.

Two more assumptions have been imposed and one yields a shrunk identification region compared to the one formed by the empirical evidence only. In this example partial identification seems to be preferable to an approach simply assuming that the missing data are ignorable, because the treatment is not generated by a randomized experiment, but determined by judges. Judges do not base their decision on random, because they probably assign the residential treatment to case-hardened criminals, because they want to protect society from those and to avoid future serious crimes.

Moreover I want to illustrate the the two step procedure of partial identification in the context of misclassification. Molinari [2008] has developed an direct missclassification approach which is based on partial identification and that I want to summarize now. By means of direct misclassification approach a relation between the distribution of the true and the misclassified variable is expressd by a linear system of simultaneous equations, in which the coefficients are described by the matrix of misclassification probabilities that converts the true variable into the observation which is potentially misclassificated. This underlying equation is similar to the basic equation (2.1) and the relation between these equations is discussed more detailed in 2.2.4. Partial identification of the coefficients is started by using the empirical evidence only and thus identification region $H^P[\Pi^*]$ is focused first, which denotes the set of Π's which fulfill the typical probabilistic constraints, like $\sum_{i=1}^{J} \pi_{ij} = 1$ and $\pi_{ij} \geq 0$. After that some assumptions on the missclassfication pattern are imposed; for example it is assumed that the probability of correct reporting is constant. As the set of matrices which fulfill the requirements that are derived from validation studies or theories from social sciences is denoted by $H^E[\Pi^*]$, in the mentioned example one would be concerned with $H^E[\Pi^*] = \Pi : \pi_{jj} = \pi$. Combining those further assumptions with the initial identification region formed by the empirical evidence, one yields $H[\Pi^*]$=$H^P[\Pi^*] \cap H^E[\Pi^*]$. More details concerning partial identification in the context of misclassification and the relation of misclassified and coarsened data can be found in Subsection 2.2.4. An additional insight into the topic of partial identification for misslcassified data can be gained by Küchenhoff et al. [2012].

Regarding these two examples, one can notice that missing (see Manski [2005]) and misclassificated (see Molinari [2008]) data represent typical sources for identification problems. Because coarse data are a generealisation of missing data, partial identification could be possible for this case as well. Therefore, I could imagine that apart from econometrics partial identification could be quite useful in several other fields of application, because as long as researcher base their findings on questionnaires, missing, missclassified and coarse data are quite common and partial identification could represent an appropriate answer to these problems.

In the beginning of this chapter it has been announced that it could be interesting to establish a connection from coarsened data to other areas, like misclassified data. This will be done by an excursus now.

2.2.4. Excursus: Relation of coarsened and misclassified data in the case of categorical data

In the previous subsection an example has shown that partial identification represents an useful approach not only for dealing with coarsened data, but also with misclassified data. Thus, it could be imaginable that there is a relation between miclassified and coarsened data. As this relation is especially direct in the case that is focused in this thesis, namely the categorical data case, it could be interesting to analyse the connection between those problems here. The similarity between the coarsened and the misclassified data problem I want to illustrate by transferring basic equation (2.1) developed in the beginning of this chapter into the context of misclassified data.

In the presence of misclassification one is concerned with a similar problem like in the case of coarsened data, namely that the variable of interest cannot be observed like requested. But instead of coarsening being the problem, here one sometimes observes wrong categories. In order to deal with this problem Molinari [2008, p. 82] introduced the direct misclassification approach that can be expressed by (see equation (1.1) of Molinari [2008, p. 82], but adjusted notation)

$$\begin{pmatrix} P(W=1) \\ \vdots \\ P(W=J) \end{pmatrix} = \begin{pmatrix} P(W=1|Y=1) & \cdots & P(W=1|Y=J) \\ \vdots & & \vdots \\ P(W=J|Y=1) & \cdots & P(W=J|Y=J) \end{pmatrix} \begin{pmatrix} P(Y=1) \\ \vdots \\ P(Y=J) \end{pmatrix}, \tag{2.7}$$

where Y again denotes the variable of interest, which cannot always be observed in a correct way. Additionaly variable W, that labels the observed variable, is introduced, where "$W = i$" means that the observed realization is i and "$Y = j$" means that the true realization of interest is j. Therefore, in case of missclassification i is unequal to j and the corresponding probability $P(W = i|Y = j)$ is greater than zero. In order to show the similarity to basic equation (2.1) of the beginning of this chapter more obviously, one can conclude from equation (2.7) the calculation simply for probability $P(W = j)$ obtaining

$$\begin{aligned} P(W = j) &= P(W = j|Y = 1) + P(W = j|Y = 2) + \quad ... + P(W = j|Y = j) + \\ &\quad ... + P(W = j|Y = J) \\ &= \sum_{i=1}^{J} P(W = j|Y = i)P(Y = i) \qquad . \end{aligned} \tag{2.8}$$

Thus, comparing basic equation (2.1) for coarsened data and equation (2.8) for misclassified data, one can note that the only difference consists of the fact that instead of having observed variable $\mathcal{Y}$, that could be coarsened, one observes variable W that could be misclassified.
This explaines the similarity between these two problems of data and why in both cases similar methods, like partial identification (see Subsection 2.2.3), can be applied.

After possible fields of application and their relation has been addressed, it is necessary to recall some formal background of partial identification. Manski has mainly shown how partial identification is able to handle missing data without making strong and potentially inadmissible assumptions like "missing at random" (MAR) and developed a promising notation for partial identification. Moreover, Manski and Tamer [2002] considered regression models in the presence of interval data and developed several assumptions under which in the framework of partial identification simple nonparametric bounds can be found.
But beside missing or interval data, it could be further imaginable that coarse data are present in a questionnaire, for example because respondents did not

want to report their answer in a precise way to preserve their anonymity. This case is not investigated quite well in the context of partial identification, but I want to refer some results of Manski [2003] for the mising data case to the coarsened data problem. But first it could be reasonable to summarize Manski [2003]'s findings concerning partial identification in the missing data context.

2.2.5. Formal background of partial identification in the missing data context

In the following I want to recall some basics concerning the formal background of partial identification. Because these fundamental notations have been developed in relation with the missing data problem, I want to address this case first. Below I will refer to the developments of Manski [2003] and Manski [2005]. Please note that it mainly will be addressed how to deal with the second kind of uncertainty and only on page 38 it will be explained how the first kind of uncertainty can be included as well.

Like already mentioned the starting point of the approach of partial identification consists of the empirical evidence only in the absence of any untestable assumptions. Being $P(Y = y)$ the parameter of interest and g the value which indicates if data are observed or missing (with $g = 1$ data being observed and $g = 0$ data being missing). Then by the Law of Total Probability one can follow the so called *identification region* (Manski 2003, p. 6 with adapted notation),

$$H[P(Y = y)] \equiv [P(Y = y|g = 1) \cdot P(g = 1) + \gamma P(g = 0), \gamma \in \Gamma_Y]. \tag{2.9}$$

All components apart from γ can be estimated by using the empirical distributions $P_N(Y = y|g = 1)$, $P_N(Y = y)$ and $P_N(g = 0)$. As no information about y being available for the missing data, γ=$P(Y = y|g = 0)$ is uninformative and can attain values within the space of probability measures Γ_Y. Manski [2003] refers this general definition of identification regions to some special population parameters of interest, for example means of functions of y and parameters that respect stochastic dominance as quantiles.

In a second step one can include plausible assumptions concerning the distribution of interest, like $H_0[P(Y = y)] \subset \Gamma_Y$, such that the identification of equation (2.9) can be shrunk to (see Manski [2003, p. 4])

$$H_1[P(Y = y)] \equiv H_0[P(Y = y)] \cap H[P(Y = y)]. \tag{2.10}$$

Manski [2003] proposes some imaginable assumptions on the distribution of interest that are able to shrink the initial identification region received by the emprical evidence only. Here a selection of those will be explained, denoted by $A1$ and $A2$. One quite simple postulate is the one that there is no difference between missing and observed data by assuming (see Manski [2003, p. 26])

$$A1: \; P(Y = y) = P(Y = y|g = 0) = P(Y = y|g = 1). \tag{2.11}$$

This assumption leads to point identification of $P(Y = y)$, because $P(Y = y|g = 1)$ can be estimated by the sampling process. Nevertheless this assumption is quite strong, because it implies that responders and nonresponders do not differ which might be wrong in several cases. Therefore, Manski [2003] decided to use instrumental variables with values v_j that are observed for every respondent j and constitute a support for identifying the distribution of interest $P(Y = y)$. For instance, he suggests the following assumption (see Manski [2003, p. 27])

$$A2: \; P(Y = y|V = v, g = 0) = P(Y = y|V = v, g = 1). \tag{2.12}$$

So this assumption is slightly weaker than $A1$, because only similar respondents and nonrespondents, namely those who exhibit the same value v, are assumed to follow the same distribution and thus this can rather be justified. For instance, if there are some Bavarian ($V =$“Bavarian”) item-nonresponder concerning the characteristic “preferred party” ($Y = y$), for example one could be interested in probability $P(Y =$ “CDU”$|V =$ “Bavarian”$, g = 0)$. According to assumption $A2$ one would decide that this probability equals the probability of electing “CDU” given the answer of Bavarian item-responder ($P(Y =$ “CDU”$|V =$ “Bavarian”$, g = 1)$. This might be more reasonable than simply using $P(Y =$ “CDU”$|g = 1)$, namely the probability of electing “CDU” given the answer of all respondents (e.g. German respondents) as it would be

assumed using assumption $A1$. In the same way covariables can reveal information in other cases as well.
Even if $P(Y = y|g = 0)$ is a priori unknown, by involving $A2$ one is able to point identify $P(Y = y)$ by (equation (2.2) of Manski [2003, p. 27])

$$P(Y = y) = \sum_v P(Y = y|V = v, g = 1)P(V = v).$$

Manski [2003] calls this assumption A2 “Assumption MAR” (Manski 2003, p. 27), but in my opinion this name could be misleading, because concerning the definition of Little and Rubin [2002] the property of MAR faces the probability of the missing conditional on the values of the variable of interest by postulating $P(g|y_{obs}, y_{mis}) = P(g|y_{obs})$ (and not turned around like here).

Until now the first kind of uncertainty has not been included and it has been assumed that the empirical estimates of the probabilities are known with certainty. In practice estimates have to be derived from finite sampling of size n and hence it is important to address the question of statistical inference. Creating confidence intervals shows a possibility to include the first kind of uncertainty and in this respect there are some ideas in the context of partial identification.
Horowitz and Manski [2000] suggest asymptotic confidence intervals that cover lower bounds ($\underline{b}$) and upper bounds ($\overline{b}$) with fixed probability and thus contain the whole identification region with given probability, namely $1 - \alpha$. Consequently their aim was to find an appropriate $z_{n\alpha}$ value such that $P(\underline{b}_n - z_{n\alpha} \leq \underline{b}, \overline{b}_n + z_{n\alpha} \leq \overline{b}) = 1 - \alpha$. For that they faced two possibilities. On the one hand they calculated z_n by an analytic way with the disadvantage of getting very complex covariance matrices. On the other hand they applied bootstrap sampling where each sample is taken to calculate a bootrap estimate for the bounds and in this way their distribution can be computed and z_n is found.
Imbens and Manski [2004] propose confidence intervals that cover the true parameter of interest with fixed probability instead of containing the whole identification region and showed that their interval represents a subset of the one of Horowitz and Manski [2000]. Because of the fact that the coverage probability does not converge to $1 - \alpha$ uniformly across various values for the length of the identification region, Imbens and Manski [2004] add a few adop-

tions to their original version. Nevertheless some assumptions are necessary in order to be able to build this interval. One of these assumptions consists of the postulation that the estimator of the nuisance parameter, namely the estimator for the length of the interval $\hat{Delta}$, is desired to be superefficient (i.e. $\sqrt{N}|\hat{\Delta} - \Delta_N| \xrightarrow{p} 0$, see Lemma 1 of Stoye [2009*a*]) at zero. Stoye [2009*a*] investigated this assumption more precisely and recognized that there are assumptions that are able to weaken this postulation of superefficiency for the nuisance parameter. Moreover he proposes a new confidence interval that can be applied without the validity of superefficiency.
After having summarized the formal backgorund for partial identification for the missing data case, it could also be interesting to reflect about the case of coarsened data, even if this area has not been investigated quite well. In this thesis there will be payed particular attention to the case of categorical data. Hence, I want to confine myself to this special case hereafter.

2.2.6. Some ideas concerning partial identificaiton in the context of coarsened data

As coarsened data can be regarded as a generalization of missing data (see 2.1), I want to apply some definitions of partial identification by Manski [2003] in the framework of coarsened data now.
For this purpose, I want to imagine a situation in which a variable of interest is able to obtain three possible categories, namely “A”, “B”, “C”, where some categories cannot be observed in a precise way such that for instance the observed categories are “A”, “B”, “C”, “A XOR B”, “A XOR C ” and “B XOR C”. This example is used to keep things simple and one can easily transfer it to a situation with more categories (e.g. categories “A” to “D”) as well as more different coarsenened observations (e.g. “A XOR B XOR C”, “A XOR D” ...). For obtaining a region for the probability of interest $P(Y = y)$, I proceed in a similar way like for the missing data case of Manski and thus using the law of

total probability for $P(Y = A)$ the following identification region results (see also dark green box of Figure 2.3):

$$\begin{aligned} H[P(Y = A)] \equiv &[\underbrace{P(Y = A|\mathcal{Y} = A)}_{1} P(\mathcal{Y} = A) + \\ &\underbrace{P(Y = A|\mathcal{Y} = (A\ XOR\ B))}_{\gamma_1} \cdot P(\mathcal{Y} = (A\ XOR\ B)) + \\ &\underbrace{P(Y = A|\mathcal{Y} = A\ XOR\ C)}_{\gamma_2} P(\mathcal{Y} = A\ XOR\ C), \\ &\forall \text{ possible } P(Y = A|\mathcal{Y} = \mathfrak{y}) = \gamma_i,\ i = 1, 2]. \end{aligned} \tag{2.13}$$

Apart from the probabilities that model the coarsening, namely $P(Y = A|\mathcal{Y} = (A\ XOR\ B)) = \gamma_1$ and $P(Y = A|\mathcal{Y} = AXORC) = \gamma_2$, all quantities (namely $P(\mathcal{Y} = \mathfrak{y})$) can be estimated by the sampling process. Using the empirical evidence only and accounting for the fact that γ_1 and γ_2 describe probabilities, γ_1 and γ_2 have to lie in the interval $[0, 1]$. Therefore, the following identification region can be obtained:

$$\begin{aligned} H[P(Y = A)] \equiv [&P(\mathcal{Y} = A), P(\mathcal{Y} = A) + P(\mathcal{Y} = (A\ XOR\ B)) + \\ &P(\mathcal{Y} = A\ XOR\ C)] \end{aligned} \tag{2.14}$$

The lower bound of equation (2.14) describes the case if the true value of coarsened observations like “A XOR B” and “A XOR C” is not “A” and the upper bound constitutes the situation if the true value of coarsened observations is always “A”.

Later on, I will show the relation between $\gamma = P(Y = y|\mathcal{Y} = \mathfrak{y})$ and $q = P(\mathcal{Y} = \mathfrak{y}|Y = y)$ (see equation (2.18)) and that there are further restrictions based on the empirical evidence for $q(\mathfrak{y}|y, \gamma)$. Because of the underlying relation the space of possible γ decreases by involving these restrictions for $q(\mathfrak{y}|y, \gamma)$ and consequently a shrunk form of the identification region in equation (2.14) might result.

According to the procedure of partial identification after having determined the identification region involving the empirical evidence only, one can consider further assumptions. Now I want to introduce two assumptions that are comparable to the ones Manski [2003] has proposed (see equation (2.11) and

(2.12)), but differ according to the underlying problem, namely coarsened data instead of missing data:

$$A1: \qquad P(Y = A|\mathcal{Y} = (A\ XOR\ B)) = P(Y = A) \tag{2.15}$$

$$A2: \quad P(Y = A|V = v, \mathcal{Y} = (A\ XOR\ B)) = P(Y = A|V = v). \tag{2.16}$$

This generally means that the true probability of a category's occurence does not depend on the observed value. Thus, in examplary assumption A1 of equation 2.15 the given observation "A or B" does not reveal any information about the probability of occurence of category "A". As in situations with more true and more coarsened categories observation "A or B" actually does give information in the sense that category "C" can not be the true category, it could be reasonable to condition a priori on all observations that are consistent with the true value of interest. In this way, here A1 expresses, that probability of occurence of category "A" is independent of the value that has been observed, namely if precise category "A", or coarse categories "A XOR B" or "A XOR C" have been observed, but other observations as "C" or "B or C" are excluded. Although the additional restriction that the observed values have to be consistent with the true value of interest is remiscent of the CAR assumption, it is important to notice that this kind of assumption differs from the CAR assumption, as within the CAR assumption it is conditioned on the true values and not on the observed values as in $A1$. $A2$ only differentiates from $A1$ by assuming that the described assumption is only satisfied for given values of an instrumental variable (see example on page 37).

Moreover contentual assumptions that seem to be plausible in this situation or can be concluded from similar studies can be imposed, that are able to shrink the indentification region based on the empirical evidence only. For instance, I could imagine that researchers are able to evaluate by contentual aspects if true category "A" is more plausible after having observed "(A XOR B)" or "A XOR C". According to this I want to derive a shrunk identification region that relies on the additional assumption $\gamma_1 = P(Y = A|\mathcal{Y} = (A\ XOR\ B)) \geq$

$P(Y = A|\mathcal{Y} = (A\ XOR\ C)) = \gamma_2$ (or vice versa). In this case only one uninformative probability is left in the underlying identification region

$$\begin{aligned} H[P(Y = A)] \equiv & [P(\mathcal{Y} = A) + \gamma_2 P(\mathcal{Y} = (A\ XOR\ B)) + \gamma_2 P(\mathcal{Y} = A\ XOR\ C), \\ & P(\mathcal{Y} = A) + P(\mathcal{Y} = (A\ XOR\ B)) + \gamma_2 P(\mathcal{Y} = A\ XOR\ C)]. \end{aligned} \tag{2.17}$$

Comparing this identification region with the one of equation (2.14) based on the empirical evidence only, one can note directly the shrunk length of the interval relying on further contentual assumptions.

$$\begin{aligned} &\text{length based on empirical evidence only :} \\ &P(\mathcal{Y} = (A\ XOR\ B)) + P(\mathcal{Y} = A\ XOR\ C) \\ &\text{length based on additional assumptions :} \\ &(1 - \gamma_2) P(\mathcal{Y} = (A\ XOR\ B)) \end{aligned}$$

Hence, the higher the value of γ_2, the shorter the length of the identification region based on this additional assumption and the more informative the assumption, i.e. the more substantial the extent of the interval's reduction. Similarly one could include assumptions about γ, for instance assuming γ_1 to lie within $[0.4, 1]$ instead of $[0, 1]$. The difference to the first proposal consists of the fact that the underlying idententification region is still dependent on both parameters γ_1 as well as γ_2.

I have derived the identification region from the missing data problem Manski [2003] has focused and so until now only the view of conditioning on the observed data has been regarded (see dark green box of Figure 2.3). In the beginning of this section the problem has been motivated from a different angle conditioning on the true variables (see bright green box of Figure 2.3). In this context the approach to solution has consisted of regarding the $q(\mathfrak{y}|y)$'s first instead of a direct analysis of $P(Y = y)$. In my opinion both perspectives are reasonable with regard to contentual justification: While conditioning on the true variable expresses the chronology of the underlying coarsening process in a proper way, namely that first the true value of the variable exists that gets coarsened in a second step, conditioning on the observed variable describes the situation the observer is concerned with, namely that some values of the

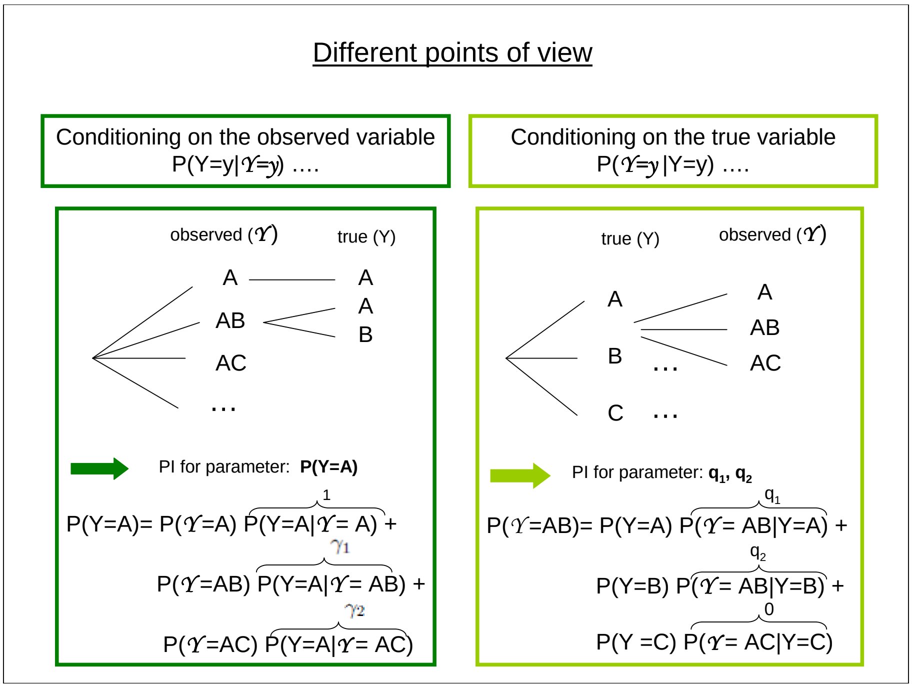


Figure 2.3.: Different perspectives as starting points for partial Identification (PI) in the context of coarsened data.

variable can only be observed in a coarsened way and therfore the task of the observer consists of concluding the true values from the observed values. The choice for one of these points of view is not of serious consequences, because there is a relation between both perspectives' underlying probabilities, which one can express in the following way:

$$\begin{aligned} \mathrm{P}(\mathrm{Y}{=}\mathrm{y}|\mathcal{Y}=\mathfrak{y}) &= \frac{P(\mathcal{Y}=\mathfrak{y}, Y=y)}{P(\mathcal{Y}=\mathfrak{y})} \frac{P(Y=y)}{P(Y=y)} \\ &= \mathrm{P}(\mathcal{Y}=\mathfrak{y}|\mathrm{Y}{=}\mathrm{y}) \frac{\mathbf{P}(\mathbf{Y}=\mathbf{y})}{\mathbf{P}(\mathcal{Y}=\mathfrak{y})}. \end{aligned} \tag{2.18}$$

To this initial situation of this chapter, namely the perspective of conditioning on the true variable, I want to link now and suggest an alternative way for partial identification in the context of coarsened data.

Transferring basic equation (2.1) from the beginning of this section to the example that is used here, one obtains for $P(\mathcal{Y} = (A\ XOR\ B))$ by the law of Total Probability (see also bright green box of Figure 2.3):

$$\begin{aligned} P(\mathcal{Y} = (A\ XOR\ B)) = \quad & P(Y = A)\, \underbrace{P(\mathcal{Y} = (A\ XOR\ B)|Y = A)}_{q_1} + \\ & P(Y = B)\, \underbrace{P(\mathcal{Y} = (A\ XOR\ B)|Y = B)}_{q_2} + \\ & P(Y = C)\, \underbrace{P(\mathcal{Y} = (A\ XOR\ B)|Y = C)}_{0} . \end{aligned}$$

In this equation several probabilities are unknown, namely q_1, q_2, $P(Y = A)$, $P(Y = B)$ and $P(Y = C)$. But partial identification of q_1 and q_2 and subsequent resolving for $P(Y = A)$ and $P(Y = B)$ could give information on those pure true probabilities and partially identify them.

The empirical evidence can not only tell that q_1 and q_2 take values within $[0, 1]$, but also some additional hints can be concluded like the fact that q_1 and q_2 cannot be simultaneously zero if there are some coarsened observations “A XOR B” and at least one $q(\mathfrak{y}|y)$ is equal to one if all observations are “A XOR B”. Similarly if the proportion of “A XOR B” is quite high (small), q_1 and q_2 cannot be small (high) at the same time. Therefore, there have to be some restrictions which can be derived from the data only without implying contentual further assumptions.

First, by rephrasing q_1, I want to derive an upper bound for q_1 that is always less than one and hence involves some information about this uninformative probability in every case:

$$\begin{aligned} P(\mathcal{Y} = (A\ XOR\ B)|Y = A) &= 1 - P(\mathcal{Y} \neq (A\ XOR\ B)|Y = A) = \\ &= 1 - P(\mathcal{Y} = A|Y = A) \\ &= 1 - \frac{P(\mathcal{Y} = A, Y = A)}{P(Y = A)} . \end{aligned}$$

If “A” is observed, it is obvious that the true value “A” is underlying and therefore the quantity $P(\mathcal{Y} = A, Y = A)$ can be simplified to $P(\mathcal{Y} = A)$. Hence, $P(Y = A)$ is the only probability that cannot be estimated from data. Because the true value “A” is only possible for observations “A” and “A XOR B”, the probability $P(\mathcal{Y} \in \{A, (A\ XOR\ B)\})$ seems to be a plausible guess. Because of

$$\begin{aligned} P(Y = A) &= P(Y = A|\mathcal{Y} = A) \cdot P(\mathcal{Y} = A) + \\ & P(Y = A|\mathcal{Y} = (A\ XOR\ B)) \cdot P(\mathcal{Y} = (A\ XOR\ B)) \\ &\leq P(\mathcal{Y} \in \{A, (A\ XOR\ B)\}) \\ &= P(\mathcal{Y} = A) + P(\mathcal{Y} = (A\ XOR\ B)), \end{aligned}$$

this proposed probability is greater than the required probability. Therefore,

$$\begin{aligned} \overline{q_1} = P(\mathcal{Y} = (A\ XOR\ B)|Y = A) &\leq 1 - \frac{P(\mathcal{Y} = A)}{P(\mathcal{Y} = A) + P(\mathcal{Y} = (A\ XOR\ B))} \\ &= \frac{P(\mathcal{Y} = (A\ XOR\ B))}{P(\mathcal{Y} = A) + P(\mathcal{Y} = (A\ XOR\ B))} \end{aligned}$$

represents an adequate upper bound whose components can be estimated. Because of the fact that in this chapter one considers the theoretical point of view, the corresponding empirical estimates will be derived in Chapter 4 when these findings will be applied.

One can directly note that this bound is always smaller than one apart from the case in which there are no precise observations availabe, i.e. $P(\mathcal{Y} = A) = 0$. This fact makes this upper bound $\overline{q_1}$ quite requirable, because in every case of coarsened data there is information that can be gained from the data only. This result seems to be quite surprising and astonishing, but the reason for this inherent information can be explained by the fact, that there are some precise observations available that are responsible for the shrinking of $q(\mathfrak{y}|y)$ to a value below one. For instance, if there are some "A" values observable in a precise way, then $q_1 = P(\mathcal{Y} = (A\ XOR\ B)|Y = A)$ will be less than one, because the value of $P(\mathcal{Y} = A|Y = A)$ is greater than zero in this case and $P(\mathcal{Y} = A|Y = A) + P(\mathcal{Y} = A|Y = A) = 1$ has to be satisfied.

Moreover I have derived an upper bound by an alternative way as appears from the appendix. It can be shown that this upper bound is always greater than the one I have proposed here (see Appendix A) and therefore it reveals less information.

Concerning the search for a nonzero lower bound of q_1, I noticed that one cannot calculate a lower bound without imposing further assumptions. As by using the data only one cannot exclude the case that all observations “A XOR B” have been produced by true values “B”, $q_1 = P(\mathcal{Y} = (A\ XOR\ B)|Y = A) = 0$ could be imaginable and therefore no lower bound that is larger than zero can be found.

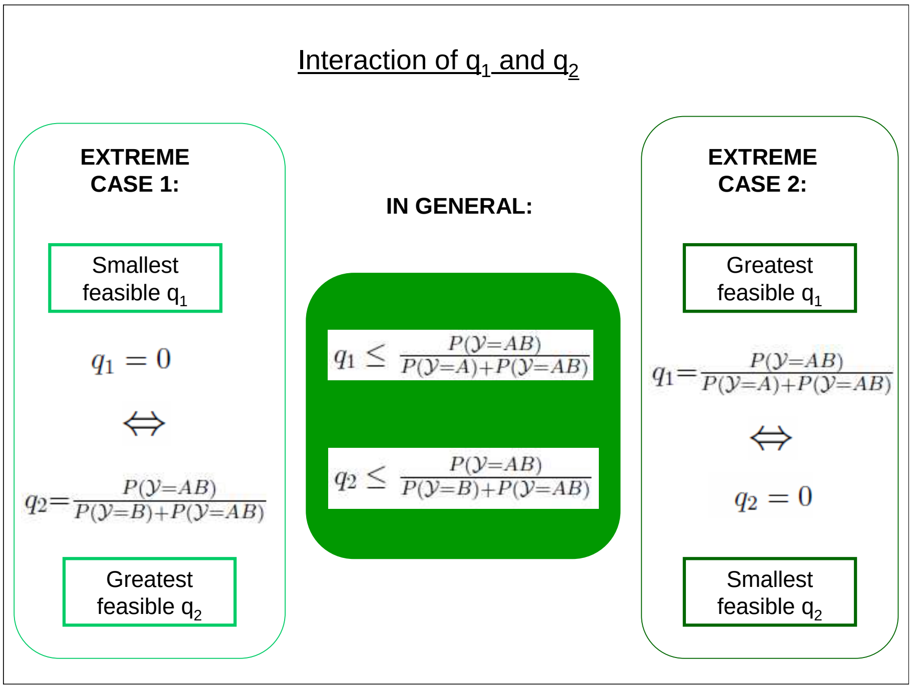

Interaction of q1 and q2
EXTREME CASE 1:
Smallest feasible q1
$q_1 = 0$
⇔
$q_2 = \frac{P(\mathcal{Y}=AB)}{P(\mathcal{Y}=B)+P(\mathcal{Y}=AB)}$
Greatest feasible q2
IN GENERAL:
$q_1 \leq \frac{P(\mathcal{Y}=AB)}{P(\mathcal{Y}=A)+P(\mathcal{Y}=AB)}$
$q_2 \leq \frac{P(\mathcal{Y}=AB)}{P(\mathcal{Y}=B)+P(\mathcal{Y}=AB)}$
EXTREME CASE 2:
Greatest feasible q1
$q_1 = \frac{P(\mathcal{Y}=AB)}{P(\mathcal{Y}=A)+P(\mathcal{Y}=AB)}$
⇔
$q_2 = 0$
Smallest feasible q2

Figure 2.4.: Interaction of q_1 and q_2 in extreme and general cases.

In Figure 2.4 the range of q_1 and q_2 and the underlying interaction between those probabilities is illustrated for given data. If true underlying values “A” are always observed in a precise way and thus for those true values no coarsened values “A XOR B” are observable, i.e. q_1=0 (see extreme case 1), the corresponding $q_2 = P(\mathcal{Y} = (A\ XOR\ B)|Y = B) = \frac{P(\mathcal{Y}=A\ OR\ B \cap Y=B)}{P(Y=B)}$ attains its upper bound, namely $q_2 = \frac{P(\mathcal{Y}=(A\ XOR\ B))}{P(\mathcal{Y}=B)+P(\mathcal{Y}=(A\ XOR\ B))}$. Explained in a contentual way probability q_2 is maximal if all observed coarsened values “A XOR B” are produced by true underlying “B” values where this fits to the described situation that all “A” values are precisely observed. In data situations in which the true value “B” is always observed in a coarsened way “(A XOR B)” (i.e. $P(\mathcal{Y} = B) = 0$), this upper bound attains its maximal possible value of one $(\frac{P(\mathcal{Y}=(A\ XOR\ B))}{P(\mathcal{Y}=(A\ XOR\ B))+P(\mathcal{Y}=B)} = 1)$. Because of the fact that precise observations

are responsible for being able to obtain an upper bound that is smaller than one, this makes sense indeed. Analogously extreme case 2 can be illustrated. In summary, even if it could be found an upper bound for q_1 that is less than one and therefore useful as long as there are coarsened observations, a lower bound cannot be found for reasons as explained.
After having involved the empirical evidence into the development of an identification region for q_1 and q_2, one can think about justified further assumptions, as "coarsening completely at random". If this quite strong assumption does not seem to be appropriate, one could introduce an instrumental variable v_j which is known for all j respondents and assume "coarsened at random", in the sense that the $q(\mathfrak{y}|y)$ values are equal conditional on the instrumental variable, namely $P(\mathcal{Y} = (A\ XOR\ B)|Y = A, v) = P(\mathcal{Y} = (A\ XOR\ B)|Y = B, v)$. Moreover, researchers might have an idea about the magnitude of q_1 and q_2. Therefore, Nordheim [1984] introduces quantity $R = \frac{q_2}{q_1}$ that reflects the relation of both unknown probabilities by factor R. In practice exact true R might be unknown, but it could be imaginable that a rough evaluation of R could be derived from contentual considerations, former studies or experiments. Introduction of auxiliary variable R describes a generalisation of the "coarsening at random" assumption, which results with R being equal to one ($R = 1 = \frac{q_2}{q_1}$ $\Leftrightarrow q_2 = 1 \cdot q1 \Leftrightarrow$ CAR). Therefore, using R is much more flexible than simply thinking about the possiblity if CAR might be satisfied.

After having summarized the advantage and formal background of partial identification, in this subsection I have proposed two possibilities for partial identification in the context of coarsend data. The first suggestion is based on the findings of Manski [2003], faces probabilities of interest directly and gradually imposes some assumptions on the probabilities conditioning on the observed values, namely γ. The second idea refers to the inital situation of the beginning of this chapter and partial identification is conducted with the probabilities conditioning on the true values, namely $q(\mathfrak{y}|y)$. Table 2.1 gives an overview of both ways of proceeding.

In Chapter 4 an approach will be presented that deals with epistemic uncertainty by means of a multinomial logit model. In this framework I want to

	Approach 1	**Approach 2**
Starting point	P(Y=y)= $= \sum_{Y} P(\mathcal{Y} = \mathfrak{y} \mid Y = y) P(Y = y)$	$P(\mathcal{Y} = \mathfrak{y}) =$ $= \sum_{\mathcal{Y}} P(\mathcal{Y} = \mathfrak{y} \mid Y = y) P(\mathcal{Y} = \mathfrak{y})$
Identification region for...	$P(Y = y)$	- $P(\mathcal{Y} = \mathfrak{y} \mid Y = y) = q$ - Deriving region for $P(Y = y)$ subsequently
Assumptions on... (point of view)	$P(Y = y \mid \mathcal{Y} = \mathfrak{y}) = \gamma$ (conditioning on the observed variable)	$P(\mathcal{Y} = \mathfrak{y} \mid Y = y) = q$ (conditioning on the true variable)
Empirical evidence	- $\gamma \in [0, 1]$ - derive further assumptions from approach 2 by using relation between those approaches	$\overline{q} \leq \frac{P(\mathcal{Y}=(A\ XOR\ B))}{P(\mathcal{Y}=A)+P(\mathcal{Y}=(A\ XOR\ B))}$ - No lower bound $\underline{q_1}$ can be found
Further assumptions	- Make plausible set-valued assumptions about γ - Evaluate by contentual aspects if $\gamma_1 > \gamma_2$ or vice versa	- CCAR - CAR - Assumption about $R = \frac{q_2}{q_1}$

Table 2.1.: Two proposed approaches for partial identification of $P(Y = y)$ in the context of coarse data.

include some ideas of partial identification by mainly using the second approach, because here the idea of epistemic uncertainty, namely a true variable exists first that is coarsened in a second step, is reflected in a proper way. I will involve assumptions derived from the empirical evidence as well as further assumptions as proposed here into the model and evaluate their benefit using simulated data. Especially it could be interesting to investigate the information that can be revealed without making contentual assumptions and to regard the corresponding bounds that can be derived by the methods described here.

All in all, I viewed partial identification as a very useful method if someone wants to analyse data without making untenable assumptions. Sensitivity analysis represents an alternative method that pursues the same goal, but proceeds from a different angle. But before explicitly comparing those two approaches in Section 2.4, I want to present some definitions, properties and examples of the procedure of sensitivity analysis first.

2.3. Sensitivity analysis

The motivation of choosing sensitivity analysis as a method of analysis is similar to that of partial identification. The analyst does not insist on point identification of estimators and prefers making justified assumptions only. Nevertheless, the procedure of sensitivity analysis differs from the one of partial identification. For giving an overview of sensitivity analysis, I first want to recall its basic idea and formal definitions that have been proposed by Vansteelandt et al. [2006]. As sensitivity analysis is used as an important tool for dealing with nonrandom missingness, I want to explain models of that kind next, that have already been proposed in literature (Molenberghs et al. 1999, Kenward et al. 2001 and Baker et al. 1992). Finally, I want to consider how basic definitions and ideas, mainly developed for the missing data problem, could be transferred to the case of coarsened data. Please note that mainly the second kind of uncertainty will be addressed and solely on page 51 it will be mentioned how the first kind of uncertainty can be incorporated.

2.3.1. Basic idea and some fundamental definitions

Instead of relying on one estimator which is derived from a special model, sensitivity analysis involves a range of estimators that can be obtained across various plausible values of the sensitivity parameter δ. The idea of a sensitivity parameter consists of the fact that even if sensitivity parameters are not identified, given a sensitivity parameter the parameter of interest θ is (Vansteelandt et al. 2006). Therefore, it is reasonable to calculate the parameter of interest for different models, namely for different values of the sensitivity parameter. Molenberghs et al. [1999] regard the missing data problem in the framework of sensitivity analysis and name the whole region of values θ that are derived from different plausible values of the sensitivity parameter δ as *ignorance region* for θ that is denoted by $ir(\theta, \Delta)$, with $\delta \in \Delta$. But before one is able to show the definition of this ignorance region, some formal notations have to be recalled which are adjusted in order to stay consistent with the previous sections. Robins et al. [1997] suggests to postulate that the underlying model class $\mathfrak{M}(\delta)$ has to be non-parametric saturated (NPS) by assuming that

$$f(Y_{\text{obs}}) = \int f(Y, G, \delta) dY_{\text{mis}}$$

has to be valid, where δ denotes the sensitivity parameter, Y denotes the variable of interest, where Y_{obs} denotes its observed and Y_{mis} its missing part, and G is 1 if Y is observed and 0 if it is missing. In words this postulate means that for each distribution of the observed data there has to be a single missing data process in the class and a unique law for the complete data such that $f(Y_{obs})$ is the marginal distribution of the observed data under the joint law $f(Y, G, \delta)$ (Robins et al. 1997). In simplified terms this means that for each imaginable value of the sensitivity parameter the class $\mathfrak{M}(\delta)$ covers an unique law that leads to the observed data. If this postulate is satisfied for each model, the parameter of interest is identified in an unique way as well (Vansteelandt et al. 2006). One can include information about the underlying missing process by constraining the possible values of the sensitivity parameter δ to the values that seem to be plausible, e.g. $\delta \in \Delta$, and thus model class $\mathfrak{M}(\delta) = \cup_{\delta \in \Delta} \mathfrak{M}(\delta)$ should be regarded (Vansteelandt et al. 2006).

Under the requirement of $\mathfrak{M}(\delta)$ being a non-parametric saturated class, Vansteelandt et al. [2006] define the ignorance region as (see equation (3.2) of Vansteelandt et al. [2006, p. 959], but adjusted notation here):

$$\begin{split} ir(\theta, \Delta) = \{\theta\{f(Y)\} : f(Y) = \int f(Y, G) dG \text{ with } f(Y, G) \in \mathfrak{M}(\delta)\} \\ \text{where } \mathfrak{M}(\delta) \text{ is NPI class.} \end{split} \tag{2.19}$$

Estimators of $ir(\theta, \Delta)$ are named *Honestly Estimated Ignorance Region (HEIR)* for θ (Vansteelandt et al. 2006). There is the property of weak consistency for point estimators that is useful in order to evaluate them. Vansteelandt et al. [2006, p. 962] developed a generalisation of the concept of weak consistency for point estimators to HEIRs by postulating weak convergence of every single point estimator $\hat{\theta}(\delta)$ to its underlying true value $\theta(\delta)$ across all $\delta \in \Delta$. Under this condition the HEIR overlies the true parameter of interest with arbitrary large probability if sample size increases. This is a quite preferable feature. An alternative concept of weak consistency can be viewed in the context of investigating if HEIR is a proper estimator of the identification region. For this purpose, one calls a HEIR $\hat{ir}_N(\theta, \Delta)$ weakly consistent for the identification region $ir(\theta, \Delta)$ if the underlying maximum distance becomes arbitrary small with increasing sample size (Vansteelandt et al. 2006, p. 967). Thus, the concept of consistency has been developed in two different ways, namely weak consistency of the HEIR for the true value as well as for the true ignorance region.

This ignorance region only accounts for uncertainty due to incompleteness (i.e. second kind of uncertainty, see Subsection 2.2.1) and neglects uncertainty that can be attributed to finite sampling (i.e. first kind of uncertainty, see Subsection 2.2.1). In Subsection 2.2.1 it has been emphasized that it is essential to distinguish between these two kinds of uncertainty and that one should not forget to incorporate the first kind of uncertainty within the analysis of missing or coarse data. There are some approaches how one could account for both kinds of uncertainty by combining the idea of confidence intervals, that are instruments for dealing with the first kind of uncertainty, and ignorance regions, that aim for the second kind of uncertainty. In this way one obtains the so-called region of uncertainty $UR_p(\theta, \Delta)$.

For calculating this region of uncertainty, the following notation might be helpful. δ_l and δ_u denote the values of the sensitivity parameter that belong to the lower and the upper bound of the ignorance region for θ, such that $ir(\theta, \Delta)$=$[\theta_l, \theta_u] = [\theta(\delta_l), \theta(\delta_u)]$. Under two assumptions, namely that for θ consistent and asymptotically normal estimators and standard errors exist according to model classes $\mathfrak{M}(\delta_l)$ and $\mathfrak{M}(\delta_u)$ (Assumption 1, see Vansteelandt et al. [2006, p. 960]) and that the observed data are independent of δ_l and δ_u as well as of θ_l and θ_u (Assumption 2, Vansteelandt et al. [2006, p. 960]), the pointwise uncertainty interval $UR_p(\theta, \Delta)$ can be constructed

$$UR_p(\theta, \Delta) = [C_L, C_U] = [\hat{\theta}_l - c_{\frac{\alpha*}{2}} se(\hat{\theta}_l), \hat{\theta}_u + c_{\frac{\alpha*}{2}} se(\hat{\theta}_u)], \qquad (2.20)$$

where the critical values $c_{\frac{\alpha*}{2}}$ can be determined by equation (4.3) of Vansteelandt et al. [2006, p. 961]. With probability of at least $(1 - \alpha)$ this region overlies $\theta(\delta)$ uniformly across all possible sensitivity parameters $\delta \in \Delta$ under the corresponding model class $\mathfrak{M}(\delta)$. From this proposition one can follow that for arbitrary $\delta_0 \in \Delta$ the pointwise uncertainty region covers the true value of the parameter of interest, namely $\theta_0 = \theta(\delta_0)$, with the requested given probability. Pointwise uncertainty regions as tools for partially identified parameters represent a generalisation of confidence intervals that can only be used for point identified parameters (Vansteelandt et al. 2006).

Another starting point for constructing uncertainty regions prefers to request that it covers the ignorance region instead of the true parameter with given probability. Thus, it can be decided to construct a rather conservative uncertainty regions $UR_s(\theta, \Delta)$ by choosing the critical value $c_{\frac{\alpha}{2}}$ to be the $(1 - \frac{\alpha}{2})$ quantile of the standard normal distribution. In this case all values within $ir(\theta, \Delta)$ are covered concurrently with given probability $(1 - \alpha)$. Apart from this described concept of strong coverage, Vansteelandt et al. [2006, p. 963–965] introduce uncertainty regions $UR_w(\theta, \Delta)$ based on weak coverage that does not cover all values of $ir(\theta, \Delta)$, but at least most of them with given probability.

Thus, because of the possibility of viewing concepts in respect to the true parameters on the one hand and to the ignorance region on the other hand, not only two versions of concepts in the context of weak consistency result, but also in the frame of construction of uncertainty regions.

After having shown the basic idea and some definitions of sensitivity analysis, I want to concentrate on one common field of application now, namely the modelling of nonrandom missingness. As missing represents a special case of coarsened data, I expect from this procedure to be able to refer some ideas to the coarsened data problem.

2.3.2. Modelling nonrandom missingness

It has already been indicated that one has to be careful with making assumptions like MAR or MCAR in the presence of missing data (resp. CAR and CCAR, in the case of coarsened data) , because in many situations this method is not justified. Therefore, a general procedure that is able to deal with different types of missing processes without the demand of any a priori information on the missingness would be desirable instead. Beside partial identification (see Section 2.2), sensitivity analysis represents a possible approach for such a procedure by formulating and comparing a number of possible non-ignorable (NI) models with the MAR model being a special case (Molenberghs et al. 2001, p. 17). But not all those NI models that can be derived seem to be appropriate. Hence, examples of Molenberghs et al. [1999] illustrate that different NI models that devote equally fits to the observed data lead to different predictions of the unobserved data. This shows the importance of incorporating plausible assumptions made on contentual grounds. Thus, corresponding to the idea of sensitivity analysis not only one special model (like a MAR model), but all NI models that are consistent with the observed data and seem to be justified should be included into the analysis.

For this purpose, in literature some suggestions have been made, where I decided to describe briefly an intuitive approach used for instance by Rubin et al. [1995], the selection model used for instance by Kenward et al. [2001] as well as the model of Baker et al. [1992].

In the framework of NI models, most literature concerns the case of contingency tables that exhibit some missing cells. For reasons of simplicity, I consider the case of a 2×2 contingency table.

Being interested in the proportion/counts of a special value of a bivariate characteristic (e.g. proportion of "yes"), forming the best-case-worst-case interval seems to be a very simple and evident approach for obtaining a range of possible parameters instead of a single one. For instance, Rubin et al. [1995] proceeded like that by calculating the proportion θ who participated in the plebiscite and elected for independence. For this, they calculate the proportion once classifying all "don't know" answers as "no" ("worst case" → θ_l) and once as "yes" ("best case" → θ_u) obtaining interval $[\theta_l, \theta_u]$. In the same way Kenward et al. [2001] calculate the best-case-worst-case interval for the proportion of HIV positive women under the presence of some women with unknown HIV status. As these intervals are frequently very wide, one rather uses this approach as a starting point and further models that shrink those initial ones might be helpful.

Selection models used by Little [1994] are a quite popular tool to face different NI models and thus include assumptions of different kinds. In the following

$\pi_{g_1 g_2,ij}$ denotes the underlying cell probabilities of a 2 × 2 contingency table, where the meassurement of ocassion 1 (with outcome categories j=1,2) and occation 2 (with outcome categories k=1,2) can either be missing (indicated by $g_1 = 0$ for occation 1 and $g_2 = 0$ for occation 2) or observed (denoted by $g_1 = 1$ for occation 1 or $g_2 = 1$ for occation 2). Selection models characterize cell probabilities $\pi_{g_1 g_2,ij}$ as the product of two components, namely p_{ij} that describes the meassurement process and $q_{g_1 g_2|ij}$ that gives some indication of the missing mechanism. Thus, the selection model can be formulated as (see equation (1) of Kenward et al. [2001])

$$\pi_{g_1 g_2,ij} = p_{ij} q_{g_1 g_2|ij}. \tag{2.21}$$

This equation shows the same contentual foundation like basic equation (2.1), but differs not only in respect of the notation, but also because here analysis is refered to contingency tables and therefore joint probabilities of bivariate (to keep things simple) characteristics are regarded. Different NI models are faced by imposing various restrictions on probability $q_{g_1,g_2|ij}$ that describes the non-response. For instance, Molenberghs et al. [1999, p. 111–113] regard three model classes. Dependend on the kind of restrictions, different degrees of freedom are obtained and thus non-saturated, saturated or overspecified models yield (see Kenward et al. [2001, p. 34]). While Model class I imposes quite strong restrictions by assuming MAR or even MCAR, in Model class II and III a nonrandom missigness process is considered, where Model II admits a variety of different dependence structures for the missing process and Model III at least postulates that the missingness of the first event's value is independent of the missingness at the second one's value. In this way a range of different models without making untenable assumptions are accounted. Nevertheless, it has been shown by Molenberghs et al. [1999] that it is reasonable to think about contentual assumptions in order to restrict the number of imaginable models.

Another suggestion concerning models that are able to face nonrandom missingness process comes from Baker et al. [1992]. Their model can be expressed as a special selection model by formulating its parameters in terms of $q_{g|ij}$ from equation (2.21) and hence the model of Baker et al. [1992] can be written as (see equation (2) by Kenward et al. [2001]):

$$\begin{aligned}\pi_{01,ij} &= & \pi_{11,ij}\alpha_{ij} \\ \pi_{10,ij} &= & \pi_{11,ij}\beta_{ij} \\ \pi_{00,ij} &= & \pi_{11,ij}\alpha_{ij}\beta_{ij}\delta \\ \text{with } \alpha_{ij} = \frac{q_{01|ij}}{q_{11|ij}},\ \beta_{ij} &= & \frac{q_{10|ij}}{q_{11|ij}},\ \delta_{ij} = \frac{q_{11|ij}q_{00|ij}}{q_{10|ij}q_{01|ij}}.\end{aligned} \qquad (2.22)$$

While parameter α models the missing of the first occasion, parameter β expresses the kind of missing of the second occasion. Parameter δ represents the additional effect that is present if at both occasions missing values are produced. Baker et al. [1992, p. 645] have shown that the latter parameter is independent from j and k. Using different dependence structures for parameter α and β by either setting them constant or admitting dependence on the first/second occasion, nine models $BRD1 - BRD9$ result that have been proposed by Baker et al. [1992]. For illustration of these models, I want to confine myself on the interpretation of a selection, namely $BRD1(\alpha, \beta)$, $BRD2(\alpha, \beta_j)$, $BRD8(\alpha_j, \beta_k)$. $BRD1(\alpha, \beta)$ can be regarded as the model that involves MCAR assumption, because missing of both characteristics is independent of the values of both characteristics. As missing at the second occasion is only dependend on the values of the first occasion, but not on the values of this occation itself, $BRD2(\alpha, \beta_j)$ represents a MAR model. An example for a nonrandom model is given by $BRD8(\alpha_j, \beta_k)$, which admitts dependence of the missingness on the value of the corresponding question (α which models missing of the first occurence is dependend of the value of the first occasion (j), the same for β respectively). Integrating some notions from sensitivity analysis, I understand parameters α and β as sensitivity parameters that determine the underlying missing model. For each value of these parameters, a different model results and thus different parameters of interest are calculated. Thus, the in this way obtained range of paremeters of interest forms the ignorance region. For more details concerning this approach see Baker et al. [1992], Molenberghs et al. [1999, p. 113], Molenberghs et al. [2001, p. 19–20] and Kenward et al. [2001, p. 35–36].

Using selection models as well as the model of Baker et al. [1992] (as a special selection model) leads to a range of models. Thus, not only one model, but some models are taken into consideration. This fact represents the fundamental idea of sensitivity analysis. In summary there are already some findings concerning modelling of nonrandom missingness by means of sensitivity analysis. As coarsening being a generalisation of missing data, it could be worth to apply some of these ideas in the coarsened data context. But before, I want to give a brief excursus concerning literature that shows how by means of sensitivity analysis maximum likelihood estimators can be found in case of partially categorized data.

2.3.3. Excursus: Sensitivity analysis of partially categorized data based on contingency tables

Sensitivity analysis in the context of missing data is often based on contingency tables (see Subsection 2.3.2). There are already existing a few approaches how maximum likelihood estimators in the presence of partially categorized data represented by a contingency table can be found. As a short excursus, I will explain the initial situation of three proposals and refer to the corresponding literature only.

Hocking and Oxspring [1974] show how one can deal with a situation in which there is given an original complete contingency table and a further table which can be derived from the origninal one for example in the sense that the first two columns have been combined. An example of application could be a questionnaire at a doctor, where some patients have to fill in an extensive questionnaire, while the majority must be classified by means of a questionnaire with limited or aggregated questions.

Blumenthal [1968] faces a situation in which the observer can classify major categories only, but distinguishing subcategories may be difficult. For example this situation can be present whether the observer prefers to make no distiction in doubtful cases or he behaves like this because of reasons of saving costs. Partial classified contingency tables are also addressed by Nordheim [1984] who also addresses the case of nonrandomly missing data. Because the parameters are not identified in the framework of maximum likelihood estimation, he introduces an additonal parameter which is the quotient of two parameters of interest.

If one is interested in one of these inital situations, please refer to the corresponding literature. Here I will try to transfer the models of Subsection 2.3.2 that have already been applied for the missing data problem to the coarsened data.

2.3.4. Refering some ideas to the problem of coarsened data

The issue of having too little a priori information about the missing process can be generalized to the coarsened data context (see Subsection 2.3.2), as determining the underlying coarsening process represents the key problem. So if the coarsening process were known, basic equation (2.1) could be resolved for the parameter of interest and the corresponding problem would vanish into the air. Thus, regarding several NI models could be reasonable for coarsened data as well.

For the case of missing data it was suggestive to regard two variables and thus to choose a depiction of data in contingency tables, because in this way for instance one could model the dependence of the missingness of one variable on the value of the other one. In the context of coarsened data I decided for reasons of simplicity to view the case of one variable only. As here the dependence of the coarsening on the underlying true value describes the center problem, this procedure might be sufficient as a first step. Thus, I modify the selection model of equation (2.21) to

$$\pi_{g,i} = p_i q_{g|i}.$$

Component $\pi_{g,i}$ represents the joint probability of the coarsened observation and the true value. This probability can be decomposed as the product of probability p_i, that can be interpreted as probability concerning the true value, and probability $q_{g|i}$, that expresses the coarsening mechanism in that way that it shows the transition from the true value of the variable of interest to its coarsened form. The latter component, namely $q_{g|i}$, will be in the center of the following considerations in order to develop a general NI model.

By analogy with the procedure by means of selection model in the missing data context, one could impose several restrictions to $q_{g|i}$ in order to obtain different models. To keep things simple, I want to face a situation, in which values “A”, “B”, “C”, “A XOR B” and “A XOR C” can be observed. Then for instance the coarsening of true value “A” to “(A XOR B)” could be of interest and thus modified selection model for

$$\pi_{(A\ XOR\ B),A} = p_A \cdot q_{(A\ XOR\ B)|A}$$

could be regarded. In this case restriction of $q_{(A\ XOR\ B)|A}$ to $q_{(A\ XOR\ B)}$ would lead to a MCAR model, because under this postulation the coarsening would be independent of the true underlying value “A” and thus $q_{(A\ XOR\ B)|A} = q_{(A\ XOR\ B)|B} = q_{(A\ XOR\ B)}$ would result. If dependency of $q_{(A\ XOR\ B)|A}$ on the true value is admitted, NCAR model and thus a NI model yields.

Please note that there is an important difference in this procedure compared to the case of missing data described in Subsection 2.3.2. Because of the fact that a value can either be missing or observed, missingness describes a binary phenomenon and hence the number of selection models that model different probabilities $\pi_{g_1,g_2,ij}$ is limited. As for instance, if you remember the situation of contingency tables in Subsection 2.3.2, three models are of interest, namely those that model target variables $\pi_{01,ij}$, $\pi_{10,ij}$ and $\pi_{00,ij}$. If one refers the model of this subsection with one variable only to the missing data case, even only one probability needs to be modeled by a selection model, namely $\pi_{0,i}$. Depending on the number of coarsened categories, there can be much more selection models in the coarsened data case describing different variables of interest. Thus, for gaining an insight into the coarsening of true values “A” in the example above, not only $\pi_{(A\ XOR\ B),A}$ of equation (2.23), but also probabilities $\pi_{A\ XOR\ C,A}$ (see equation (2.23)) and $\pi_{A,A}$ have to be explained by an own selection model. Therefore, I expect increased complexity by the application of selection models in the context of coarsened data compared to missing data.

Nevertheless there are some predetermined laws that have to be incorporated and thus the number of different models can be reduced. So if one focuses on the modified selection model

$$\pi_{A\ XOR\ C,A} = p_A \cdot q_{(A\ XOR\ C)|A}$$

as well, one has to account for the restriction that the sum of $q_{(A\ XOR\ B)|A}$ from equation (2.23) and $q_{A\ XOR\ C|A}$ from equation (2.23) has to be less or equal one, because $q_{A|A} + q_{(A\ XOR\ B)|A} + q_{(A\ XOR\ C)|A} = 1$ has to be valid. Thus, $q_{(A\ XOR\ C)|A}$ is a priori restricted by $q_{A\ XOR\ C|A} \leq$ 1-$q_{(A\ XOR\ B)|A}$. Additionaly one has to pay attention because of further restrictions that have been investigated in the context of partial identification (see Subsection 2.2.6, approach 2). There one could notice that there is a relation between $q_{(A\ XOR\ B)|A}$ and $q_{(A\ XOR\ B)|B}$ in the way that if $q_{(A\ XOR\ B)|A}$ is maximal then $q_{(A\ XOR\ B)|B}$ has to be zero and vice versa. Both aspects have to be accounted into the requirements for $q_{g|i}$.

After having calculated all possible models (e.g. for $\pi_{(A\ XOR\ B),A}$) under different plausible assumptions (MAR, NMAR) obtained by different restrictions on $q_{g|i}$, for each model an estimator for the parameter of interest (e.g. for p_A) can be concluded. By collecting all those estimators one could obtain the ignorance region.

Sensitivity analysis and partial identification presented in Section 2.2 can be considered as very similar approaches. Thus it could be intersting to compare those methods in respect of several aspects like their objective or their basic procedure. In this way a better understanding for the approach specific peculiarities can be developed and the decision for one of those can be facilitated. Therefore, comparison of these two approaches will be the main content of the next section.

2.4. Comparison of these approaches

In order to summarize the main aspects described in this chapter, it could be enlightening to establish a connection to the starting problem and to contrast the proposed approaches.

Recalling the initial situation of the beginning of this chapter, one can notice that the concepts and approaches that have been explained in the course of this chapter, namely ignorability, partial identification and sensitivity analysis, constitute possibilities to react to this problem. By means of basic equation (2.1) the basic problem of neither knowing the probability of interest $P(Y = y)$ nor the probability that describes the coarsening process $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ has been shown.

Partial identification and sensitivity analysis deal with this problem in a similar way compared to the approach associated with the concept of ignorability. Consequently the comparison of partial identification and sensitivity analysis is expected to be more insightful than the one of these two approaches and the concept of ignorability. Thus, I now want to start with shortly mentioning basic aspects only that characterize the comparison of ignorability and the other two approaches, before I mainly want to concentrate on working out the differences and similarity of partial identification and sensitivity analysis.

While the concept of ignorability rather deals with the initial problem by imposing assumptions on $P(\mathcal{Y} = \mathfrak{y}|Y = y)$, partial identification and sensitivity analysis insist on avoiding untenable assumptions and thus represent approaches that incorporate some uncertainty of the second kind. In this way simply assuming properties that lead to ignorability, namely CCAR or CAR plus distinctness of parameters, contradict the basic idea of partial identification and sensitivity analyis. Nevertheless, if further considerations on the plausibility of ignorability are made, assumptions like CCAR and CAR can be embedded within partial identification as well as sensitivity analysis. For shrinking the underlying identification region, it can be reasonable to add ignorability underlying assumptions to partial identification, like applied by approach 2 of Subsection 2.2.6. Sensitivity analysis can support the idea of ignorability if CAR or CCAR models are involved into the range of models that form the ignorance region, like explained in Subsection 2.23.

Thus, as long as ignorability is not seen as a concept that is generally able to deal with the initial problem, but rather as an instrument for improving identifiability within the framework of partial identification as well as sensitivity analysis, this concept can be brought into accordance with those two general concepts. Unfortunately, in practice analysts often take ignorability as a general approach for dealing with incompleteness and rely on assumptions like CAR or CCAR (or MAR and MCAR) just to simplify the problem and without considering if this is justified indeed.

Concerning the comparison of partial identification and sensitivity analysis in my opinion the most important commonality is reflected by the same objective that is pursued by these approaches. Thus, partial identification as well as sensitivity analysis account for the second type of uncertainty by making plausible assumptions only and do not insist on point identification. Therefore, in both approaches identifiability is no longer seen as a binary event and partial identified estimators in terms of intervals are admissible as an appropriate result.

Nevertheless, the way of achieving this objective differs enormously. Partial identification first uses the empirical evidence only by using information that can be revealed from the data generating process without making any further assumptions formed by contentual aspects. Only then in a second step assumptions that seem to be justified in this situation are imposed by degrees yielding more precise results. Thus, partial identification starts from total uncertainty (second type), which is reduced gradually by adding further assumptions. In contrast, sensitivity analysis begins by estimating point identified parameters derived by different models (for different sensitivity parameters) that seem to be plausible. In this way a set of point identified parameters results, where every point in this set is deduced by a different model that is consistent with the observed data. Hence, sensitivity analysis proceeds by starting with point identification and including the second kind of uncertainty by the union of several plausible models. Altogether one can determine that while partial identification starts with a rather high degree of the second kind of uncertainty, there is a precise point estimator in sensitivity analysis at the beginning. Thus, a main difference between partial identification and sensitivity analysis can be described by the direction of the procedure, which is illustrated by Figure 2.5 as well.

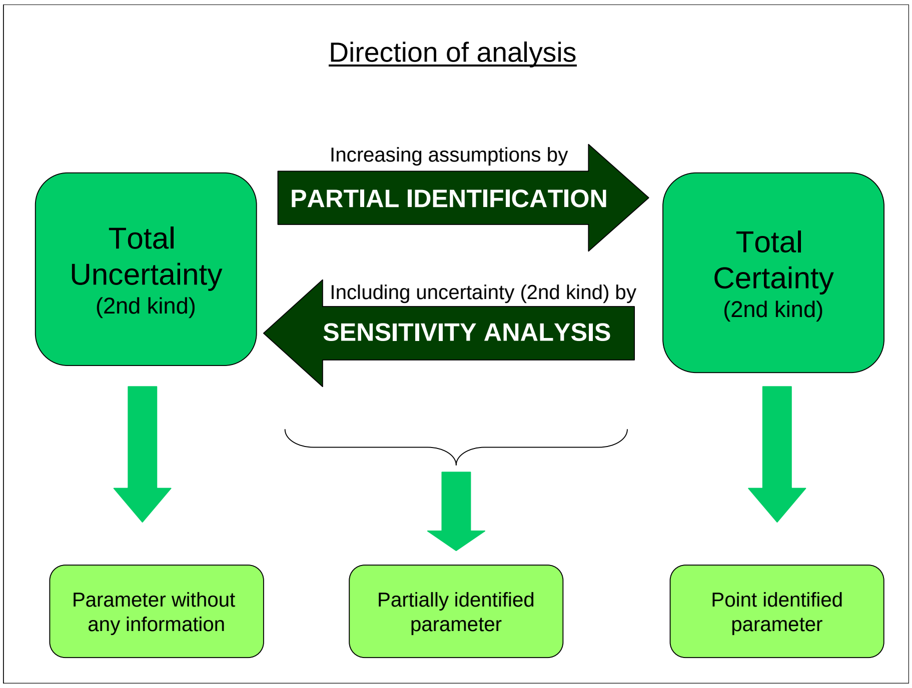

Direction of analysis
Increasing assumptions by
PARTIAL IDENTIFICATION
Total Uncertainty (2nd kind)
Including uncertainty (2nd kind) by
SENSITIVITY ANALYSIS
Total Certainty (2nd kind)
Parameter without any information
Partially identified parameter
Point identified parameter

Figure 2.5.: During partial identification reduces the second kind of uncertainty by gradually adding plausible assumptions, sensitivity analysis starts with point identified parameters and involves the second kind of uncertainty by implying a range of plausible models.

Even if their way of proceeding differs, because of focusing on the same objective, the result is depicted in a similar way. Hence, partial identification expresses the result by means of identification regions and the result that can be concluded by sensitivity analysis is named ignorance region, where both intervals cover all values of the parameter of interest that could be imaginable. The similarity of those presentations of the result especially can be noted if one views the approaches that have been developed in Subsection 2.2.6 and Susection 2.3.4 for the case of coarsened data.

Regarding approach 2 of partial identification in the context of coarsened data, one can notice that probability $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ is the center of analysis and restrictions on this probability are used in order to shrink the resulting interval for $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ and thus the length of the interval that will result for the parameter of interest will be decreased. A similar approach has been proposed here within the framework of sensitivity analysis, where one differentiates between various plausible models by restricting probability $q(g|i)$ of the selection model, which is comparable to $P(\mathcal{Y} = \mathfrak{y}|Y = y)$. The choice of some plausible models for $q(g|i)$ determines some resulting potential probabilities of interest by appliying the selection model and the union of those yields the ignorance region. Thus, if someone includes the same underlying dependence structures on $P(\mathcal{Y} = \mathfrak{y}|Y = y)$ (e.g. CAR) into a partial identification or sensitivity analysis based approach, at least similar results should yield, where this emphasizes the insignificance of the decision for one of those approaches.

Another similarity between partial identification and sensitivity analysis can be described by the fact that in both approaches there are ideas concerning the inclusion of the first kind of uncertainty. Additionaly in both approaches two types of confidence intervals have been developed. While the first type of confidence interval covers the parameter of interest with given probability $1 - \alpha$, the second type focuses on the coverage of the identification region in case of partial identification and of the ignorance region in case of sensitivity analysis. In the area of sensitivity analysis these confidence intervals are termed uncertainty regions.

Generally both approaches are mostly used in the framework of the missing data problem and because of that reason the underlying theoretical background is mainly formulated in this context. Additionaly theses basics have been related to the problem of coarsened data in this chapter. Moreover an example has shown that partial identification is able to deal with misclassified data which could be also imaginable for sensitivity analysis as selection model could include the misclassification mechanism of Molinari [2008]'s direct misclassification approach in the same way like the missing process $q_{g|i}$. Thus, partial identification and sensitivity analysis exhibit similar fields of application, like dealing with missing, misclassified and coarsened data.

In summary it has been revealed that there are some aspects like the underlying objective, the way of including the first kind of uncertainty and common fields of application that show the contentual closeness of partial identification and sensitivity. Additionaly, the depiction of the result in terms of intervals is very similar. Nevertheless, the direction of the procedure represents an important difference between those approaches and leads to different corresponding basic ideas.

Table 2.2 gives an overview to the similarity and differences in respect to these aspects.

	Partial identification	**Sensitivity analysis**
Objective	- Account for incompleteness by making justified assumptions only - Does not insist on point identification	
Way of procedure	- First use empirical evidence only - Then add justified assumptions → Direction: Getting more precise by adding assumptions	- Estimate parameter of interest under different models - Regard the union of all models in order to incorporate the second kind of uncertainty → Direction: Adding first kind of uncertainty by regarding several precise models
Depiction of the result	Identification region	Ignorance region
Including statistical imprecision	Two types of confidence intervals available * Coverage of parameter of interest * Coverage of identification/ignorance region	
Fields of of application	- Missing data - Coarsened data - Misclassified data	

Table 2.2.: Comparison of two approaches: Partial identification versus Sensitivity analysis.

In this chapter several possibilities have been shown that can be helpful to deal with initial situation of epistemic uncertainty described by basic equation (2.1), namely to get an insight about probability $q(\mathfrak{y}|y)$ that characterizes the coarsening process. Generally, one can be concerned with two situations, namely the case of a known or an unknown coarsening process. In the presence of a known coarsening process, on the one hand CAR or $CCAR$ assumptions could be reasonable and thus the corresponding likelihood could be simplified and on the other hand one could try to model the coarsening process (see Section 2.1). In the case that the coarsening is unknown, partial identification (see Section 2.1) and sensitivity analysis (see Section 2.3) have been shown

as approaches that involve some justified assumptions in order to be able to reveal some information about the underlying coarsening.
Having given an overview about possible approaches that are able to deal with epistemic uncertainty in this chapter, I want to concentrate on the presence of ontologic uncertainty in Chapter 3 and think about different methods that address this type of uncertainty.

3. Distribution on the power set as an approach for dealing with coarse categorical data under ontologic uncertainty

In Chapter 1 not only epistemic uncertainty has been presented as a reason for coarse data, but also ontologic uncertainty. At this point it has been worked out that ontologic uncertainty is present in situations that are described by indecision, such that coarse data of that kind do not show precise true values, but these coarse values already represent the truth. It is obvious that approaches as discussed in Chapter 2 are not appropriate for dealing with ontologic uncertainty, wherefore in this chapter I will consider some approaches that rely on the nature of this type of uncertainty and are able to include it. First it will be explained, why the procedure of presenting those approaches will be different compared to the previous chapter that focused on possibilities to deal with epistemic uncertainty. In Chapter 2 the initial situation in terms of equation (2.1) has been shown first and in the course of this chapter some approaches have been explained that focus on the corresponding problem. Under epistemic uncertainty there are true values that potentially can not be observed in a precise way. Thus, the task of those approaches was to try to get an idea about these true values and thus to calculate or at least partially identify the corresponding probabilites (e.g. $P(Y = A)$). As in the presence of ontologic uncertainty coarse observations represent the truth, methods that are able to give some information on the probabilities of the precise values can not be the main objective. Instead it is of peculiar interest how analysis of data and its fundamental notions change in the presence of ontologic uncertainty.

There are already some approaches that are able to depict ontologic uncertainty and provide a formal background as the random set theory and the Dempster-Shafer theory. While random set theory exhibits a well-defined mathematical framework for data of that kind, the Dempster-Shafer theory is helpful in context of interpretation of basic concepts as well as prediction of probabilities, when decisions have been made.

Section 3.1 will cover the random set theory, where the basic theory of random closed sets and the foundations of finite random sets, which are of peculiar interest here, will be addressed first. Afterwards considerations concerning coarse categorical data under ontologic uncertainty will be made by applying some ideas of random set theory. Thereby, general analysis in the presence of coarse categorical data under ontologic uncertainty is considered and the main difference compared to commonly used probability theory will be explained.

Section 3.2 proceeds in a similar way with regard to the Dempster-Shafer theory. After having described some basic ideas of the Dempster-Shafer theory, some conceptions will be used in order to extend the framework for analysing coarse categorical data under ontologic uncertainty. Some concepts of random sets are also available in context of the Dempster-Shafer theory, where these can be easily interpreted in the latter context. But the Dempster-Shafer theory will not only be useful in order to involve contentual aspects, but also shows the relation between commonly utilized probability theory and probability theory that will be used in order to include ontologic uncertainty. Notions of the Dempster-Shafer theory are applied for the case that decisions have already been made and that one is interested in the predicition of probabilities.

Both goals within the analysis of coarse categorical data under ontologic uncertainty, namely its general representation as well as prediction when decision has been made, will be focused in Section 3.3. Thereby, general conceptions in the presence of data of that kinds will be summarized in terms of the herefore introduced $\star$-notation. Moreover, a summary of the main differences of the conceptions of the $\star$-notation and the commonly used probability theory, which will be called classical probability theory, will be given. In order to be able to draw an appropriate comparison between those frameworks, it is important to determine the term “classical probability”.

In the following classical probability $P(A)$ will be defined as follows (Meintrup and Schäffler [2007], adjusted notation):

Definition 2. *Let Ω be the sample space, $\mathcal{F}$ a sensible σ-algebra, $(\Omega, \mathcal{F})$ a measuring space and $P : \mathcal{F} \to [0,1]$ a measure on $(\Omega, \mathcal{F})$ with $P(\Omega) = 1$. Then P is called probability measure and assigns to every event $A \in \mathcal{F}$ its probability $P(A)$.*

Classical probability has been defined as in definition 2 in order to ensure that axioms from Kolmogorov are valid (Kolmogorov 1933), which will be of importance at several points later on and enumerated in the following way (Narens 2007, p. 8):

1. $P(\emptyset) = 0$
2. $P(\Omega) = 1$
3. If A_i is a sequence of pairwise disjoint sets, then $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$ (σ-additivity).

There are some rules for calculation that will be further needed and that already result from general measures μ and by the additional condition $P(\Omega) = 1$ they are simplified only (Meintrup and Schäffler 2007, p. 61). In this way,

a. $P(A^c) = 1 - P(A)$

b. $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} (-1)^{i+1} \sum_{i \leq j_1 < \ldots < j_i \leq n} P(\cap_{k=1}^{i} A_{j_k})$

c1. continuity from below: if $A_n \nearrow A$ then $P(A_n) \nearrow P(A)$

c2. continuity from above: if $A_n \searrow A$ then $P(A_n) \searrow P(A)$

is valid, if $(\Omega, \mathcal{F}, P)$ is a probability space and $A, B, A_n \in \mathcal{F}, n \in N$. The second rule is often called sieve formula of Poincaré and Sylvester.

Apart from this mathematical definition, there are some interpretations of "probability", where the frequentist (Neyman 1977) and subjectivist interpretation (De Finetti 1977) are the most popular ones. While frequentist probability interpretes probability $P(A)$ as relative frequency of event A when random experiments have been realized infinitely times independetly of each other, the personal evaluation of probability by the observer is central in subjectivist interpretation. As the frequentist interpretation addresses mainly randomness and the subjectivist one concernes personal uncertainty, the latter case is of special importance here so that within this chapter the term “classical probability” will be interpreted in this way. Determination of subjectivist probability can be described as the consideration of the “betting odds”, which is calculated by the ratio of personal stake and profit in the context of a bet on the occurrence of event A.

In the presence of ontologic uncertainty analysts often proceed by excluding all values of variables that are in a coarse form and thus ontologic uncertainty is ignored. By doing this, it is assumed implicitly that coarse data do not reveal any information. But especially in the presence of ontologic uncertainty coarse observations do contain information, because these coarse data reflect the truth in the sense that there is no single true value. Additionaly coarse observations are informative, because knowing for instance that a person belongs to category “A” or category “B” implies that this person does not belong to category “C” in the case that three categories “A”, “B” and “C” basically can be possible. Hence, involving coarsened observations that come from ontologic uncertainty instead of ignoring them can improve results.

Therefore, the main goal of this chapter will be to show ways how analysis changes in presence of ontologic uncertainty and to transfer already existing foundations as random set theory and the Dempster-Shafer theory to the case of ontologic categorical coarse data. Thereby, the second kind of uncertainty (see Subsection 2.2.1) in the ontologic case is addressed exclusively, so that it is not accounted for sampling variability.

3.1. Theory of random sets

Even if Kolmogorov [1933, p. 46] had introduced random sets indirectly by regarding a “measurable region of the plane whose shape depends on chance”, before the 1970s there was hardly taken note of random sets (Stoyan 1998, Molchanov [2005]). Only when Matheron [1975] has defined the concept of random closed sets and discovered some fundamental mathematical background, more importance has been ascribed to random sets so that applications and extensions followed.

From then on random sets appeared in a variety of different fields of application. Closely followed to the initial idea of Matheron [1975], random sets can be helpful in the framework of geometrical statistics and image analysis. In this field random sets form the centerpiece and are regarded as stochastic models of geometrical structure (Stoyan 1998, p. 2). But random sets are also used in completely different areas as econometrics, where for example the demand of sets of consumers is modeled by random sets (Stoyan 1998, p. 1). Even within statistics the usage of random sets is varied so that random sets are used within survey statistics to determine the sampling design (Nguyen 2007, p. 137), set-valued stochastic processes can be understood as random sets and there is application of random sets in the context of grouped, censored or generally coarsened data as well (Nguyen 2006, p. 24).

In this section I will especially concentrate on the latter application, namely using random sets in the presence of coarse data (see Subsection 3.1.3). But before I want to recall theory of random sets in Subsection 3.1.1 and Subsection 3.1.2 by addressing the mathematical background of random closed sets in general as well as finite random sets. The special case of finite random sets will be covered as well because of its particular interest in the context of considering some ideas how the concept of random sets can be transferred to the situation of categorical data under ontologic uncertainty.

3.1.1. Random closed sets

Random sets can be regarded as a generalization of the concept of random variables. While random variables X can be described by a measurable mapping (e.g. $X : \omega \rightarrow \mathbb{R}$) that assigns to every elementary element ω a single value (e.g. arising from the real numbers $\mathbb{R}$) which is equipped with a proper σ-algebra (e.g. the Borel σ-algebra), in case of random sets to every elementary event of a field of probability a measurable region is assigned (Kolmogorov 1933, p. 46). Even if in the context of random variables measurability of the corresponding mapping represents the essential requirement, in the framework of random sets there has not been found a suitable σ-algebra if one concentrates on all measurable bounded subsets of the given space. Instead Matheron [1975] as well as most of the researchers who addressed random sets after him based their considerations on random closed sets, where in this case an appropriate σ-algebra is the Borel σ-algebra that is consistent with the Hausdorff metric[1] in the system $\mathcal{K}$ of all compact subsets ([Stoyan, 1998, p. 3]). Consequently Matheron [1975] concluded the following definition of random closed sets:

Definition 3. *A map $X : \Omega \rightarrow \mathcal{F}$ from a probability space (Ω, $\mathfrak{F}$, P) to the family $\mathcal{F}$ of closed subsets of locally compact seperable Hausdorff space E^2 is called a random closed set if $\{X \cap K \neq \emptyset\} \in \mathfrak{F}$ for every K from the family $\mathcal{K}$ of a compact subset of E.*

Thus, the main idea of Matheron's definition is not described by measures, but by hit-or-miss events instead and hence $\mathcal{F}$ is sometimes called hit-or-miss topology. The nature of this topology can be illustrated by means of an example depicted in Figure 3.1.

In this example policemen are on search of drugs which are burried deeply within the soil. For this purpose they are sending dogs within different areas (marked by the rectangles). The dogs are able to detect the drugs and bark everytime they are directly above the bag of drugs such that the policemen can increase their evidence bit by bit. Hence, it is observable whether random set X, namely the bag of drugs, hits or misses the selected testing sets K that are the compact subsets within the whole space E in which the dogs are sent. Having understood the contentual conception of this hit-or-miss topology, the definition of the distribution of random sets, which is given by the probabilities T(K)= $P(X \cap K \neq \emptyset)$ with $K \in \mathcal{K}$, can be easy comprehensible (Molchanov 2005, p. 3). This functional is called *capacity functional* of X and can be characterized by the following three fundamental properties (Nguyen 2006, p. 118, Nguyen [2007, p. 138]):

Definition 4. *A set-function $T : \mathcal{K} \rightarrow [0, 1]$ is called a capacity functional if it satisfies:*

[1] Matheron [1975]: Topology of space E is defined by a metric $d : E \times E \rightarrow \mathcal{R}_+$. Hausdorff metric ρ on $\mathcal{K}' = \mathcal{K} \setminus \{\emptyset\}$ is defined as $\rho(K, K') = max\{sup_{x \in K}\ d(x, K'),\ sup_{x' \in K'}\ d(x', K)\}$

[2] Matheron [1975, p. 1]: each point in E admits a compact neighborhood, and the topology of E admits a countable base

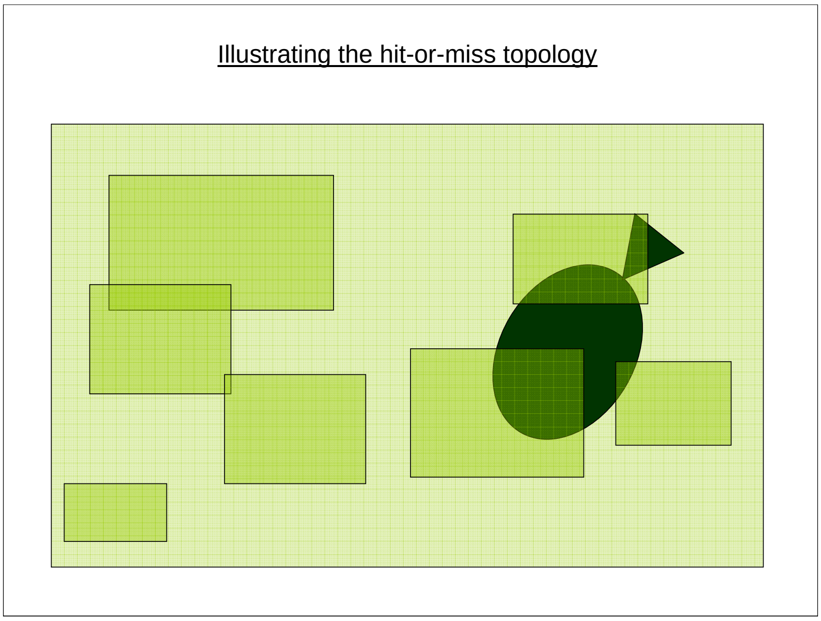


Figure 3.1.: The hit-or-miss topology: Calculating the probability that there is a non-empty intersection of testing set K (green rectangles) and random set X (dark-green area, representing a drug bag) within space E (big bright green rectangle), namely $P(X \cap K \neq \emptyset)$ (illustration similar to that of Davidson et al. [1974, p. 324])

1. $0 \leq T \leq 1$, $T(\emptyset) = 0$.

2. *T is alternating of infinite order.*

3. *T is upper semicontinous on $\mathcal{K}$.*

While the first postulation is equal to the foundation of probability theory (compare to first axiom on page 69), the second one means that it has to be satisfied for every $K_i \in \mathcal{K}$ ($i = 1, 2, ..., n \geq 2$) that

$$T(\cap_{j=1}^{k}) \leq \sum_{\emptyset \neq I \subseteq \{1,2,...,n\}} (-1)^{|I|+1} T(\cup_{i \in I} K_i)$$

such that, for instance, in case of $k = 2$ one obtains

$$T(A \cap B) \leq T(A) + T(B) - T(A \cup B).$$

It is clearly evident that this requirement differs from the corresponding rule of classical probability theory, namely the sieve formula of Poincaré and Sylvester (see page 69, case b.), as here inequality instead of just equality is required. Additionaly, one can conclude by the second requirement that the corresponding functional T is monotone (Nguyen 2006, p. 40). The third requirement means that if $K_n \searrow K$, then $T(K_n) \searrow T(K)$ (Matheron 1975), which corresponds to the continuity from above from classical probabiliy (see equation 69). Moreover, under this definition of capacity functionals Choquet-Kendall-Matheron theorem on $\mathcal{F}$ (Molchanov 2005) states that random closed set with capacity functional T is unique. Thus, following Choquet-theorem (Nguyen 2007, p. 138) there exists uniquely a probability measure P on $\sigma(\mathcal{F})$ such that $P(\mathcal{F}_K) = T(K)$ for all $K \in \mathcal{K}$.

Further properties have been investigated as for instance stationarity of random closed sets, i.e. invariance of its distribution in respect of translation in the sense that

$$T(K) = T(K + h)$$

is satisfied for all $K \in \mathbb{K}$ and all $h \in E$ (Stoyan 1998, p. 4).

But for the purpose of this thesis this basic foundation of random closed sets is sufficient. For considering how coarse categorical data under ontologic uncertainty could be modeled, it is more important to regard finite random sets in more detail, which will be central in the next subsection.

3.1.2. Finite random sets

The framework of finite random sets simplifies some findings of the general concept of random closed sets and thus can be regarded as an illustration of some results already obtained in Subsection 3.1.1 (Nguyen [2006, p. 109]). As in this subsection simply some definitions have to be applied in the special case of finite spaces, in this context I want to give some further information, namely recalling the distribution function of finite random sets additionally to the capacity functional already defined in the previous subsection. Apart from the capacity function, the distribution function of finite random sets will be important in the framework of deriving some concepts based on the analysis on the power set in Subsection 3.1.3, 3.2.2 and Section 3.3. Because finite random sets will be of peculiar interest with regard to dealing with coarse categorical data under ontologic uncertainty (see Subsection 3.1.3), it is generally worth to show some concepts for finite spaces in more detail as it will be done in the following.

It is reasonable to recall the general definition of finite random sets first, where Ω will denote a finite set from now on, $\mathcal{P}(\Omega)$ its power set and $(\Omega, \mathcal{A}, P)$ the corresponding proabability space (Nguyen 2006, p.37, adjusted notation):

Definition 5. *A finite random set with values in* $\mathcal{P}(\Omega)$ *is a map* $X : \Omega \to \mathcal{P}(\Omega)$ *such that* $X^{-1}(\{A\}) = \{\omega \in \Omega : X(\omega) = A\} \in \mathcal{A}$*, for any* $A \subseteq \Omega$*.*

Even if finite random sets are characterized by a similar mapping, it is obvious that there are some differences compared to random closed sets that are described in general definition 3. For characterizing the difference between the underlying random sets of these two definitions, in my opinion two aspects are of particular importance. While in the definition of finite random sets the corresponding mapping simply takes values within the power set $\mathcal{P}(\Omega)$, definition of random closed sets requires that mapped values come from the family $\mathcal{F}$ of closed subsets of locally compact seperable Hausdorff space E. This point can be caused by the fact that in the framework of general random closed sets an infinitely uncountable power set would result, wherefore this simplified definition 5 only can be applied for finite spaces. A second difference between definition 3 and 5 can be decribed by the underlying point of view. While in the context of random closed sets the idea of the hit-or-miss topology is included into the definition, finite random sets are defined in a rather original way by requesting measurability[3] on X as in the definition of finite random variables. Measurability is not sufficient in the framework of random closed sets as here one has to involve further constrains as restriction to the space of compact subsets.

As already mentioned finite random sets are not recalled in line with the hit-or-miss topology and thus capacity functional will be explained later on and

[3] measurable mapping: $(\Omega_1, \mathcal{F}_\rangle), i = 1, 2$ are two measuring spaces. A mapping $f : \Omega_1 \to \Omega_2$ is called $\mathcal{F}_1 - \mathcal{F}_2$-measurable, if $f^{-1}(\mathcal{F}_2) \subset \mathcal{F}_1$ Meintrup and Schäffler [2007]

first the *distribution function on the power set* $F(A)$ will be defined (Nguyen 2006, p. 38, adjusted notation):

Definition 6. *If* $F : \mathcal{P}(\Omega) \to [0,1]$ *is such that*

1. $F(\emptyset) = 0,\ F(\Omega) = 1$
2. *For any* $k \geq 2$, *and* $A_1, A_2, ..., A_k$ *subsets of* Ω, $F(\cup_{i=1}^{k} A_j) \geq \sum_{\emptyset \neq I \subseteq \{1,2,...,k\}} (-1)^{|I|+1} F(\cap_{i \in I} A_i)$.

then $\forall A \subseteq \Omega,\ F(A) = \sum_{B \subseteq A} f(B)$
where $f : \mathcal{P}(\Omega) \to [0,1]$ *is such that* $f(\cdot) \geq 0$ *and* $\sum_{B \subseteq \Omega} f(B) = 1$.

A function $F : \mathcal{P}(\Omega) \to [0,1]$ that satisfies these two requirements of definition 6 is called distribution function on the power set $F(A)$ (Nguyen 2006, p. 37) and can be calculated using probability densitiy function on the power set f that satisfies properties that are commonly known from probability theory, namely that f is non-negative and the sum of all possible densities, which are here defined by the densities of all feasible subsets ($B \subseteq \Omega$) (instead of all feasible singletons), is equal to one. The calculation of the distribution function on the power set $F(A)$ as well as its underlying idea will be comprehensible in the context of the Dempster-Shafer theory in Subsection 3.2.
Comparing the conditions of definition 6 and the definition of the capacity functional in the context of random closed sets (defintion 4), one notes that the first ones are equal respectively. Because of the fact, that Ω denotes the finite set that contains all elements, $F(\Omega)$ has to attain the maximal value which is 1 according to the first condition of defintion 6. Thus, one can conclude that $0 \leq F(A) \leq 1$, so that both conditions are equal. While the third condition of definition 4 is not needed for the finite case, the second conditions of those two definitions differ in the sense that in definition 4 T has to be of infinite alternating order and in definition 6 the so-called property of ∞-monotonicity has to be valid. Only if one restricts to nonempty finite random sets, which is required by $F(\emptyset) = 0$, 2-monotone distribution functions

$$F(A \cup B) \geq F(A) + F(B) - F(A \cap B),$$

i.e. ∞-monotonicity with $k = 2$ which is usually compatible with interval probabilities, but with classical probabilities not, are also monotone in the sense that (Nguyen 2006, p. 36)

$$A \subseteq B \Rightarrow F(A) \leq F(B)$$

is valid.

Apart from distribution function, capacity functional can be defined for finite random sets as well. But compared to the corresponding definition of random closed sets, the third condition (continuity) does not have to be satisfied in the context of finite random sets and thus one obtains the following defintion for the capacity functional (Nguyen 2006, p. 40):

Definition 7. *A set function $T : \mathcal{P}(\Omega) \to [0,1]$ is a capacity functional of some random set if it satisfies*

1. $T(\emptyset) = 0$*,* $T(\Omega) = 1$

2. *For any $k \geq 2$, and $A_1, A_2, ..., A_k$ in the power set,* $T(\cap_{i=1}^{k} A_j) \leq \sum_{\emptyset \neq I \subseteq \{1,2,...,k\}} (-1)^{|I|+1} T(\cup_{i \in I} A_i)$*.*

Capacity functionals represent the dual concept of distribution functions on the power set $F(A)$ and both concepts can be easily converted to each other by

$$T(A) = 1 - F(A^c).$$

Thus, the dual of ∞-monotonicity is the property of alternating of infinite order of definition 4 or definition 7 respectively and T is monotone like F (Nguyen 2006, p. 40).

The contentual differene between capacity functional and distribution function as well as their interpretation will be explained in the framework of the Dempster-Shafer theory in Section 3.2, where these are called plausibility function and belief function respectively.

3.1.3. Applying some ideas of finite random sets in the context of coarse categorical data under ontologic uncertainty

As already mentioned in the beginning of Section 3.1, grouped, censored or generally coarse data represent a possible field of application of random sets. Thus, there are already some first ideas how to use random sets in order to deal with coarse data (Schreiber 2000, Nguyen 2006, p. 42-50, p.186-190, Nguyen 2012), for which a short overview with regard to the central idea is given first. Nevertheless, these methods concern data under epistemic uncertainty, wherefore some considerations concerning approaches that try to deal with ontologic uncertainty by means of random sets will be made after that. As in this thesis categorical data are of peculiar interest, whose power set in practice mostly is of finite cardinality, thereby I will generally concentrate on finite random sets.

In most of the approaches that deal with coarse data by means of random sets, it is clearly evident that a situation of epistemic uncertainty is addressed. In this way, these approaches focus on random variables W whose outcome W_j ($j = 1, 2, ..., n$) is unobservable. Instead of observing this random variable, the outcomes X_j ($j = 1, ..., n$) which are an *iid* sample of random set X are observable. Moreover, it is assumed that the space that contains all possible observations, namely the space of all nonempty subsets of the sample space of Ω_W ($\mathcal{P}(\Omega_W) \setminus \emptyset$), always covers the real outcomes. For instance, if $\{a, b\}$ has been observed first (such that X_1=$\{a, b\}$), either a or b represents the true outcome (W_1=a or $W_1 = b$). In other words, random element W has to belong to random set X almost surely (with probability 1), wherefore in this case W is called to be an almost sure selector of X (Nguyen 2006, p. 25). Keeping this framework in mind, the goal of researchers who use random sets in the analysis of coarse data consists of considering how to conclude some information from these coarse observations and thus how to investigate the distribution of the random set's selector W (Schreiber 2000, p. 223).

In the course of this purpose Schreiber [2000, p. 223-p.227] as well as Nguyen [2006, p. 42-48] try to restrict potentially qualified probability measures μ on W. Thereby, Schreiber [2000] bases his considerations on the capacity functional and shrinks the space of possible probability spaces P by requiring that all possible $\mu \in P$ are dominated by capacity functional T ($\mu \preccurlyeq T$). This means that a probability space has to exist that involves versions of the random set X and a random variable W with distribution μ such that W is an almost sure selector. This resulting class of probability measures is called core (Schreiber 2000, p. 225):

$$core(T) = \{\mu \in P | \mu \preccurlyeq T\}.$$

Nguyen [2006] analogously focuses the core of the distribution function on the power set F and further restricts it by showing that F is the lower envelope of its core, which implies that (Nguyen 2006, p. 46)

$$F(A) = \inf\{\pi(A) : \pi \in core(F)\},$$

where A is a subset of the finite space of W and $\pi(A)$ its corresponding probability. Hence, $F(A)$ can be regarded as a kind of lower bound for $\pi(A)$. Subsequently it is shown how by means of CAR assumption (see Section 2.1) an unique CAR probability can be found.

As in the corresponding framework of these methods it is assumed that on the one hand true values, namely the values of random variable W, are underlying and on the other hand the chronology can be characterized by the procedure that these true values of W is given first and after that coarsened observation X are observed in a coarsened way, it is obvious that the case of epistemic uncertainty is considered. By contrast, under ontologic uncertainty coarse observations are not coarsened by a coarsening process, but coarse by nature, such that there are no true values underlying and for instance the coarse observed values X_j are induced by indecision. Therefore, the goal of the described analysis of Schreiber [2000] and others, namely finding statistical procedures for estimating the actual distribution of W, is improper in context of ontologic uncertainty.

Because of the inappropriateness of these procedures for the case of ontologic uncertainty, it is reasonable to make some considerations how one can represent coarse categorical data under ontologic uncertainty by means of finite random sets. Finite random sets are defined as measurable mappings $X : \Omega \rightarrow \mathcal{P}(\Omega)$

(see definition 5) with values within the power set $\mathcal{P}(\Omega)$. Furthermore, in this context F and T are defined as functions on the power set in the sense that $F : \mathcal{P}(\Omega) \to [0,1]$ and $T : \mathcal{P}(\Omega) \to [0,1]$. This random set based idea of focusing on the whole power set instead of single elements as in classical probability theory could be useful in presence of ontologic uncertainty, because here apart form precise observations as $\{a\}$ also set-valued coarse observations as $\{a,b\}$ are available which represent an element of the power set. Because of the absence of a true underlying value in case of observing coarse observations and the fact that these observations are coarse by nature, this realizations can be considered as own possible observations.

This thought and its consequences in context of foundations of classical probability theory shall be illustrated by an example, which contrasts classical probability theory (case 1) and ideas that include ontologic uncertainty (case 2). I want to emphasize that this example only raises some elected aspects of probability theory, where some further aspects will be addressed later on in Subsection 3.2.2 and conclusions will be summarized in Section 3.3. All commonly used notations, as for instance Ω for the sample space and ω for its elements, are equipped with a star in the presence of ontologic uncertainty. The star $\star$ is generally used in this thesis in order to mark the notations which are suggested for representing ontologic uncertainty.

Example 3
Case 1: Exclusive availability of precise possibilities of answer:
Ω=$\{a,b,c\}$, event of interest: $A = \{\text{a, b}\}$
Case 2: Availability of precise and coarse possibilities of answer:
$\Omega^\star$=$\mathcal{P}(\Omega) \setminus \emptyset$=$\{\{a\},\{b\},\{c\},\{a,b\},\{a,c\},\{b,c\},\{a,b,c\}\}$
event of interest: $A^\star = \{\omega^\star | \text{no singleton}\}$

By means of this example I go into the difference between case 1 and case 2 with regard to the definition of probability and basic probability calculation for the easy case of imposing the assumption that every outcome is equally probable with the result that Laplace probabilities are applicable. Please note that this assumption is quite strong and that there are many cases in which it is not justified. Nevertheless, for the purpose of illustration it is appropriate to impose this assumption and it will be obvious that general calculations can be realized analogously.

In classical probability theory (case 1) probability is defined as in definition 2, where a usually used σ-algebra is represented by the power set of Ω, which assignes to every event its underlying probability, such that

$$\begin{aligned} P : \mathcal{P}(\Omega) &\rightarrow \mathbb{R} \\ A &\rightarrow P(A). \end{aligned}$$

Refering this equation to the example and implying that every outcome is equally probable, probability $P(A)$ can be calculated by Laplace $P(A) = \frac{|A|}{|\Omega|} = \frac{2}{3}$ so that a probability of $\frac{2}{3}$ can be assigned to event A.

Extending this classical definition to the case of the availability of some coarse categorical data as well (case 2), sample space Ω changes to $\Omega^{\star}$=$\mathcal{P}(\Omega) \setminus \emptyset$. The empty set is excluded in order to ensure desirable properties as monotonicity as already mentioned in the context of finite random sets (see Subsection 3.1.2). In practice this means that every respondent has answered this question and reports at least one of the possible categories. Because of the fact that in case of implying ontologic uncertainty the outcomes within the sample space do not have to be singletons, the mapping that defines probability $P^{\star}$ is characterized as

$$\begin{aligned} P^{\star} : \mathcal{P}(\Omega^{\star}) = \mathcal{P}(\mathcal{P}(\Omega) \setminus \emptyset) &\rightarrow \mathbb{R} \\ A^{\star} &\rightarrow P^{\star}(A^{\star}). \end{aligned}$$

Requiring again the applicability of Laplace probability, $P(A^{\star}) = \frac{4}{7}$ can be assigned to event $A^{\star}$, so that respondents are indecisive between at least two answers with this probability. Apart from the fact that the classical sample space is replaced by its power set without the empty set in order to be able to include coarse observations and thus probability is no longer described as a mapping from the power set of Ω, but from the power set of the power set of Ω without the empty set, no big changes compared to classical probability theory seem to result. As the conception of ontologic uncertainty implies that coarse observations can be interpreted as own outcomes, this extention can be solely characterised by an alternative sample space $\Omega^\star$ (instead of Ω), such that this result is not suprising. Therefore, interpretation and properties of $P^\star$ can be compared to that of P with the only difference that the former relys on $\Omega^\star$ and the latter on Ω.

In this subsection only first considerations concerning a possibility to incorporate ontologic uncertainty have been made. It has been derived from the basic idea of random sets that a probability definition on the power set $P^\star : \mathcal{P}(\Omega^\star) \rightarrow \mathbb{R}$ could be appropriate in the presence of coarse categorical data under ontologic uncertainty. As under ontologic uncertainty coarse data represent the truth, it has been concluded that every element of the corresponding power set without the empty set forms an own category so that $\Omega^\star{=}\mathcal{P}(\Omega) \setminus \emptyset$. $P^\star$ will be the only notion of the *-framework that can be interpreted in terms of probabilities. Results that will be derived from the Dempster-Shafer theory concern conceptions that rely on evidences and beliefs instead and will be used in order to make predicitons that involve ontologic uncertainty.

Even if the distribution function does not play an important role in presence of categorical data because of the underlying nominal scale of measurement, it will be shown that the distribution function on the power set as well as the capacity functional can be interesting for including ontologic uncertainty into results when decision has already been made. As the interpretation of these functions and their difference will be more comprehensible in context of the Dempster-Shafer theory, it will be addressed in Subsection 3.2.2. Before it is reasonable to give summary of the basic idea of the Dempster-Shafer theory.

3.2. The Dempster-Shafer theory

Although 1665 already Leibniz suggested a numerical assignment on the scale of $[0, 1]$ in order to formalize the “degrees of proof” and Bernoulli [1713] thought about pure and mixed evidence, Shafer [1976] who relies on Dempster’s idea of lower and upper probabilities (Dempster, A. 1967) and thus the so-called Dempster-Shafer theory (DST) gave rise to more analysis of beliefs (Fine 1977). Generally, there are different points of view how one can consider DST, as one can either interprete it in relation to probability theory or regard it as an autonomous theory. This distinction is of particular importance, if the combination rule is applied as here one has to decide if impossible events (i.e. the empty set) shall be excluded or not. As this rule of combination will not be of peculiar interest here, it will be briefly illustrated only and thus this differentiation is not very important. Nevertheless, DST will be regarded rather as an autonomous approach, as it can be noted from the exlusion of the empty set within the $\star$-notation, but at some points comparisons to probability theory will be drawn in order to get a better understanding of the underlying notations with regard to their generality.

Even if some mathematical foundations of the DST overlap with the theory of random sets (see Section 3.1), as the definition of capacity functionals, which are called “plausibility function” in terms of DST, and the distribution functions on the power set, for which the notion “belief function” is used, show, in this section foundations of DST will be presented mostly without mentioning any relations to random sets. Instead, the basic goal will be to conclude some further ideas from DST for dealing with coarse categorical data under ontologic uncertainty.

For this purpose, some general aspects of DST will be presented in Subsection 3.2.1, where especially the basic conception and their underlying interpretiation of essential notions as “belief functions” and “plausibility functions” will be addressed. Keeping foundations of DST in mind and considering how one could deal with coarse categorical data under ontologic uncertainty, findings of Subsection 3.1.3 that concerned this kind of uncertainty by applying some results of random set theory will be extended in Subsection 3.2.2. Thereby, a generalized framework of probability will be recalled and it will be explained how intervals can be constructed that are able to represent ontologic uncertainty.

3.2.1. Foundations of Dempster-Shafer theory

First some basic terms and conceptions will be illustrated by means of an example that is similar to that of Zadeh [1986].

> **Example 4** Six persons, who are not sure about their exact weight, report it in terms of an interval (see Table 3.1). More-

over, it is assumed that there are no effects of social desirability so that the true weight is within the interval indeed. The fraction of respondents whose weight is within the query set $Q = [75, 80]$ will be of interest.

person no.	1	2	3	4	5	6
weight w_i [*in kg*]	$[55, 57]$	$[56, 61]$	[75, 82]	[78, 80]	[77, 81]	[98, 101]

Table 3.1.: Explaining baisc terms of DST by means of Example 4.

As there are some answers that overlap with the query set, but are not definitely within the query set, it is difficult to determine the required fraction, if there is no rule available. DST provides two notions that can help to find a satisfying answer in questions of that kind.

1. *Measure of belief* includes all weights w_i that are fully contained within the query set (i.e. $w_i \subseteq Q$).
2. *Measure of plausibility* involves all weights w_i that intersect the query set Q (i.e. $w_i \cap Q \neq \emptyset$).

Thus, referred to the example the fraction of respondents whose weight is within the query set is equal to $\frac{1}{6}$ using the measure of belief as only $[78, 80]$ is covered completely by the query set. Since the answers $[75, 82]$, $[78, 80]$ and $[77, 81]$ exhibit some values that coincide with the values of the query set, a fraction of $\frac{1}{2}$ results when applying the measure of plausibility. In order to express these results in words, one can state that the fraction is certainly $\frac{1}{6}$ and possibly $\frac{1}{2}$ and thus the answer can be represented as an interval whose lower bound is described by the measure of belief and the measure of plausibility forms its upper bound.

Having gained an insight into the contentual idea of the measure of belief as well as the measure of plausibility which both play an important part in the framework of DST, it is reasonable to proceed with more formal definitions. For this purpose, the following notion of the basic probability assignement is of particular significance, because it can be regarded as a building block for the construction of the belief function as well as the plausibility function (Beynon et al. 2000, p. 40):

Definition 8. *A basic probability assignement is a function* $m : \mathcal{P}(\Omega) \rightarrow [0,1]$ *such that:* $m(\emptyset) = 0$ *and* $\sum_{A \subseteq \Omega} m(A) = 1$.

It is important that one distinguishes between probabilities that evaluate how probable events occur and those masses $m(A)$ that concern the confidence that can be exactly commited to A.
Using the basic probability assignement belief function as well as plausibility function can be defined as follows (Shafer 1976):

Definition 9. *A function* $Bel : \mathcal{P}(\Omega) \rightarrow [0,1]$
with $Bel(\Omega) = 1$, $Bel(\emptyset) = 0$ *that is* ∞*-monotone is called belief function and can be calculated by* $Bel(Q) = \sum_{A \subseteq Q} m(A)$.

A function $Pl : \mathcal{P}(\Omega) \rightarrow [0,1]$
with $Pl(\Omega) = 1$, $Pl(\emptyset) = 0$ *that is alternating of infinite order is called plausibility function and can be calculated by* $Pl(Q) = \sum_{A \cap Q \neq \emptyset} m(A)$.

Consequently the belief function is the sum of the masses of all subsets that imply the query set Q (hypothesis) and in accordance with the illustrating example 4 can be interpreted as a lower bound for the confidence of a hypothesis. By contrast, plausibility forms an upper bound by accounting for all sets that could possibly support hypothesis Q in the sense that the hypothesis and the subsets of the frame of discernment only have to intersect in order to be included. Therefore, plausibility expresses the extent to which one fails to disbelieve hypothesis Q (Beynon et al. 2000, p. 40)
Concerning the mentioned conditions within definition 9 it follows that both functions range from 0 to 1, where $Bel(Q)$ and $Pl(Q)$ are zero in case of the query set Q being the empty set and the maximal value of one is attained if Q is equal to Ω. For understanding the additional requirements, namely that $Bel(Q)$ is ∞-monotone and $Pl(Q)$ is an alternating of infinite order (Wu and Mi 2008, p. 77), I refer to Section 3.1 where these properties have proposed in the framework of capacity functionals as well as distribution on the power set. Having proposed the measure of belief and the measure of plausibility as

lower and upper bound respectively, it is obvious that $Bel(Q) \leq Pl(Q)$. Both functions can be converted into each other by (Beynon et al. 2000, p. 40)

$$Pl(Q) = 1 - Bel(Q^c),$$

where Q^c is the complement of set Q.

Even if the foundations of the conceptions belief and plausibility are more important here, the second aspect of DST, namely the rule of combination that prescribes how evidence from two or more independent sources can be combined as well as the associated problem of normalization, will be shortly addressed for the sake of completeness.

Generally combined beliefs or plausibilities can be calculated by the original way (see definition 9), but the rule of combination has to be used for the involved basic probability assignements m_1 (of first source) and m_2 (of second source) (Beynon et al. 2000, p. 41):

$$[m_1 \otimes m_2]\,(Q) = \begin{cases} 0, & \text{if } Q = \emptyset \\ \frac{\sum_{A \cap B = Q} m_1(A) \cdot m_2(B)}{1 - \sum_{A \cap B = \emptyset} m_1(A) \cdot m_2(B)}, & \text{if } \ Q \neq \emptyset \end{cases} . \tag{3.1}$$

For illustration of this rule, example 5 might be helpful.

> **Example 5** The goal is to find out who had filled the role of Santa Claus this year. Child 1 (C_1) guesses that Santa Claus is played by his uncle (U), where child 2 (C_2) guesses that her father (F) is Santa Claus. The credibilities of both children are 0.6 and 0.7 respectively. Both children are brother and sisters so that they cannot be right at the same time.

	C_2 **credible (0.7)**	C_2 **not credible (0.3)**
C_1 **credible (0.6)**	impossible! 0.56	U is Santa: 0.18
C_1 **not credible (0.4)**	F is Santa: 0.28	uncertain Santa: 0.12

Table 3.2.: Explaining combination rule of DST by means of Example 5.

Table 3.2 shows that the combination of these two independent sources (C_1 and C_2) follows by calculating the product of corresponding evidences, which equals the numerator of equation (3.1). Because of the possibility of non-combinable events, the resulting product has to be normalized by involving those cases only that do not lead to this conflict. This normalization is realized by the denominator of equation (3.1). Thus, in this example one obtains combined mass probability assignement according to equation (3.1), so that

$$[m_1 \otimes m_2](U) = \frac{0.6 \cdot 0.3}{1 - 0.56} \approx 0.41.$$

Because of the involved normalization, this rule of combination can lead to counterintuitive conclusions (Zadeh 1984, p. 82) in case that DST is considered with regard to probability theory. If for instance, C_1 is 98 percent sure that person A has played Santa Claus and 2 percent sure that it has been real Santa Claus R, while C_2 is 99 percent sure that Santa has been played by person B and 1 percent sure that real Santa Claus R has been there, then rule of combination would conclude that $Bel(R) = 1$ even if both children were highly unlikely that it has been real Santa Claus (normalization excludes all other cases as they lead to empty sets). Other examples that show counterintuitive results can be easily constructed by applying this rule of combination (see for instance Zadeh [1986, p. 89]). Otherwise, if DST is regarded as autonomous theory as here, one is not concerned with counterintuitive results, as in this case interpretations do not have to be made in terms of probabilities, but terms as “confidence” or “belief” instead.

3.2.2. The Dempster-Shafer theory as an instrument in order to include ontologic uncertainty

After having acquired some knowledge of DST, ideas concerning tools that are able to deal with ontologic uncertainty can be derived. In Subsection 3.1.3 some first considerations of that kind have been made by involving aspects of random set theory. At this point two findings have been noted: Firstly, instead of regarding all elements of the sample space Ω, focusing on the power set ($\Omega^\star = \mathcal{P}(\Omega) \setminus \emptyset$) should be preferred as coarse data under ontologic uncertainty already represent the truth and thus secondly probability on the power set of the power set without the empty set $P^\star : \mathcal{P}(\Omega^\star) \rightarrow \mathbb{R}$ seemed to be a suitable way in order to represent ontologic uncertainty. Thus, no change apart from an enlargement of the sample space has been resulted if one includes coarse data under ontologic uncertainty.

In this section considerations will rely on the central idea of using the power space of Ω without the empty set as new sample space $\Omega^\star$ again, but at this point mainly distribution functions of this probability theory on the power set will be addressed. In this thesis the categorical case will be of special interest, so that distribution functions do not play an important role at the first glance. But as summation of several basic probability assignements $m(A)$ that support hypothesis of interest Q (either by using certain sets only ($A \subseteq Q \rightarrow$ belief function) or possible sets as well ($A \cap Q \neq \emptyset \rightarrow$ plausibility function)) can be regarded as general distribution functions, distribution functions on the power set are an important tool even in the presence of categorical data. The mentioned fact that these distribution functions on the power set can be regarded as generalization of classical distribution function will be explained in this section.

The class of all possible distribution functions on the power set will be denoted by $\Pi^\star$ and its elements by $F^\star$, while distribution function from classical probability theory will be labeled by F. As by DST the interpretation of special distribution functions on the power set $F^\star$, namely the belief function $Bel(A)$ as well as the plausibility function $Pl(A)$, is easily comprehensible, some contentual aspects can be derived for dealing with data under ontologic uncertainty. Please note, that findings from random set theory that have been presented in Subsection 3.1.3 and conclusions from DST that will be drawn here, pursue completely different goals. While the former results rather concern the general framework of probability theory and basic idea concerning dealing with ontologic uncertainty, derivations of the latter concentrate on prediction after a decision has been made and data are not coarse anymore so that it does not directly attend to ontologic uncertainty itself. This will be emphasized in Section 3.3 as well, where both approaches that involve analysis on the power set without the empty set will be summarised.

In this section first I will specify the relation between notions from classical probability theory and probability distributions on the power set $F^\star$ in more detail, so that one can gain a better understanding about the latter one and embed it into usual foundations. After that further thoughts are given to the way how prediction could be made after one has come to a decision that incorporates ontologic uncertainty.

DST exhibits distribution functions on the power set, namely the belief function $Bel : \mathcal{P}(\Omega) \rightarrow [0, 1]$ as well as the plausibility function $Pl : \mathcal{P}(\Omega) \rightarrow [0, 1]$. These two functions come from a family of possible distributions on the power set $\Pi^\star$, where belief function and plausibility function mark the range of possible probability distributions $F^\star \in \Pi^\star$ in the sense that (Beierle and Kern-Isberner 2005, adjusted notion)

$$\begin{aligned} Bel(A) &= \inf\{F^\star(A)|F^\star \in \Pi^\star\} \\ Pl(A) &= \sup\{F^\star(A)|F^\star \in \Pi^\star\}, \end{aligned}$$

so that $Bel(A)$ and $Pl(A)$ represent lower and upper bound (Zadeh 1986, p. 86) respectively as already mentioned in Subsection 3.2.1.

Lack of knowledge concerning underlying probabilities is expressed by this range between $Bel(A)$ and $Pl(A)$ and thus total knowledge is available in case of both functions being equal (Beynon et al. 2000, p. 41). Under this condition classical probability theory and probabilities from DST $F^\star$ coincide. Thus, the larger the difference between belief and plausibility, the more one removes oneself from classical probability theory in the sense that uncertainty is involved by means of admitting a range of possible distribution functions limited by the belief and the plausibility function. In this way distribution functions on the power set from DST extend notions from classical probability theory in the sense that the lack of knowledge can be specified instead of simply assuming that there is total knowledge about probabilities available (Bellenger and Gatepaille 2011). On the one hand in the case that probabilities are available indeed, analyses of classical probability theory generally leads to more precise results (Kohlas and Monney 1995, p. 8). But on the other hand by using notions of DST one can avoid the problem that sometimes one does not have any information about underlying probabilities so that no assumptions that potentially are not justified need to be imposed (Beynon et al. 2000, p. 39). Moreover, one can regard the rule of combination in context of DST (see Equation 3.1) as generalization of Bayes' rule $P(A|B) = \frac{P(A \cap B)}{P(B)}$ (Dempster 2008, p. 5) that concernes the calculation of conditional probabilities, as the intersection of the numerator reflects the product of (independent) sources and the denominator represents a normalization.

These first arguments for the fact that classical probabilities can be regarded as a special case of the more general notions from DST that rely on the power set shall be further extended by comparing some more methods and properties of those two conceptions hereafter.

Therefore, the following three selected properties of classical probability theory

1. $P(\emptyset) = 0, P(\Omega) = 1$

   (see first and second equation on Page 69
2. $P(A) = 1 - P(A^c)$

   (see on Page 69, case a.)
3. $P(\cup_{i=1}^{n} A_i) = \sum_{i=1}^{n} (-1)^{i+1} \sum_{i \leq j_1 < \ldots < j_i \leq n} P(\cap_{k=1}^{i} A_{j_k})$

   (see on Page 69, case b., sieve formula)

that have already been presented as axioms of Kolmogorov or rules for calculation in the beginning of this chapter on Page 69 shall be compared with properties for distribution functions on the power set.

Belief function as well as plausibility function represent elements of the family of possible distributions on the power set $\Pi^\star$ by forming the limits of imaginable distributions $F^\star$ (compare equation (3.2)). Thus, properties that have to be valid for belief function and plausibility function are properties that are valid for all elements within $\Pi^\star$ and thus for all distributions on the power set. Hence, one can compare properties of belief function and plausibility function which have been presented in Subsection 3.2.1 with properties for classical probability theory of equation (3.2).

While the first property of equation (3.2) has to be satisfied for $F^\star$ as well (see definition 9), the second property is not generally satisfied for $F^\star$, which shall be illustrated by means of a narrowly mixed example 5.

Uncle U, father F and real Santa Claus R are potential candidates for the role of Santa Claus, so that $\Omega = \{U, F, R\}$ is given. C_1 is 90 percent sure that

Santa Claus has been played by U or R, such that basic probability assignement of 0.9 is given to the set $\{U, R\}$, i.e. $m(\{U, R\}) = 0.9$. Since there is no knowledge about the remaining probability, it is assigned to whole Ω, i.e. $m(\{U, F, R\})$. Against this, in context of classical probability theory, probability of 0.9 would not describe the certainty of C_1's statement, but one would assume that C_1 has total knowlede and thus probability $P(\{U, R\} = 0.9$ would be used. Moreover, in this case the remaining probability of 0.1 would be allocated to the complement of set $\{U, R\}$, i.e. $P(\{F\} = 0.1$, as there is no lack of knowledge included. As the components of $F^\star$, namely basic probability mass assignements, do not satisfy property two of equation (3.2), generally one has to conclude that

$$F^\star(A) \neq 1 - F^\star(A^c).$$

The third property of equation (3.2) is not generally valid as well, because for the belief function ∞-monotonicity defined in equation (6) is required and the plausibility function is alternating of infinite order (see (4)), which both do not exclusively allow for equality as in equation (3.2), but also admitt inequality. Thus, comparison according to these three properties has been shown that distributions on the power set $F^\star$ can be regarded as a more general concept, where notions from classical probability theory represent a special case, namely the case without lack of knowledge.

Moreover, the general nature of analysis on the power set compared to classical probability theory can be shown by the fact that the former addresses probabilities of events composed of singletons, where the latter concernes events that consist of elements on the power set, so that probabilities of several multivalued sets can be calculated. This has been illustrated in the framework of random sets in Subsection 3.1.3, where for instance $P(A^\star\text{: no singleton})$ has been calculated, i.e. the probability that a person is indecisive between several answers.

Now, after the general nature of distributions on the power set $F^\star$ has been described by comparing these notions with those of classical probability theory, considerations will be made how this kind of uncertainty can be involved into results of prediction and how these distributions on the power set could be used as instruments in order to represent the extent of ontologic uncertainty. In the framework of prediction of the confidence of a certain question set $Q^\star$ when a decision has been made, belief functions as well as plausibility function represent important parts as they form lower and upper bound respectively, so that

$$F^\star(Q^\star) = [Bel(Q^\star),\ Pl(Q^\star)].$$

The representation of the confidence in terms of an interval from equation (3.2) allows to make a prediction without imposing further assumptions and includes total ontologic uncertainty. The reason for taking $Bel(A)$ and $Pl(Q)$ as interval limits can be comprehensible by means of initial Example 4, where this framework of predicition will be illustrated in Section 3.3, where all important notions will be summarized and explained within a detailed example. As the difference between the values of the belief function and the plausibility function represents the lack of knowledge, it is obvious that the length of this interval can be interpreted as the extent of underlying ontologic uncertainty.
Conclusion from random set theory made in Subsection 3.1.3 as well as these deductions from DST, will be summarised in the next section.

3.3. Analysis on the power set in case of coarse categorical data under ontologic uncertainty

Some considerations regarding the presence of coarse categorical data under ontologic uncertainty have been made in the course of this section, whereby ideas and notions from random set theory as well as DST have been applied in this context. The notions used in random set theory as well as in DST are quite similar. Dual conceptions like distribution on the power set as well as capacity functional in the framework of random sets are reflected by belief functions and plausibility functions respectively in context of DST as all notions are distributions $\mathcal{P}(\Omega) \rightarrow [0, 1]$ and show same properties and way of calculations respectively (compare definition 6 and 7 to definition 9).
Nevertheless, both areas impress with different aspects. While random set theory exhibits a well-elaborated mathematical foundation, in many cases DST

proceeds in a more practical way and interpretations of the basic notions are central. Therefore, different conclusions concerning the dealing of coarse categorical data under ontologic uncertainty have been made from these two approaches. Thereby deductions from random sets concentrated on the basic dealing with data of that kind, so that the idea of regarding probabilities on the power set of the sample space has been derived. Against this conclusions from DST focused on the situation when there is no ontologic uncertainty anymore, because decisions have been made, and thus one wishes to report results about probabilities of singletons, or at least more precise sets, without making any assumptions. Both approaches, random set theory as well as DST, provided some concepts that can be used to introduce a notion for a power set based theory that is applicable in the framework of coarse categorical data under ontologic uncertainty. All components that belong to this theory have been marked with a star in the previous subsections. Basic findings concerning dealing with coarse categorical data under ontologic uncertainty are summarized in Figure 3.2. Thereby, the first part of this box that concernes the general idea of analysis has been derived from random set theory, whereas the second part which addresses prediction of the confidence of particular questionary sets $Q^\star$, has been concluded from DST.
In order to illustrate conclusively the role of the most important *-notation instruments within an analysis in the framework of ontologic uncertainty, namely $P^\star(A^\star)$, $m^\star(A^\star)$ as well as $F^\star(Q^\star)$, the following example might be helpful.

Example 6 For reasons of simplicity there are only three parties, namely A, B, and C that can be elected. Before election day some respondents are indecisive between several parties.

The example faces a situation of coarse categorical data, whereby ontologic uncertainty is present, because coarse observations are coarse by nature so that it is no single true value available as the corresponding respondent is indecisive. The framework by means of the *-notation suggests an analysis on the power set of $\Omega = \{A, B, C\}$ without the empty set, namely

$$\Omega^\star = \{\{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{A, B\}, \{A, B, C\}\},$$

and regarding every element within $\Omega^\star$ as own outcome that could be possible. In this way the sample space simply has been extended by involving coarse data. Thus, probabilities of events that are composed by sets easily can be calculated, as for instance probability of event $E^\star$:"Being indecisive between at least two parties" ($P^\star(E^\star)$) can be calculated as adding up the corresponding probabilities, so that

$$P^\star(E^\star) = P^\star(\{A,B\}) + P^\star(\{A,C\}) + P^\star(\{A,B,C\}).$$

As coarse observations can be treated as typical outcomes, probabilities $P^\star(E^\star)$ with $E^\star \in \mathcal{P}(\Omega^\star)$ can be regarded as classical probabilities as there are no differences between classical probabilities P and $P^\star$ except of the modified sample space.

Against this, basic conceptions and properties are more general than those of classical probability theory when decisions have been made and one is interested in the confidence for a special query set $Q^\star$ as for instance $Q^\star = \{B,C\}$. The generality results from the fact that uncertainty about probabilities can be included compared to classical probability theory. As the case of categorical data is faced, distribution functions $F^\star(Q^\star)$ do not pursue the main goal of classical probability theory, namely the calculation of probabilities $P(X \leq x)$, but are needed to add up several evidences $m(A^\star)$ with $A^\star \subseteq Q^\star$ of different sets of categorical data and to predict the confidence of a special query set $Q^\star$ in this way. Even if the basic structure of $F^\star(Q^\star)$ equals the one of distribution function of classical probability theory, underlying properties are more general as explained in Subsection 3.2.2 and distribution function from classical probability theory can be regarded as a special case of $F^\star(Q^\star)$. As $F^\star(Q^\star)$ is able to incorporate ontologic uncertainty, it is an element of all distribution functions on the power set $\Pi^\star$, where lower and upper bound are defined by $\overline{F^\star}(Q^\star)$ and $\underline{F^\star}(Q^\star)$, whereby in case of $Q^\star = \{B,C\}$ they can be calculated by

$$\begin{aligned}
\underline{F^\star}(\{B,C\}) &= m(\{B\}) + m(\{C\}) + m(\{B,C\}) \\
\overline{F^\star}(\{B,C\}) &= m(\{B\}) + m(\{C\}) + m(\{A,B\}) \\
&+ m(\{A,C\} + m(\{B,C\}) \\
&+ m(\{A,B,C\}).
\end{aligned}$$

Thus, the confidence of query set $Q^\star = \{B, C\}$ can be predicted in terms of interval

$$F^\star(\{B,C\}) = [\underline{F}^\star(\{B,C\}), \overline{F^\star}(\{B,C\})]$$

and the length of this interval represents the extent of ontologic uncertainty.
Furthermore, a relation between this theory based on the power set and classical probability theory has been established. It has been shown that classical probability theory can be regarded as a special case of the power set based theory, as several methods and properties are generalized in the latter one. Both theories coincide in the case that total knwoledge about the underlying probabilities is available. This relation can also be noted by means of Table 3.3 that compares both conceptions with regard to different aspects.
Random set theory as well as DST are not the only areas that are based on distribution functions on the power set, as for instance the theory of hints as well as the as transferable belief model represent similar approaches. While the transferable belief model provides a framework for considering quantified beliefs that differs from DST according to a few aspects as for instance by distinguishing between the credal and the pignistic level, the theory of hints develops DST as a theory with uncertain arguments. In my opinion in particular it could be worth to look closely at the theory of hints and extend the proposed *-notion by some underlying ideas. For more details concerning the transferable belief modell and theory of hints see Smets and Kennes [1994] and Kohlas and Monney [1995], respectively.
In summary, some approaches for dealing with epistemic and ontologic uncertainty have been explained in Chapter 2 and this chapter by presenting some already used methods of other areas first and applying them in the framework of the corresponding uncertainty in a second step.

	Classical Probabilities	**Probabilities on $\mathcal{P}(\Omega)$**
Sample space	$\Omega = \{\omega_1, ...\omega_n\}$ contains outcomes that are singletons	$\Omega^\star = \mathcal{P}(\Omega) \setminus \emptyset$ contains all subsets of Ω without $\emptyset$ $\rightarrow$ coarse data can be represented as well
Probability	$P : \mathcal{P}(\Omega) \rightarrow [0, 1]$ $A \rightarrow P(A)$ where $A \subseteq \Omega$	$P^\star : \mathcal{P}(\Omega^\star) = \mathcal{P}(\mathcal{P}(\Omega)) \rightarrow [0, 1]$ $A^\star \rightarrow P^\star(A^\star)$ where $A^\star \subseteq \Omega^\star$
Probability function	$p : \Omega \rightarrow [0, 1]$ with $p(\emptyset) = 0$ and $\sum_{\omega \in \Omega} p(\omega) = 1$ where $p(\omega)$ indicates	$m : \mathcal{P}(\Omega) \rightarrow [0, 1]$ with $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$ where $m(A)$ reflects the confidence of statement A
Distribution function	$F : \Omega \rightarrow [0, 1]$ not important in case of categorical data	$F^\star : \mathcal{P}(\Omega) = \Omega^\star \rightarrow [0, 1]$ in order to add up masses of several sets according to criterion of interest (e.g. $A \subseteq Q$ or $A \cap Q \neq \emptyset$)
Fundamental properties	$P(\emptyset) = 0$ $P(\Omega) = 1$ $P(A) = 1 - P(A^c)$ if not A then A^c (total knowledge) sieve formula of Poincaré and Sylvester	$F(\emptyset) = 0$ $F(\Omega) = 1$ $F^\star(A) \neq 1 - F^\star(A^c)$ no knowledge about remaining probability of A^c $\overline{F^\star}(A \cup B) \geq F(A) + F(B) - F(A \cap B)$ $\underline{F}^\star(A \cup B) \leq F(A) + F(B) - F(A \cap B)$
Conditioning	Bayes' rule	Rule of combination of DST
Both cases coincide if...	...there is no lack of knowledge about probabilities $\Rightarrow$ $\Omega^\star = \Omega$ contains singletons only $\Rightarrow$ F(A)=Pl(A)=Bel(A) $\Rightarrow$ all properties of classical probability theory are valid	

Table 3.3.: Comparison of classical probability theory and probabilities on the power set.

Analysis of coarse categorical data under ontologic uncertainty

Let X be a categorical random variable and $\Omega = \{A, B, C, ...\}$ its sample space.
Moreover, let $X^\star$ be a random variable that attains coarse values as well $\Rightarrow \Omega^\star = \{\{A\},\{B\},\{C\},\{A,B\},\{A,C\},\{A,B,C\},...\}{=}\mathcal{P}(\Omega) \setminus \emptyset$.
The star * marks analysis in the framework of ontologic uncertainty.

General Analysis by means of $P^\star$:
As coarse observations can be regarded as own outcomes in the presence of ontologic uncertainty, analysis on the power set is applicable.
Probabilities $P^\star(A^\star)$ on $\Omega^\star$ for $A^\star \subseteq \Omega^\star$:

$$\begin{aligned} P^\star : \Omega^\star \quad &\to [0,1] \\ A^\star \quad &\to P^\star(A^\star), \end{aligned}$$

where $P^\star(\emptyset) = 0$ and $\sum_{A^\star \subseteq \Omega^\star} P^\star(A^\star) = 1$.

Prediction of $F(Q^\star)$ without additional assumptions:
Confidence of a certain questionary set $Q^\star$ shall be predicted by means of $F^\star(Q^\star) \in \Pi^\star$, where $\Pi^\star$ is the family of distributions on the power set. Calculation by adding up basic probability mass assignement $m : \mathcal{P}(\Omega) \to [0,1]$ with $m(\emptyset) = 0$ and $\sum_{A \subseteq \Omega} m(A) = 1$, where $m(A)$ reflects the confidence that exactly can be committed to A.

$$\begin{aligned} F^\star : \Omega^\star \quad &\to [0,1] \\ \underline{F}^\star(Q^\star)^\star \quad &= \quad \sum_{A \subseteq Q} m(A) = \inf\{F^\star(A^\star) | F^\star \in \Pi^\star\} \\ \overline{F^\star}(Q^\star) \quad &= \quad \sum_{A \cap Q \neq \emptyset} m(A) = \sup\{F^\star(A^\star) | F^\star \in \Pi^\star\} \\ \Rightarrow F^\star(Q^\star) \quad &= \quad [\underline{F}^\star(Q^\star), \overline{F^\star}(Q^\star)] \end{aligned}$$

where the length of the interval indicates the extent of ontologic uncertainty.

Figure 3.2.: Summary of the conceptions concerning the *-notation

4. A multinomial logit model based approach under epistemic uncertainty

After having addressed some general approaches for the analysis of coarse data until now, in Chapter 4 and Chapter 5 it will be analysed how one can account for epistemic and ontologic uncertainty within the dependent categorical variable of a regression model. Thereby, the proceeding will be structured in the way that this chapter will concentrate on some general information concerning the basic model and its extension in case of epistemic uncertainty, whereas Chapter 5 will investigate how to include ontologic uncertainty.

In this chapter some general considerations concerning the basic model will be made first. In the situation of a categorical response variable a multinomial logit model is appropriate, whose foundations will be explained and whose applicability will be shortly discussed for the case which is addressed here. Afterwards this basic model will be extended by accounting for epistemic uncertainty. For this purpose some comments on the data generating process in the presence of epistemic uncertainty will be made first. Subsequently, it will be considered how this epistemic uncertainty can be incorporated into a multinomial logit model, where the model without covariates will represent the starting point (model 1), which will be extended by including two covariates (model 2) in a second step. In this context it will not only be explained how the likelihoods can be derived for model 1 and 2 respectively, but also the problem of identifiability as well as possible solutions by assumptions like CAR or by partial identification will be investigated. Furthermore, an alternative imputation based approach will be sketched.

4.1. Multinomial logit model in the precise case

As considerations concerning extensions of the multinomial logit model will be made, it is reasonable to summarize the basic idea and formal notation of the classical multinomial logit model that concernes the precise case first. In this connection I want to justify why the multinomial logit model is appropriate for the purpose that is pursued here as well as show the limits that result by using this model. In this section it is mainly referred to Fahrmeir et al. [2007, p. 238-241].

The multinomial logit model is generally used if the dependent variable Y represents a characteristic that is of nominal scale, which means that the categories of Y either can not be ordered or at least that the underlying order has no relevance. Thus, as for instance familiy status or political party preference show categorical charactericstics that are of nominal scale, possible fields of application for the multinomial logit model are survey statistics or psephology. As coarsening of the dependent variable will be of peculiar interest as described in the introduction to this chapter and generally categorical data are focused in this thesis, the multinomial logit model is suitable. Nevertheless, the fact that the multinomial logit model requires response variables that are of nominal scale shows the limit of the analysis by means of this model as at this point categories that imply an order can not be analysed. But in order to maintain a clear scope of this thesis, only the multinomial logit model and thus only coarsened categorical dependent variables of nominal scale will be addressed.

In the following the underlying category is denoted by r and there are c categories such that $Y_i \in \{1, ..., c\}$. The probability of occurrence for category $r = 1, ..., c-1$ for given covariates $\mathbf{x}_i$ is determined by

$$P(Y_i = r|\mathbf{x}_i) = \pi_{ir} = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{c-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}. \tag{4.1}$$

As all probabilities add up to one, the corresponding probability for reference category c can be calculated by

$$P(Y_i = c|\mathbf{x}_i) = \pi_{ic} = 1 - \pi_{i1} - ... - \pi_{ic-1} = \frac{1}{1 + \sum_{s=1}^{c-1} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_s)}.$$

This procedure corresponds to involving the constraint that the beta coefficient of the last category is zero, i.e. $\boldsymbol{\beta}_c^T = (0, ..., 0)$, because $\exp(\mathbf{x}^T\boldsymbol{\beta}_c) = \exp(0) = 1$. In order to ensure identifiability, in the basic model it is important to involve a constraint of that kind. Alternatively one can choose any other category as reference category or choose a symmetric type of constraint by requesting that the sum of all category specific coefficients $\boldsymbol{\beta}_s$ amounts to zero, i.e. $\sum_{s=1}^{c} \boldsymbol{\beta}_s^T = (0, ..., 0)^T$, where the underlying parameters have to be interpreted as deviation from the mean response (Tutz [2000, p. 163-164]).
Solving equation (4.1) for the linear predictor $\eta = \mathbf{x}_i^T\boldsymbol{\beta}_r$, one obtains

$$\log \frac{\pi_{ir}}{\pi_{ic}} = \mathbf{x}_i^T\boldsymbol{\beta}_r,\ r = 1, ..., c, \tag{4.2}$$

i.e. the logarithmised chance, or

$$\frac{\pi_{ir}}{\pi_{ic}} = \exp(\mathbf{x}_i^T\boldsymbol{\beta}_r),\ r = 1, ..., c, \frac{\pi_{ir}}{\pi_{ic}} = \exp(\mathbf{x}_i^T\boldsymbol{\beta}_r),\ r = 1, ..., c, \tag{4.3}$$

i.e. the relative chance of category r and reference category c. Consequently, the exponential of the coefficients of the multinomial logit model expresses how the chance for category r compared to the reference category changes if the value of a particular covariate x_j is increased by one unit in the case of metric covariates or if x_j is taken instead of reference category x_J in the case of categorical covariates. Thereby one has to be attentive as an increase of this chance does not mean that the chance of the corresponding category r compared to other categories apart from the reference leads to an increase as well. Equations (4.2) and (4.3) show that every category r $(r = 1, ..., c - 1)$ exhibits its own linear predictor η_{ir} as well as its own coefficent.
The described interpretation as well as the representation of the linear predictor $\eta = \mathbf{x}_i^T\boldsymbol{\beta}_r$ in terms of equation (4.2) and (4.3) is reminiscent of the logit model that is described by

$$\log \frac{\pi}{1-\pi} = \mathbf{x_i}^T\boldsymbol{\beta} \quad \text{or} \quad \pi = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T\boldsymbol{\beta})}.$$

The logit model concernes a binary dependent variable Y such that the corresponding interpretation of the coefficients refers to the chance of the occurrence of Y compared to the non-occurence. Thus, the logit model can be regarded as a special case of the multinomial logit model, namely the one that exhibits two categories only. At many points later on the logit model will form the basic model, as the graphical illustration is simplified by considering two categories only and transferring the proposed extended models to cases of more categories will be obvious.

4.2. The data generating process

For persuing the goal of constructing a model that is able to deal with a dependent categorical variable that attains values that are coarsened induced by epistemic uncertainty, first a quite simple model will be introduced that relies on the *iid* assumption and does not include any covariates. In a second step this model will be extended by regarding a model that is based on the commonly known multinomial logit model and thus incoporates covariates. While within the *iid* model the probability of category “A” and of category “B” is the same for all individuals, the model with covariates implies probabilities that depend on the corresponding covariates x_{i1} and x_{i2} of individual i. Hence, different data generating processes (DGP) are needed, which will be described in more detail next.

As in the majority of cases real datasets that contain coarsened observations do not give any indication of the true values of these coarse observations, it has been decided to base the analysis on simulated data. The categorical variable that will be considered has two true categories only, so that the formal description of the multinomial logit model will equal the special case of a logit model (see equation (4.4)). Hence, in this case the requested data should contain a categorical variable Y with two categories “A” and “B”, variable Y_{coarse} which reflects the observed values, namely “A”, “B” or “A XOR B”, and in case of regarding the model with covariates the data has to enclose variables X as well.

The parameters that have been involved within the data generating processes can be infered from Table 4.1, where the processes itself will be described in more detail in the following.

Because of the *iid* assumption within the model without covariates, in this

General parameters (both models)	number of observations per dataset: n=10000 number of datasets : $M = 100$ number of categories of Y: $c = 2$
Parameters for DGP of *iid*- model	$\pi_A = 0.67$ $\Rightarrow \pi_B = 0.33$
Parameters for DGP of model with covariates	number of covariates: $p = 2$ $X_1 \sim Po(\texttt{lambda} = \texttt{3})$, $X_2 \sim \mathcal{N}(\texttt{mean} = \texttt{0},\ \texttt{sd} = \texttt{2})$ $\beta_{A0} = -0.3$, $\beta_{A1} = 0.6$, $\beta_{A2} = 1.5$

Table 4.1.: Parameters of data generating processes under epistemic uncertainty.

model the probability of category "A" and of category "B" is the same for all individuals i $(i = 1, ..., n)$, i.e. $\pi_{iA} = \pi_A$ and $\pi_{iB} = \pi_B$. Here category "A" has been sampled with probability 0.67, as this value roughly equals the mean probability of occurence of category "A" that will result in the model with covariates. This choice reflects the intention to have two comparable models. Consequently category "B" will be sampled with probability $1 - \pi_A = 0.33$. As one is interested in involving the coarsened values of Y instead of its true precise values, data additionaly have to contain variable Y_{coarse}. In order to generate Y_{coarse}, different values of $q_1 = P(Y_{coarse} = \text{“(A XOR B)”} | Y = \text{“A”})$ and $q_2 = P(Y_{coarse} = \text{“A XOR B”} | Y = \text{“B”})$ have been applied (notation see page 9). For this purpose q_1 and q_2 varied between 0.1 and 0.9 by increments of 0.1 ($\rightarrow q_1 = 0.1, 0.2, 0.3, ..., 0.9,\ q_2 = 0.1, 0.2, 0.3, ..., 0.9$) and all 81 combinations of q_1 and q_2 were used. Thus, 81 Y_{coarse} variables of dimension $n = 10000$ were created by sampling category "A XOR B" with probability q_1 (resp. q_2) in case that the true category of Y is "A" (resp. "B"). Thus, given true category "A" and given true category "B" those categories are observed in a precise way with conditional probability $(1 - q_1)$ and $(1 - q_2)$ respectively. The values of the true variable Y, the values of these 81 observed variables Y_{coarse} as well as one further variable $Y_{coarse82}$, which will be needed in Subsection 4.5 in context of partial identification, are depicted in Table 4.2 and complete the simulated dataset that will be used in context of the model without covariates.

In order to be able to obtain more convincing results and to calculate evaluating measures as the median relative bias or the variance of that bias, not only one dataset of that kind, but 100 ones have been simulated, which have

Y	Ycoarse1	Ycoarse2	$\cdots$	Ycoarse82
B	A XOR B	B	$\cdots$	B
A	A	A	$\cdots$	A XOR B
B	B	B	$\cdots$	A XOR B
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A	A	A	$\cdots$	A XOR B

Table 4.2.: Structure of simulated datasets that are used for the model without covariates.

10000 observations respectively. These datasets differ by renewed sampling of the true categories of Y, such that different Y_{coarse} can result as well.

As in the model with covariates the probabilities of occurence $P(Y_i = r|\boldsymbol{x_i}) = \pi_{ir}$ of category $r = A, B$ represent probabilities conditional on the values of the covariates X, in this context covariates are generated first. In order to be able to gain insight into the consequences of discrete as well as continuous covariates, one Poisson distributed variable X_1 with λ equal to three ($X_1 \sim Po(3)$) and one normal distributed variable X_2 with a mean of zero and a standard deviation of two ($X_2 \sim \mathcal{N}(0, 4)$) have been used. By means of these covariates probabilities of occurence could be calculated according to equation (4.1), which in turn where used in order to determine the true categories of Y. Thereby, for every individual i specific probabilities of occurence exist that rely on the corresponding values x_{i1} and x_{i2}, where the true categories of variable Y, namely category "A" and "B", have been sampled with probability π_{iA} and $\pi_{iB} = 1 - \pi_{iA}$, respectively.

In this way data as sketched in the first three columns of Table 4.3 result, which rely on the following precise multinomial logit model with two covariates and two possible true categories of the dependent variable Y:

$$P(Y_i = A|\boldsymbol{x_i}) = \pi_{iA} = \frac{\exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}.$$

As one can notice from Section 4.1, the multinomial logit model is described by $c - 1$ category specific coefficients. Nevertheless, here only one $\boldsymbol{\beta}$, namely $\boldsymbol{\beta_A} = (\beta_{A0}, \beta_{A1}, \beta_{A2})$, has to be estimated as there are two categories ($c =$

$2 \Rightarrow c - 1 = 1$) only and category B will be the reference category with corresponding probabilities

$$P(Y_i = B|\boldsymbol{x_i}) = \pi_{iB} = \frac{1}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}.$$

This precise model shall be extended in Subsection 4.3 by including a coarsened variable Y_{coarse} instead of Y and the goal of this section will be to investigate under which conditions the model that relies on the coarsened values is able to generate appropriate estimators $\hat{\boldsymbol{\beta}}$ for parameter $\boldsymbol{\beta}$. Different kinds of coarsening processes are determined by the same procedure as described in context of the data on which the model without covariables will be based on, such that again 82 variables Y_{coarse} of dimension $n = 10000$ are generated. An extraction of the data that will be used in the framework of the model with covariates is illustrated by Table 4.3.

As in the case of data in context of the model without covariates, not only one dataset, but 100 datasets of that kind are generated. These datasets differ by renewed sampling of X_1 and X_2, such that different values of Y and Y_{coarse} can yield.

From both datasets it is expected that one is able to analyse the change of $\hat{\theta}$, the estimator of the parameter of interest θ, when extended models that incorporate the imprecision of the observed data Y_{coarse}, formed by different coarsening mechanisms, are applied. But before, these models that extend the commonly known precise models have to be introduced.

Y	X1	X2	Ycoarse1	Ycoarse2	...	Ycoarse82
A	7	0.2456983	A	A XOR B	···	A
A	1	1.7636975	A	A	···	A
A	5	0.8042766	A	A	···	A XOR B
B	2	0.5196141	B	B	···	B
⋮	⋮	⋮	⋮	⋮	⋮	⋮
B	3	-5.134471	B	A XOR B	···	A XOR B
A	1	-0.7402479	A	A	···	A
A	2	2.448102	A	A	···	A XOR B

Table 4.3.: Structure of simulated datasets that are used for the model with covariates.

4.3. The model without and with covariates

In the course of this chapter analyses are based on two models, namely the model without and and the model with covariates, where these models will be called “model 1” and “model 2” respectively. After having addressed the corresponding optimization problems of these two models in Subsection 4.3.1, some implications of the estimation problem (see Subsection 4.3.2) as well as some general first results will be shown (see Subsection 4.3.3). Thereby, all analyses will be shown for model 1 first, where analogous investigations will be made for model 2 afterwards. While this section considers estimation more generally, Section 4.4 and 4.5 focus on the case of CAR and the way how partial identification can be applied.

4.3.1. The optimization problem

The *iid* assumption that will be implied by model 1 simplifies a lot in the sense that the probability of occurence of a particular category r $(r = 1, ..., c)$ is the same for all individuals i $(i = 1, ..., n)$. In order to keep things simple, the case will be addressed that the dependent variable Y exhibits two true categories only, namely “A” and “B”, which can be observed in a coarsened way with probabilities $q_1 = P(\mathcal{Y} = (A\ XOR\ B)|Y = A)$ and $q_2 = P(\mathcal{Y} = (A\ XOR\ B)|Y = B)$, respectively (see page 9). Thus the corresponding likelihood is described by

$$
\begin{aligned}
L(q_1, q_2, \pi_{iA}) = & \prod_{\mathcal{Y}_i} P(\mathcal{Y}_i = \mathfrak{y}) \\
= & \prod_{i:\mathcal{Y}_i = A} P(\mathcal{Y}_i = A|Y_i = A) \cdot \pi_{iA} \cdot \prod_{i:\mathcal{Y}_i = B} P(\mathcal{Y}_i = B|Y_i = B) \cdot (1 - \pi_{iA}) \cdot \\
& \prod_{i:\mathcal{Y}_i = (A\ XOR\ B)} P(\mathcal{Y}_i = (A\ XOR\ B)|Y_i = A) \cdot \pi_{iA} + P(\mathcal{Y}_i = (A\ XOR\ B)|Y_i = B) \cdot (1 - \pi_{iA}) \\
= & \prod_{i:\mathcal{Y}_i = A} (1 - q_1) \cdot \pi_{iA} \cdot \prod_{i:\mathcal{Y}_i = B} (1 - q_2) \cdot (1 - \pi_{iA}) \cdot \prod_{i:\mathcal{Y}_i = (A\ XOR\ B)} q_1 \cdot \pi_{iA} + q_2 \cdot (1 - \pi_{iA}).
\end{aligned}
\tag{4.4}
$$

and the log-likelihood by

$$\begin{aligned}
l(q, \pi_{iA}) &= \ln(L(q, \pi_{iA})) = \sum_{i:\mathcal{Y}_i = A} (\ln(1-q_1) + \ln(\pi_{iA})) + \\
&\sum_{i:\mathcal{Y}_i = B} (\ln(1-q_2) + \ln(1-\pi_{iA})) + \\
&\sum_{i:\mathcal{Y}_i = AB} \ln(q_1 \pi_{iA} + q_2(1-\pi_{iA})) \\
&\overset{\text{iid}}{=} n_A \cdot [\ln(1-q_1) + \ln(\pi_A)] + n_B \cdot [\ln(1-q_2) + \ln(1-\pi_A)] \\
&+ n_{AB} \cdot [\ln(q_1\pi_A + q_2(1-\pi_A))],
\end{aligned}$$

where n_A, n_B, and n_{AB} denote the number of cases in which categories "A", "B" and "A XOR B" have been observed, respectively.

By optimization of the loglikelihood the following three estimation equations result:

$$\begin{aligned}
\text{I.)}\ \frac{\partial}{\partial q_1} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}\pi_A - \frac{n_A}{1-q_1} \overset{!}{=} 0 \\
\text{II.)}\ \frac{\partial}{\partial q_2} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(1-\pi_A) - \frac{n_B}{1-q_2} \overset{!}{=} 0 \qquad (4.5) \\
\text{III.)}\ \frac{\partial}{\partial \pi_A} &= \frac{n_{AB}}{q_1\pi_A + q_2(1-\pi_A)}(q_1 - q_2) + \frac{n_A}{\pi_A} - \frac{n_B}{1-\pi_A} \overset{!}{=} 0.
\end{aligned}$$

In the precise case only the third estimation equation with $q_1 = q_2 = 0$ is of importance and hence one obtains

$$\begin{aligned}
\frac{n_A}{\pi_A} - \frac{n_B}{1-\pi_A} &\overset{!}{=} 0 \\
\Leftrightarrow \pi_A &= \frac{n_A}{n_A + n_B},
\end{aligned}$$

which means that the probability of occurrence of category "A" is estimated by the relative proportion of values "A".

But intuitive unique results as in the precise case do not result if one focuses on the imprecise case ($q_1 \neq 0$, $q_2 \neq 0$) when additionaly to π_A, parameters q_1 and q_2 are unknown and requested to be estimated. This can be explained by the relation between q_1 and q_2 which is induced by the fact that both probabilities characterize the mechanism of creating coarsened observations "A XOR B".

Thus for instance for a given medium number of coarsened values and a rather high value of q_1, the value of coarsening parameter q_2 is restricted to be relative small compared to q_1 as only a fixed medium number of "A XOR B" values has been generated. This relation will be shown analytically later on in context of partial identification in Section 4.5. The dependence between q_1 and q_2 is also reflected in the first two estimation equations of (4.5) and hence there are only two independent equations, such that two parameters can be estimated only. Therefore, one has to think about identifying restrictions according to q_1 and q_2, as it will be done in the course of this section.

Apart from this model without covariates, a model with covariates will be addressed which differs by the fact that the probabilities of occurence depend on the values of the covariates in the way as described in (4.1). Thus, one can replace this dependence structure for every π_{iA} of the likelihood of equation (4.4) and one obtains the following likelihood for the model with two covariates which refers to a data structure that is ordered in the sense that individuals $1, ..., N_1$ exhibit observed value "A", value "B" is observed for individuals $N_1 + 1, ..., N_2$ and individuals $N_2 + 1, ..., N$ are the ones that show coarsened values "A XOR B":

$$\begin{aligned} L(q_1, q_2, \beta_A) &= \prod_{i=1}^{N_1} (1 - q_1) \frac{\exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})} \\ & \prod_{i=N_1+1}^{N_2} (1 - q_2) \frac{1}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})} \\ & \prod_{i=N_2+1}^{N} (q_1 \frac{\exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})} + \\ & \qquad \frac{q_2}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}). \end{aligned}$$

The last part is dropped if one addresses the precise case with $q_1 = q_2 = 0$, such that the likelihood for the multinomial logit model (see Section 4.1) follows. By optimization of the general log-likelihood

$$\begin{aligned} l(q_1, q_2, \beta_A) \quad = \quad & \sum_{i=1}^{N_1} \ln((1-q_1)\frac{\exp(\beta_{A0}+x_{i1}\beta_{A1}+x_{i2}\beta_{A2})}{1+\exp(\beta_{A0}+x_{i1}\beta_{A1}+x_{i2}\beta_{A2})}) + \\ & \sum_{i=N_1+1}^{N_2} \ln((1-q_2)\frac{1}{1+\exp(\beta_{A0}+x_{i1}\beta_{A1}+x_{i2}\beta_{A2})}) + \\ & \sum_{i=N_2+1}^{N} \ln(q_1\frac{\exp(\beta_{A0}+x_{i1}\beta_{A1}+x_{i2}\beta_{A2})}{1+\exp(\beta_{A0}+x_{i1}\beta_{A1}+x_{i2}\beta_{A2})} + \\ & \qquad \frac{q_2}{1+\exp(\beta_{A0}+x_{i1}\beta_{A1}+x_{i2}\beta_{A2})}), \end{aligned}$$

one can draw the same conclusions as in the context of the model without covariates, as the dependence of q_1 and q_2 and their underlying estimation equations leads to an identification problem in the sense that additional assumptions concerning q_1 and q_2 are necessary.

4.3.2. Implications of the optimization problem

Generally, one can distinguish between the following four cases in the framework of dealing with the identification problem arising in model 1 and model 2 (see Subsection 4.3.1):

a) The precise case
b) Known coarsening mechanism
c) Assuming CAR
d) Involving ideas of partial identification.

These cases are also illustrated by means of Figure 4.1.

Case *a*) and *b*) address cases in which the coarsening is known, because the precise case can be regarded as a special case of case *b*) as in this case coarsening parameters are known to be $q_1 = q_2 = 0$, i.e. that there is no coarsening.

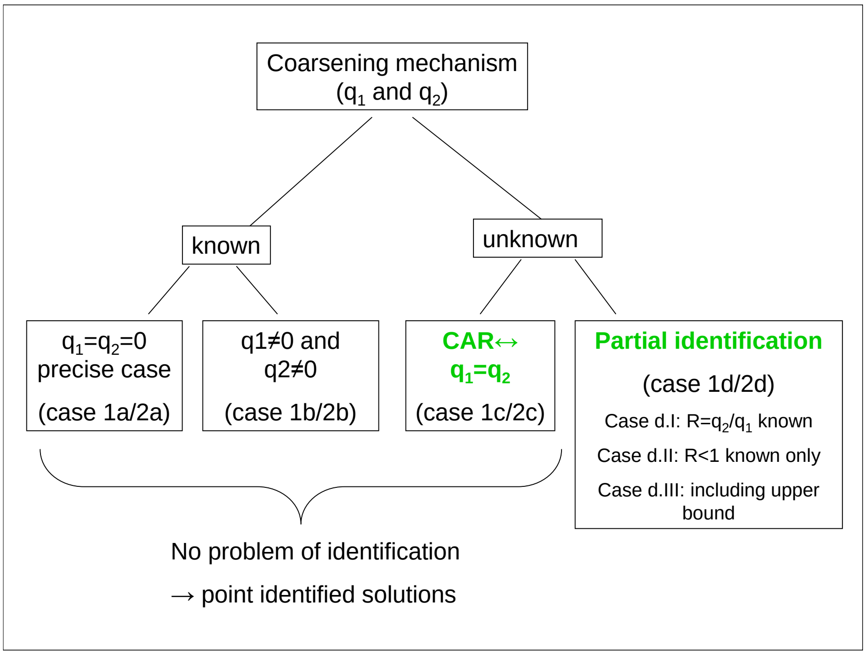


Figure 4.1.: Cases a) to d) that are implied by a known and an unknown coarsening mechanism.

Thereby, these both cases can be regarded as rather easy cases, because there are as many independent estimation equations as parameters that have to be estimated (model 1: one estimation equation and only π_A has to be estimated, model 2: three estimation equations and β_{A0}, β_{A1} and β_{A2} are the parameters that have to be estimated).
When the coarsening is unknown, parameters q_1 and q_2 have to be estimated, which leads to dependent estimation equations and the need for identifying restrictions. These restrictions can be derived by including the CAR assumption or aspects of partial identification and thus case c) and case d) represent methods that show possibilities how one can deal in situations when the coarsening is unknown.

4.3.3. Some general first results

Some particular circumstances of these four cases will be focused for both models in a general way in the framework of this subsection first. More details concerning the analysis in case of imposing the assumption of CAR as well as investigations that rely on aspects of partial identification will be considered in Section 4.4 and 4.5 respectively.

In the context of the following analyses it is of peculiar interest to evaluate the empirical estimators that result by optimization of the loglikelihood of model 1 and model 2. Thereby the relative empirical bias (or short: relative bias) will play an important part, which is calculated by

$$Bias_{\mathrm{rel}} = \frac{\hat{\theta} - \theta}{|\theta|},$$

where θ is the parameter of interest and $\hat{\theta}$ its empirical estimator. The relative bias indicates the percentual deviation of the empirical estimator from the true estimator and its preference consists of the independence of the magnitude of the true parameter. The calculation of the relative bias will be based on the 100 datasets of the simulated data proposed in Section 4.2, where it will be started with some first results of model 1.

Model 1:

Parameter π_A represents the parameter of main interest in model 1, so that the relative bias for this estimator will be considered for particular simple situations of cases a) to d), where Figure 4.2 showes the corresponding results. As the results concerning the realtive bias of these four cases are depicted within the same plot, comparisons can easily be made.

The underlying estimator of the first boxplot that concernes the precise case in the sense that precise variable Y instead of Y_{coarse} has been involved into the model, can be regarded as unbiased with a median relative bias of `1.228e-05` and a very small standard deviation of `2.625542e-07`.

From the second boxplot it can be noted that a minimal underestimation with a median relative bias of `-0.0004090` and an increasement of standard deviation to `0.004756693` result when the liklihood is optimized which here implies a constant known coarsening of $q_1 = 0.2$ and $q_2 = 0.4$ (i.e. $Y_{coarse13}$ has been implied). This additional uncertainty in case b) compared to case a) can be ascribed to the underlying sampling process in the framework of the coarsening.

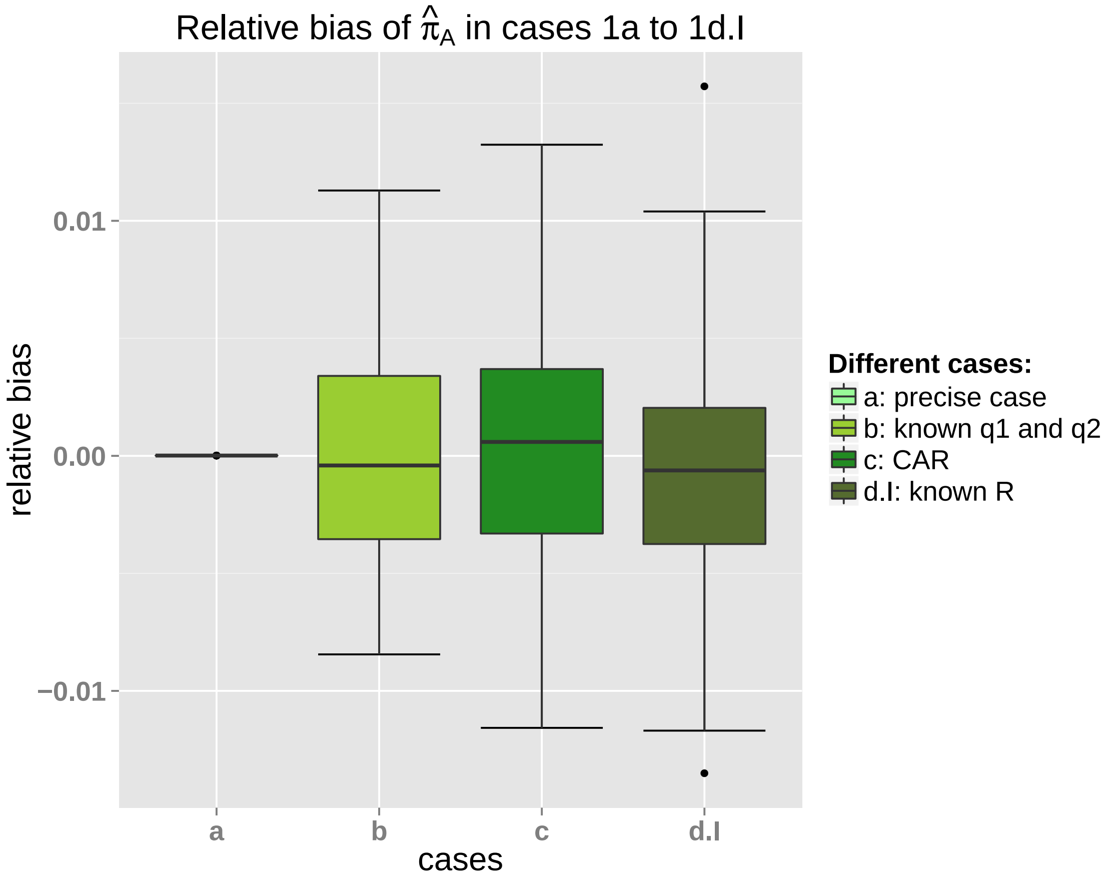


Figure 4.2.: Case 1a-1d.I: Boxplot showing the relative bias of $\hat{\pi}_A$.

In the sense the true values $q(\mathfrak{y}|y)$ are not exactly reflected by the sampled data.

Under the assumption of CAR (case c)) estimator $\hat{\pi}_A$ is point identified, as it will be explained in Subsection 4.4.1. Here this CAR assumption is involved within the estimation and data are underlying that are described by coarsening parameters $q_1 = q_2 = 0.3$ ($Y_{coarse21}$ has been implied), so that CAR is valid indeed. Comparing the boxplot of case c) to the one of case b), the standard deviation of `0.005218006` is only neglegible higher and the median bias of `0.0005926` is very small. Thus, the empirical estimator under CAR seems to be quite good. Nevertheless, here the case has been considered that CAR has rightly been assumed, but in practice deviations of CAR are expected. Thus it could be interesting to analyse the impact concerning the relative bias if

the CAR assumption is incorporated within the estimation, but it is not valid indeed. This will be addressed in Subsection 4.4.2
The last boxplot depicts a very easy situation in the framework of partial identification, which is denoted by case *d.I* in Figure 4.2. In this case the relation of the coarsening parameters $R = \frac{q_1}{q_2}$ is known and rightly involved into the estimation. In this way one parameter less has to be estimated and thus there is no identification problem anymore. For illustrating this situation, analysis has been based on Y_{coarse82} (see Subsection 4.2) of the 100 simulated datasets, where these observed variables have been coarsened by using a coarsening process that is described by $q_1 = 0.4$ and $q_2 = 0.32$. Thus $R = \frac{q_2}{q_1} = \frac{0.32}{0.4} = 0.8$ is valid. The last boxplot shows the result concerning the realtive bias of $\hat{\pi}_A$, if this value of $R = 0.8$ is known and applied within the estimation. One can note that minimum and maximum relative bias of `-0.0135100` and `0.0157200` as well as median and standard deviation of the relative bias of `-0.0006198` and `0.004991033` are very similar to the corresponding values of the case that applies the CAR assumption. This is not surprising, as CAR is simply a special case of $q_2 = R \cdot q_1$ with $R = 1$. Although one is able to make more flexible assumptions by means of this generalization, in many situations – especially in areas where there has been sparse research – the problem of having no idea about the relation of q_1 and q_2 and thus about the true value R endures. This problem will further be addressed in Section 4.5.
Here in all situations true coarsening probabilities that are of rather small or middle amount are involved ($q(\mathfrak{y}|y)$ values between 0.2 and 0.4) in order to obtain results that are fairly comparable. Generally, one expects a comparable greater relative bias, the greater the amount of coarse observations, i.e. the greater the values of the true coarsening parameter $q(\mathfrak{y}|y)$, because the implied uncertainty is increased under these circumstances.
Analgous investigations will be made for model 2 next.

Model 2:

In the framework of model 2 estimation of parameters β_0, β_1 and β_2 is of main interest. The boxplots in Figure 4.3 concern the relative bias of the corresponding estimators for particular simple circumstances of cases a) to d).

Thereby, the same situations are addressed as in model 1 (e.g. truely assumed value of $R = 0.8$).
For all four cases it can be noted that the median relative bias of $\hat{\beta}_0$ and its standard deviation are relatively large compared to median relative bias and standard deviation of $\hat{\beta}_1$ and $\hat{\beta}_2$ (e.g. median relative bias of $\hat{\beta}_0$ for cases a, b, c, d, resp.: `-0.02260`, `-0.02895`, `0.03795`, `0.01297`, median relative bias of $\hat{\beta}_1$ for cases a, b, c, d, resp: `-0.0020950`, `-0.005533`, `0.004951`, `0.006704`). The fact that estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ are nearly unbiased is in accordance with some further results that will be shown later on.
Moreover, comparing the boxplot of the different cases one can conclude that the median relative bias of the precise case as well as its standard deviation is lower than in case b) and case c). Furthermore, one can note from comparing case b), c) and d) that similar median relative bias as well as a similar standard deviations of the relative bias result. These results correspond to the investigation of the relative bias of $\hat{\pi}_A$ in the model without covariates and explanations from there can be referred to this case.
As here the simplest circumstances of case c) and d) have been regarded, namely that CAR is truely assumed and that the true value of R is known, further considerations concerning cases of wrong or less information have to be made concerning the assumtpion of CAR in Section 4.4 and partial identification in Section 4.5.

4.4. The assumption of “coarsening at random”

Firstly, the general meaning of the assumption of CAR in the context of the modelling approach as well as its inclusion into the optimization problem will be regarded in Subsection 4.4.1. Afterwards the problem of wrongly assuming CAR will be investigated for model 1 and model 2 in Subsection 4.4.2.

4.4.1. The optimization problem under the assumption of CAR

As already extensively explained in Section 2.1, “coarsening at random” (CAR) means that the conditional probability of obtaining a particular coarsened

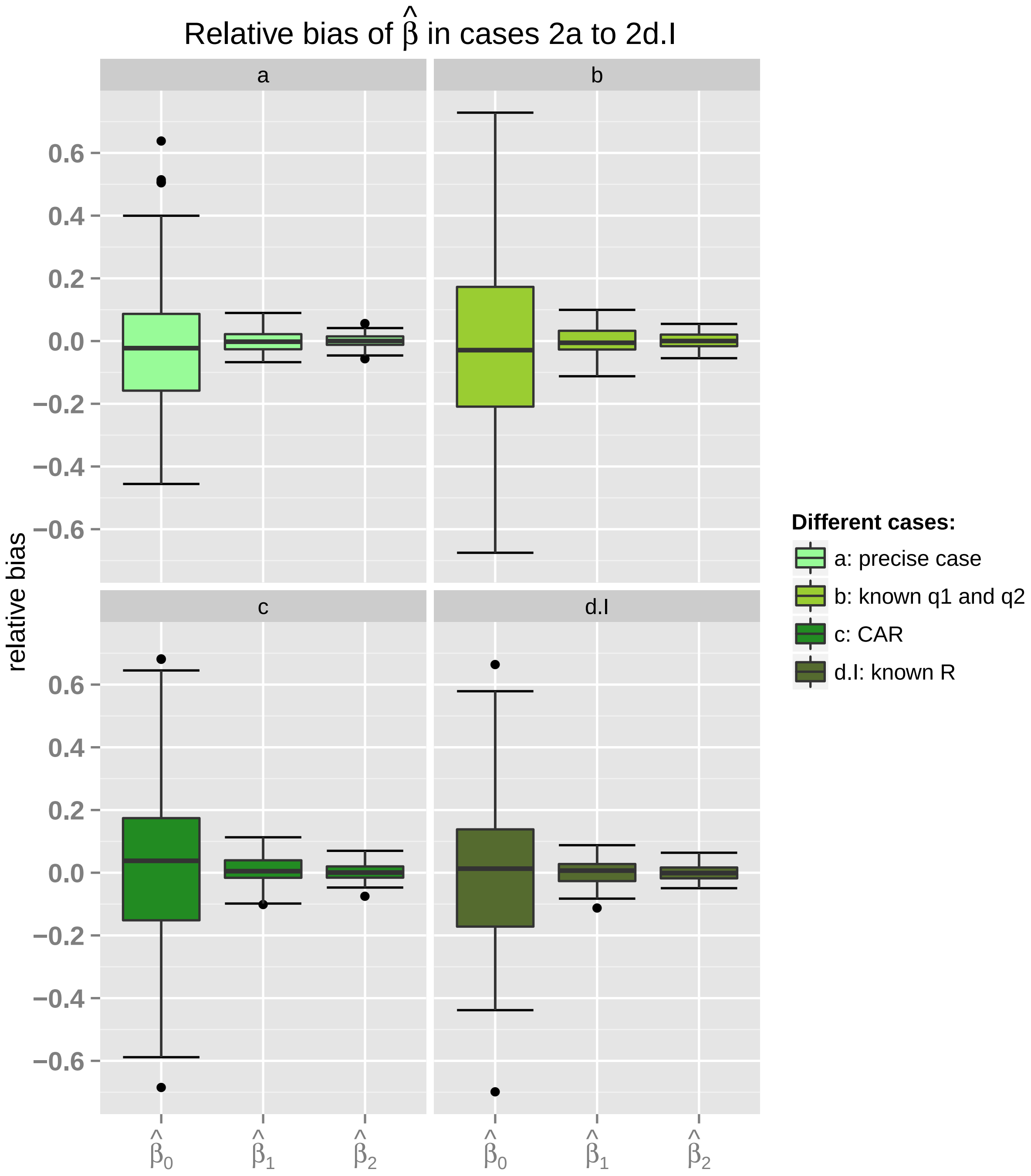


Figure 4.3.: Case 2a-1d.I: Boxplot showing the relative bias of the β estimators.

observation for given true values (i.e. $P(\mathcal{Y} = \mathfrak{y}|Y = y)$) takes the same value for all true values y that correspond to the observed data. Referring to the simple case of regarding two true categories only, namely “A” and “B”, under CAR the probability that determines the coarsening mechanism $P(\mathcal{Y} = (A\ XOR\ B)|Y = y)$ is constant, no matter which true categories are underlying as long as they are consistent with the observed values, i.e. the true value y can only be “A” or “B”, but not “C”. Thus, both categories are coarsened to “A XOR B” with the same probability and hence $q_1 = P(\mathcal{Y} = (A\ XOR\ B)|Y = A) = P(\mathcal{Y} = (A\ XOR\ B)|Y = B) = q_2$.
For instance, if some categories are socially undesirable this assumption does not seem to be justified, as respondents that belong to these categories might answer in a coarsened way more probable. Therefore, it is important to reflect about the justification of the assumption of CAR before involving it into the model.

In case that CAR can be regarded as a reasonable assumption, the corresponding relation of $q_1 = q_2 = q$ leads to a solution of the identification problem already addressed in Subsection 4.3, because one parameter less, namely only q instead of q_1 and q_2, has to be estimated (notatation q: see page 2). Although model 1 and model 2 differ by the fact that parameter π_A in model 2 is not constant, the structure of the corresponding likelihoods is similar, wherefore it will only be analysed how estimators can be derived in model 1 involving the CAR assumption $q_1 = q_2 = q$.

For this purpose, one can find an estimator for $q(\mathfrak{y}|y)$ by summing up the first and the second estimation equation of Equation 4.5, so that

$$\begin{aligned}
I.) + II.) : \quad & \frac{n_{AB}}{q \cdot \pi_A + q \cdot (1 - \pi_A)} \cdot \pi_A - \frac{n_A}{1-q} + \\
& \frac{n_{AB}}{q \cdot \pi_A + q \cdot (1 - \pi_A)} \cdot (1 - \pi_A) - \frac{n_B}{1-q} = 0 \\
& \Leftrightarrow \frac{n_{AB}}{q} - \frac{n_A}{1-q} - \frac{n_B}{1-q} = 0 \\
& \Leftrightarrow \frac{n_{AB}}{q} = \frac{n_A + n_B}{1-q} \\
& \Leftrightarrow n_{AB} \cdot (1-q) = q \cdot (n_A + n_B) \Leftrightarrow \frac{1-q}{q} = \frac{n_B + n_B}{n_{AB}} \\
& \Rightarrow \hat{q} = \frac{1}{\frac{n_A + n_B}{n_{AB}} + 1} = \frac{n_{AB}}{n_A + n_B + n_{AB}}
\end{aligned}$$

results. Moreover an empirical estimator for the probability of occurence of category “A”, namely $\hat{\pi}_A$, can be derived by simply solving the third estimation equation of Equation (4.5) for π_A, in which the first part is dropped, as the factor $(q_1 - q_2)$ is equal to zero in case of CAR:

$$\begin{aligned}
\text{from III.)} \quad & \frac{n_A}{\pi_A} - \frac{n_B}{1-\pi_A} \Leftrightarrow n_A(1 - \pi_A) = n_B \pi_A \\
& \Leftrightarrow n_A - n_A \cdot \pi_A = n_B \cdot \pi_A \Leftrightarrow n_A = \pi_A \cdot (n_A + n_B) \\
& \Rightarrow \hat{\pi}_A = \frac{n_A}{n_A + n_B}.
\end{aligned}$$

Therefore, the probability that describes the coarsening mechanism $q(\mathfrak{y}|y)$ is estimated by the proportion of observed “A XOR B” values and the empirical estimator for the probability of occurence of category “A” equals the proportion of observed “A” values if all coarsened values are ignored.

Both results can be illustrated by means of two CAR situations which are depicted in Table 4.4 and 4.5. Thereby, it has to be noted that the estimation of parameters like $q(\mathfrak{y}|y)$ and π_A based on six (situation 1) or nine (situation 2) observations can hardly be justified. But as these examples only serve as illustration of the estimators’ interpretation, this should not be too problematic.

In case 1 (case 2) for given true category “B” the estimated probability of observing the coarsened value “A XOR B” is $\hat{q}_2 = \frac{2}{4} = \frac{1}{2}$ $(\hat{q}_2 = \frac{2}{6} = \frac{1}{3})$ and

Y	A	A	B	B	B	B
$\mathcal{Y}$	A	A XOR B	B	A XOR B	A XOR B	B

Table 4.4.: Example 1: CAR.

Y	A	A	A	B	B	B	B	B	B
$\mathcal{Y}$	A	A	A XOR B	A XOR B	B	B	B	B	A XOR B

Table 4.5.: Example 2: CAR.

thus equals the estimated probability $\hat{q}_1$ of observing "A XOR B" for given true category "A". Therefore, the probability of observing coarsened observation "A XOR B" is constant no matter which true category is underlying and hence CAR is valid. Under CAR the empirical estimator of Equation (4.6) is applicable according to which the probability of observing coarsened value "A XOR B" for given true values can be estimated by $\hat{q}{=}\frac{n_{AB}}{n_A+n_B+n_{AB}}$ and hence $\hat{q} = \frac{3}{6} = \frac{1}{2}$ in case 1 ($\frac{3}{9} = \frac{1}{3}$ in case 2). Thus, it could be illustrated that under CAR there is no difference between conditioning only on the true "A" values, only on the true "B" values or conditioning on all values when one is interested in the estimated probability of observing coarsened values, i.e. $\hat{q}_1 = \hat{q}_2 = \hat{q}$. In this way the reason for conditioning on all observations $(n_A + n_B + n_{AB})$ and regarding all coarsened observations (n_{AB}) in the empirical estimator for $q(\mathbf{y}|y)$ can be comprehensible.

Moreover, the estimator for the probability of occurence of category "A", namely $\hat{\pi}_A$, can be illustrated by means of this example. As under CAR the conditional probability of a particular coarse observations for given true values is constant no matter which true category is underlying, this coarsening by random leads to the conclusion that these coarse observations can be simply ignored. Thus, in case 1 one obtains an empirical estimator of $\hat{\pi}_A = \frac{n_A}{n_A+n_B} = \frac{1}{3}$ (in case 2 of $\hat{\pi}_A = \frac{2}{6} = \frac{1}{3}$), which equals the empirical estimator calculated by means of the true Y values. Although the derived estimator for π_A under CAR looks equally as the estimator in the precise case (see equation (4.6)), it has to be noted that these estimators differ in the sense that the former one drops all coarsened observations, while the later does not exhibit coarsened values at all ($q_1 = q_2 = 0$).

Solving estimation equations (see Equation (4.5)) for parameter $q(\mathbf{y}|y)$ and

π_A under CAR has shown that the first two estimation equations have been helpful for determining an estimator for $q(\mathfrak{y}|y)$, where the third equation was applied in order to obtain $\hat{\pi}_A$. Thus, the first two equations are dependent and are only able to estimate one parameter, namely q, instead of two parameters q_1 and q_2. That is why unique empirical estimators could be found under the assumption of CAR. Analogous findings result when CAR assumption is involved within the analysis of deriving emprical estimators for $q(\mathfrak{y}|y)$, β_0, β_1, and β_2 in the framework of model 2.

With regard to rightly assuming CAR in case of coarsening parameters $q_1 = q_2 = 0.3$ the relative bias for the parameters of interest, namely $\hat{\pi}_A$ for model 1 and $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ for model 2, already has been shown in Figure 4.2 and 4.3. For the same setting boxplots depicting the relative bias of $\hat{q}$ can be found in the Appendix A.

But in practice analysts often do not know, if they face a situation of CAR or not. Referring to this, further considerations and investigations concerning the consequences that are connected with wrongly assuming CAR will be made in the next subsection.

4.4.2. Analysis if CAR is wrongly assumed

Here it will be of peculiar interest how the relative bias increases if one involves the assumption of CAR into the estimation in case that it actually is not valid.

Model 1:

This can be investigated by regarding Figure 4.4, which showes the relative bias of $\hat{\pi}_A$ if the true coarsening mechanism is characterized by some different combinations of q_1 and q_2 and the CAR assumption is still involved into the estimation by setting $q_1 = q_2 = q$. Thereby, the median of the relative bias based on all 100 simulated datasets is depicted. One can note that the relative median bias increases the more one deviates from the case of CAR. Thus, the combination of q_1 and q_2 values that differs most from the case of CAR, namely $q_1 = 0.9$ and $q_2 = 0.1$, causes the largest median relative bias of `0.7246411`.

It can be noted from the extent of the $q_1 = 0.1$ and $q_2 = 0.9$ combination underlying median relative bias of `0.4151588`, that one does not face a symmetric problem. This can be explained by the fact that in the simulated data the number of true “A” values exceeds the number of true “B” values and thus for instance a situation I.) with $q_1 = P(\mathcal{Y} = (A\ XOR\ B)|Y = A) = 0.9$ and $q_2 = P(\mathcal{Y} = (A\ XOR\ B)|Y = B) = 0.1$ absolutely creates more coarsened values as ninety percent of the large number of true “A” values are going to be coarsened, where in situation II.) with $q_1 = P(\mathcal{Y} = (A\ XOR\ B)|Y = A) = 0.9$ and $q_2 = P(\mathcal{Y} = (A\ XOR\ B)|Y = B) = 0.1$ absolutely less “B” values are observed as “A XOR B”. Hence, implied uncertainty is larger in situation I.) and thus the underlying median relative bias is expected to be larger.

Model 2:

Analogous analyses have been made concerning model 2, where the median bias of estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ will be depicted here. As the order of magnitude of the relative bias of $\hat{\beta}_0$ and the one of $\hat{\beta}_1$ and $\hat{\beta}_2$ is quite different, a seperate plot for $\hat{\beta}_0$ has been created.

Figure 4.5 showes that the larger the deviation from CAR, the larger the median relative bias. Thus the maximal median relative bias for $\hat{\beta}_0$ results if the empirical estimator that involves the CAR assumption is applied in case that $q_1 = 0.1$ and $q_2 = 0.9$ (relative bias of: `7.29290594`) as well as $q_1 = 0.9$ and $q_2 = 0.1$ (median relative bias of: `-7.27332235`). Thereby, the situation looks fairly symmetric compared to the one concerning model 1 (see Figure 4.4). This could be reasoned by the fact that here more aspects, as for instance the interaction of the bias of the three parameters, have to be involved compared to model 1 in which the bias of just one parameter, namely $\hat{\pi}_A$, has been investigated.

Against this, one can note in Figure 4.6 that this structure of observing an increasing bias the more one deviates from CAR, can not be made for the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$. Here the median relative bias is relatively small for all combinations of q_1 and q_2 and even if there is a quite strong deviation from the CAR assumption, the calculation of empirical estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ by involving the CAR assumption seems to be justified.

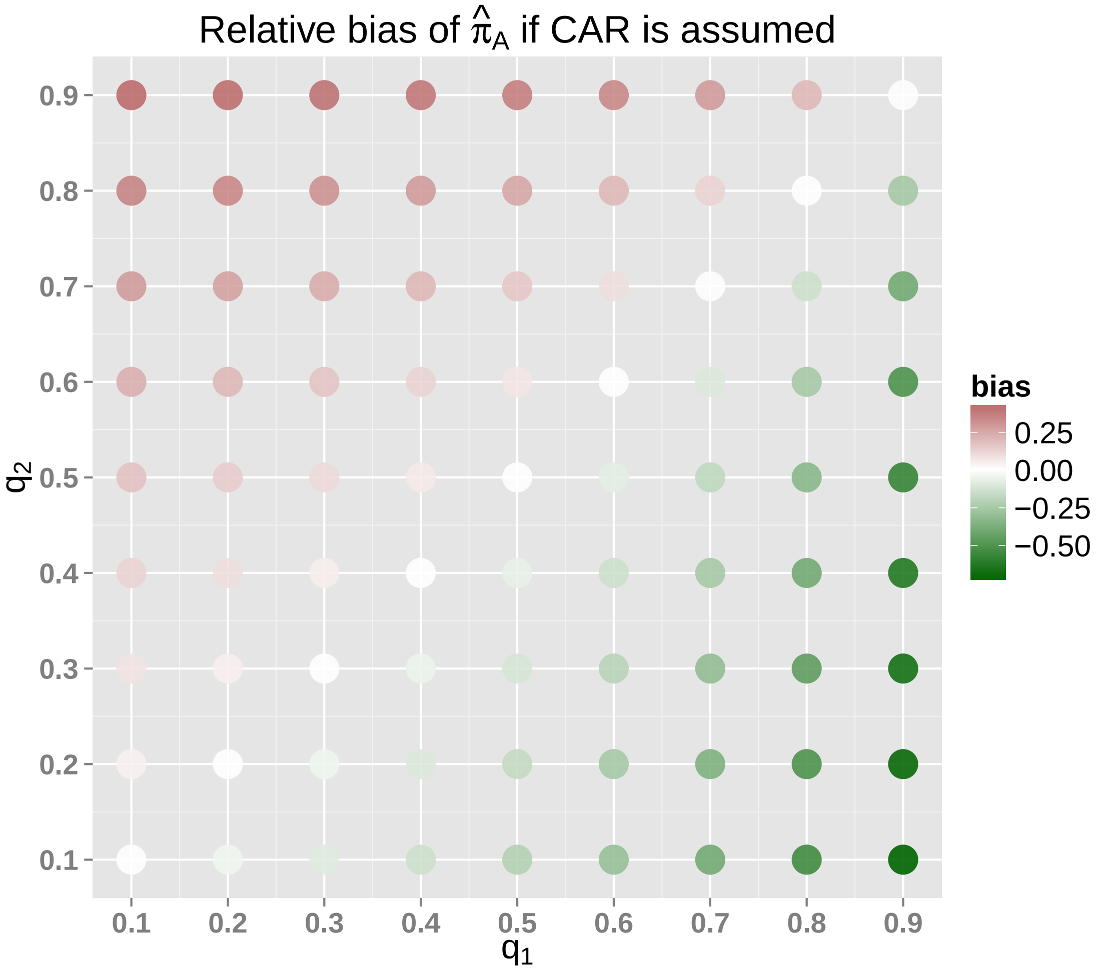


Figure 4.4.: Case 1c: Consequences for the median relative bias of $\hat{\pi}_A$ if there is a deviation of CAR.

In summary, it has been shown that point identified results can be obtained in the case of CAR. Nevertheless, it is of prime importance to check if CAR assumption is justified indeed, as otherwise quite large deviations for emprical estimators $\hat{\pi}$ in model 1 and $\hat{\beta}_0$ in model 2 resulted. This aspect clarifies the necessity to find a solution how one can deal with the identification problem proposed in Section 4.3, in case that CAR assumption has to be neglected. This problem will be addressed in the next section.

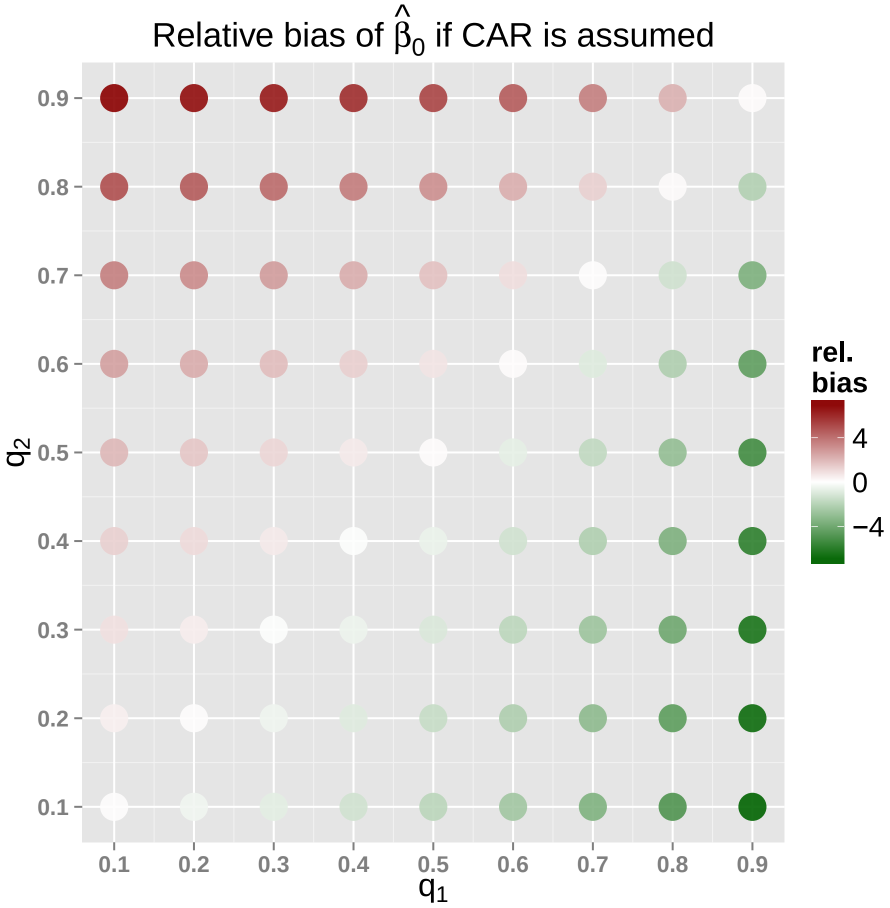


Figure 4.5.: Case 2c: Consequences for the median relative bias of $\hat{\beta}_0$ if there is a deviation of CAR.

4.5. Involving ideas of partial identification

Solving estimation equations of model 1 and 2 leads to an identification problem (see Subsection 4.3) in the sense that a set of possible $\hat{\pi}_A$ in model 1 and possible $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ in model 2 are available for different values of q_1 and q_2 that are determined by the empirical evidence. Investigations can either incorporate the empirical evidence that is described by the traditional assumption of $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$ or further restrict it by the upper bounds that have been derived in Subsection 2.2.6 and do not rely on any contentual assumptions.

On the one hand one pursues the goal to address this identification problem by

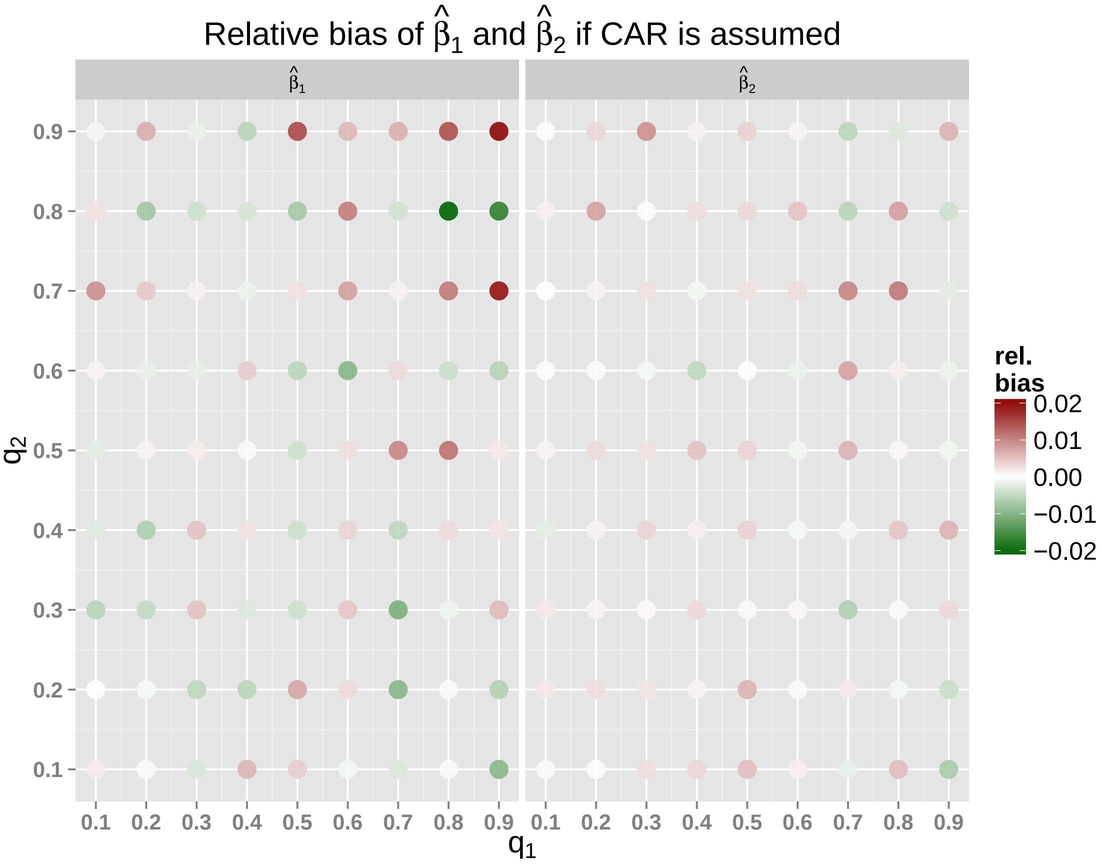


Figure 4.6.: Case 2c: Consequences for the median relative bias of $\hat{\beta}_1$ and $\hat{\beta}_2$ if there is a deviation of CAR.

strong assumptions for q_1 and q_2 in order to obtain precise empirical estimators $\hat{\pi}_A$ or $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. But on the other hand, strong assumptions that are not justified should not be imposed as it has been investigated in the framwork of the CAR assumption in Subsection 4.4.2.

Approaches which face this tradeoff are partial identification and sensitivity analysis that have been proposed in Section 2.2. As it will be of peculiar interest how those sets of possible empirical estimators can be reduced if different assumptions for q_1 and q_2 are involved, the direction of analysis equals the procedure of partial identification which thus seems to be more appropriate here (see for instance Figure 2.5). Therefore, some ideas that have been made in the framework of partial identification will be implied in context of the extended multinomial logit model now.

While the inclusion of the relation in case of a rightly assumed R (case *d.I.*) has already been considered in Subsection 4.3.3, here two further suggestions that have been made in Section 2.2 will be addressed, namely

d.I relaxed assumption concerning R, $R < 1$

d.II implying an upper bound.

Both ideas are denoted by d) according to the notation of cases presented in Subsection 4.3.2 and will be regarded more in detail in Subsection 4.5.1 and 4.5.2 respectively.

4.5.1. Implying a relaxed assumption concerning R

In many cases one does not know the exact value of R as it has been assumed in case *d.I.* (see Subsection 4.3.3). Because of this, it is interesting to investigate the impact of implying a factor R that is roughly known only, as for instance that $R < 1$ (case 1d.II). This means that $q_1 > q_2$ and thus category "A" is coarsened to "A XOR B" more probable than category "B". Again results with regard to model 1 will be shown first, where those of model 2 will follow afterwards.

Model 1:

For analysing this question, Figure 4.7 is helpful. Here empirical estimators $\hat{\pi}_A$ have been calculated for different assumptions of $R = 0.1,\ 0.2,\ 0.3,\ \ldots\ , 0.8,$ $0.9,\ 1$, where these points have been used for a interpolation. Thereby, cases $R = 0$ and $R = 1$ are equal to the assumed precise case and the assumed CAR case, respectively. In order to avoid confusion, results from the first ten simulated datasets and thus only ten lines are depicted, where it has been noted that results from the remaining datasets are similar. The green dashed line marks the true value of R and the true π_A that has been involved within the data generating process. By implying this vague assumption of $R < 1$, one obtains a set of possible estimators of π_A, namely $\hat{\pi}_A \in [0.64, 0.78]$ if the mean start ($R = 0$) and end ($R = 1$) values of the 100 lines are involved. As without this assumption the length of this interval is almost doubled, namely $\hat{\pi}_A \in [\frac{n_A}{n_A+n_B+n_{AB}}, \frac{n_A+n_{AB}}{n_A+n_B+n_{AB}}] = [0.40, 0.77]$, implying this assumption can be

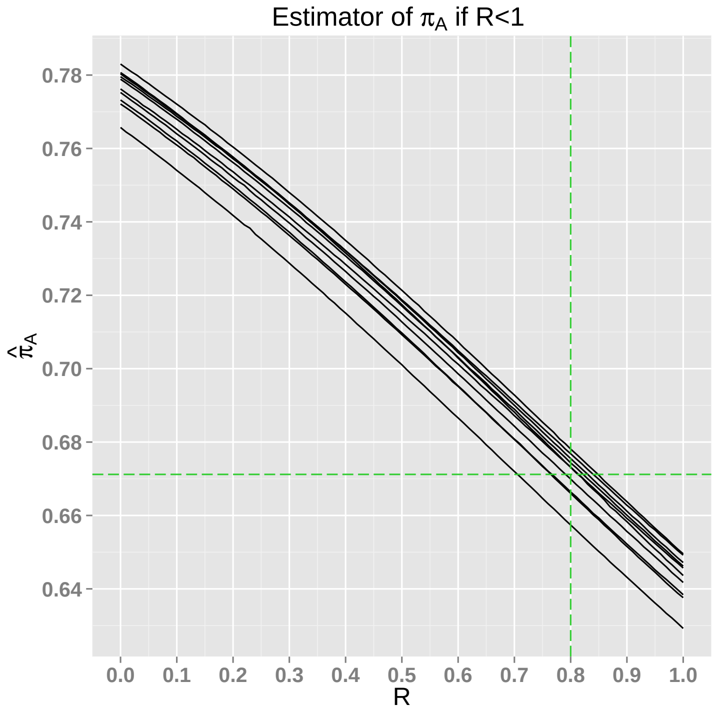


Figure 4.7.: Case 1d.II: Resulting estimators for π for different values of assumed $R < 1$ (interpolation).

very useful, if it is satisfied indeed.

Regarding the boxplots in Figure 4.8 apart the range of the median relative bias of `-0.05577917` and `0.17631857`, which can also be derived from the finding that $\hat{\pi}_A \in [0.64, 0.78]$, one can note that under each assumed R a similar standard deviation of about `0.005` results and thus the only impact of assuming a wrong R are biased estimators the more one deviates from true R as there is no additional uncertainty implied.

Model 2:

Until now, analyses in context of weaker assumptions concerning R have been considered for model 1 only, wherefore some results for model 2 will be presented next. Thereby, it has to be noted, that in model 2 three (β_0, β_1, β_2) instead of only one parameter (π_A) as in model 1 have to be estimated under different assumptions of q_1 and q_2.

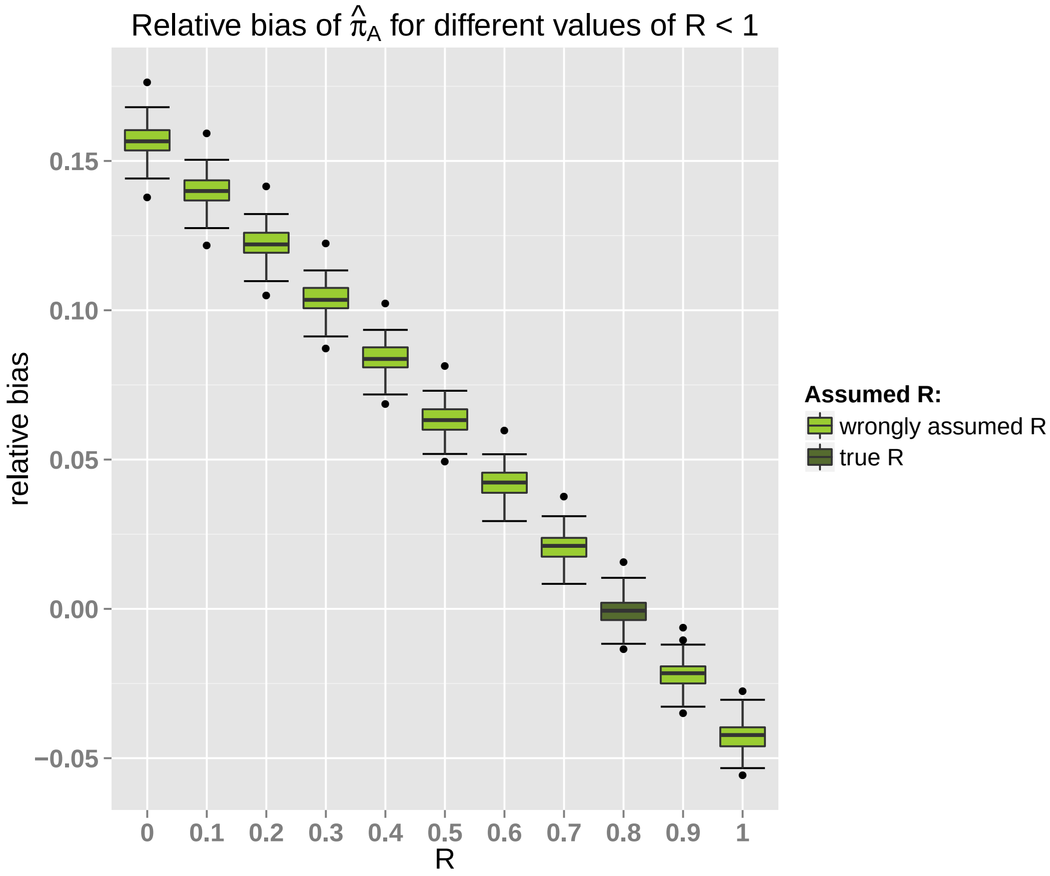


Figure 4.8.: Case 1d.II: Boxplots showing the relative bias for different values of assumed $R < 1$.

The lines in Figure 4.9 show how the estimators change for different values of $R < 1$, where the true R that is equal to 0.8 roughly corresponds to the true beta values, which is illustrated by the green dashed line. As in model 1, corresponding parameters have been estimated for different assumed values of $R = 0,\ 0.1,\ 0.2, ...,\ 0.9,\ 1$ only and lines are obtained by interpolations. Furthermore, results from the first ten datasets are depicted only again. One can note that implying the assumption $R < 1$ leads to $\hat{\beta}_0 \in [-0.43, 0.79]$, $\hat{\beta}_1 \in [0.35, 0.60]$ and $\hat{\beta}_2 \in [0.85, 1.50]$, where the bounds represent the mean estimator for the starting and ending point of the lines.

From the estimator specific intervals under the assumption of $R < 1$ one can note the comparably large bias of $\hat{\beta}_0$, wherefore seperated plots for this and the other estimators have been created in order to investigate the relative bias.

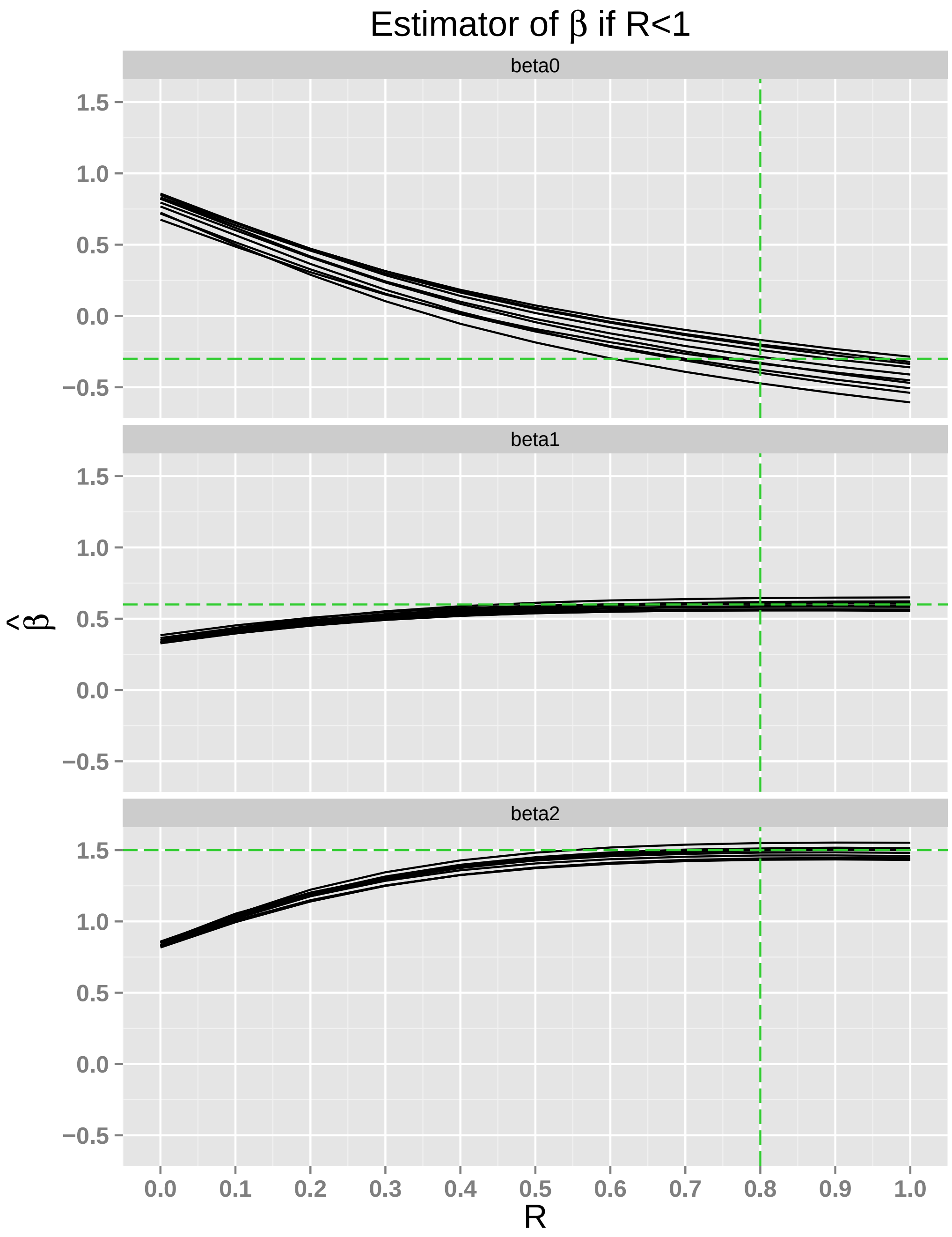


Figure 4.9.: Case 2.d.II:Resulting estimators for β_0, β_1 and β_2 for different values of assumed $R < 1$ (interpolation).

From Figure 4.10 one can see that large deviations of the true $R = 0.8$ can lead to an substantial relative bias of $\hat{\beta}_0$. Thus, under the assumption of $R = 0$ a maximal median relative bias of `3.64` results, which means that β_0 is estimated more than three times larger than it actually is. Nevertheless, the standard deviation increases minimally from `0.20` under assumed $R = 0$ to `0.25` for assumed $R = 1$.

In Figure 4.11 an essentially smaller median relative bias is depicted for estimators $\hat{\beta}_1$ and $\hat{\beta}_2$, where maximal values of `-0.4159303` and `-0.4319532` respectively are attained implying the assumption of $R = 0$. Even if the relative bias of $\hat{\beta}_1$ and $\hat{\beta}_2$ is quite similar, a slightly smaller standard error of about `0.024` can be noted for the relative bias of $\hat{\beta}_2$ compared to the one of $\hat{\beta}_1$, which is around `0.025`. In both cases the standard error increases minimally with increasing assumed values of R.

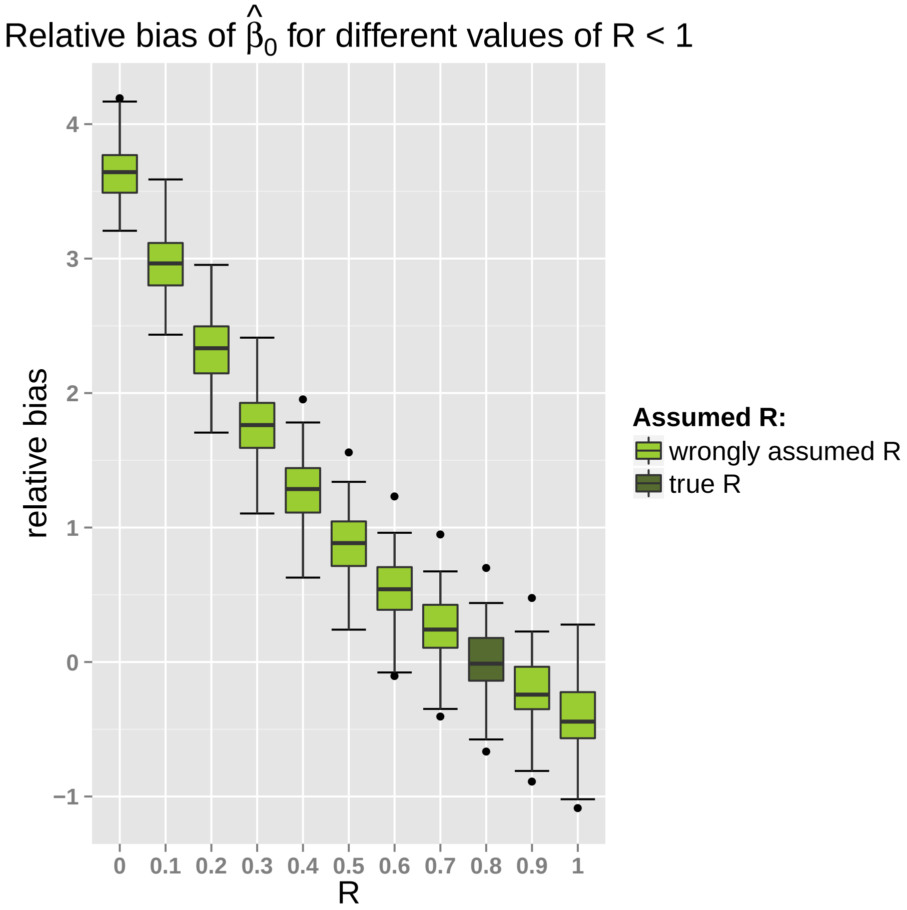


Figure 4.10.: Case 2.d.II: Boxplots showing the relative bias of $\hat{\beta}_0$ for different values of assumed $R < 1$.

In order to gain an insight into the importance of the assumption $R < 1$, it

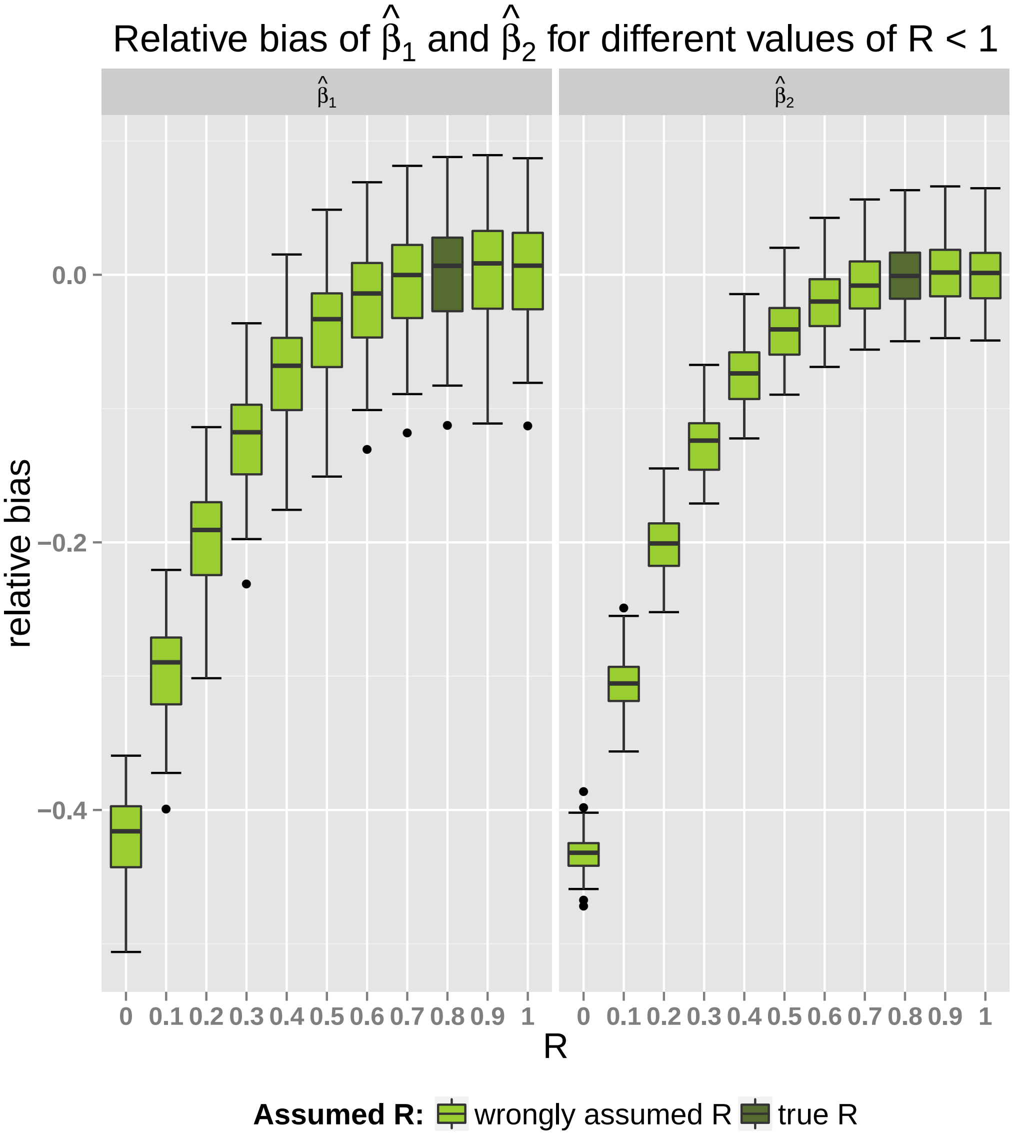


Figure 4.11.: Case 2.d.II: Boxplots showing the relative bias of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{q}$ for different values of $R < 1$.

is reasonable to compare the relative bias obtained under this assumption and the one without any assumption, i.e. including the empirical evidence only. In the latter case the relative bias of the empirical estimators $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ has been calculated that have been obtained by optimization of $\boldsymbol{\beta}$ for different given combinations of q_1 and q_2 that are between 0 and 1 (i.e. here the classical empirical evidence is used instead of the upper bound → see Subsection 4.5.2). The median relative bias in each of the 100 simulated datasets has been calculated for both cases, respectively, and the corresponding boxplots are depicted in Figure 4.12 and 4.13. Because of the different scales, results for the relative bias of $\hat{\beta}_0$ as well as of $\hat{\beta}_1$ and $\hat{\beta}_2$ are shown in seperated plots. Thereby, one can note that incorporating the assumption $R < 1$ leads to a smaller bias compared to applying the empirical evidence only, and in case of $\hat{\beta}_0$ additionaly a substantially decreased standard deviation results. Thus, if one is able to make assumptions as $R < 1$, one should include them here, even if they are of a very vague nature.

4.5.2. Implying an upper bound

In Section 2.2 upper bounds for q_1 and q_2 could be derived which do not require any contentual assumption and simply are possible because of the information that is generated by some precise observations (details see Subsection 2.2.6). According to this, these upper bounds can also be regarded as a kind of empirical evidence that further restricts the commonly used empirical evidence, namely $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$. This raises the question (case 1d.III.) to which extent the upper bounds are able to restrict and outperform the feasible solutions of $\hat{\pi}_A$ that result from incorporating $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$ only. In Subsection 2.2.6 the upper bounds $\overline{q_1}$ and $\overline{q_2}$ have been proposed as probabilities, which here have to be estimated. If one approximates the containing probabilities by implying their corresponding empirical estimators, one obtains upper bounds for estimators $\hat{q}_1$ and $\hat{q}_2$ as follows:

$$\overline{\hat{q}_1} = \frac{\frac{n_{AB}}{n}}{\frac{n_A}{n} + \frac{n_{AB}}{n}} = \frac{n_{AB}}{n_A + n_{AB}} \quad \text{and}$$
$$\overline{\hat{q}_2} = \frac{\frac{n_{AB}}{n}}{\frac{n_B}{n} + \frac{n_{AB}}{n}} = \frac{n_{AB}}{n_B + n_{AB}}.$$

Thereby, n_A, n_B and n_{AB} denote the number of cases being observed as “A”, “B” and “A XOR B” respectively and $n = n_A + n_B + n_{A\ XOR\ B}$. In this

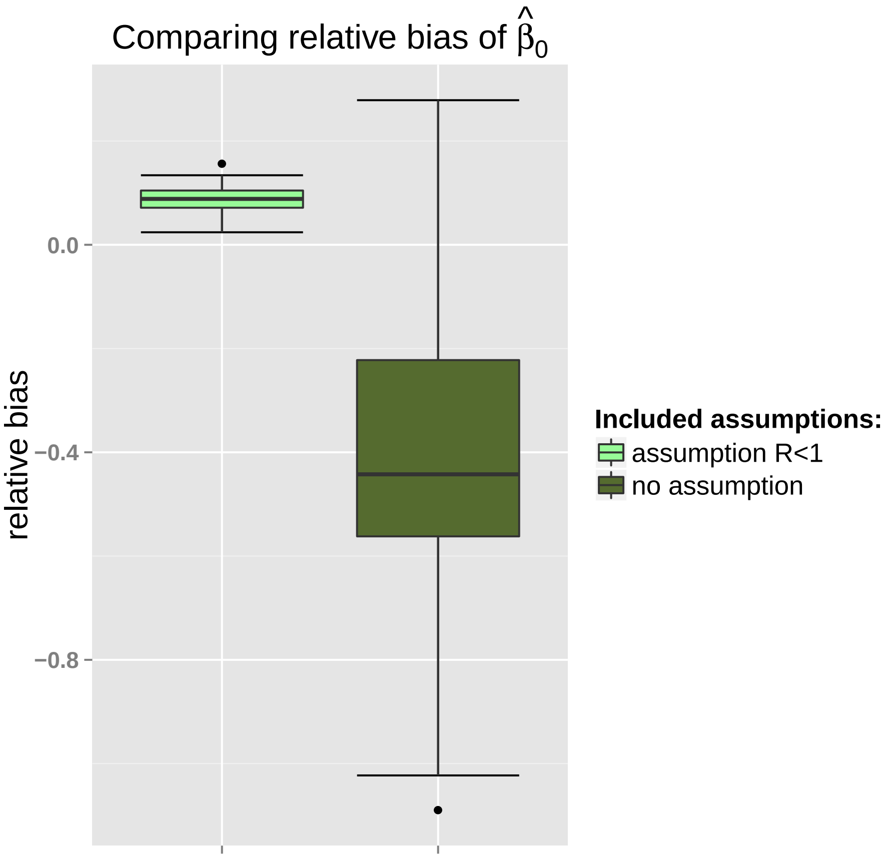


Figure 4.12.: Comparison of relative bias of $\hat{\beta}_0$ if 1.) including assumption R<1 and 2.) empirical evidence is used only.

case applying these estimators can be considered as unproblematic, as the underlying sample space of $n = 10000$ can be regarded as quite large. Against this, in situations that show a rather small n, it is important to account for the kind of uncertainty that is induced by sampling variability. As here a proportion, namely the proportion of the coarse observations conditioned on the observation of category "A" and the observation of category "A XOR B", is considered, the confidence interval for the estimator of the proportion (for large sample sizes and medium sizes of proportions) can be applied (Kauermann, G. and Küchenhoff, H. [2011, p. 31]). Here an estimator for the upper bound is requested, wherefore the upper bound of the confidence interval has to be

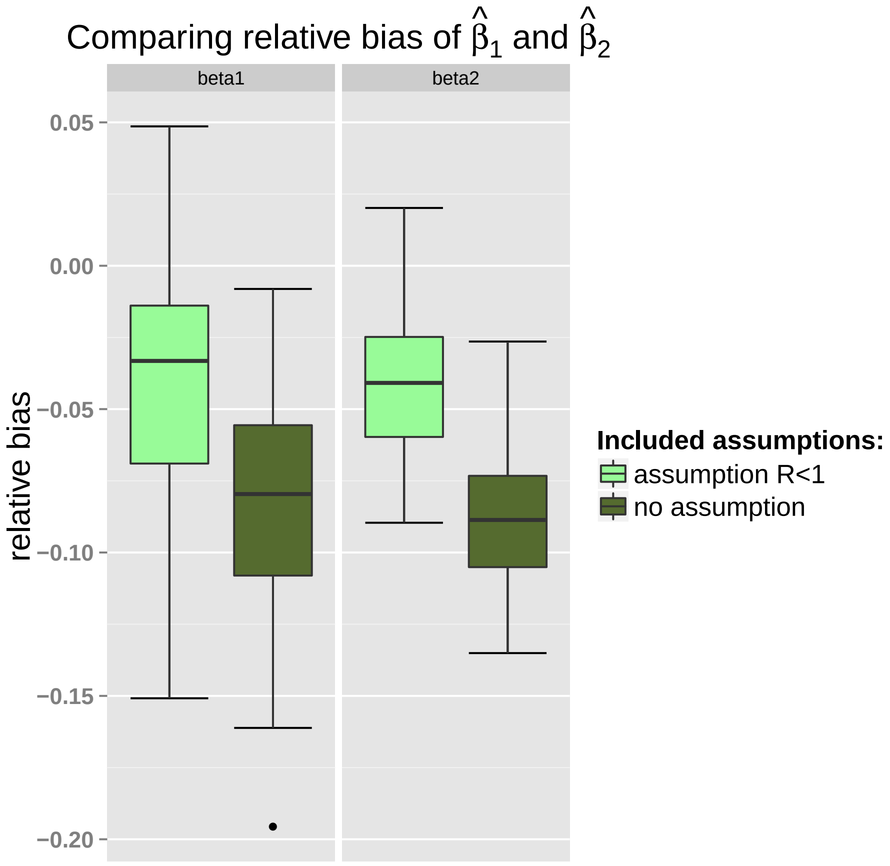


Figure 4.13.: Comparison of relative bias of $\hat{\beta}_1$ and $\hat{\beta}_2$ if 1.) including assumption R<1 and 2.) empirical evidence is used only.

incorporated only, so that modified upper bounds $\overline{\hat{q}_1}^*$ and $\overline{\hat{q}_2}^*$ that account for sampling variability can be determined as

$$\begin{aligned}
\overline{\hat{q}_1}^* &= \overline{\hat{q}_1} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{\hat{q}_1} \cdot (1 - \overline{\hat{q}_1})}{n-1} \cdot \frac{N-n}{N}} \\
\overline{\hat{q}_2}^* &= \overline{\hat{q}_2} + z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\overline{\hat{q}_2} \cdot (1 - \overline{\hat{q}_2})}{n-1} \cdot \frac{N-n}{N}},
\end{aligned}$$

where N is the size of the population and $z_{1-\frac{\alpha}{2}}$ the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution, which can be applied in case of large sample sizes. In case of large sample sizes $\frac{N-n}{n}$ is neglegible.

But incorporating upper bounds only, there are still a lot of useless solutions included. For instance if there is a relatively high proportion of coarsened observations “A XOR B”, estimators of π_A that result from involving pretty small values of q_1 as well as q_2 should not be involved. Thus, a procedure that not only restricts the number of possible solutions of $\hat{\pi}_A$ by incorporating the derived upper bounds is required, but that also accounts for the relation between q_1 and q_2 and thus ensures that the observed number of “A XOR B” can be generated under these values of q_1 and q_2. Accounting for the relation between q_1 and q_2 does not include any contentual assumptions and thus this idea can be considered as a kind of empirical evidence as well. In the context of this relation again empirical estimators are involved that do not account for sampling variability, which does not seem to be a problem for reasons of the large sample size of $n = 10000$.

Thus, the following method has been applied, which is also illustrated by Figure 4.14:

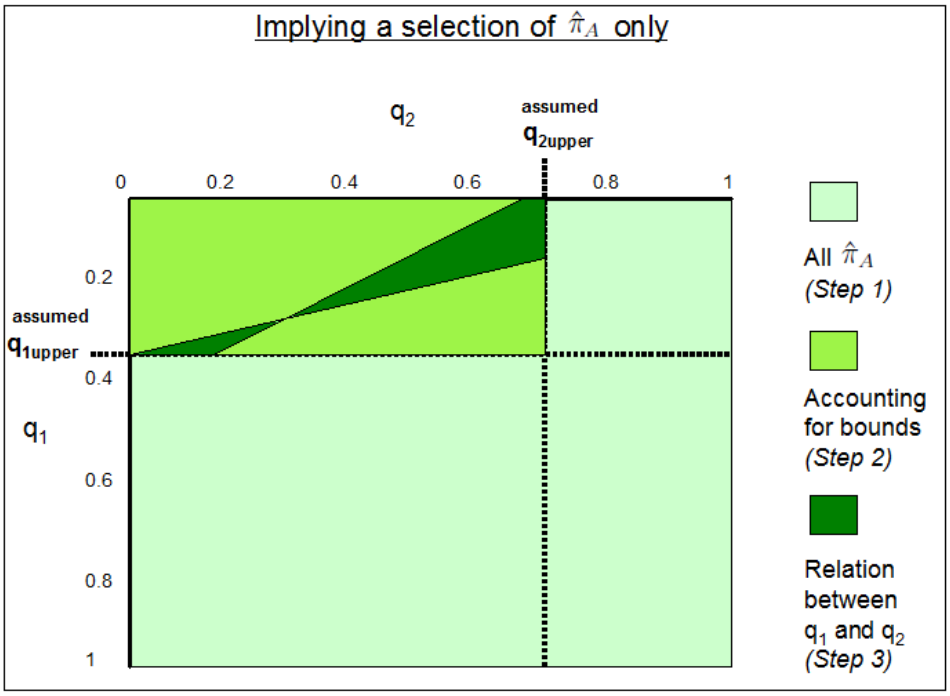

Implying a selection of $\hat{\pi}_A$ only
q2
assumed q2upper
0
0.2
0.4
0.6
0.8
1
0.2
assumed q1upper
0.4
q1
0.6
0.8
1
All $\hat{\pi}_A$ (Step 1)
Accounting for bounds (Step 2)
Relation between q1 and q2 (Step 3)

Figure 4.14.: Case 1.d.III: Graphical depiction of proposed procedure finding a selection of possible $\hat{\pi}_A$.

- *Step 1:* Calculate all possible solutions if $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$, is used only, i.e. optimize loglikelihood by π_A for different given assumed conceivable values of q_1 and $q_2 \Rightarrow$ matrix M_1 that includes all $\hat{\pi}_A$

- *Step 2:* Restrict the solutions from matrix M_1 by only including those whose corresponding coarsening estimators $\hat{q}_1$ and $\hat{q}_2$ values do not exceed their upper bound, i.e. $\hat{q}_1 \leq \frac{n_{AB}}{n_{AB}+n_A}$ and $\hat{q}_2 \leq \frac{n_{AB}}{n_{AB}+n_B} \Rightarrow$ matrix M_2 that includes all $\hat{\pi}_A$ that are valid under this upper bound based condition

- *Step 3:* Select only those entries of matrix M_2 that additionally satisfy the relation between q_1 and q_2 values induced by the law of total probability:

  $$\begin{aligned} P(\mathcal{Y} = AB) &= P(Y = A) \cdot P(\mathcal{Y} = (A \ XOR \ B)|Y = A) \\ &\quad + P(Y = B) \cdot P(\mathcal{Y} = (A \ XOR \ B)|Y = B). \end{aligned}$$

  Replacing probabilities $P(\mathcal{Y} = (A \ XOR \ B))$, $P(\mathcal{Y} = (A \ XOR \ B)|Y = A)$ and $P(\mathcal{Y} = (A \ XOR \ B)|Y = B)$ by their empirical estimators, one obtains:

  $$\frac{n_{AB}}{n} = \pi_A \hat{q}_1 + (1 - \pi_A) \cdot \hat{q}_2.$$

  Solving for $\hat{q}_1$ and $\hat{q}_2$, it follows

  - *Step 3a):*
    $$\hat{q}_1 = \frac{n_{AB} - n \cdot \hat{q}_2 \cdot (1 - \pi_A)}{n \cdot \pi_A}.$$

  - *Step 3b):*
    $$\hat{q}_2 = \frac{n_{AB} - n \cdot \hat{q}_1 \cdot \pi_A}{n \cdot (1 - \pi_A)}.$$

  According to *Step 3a)* for given $\hat{q}_2$ all possible $\hat{q}_1$ are calculated that should be involved and analogously values of $\hat{q}_2$ are found by means of *Step 3b)*. As π_A is unknown, this is done for all $\hat{\pi}_A \in [\frac{n_A}{n}, \frac{n_A+n_{AB}}{n}] \Rightarrow$ matrix $M_3 = M_{3a} \cap M_{3b}$ that contains all entries of M_1 that fullfill the upper bound restriction as well as the relation between q_1 and q_2

- *Step 4:* This procedure is realized by means of all coarsened variables `Ycoarse1` to `Ycoarse81` of the simulated data and hence for the 81 true combinations of true coarsening parameters q_1 and q_2. For every dataset the median relative bias has been calculated based on the set of all resulting $\hat{\pi}_A$ values, which has been done for every combination of q_1 and q_2. Thereby, it has been investigated that similar results can be concluded from these different datasets and thus it can be justified to illustrate the result from the first dataset only in order to keep things simple.

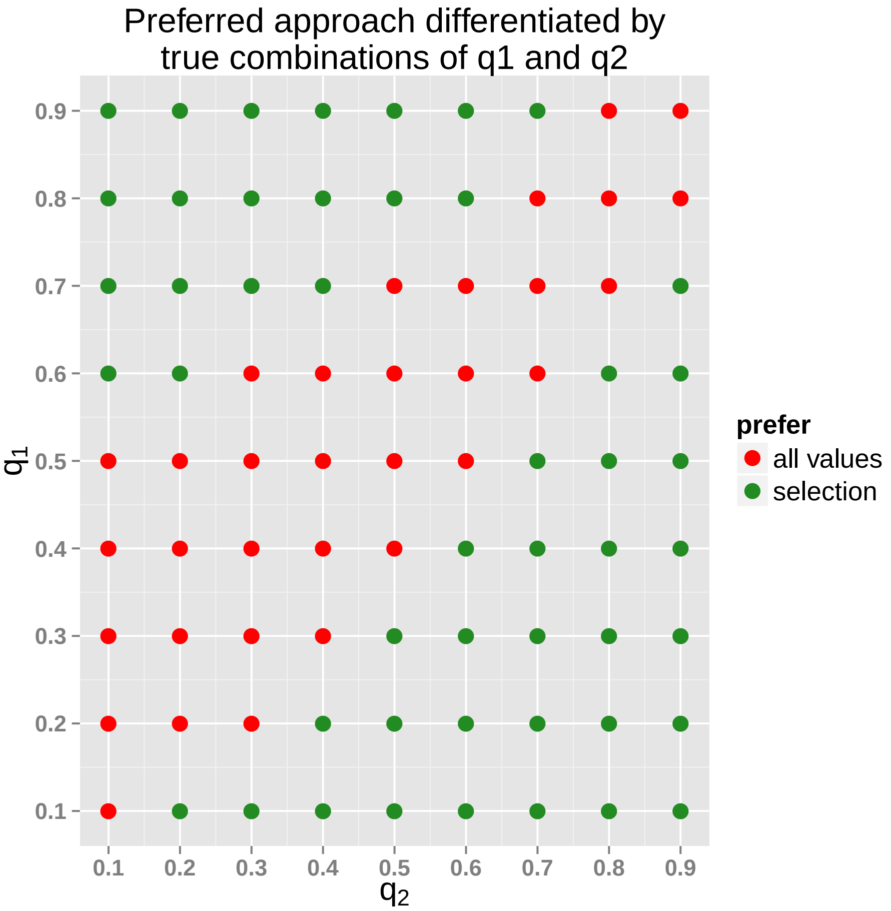


Figure 4.15.: Case 1.d.III: Evaluation of methods.

In order to evaluate this procedure, median relative bias of the estimators of π_A within matrix M_1 and M_3 are compared for the first dataset. In this connection Figure 4.15 showes by the green points for which true underlying combinations of q_1 and q_2 the described procedure outperforms the one which simply includes all $\hat{\pi}_A$ that result by assuming $0 \leq q_1 \leq 1$ and $0 \leq q_2 \leq 1$ only. Thus, the procedure that relys on a selection of $\hat{\pi}_A$ should only be applied in cases which are described by very different underlying true q_1 and q_2 values, i.e. situations which strongly differ from CAR. In order to be able to understand, why this procedure is appropriate in these cases only, the relation of *Step 3* is investigated for the underlying assumption of CAR, namely $q_1 = q_2 = q$. If CAR is valid this equality roughly has to be valid for the corresponding empirical estimators as well (at least if one faces a large sample size as here), so that

$$\hat{q}_1 = \hat{q}_2 = \hat{q} \Leftrightarrow \frac{n_{AB} - \hat{q} \cdot n \cdot (1 - \pi_A)}{n\dot{\pi}_A} = \frac{n_{AB} - \hat{q} \cdot n \cdot \pi_A}{n \cdot (1 - \pi_A)}.$$

This is only satisfied in the case of $(1 - \pi_A) = \pi_A \Leftrightarrow \pi_A = 0.5$, which corresponds to the fact already investigated on page 120, namely that one faces a symmetric problem only if the proportion of true “A” and true “B” values is the same, i.e. $\pi_A = 0.5$. In simulated data one is concerned with parameter $\pi_A = 0.67 \neq 0.5$, so that the described procedure is inappropriate in case of CAR or situations that are similar to CAR in this situation of data. As for situations of CAR or situations that are similar to the case of CAR (see Figure 4.4) two parameters have to be estimated from the two independent estimation equations (see equation (4.5)) only and thus point identified estimators can be derived, this does not have to be regarded as problematic.

Critique of the described procedure

Nevertheless, the described procedure that tries to restrict the set of possible π_A without real contentual assumptions, but by implying the upper bounds as well as the relation between q_1 and q_2 only, offers a problem that should not be ignored. As only one parameter has to be identified (either q_1 or q_2 or π_A) in order to be able to determine the other two from the estimation equations, only one parameter should be considered as given. But in the described procedure both parameters, q_1 and q_2, are treated as given and represent a kind of two dimensional sensitivity parameter as they are required to satisfy particular assumptions and for each choice of q_1 and q_2 a different precise model follows. In accordance to this, only one parameter, namely π_A, is estimated for all given imaginable combinations of q_1 and q_2. In this way, as an actually available restriction about a further parameter is not included, the set of possible $\hat{\pi_A}$ is too large and involves values that actually do not represent the optima, such that one is concerned with an overidentified identification problem. Hence, only a subset of the darkgreen area of Figure 4.14 should be selected. In order to further restrict this area, one could imply either instead of or additionally to the restrictions of Step 3) the restriction of the original optimization problem solved for q_1 and q_2, and again calculate all possible q_1 values that correspond to the underlying restriction given all possible q_2 values that result from their corresponding restriction (and vice versa). As these restrictions are dependent on the parameter π_A, this should be done for all $\hat{\pi}_A \in [\frac{n_A}{n}, \frac{n_A+n_{AB}}{n}]$ for a start. As further research will be necessary concerning this procedure, the case d.III. will not be considered in context of model 2.

Until now it has been shown that accounting for epistemic uncertainty within a multinomial logit model leads to an identification problem, which can be solved by implying assumptions as CAR. If weak assumptions that do not induce point identification of the parameters of interest can be imposed only, partial identification represents an instrument that is able to restrict the possible solutions. Across all analyses a quite interesting result has been noted, namely that estimators $\hat{\beta}_1$ as well as $\hat{\beta}_2$ are nearly unbiased concerning all analyses, even if CAR is wrongly assumed or wrong assumptions concerning R (as long as $R < 1$) are included within the estimation.

In cases in which point identification is required and valid estimators are of peculiar interest instead of finding the right true categories of coarsened observations another approach that is based on imputation, a method that is commonly used in the context of the missing data problem, could be useful. Some ideas concerning this approach will be suggested in the next subsection.

4.6. Imputation as an alternative approach

Considering the case of three observed categories only, namely “A”, “B” and “A XOR B”, the correspondence to the missing data problem is obvious if one recalls the findings from Subsection 2.1.6. There, the missing data problem has been described as a mapping from the sample space either into a singelton, which corresponds to the precise observations “A” and “B” here, or into the full power set and thus the sample space, which is in accordance with the coarsened observation “A XOR B”. Thus, as there is no additional restriction in the sense that particular elements of the power set represent the potentially true values only, but all elements of the sample space $\Omega = \{A, B\}$ are possible, coarsened data “A XOR B” can be interpreted as missing.

For that reason commonly used methods in the framework of missing data, as imputation, can be applied in the context of coarsened data as well in order to determine quantities of interest. Here these quantities of interest will be the proportion of cases that belongs to category “A”, namely $\hat{\pi}_A$, as well as the coarsening mechanisms $\hat{q}_1$ and $\hat{q}_2$.

Imputation pursues the goal to represent observed information in a way that valid inferences can be obtained instead of requiring a good predicition of the missing values. Generally imputation is appropriate if the underlying missing mechanism is MCAR or MAR (see Subsection 2.1.6), i.e. if the missing does not depend from the missing variable itself and thus is ignorable. Consequently, the fact whether observations of variable Y are coarsened or not should not depend on the values of the true underlying categories of Y and hence CAR should be valid, i.e. category “A” and “B” should be coarsened to “A XOR B” with the same probability $q_1 = q_2 = q$. Nevertheless, it is interesting to gain an insight into the apropriability of this method by means of an example of the case when CAR is not valid. Here multiple imputation will be considered, where $M > 1$ datasets will be generated to assess the additional uncertainty from imputation (Little and Rubin 2002, p. 85–90).

In order to perform multiple imputation in the described situation of coarsened data, coefficients β_0^p, β_1^p and β_2^p have been estimated by including the cases only that exhibit precise observations of Y, where the superscript p signifies “precise”. In a second step these estimated coefficients have been used in order to estimate the probability of sampling “A” in the observed imprecise cases “A XOR B”. These cases and their corresponding quantities are marked by superscript *imp* for “imprecise”. Hence, one obtains

$$\hat{\pi}_{iA}^{imp} = \frac{\exp(\hat{\beta}_0^p + \hat{\beta}_1^p x_{i1}^{imp} + \hat{\beta}_2^p x_{i2}^{imp})}{\exp(1 + \hat{\beta}_0^p + \hat{\beta}_1^p x_{i1}^{imp} + \hat{\beta}_2^p x_{i2}^{imp})}. \tag{4.6}$$

By sampling categories “A” and “B” with probabilities $\hat{\pi}_{iA}^{imp}$ and $1 - \hat{\pi}_{iA}^{imp}$, respectively, originally coarsened observations “A XOR B” could be replaced by precise categories, namely either “A” or “B”. According to the idea of multiple imputation, this sampling process has been conducted multiple times (here $M = 5$) and hence variables $Yimputed.1, ..., Yimputed.5$ result.

In order to incorporate results of all five imputations, the combining rule (Little and Rubin 2002, p. 86) has to be applied, according to which aggregated $\bar{\hat{\pi}}_A$ can be computed by

$$\bar{\hat{\pi}}_A = \frac{1}{M} \sum_{m=1}^{5} \hat{\pi}_A^{(m)},$$

where M is equal to five and the $\hat{\pi}_A^{(m)}$ has been estimated by the proportion of “A” values within the m-th imputed variable. In the same way results concerning $\hat{q}_1$ and $\hat{q}_2$ can be aggregated by

$$\bar{\hat{q}} = \frac{1}{M} \sum_{m=1}^{5} \hat{q}_1^{(m)} \text{ and } \bar{\hat{q}} = \frac{1}{5} \sum_{m=1}^{5} \hat{q}_2^{(m)},$$

where $q_1^{(m)}$ ($q_2^{(m)}$) is estimated by the proportion of all “A” (“B”) values within the m-th imputed variable.

Following this procedure illustrative imputations have been made for variable $Ycoarse11$, whose underlying coarsening mechanism is determined by $q_1 = q_2 = 0.2$ such that CAR is valid, as well as for variable $Ycoarse13$, which is characterized by $q_1 = 0.2$ and $q_2 = 0.4$, i.e. a mechanism in which CAR is not valid.

The relative bias if π_A is estimated by means of the true categories Y as well as relative bias $\bar{\hat{\pi}}_A$ based on the imputed variables $Ycoarse11$ (CAR) as well as $Ycoarse13$ (not CAR) are depicted in Figure 4.16.

One can easily note that from imputation in case of CAR an almost unbiased

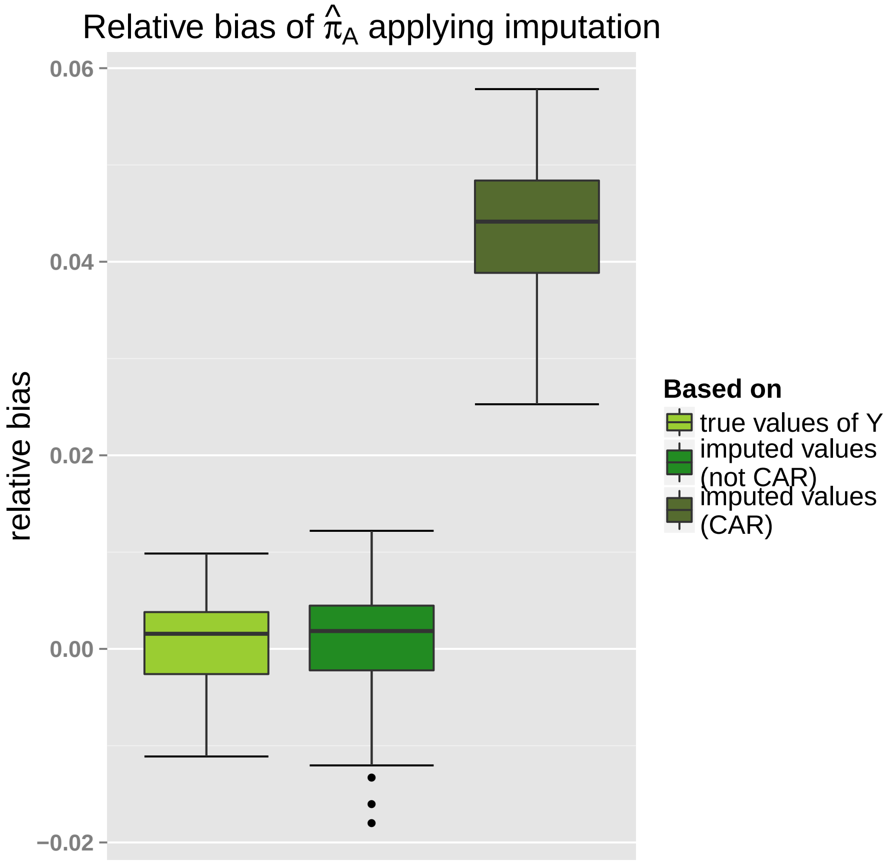


Figure 4.16.: Evaluation of $\hat{\pi}_A$ if multiple imputation based on five imputations for Y_{coarse11} (CAR) and Y_{coarse13} (not CAR) are conducted.

estimator $\bar{\hat{\pi}}_A$ with median relative bias of `0.001320` results, which additionally impresses because of its quite small standard error of `0.005164195`. In this way, apart from some outliers, the boxplot in case of CAR looks very similarly to the one in which the relative bias of the estimator based on the true values of Y is regarded, which shows a median relative bias of `0.0015620` and a standard deviation of `0.004514712`. Although the relative bias for $\bar{\hat{\pi}}_A$ in the case in which CAR is not valid is considerably larger by involving a median relative bias of `0.04429`, this amount of relative bias generally can be classified as still small and acceptable. Moreover, the underlying standard deviation of `0.006105301` is fairly larger compared to the other two cases.

Considerung Figure 4.17 tendency of results seems to be similar according to

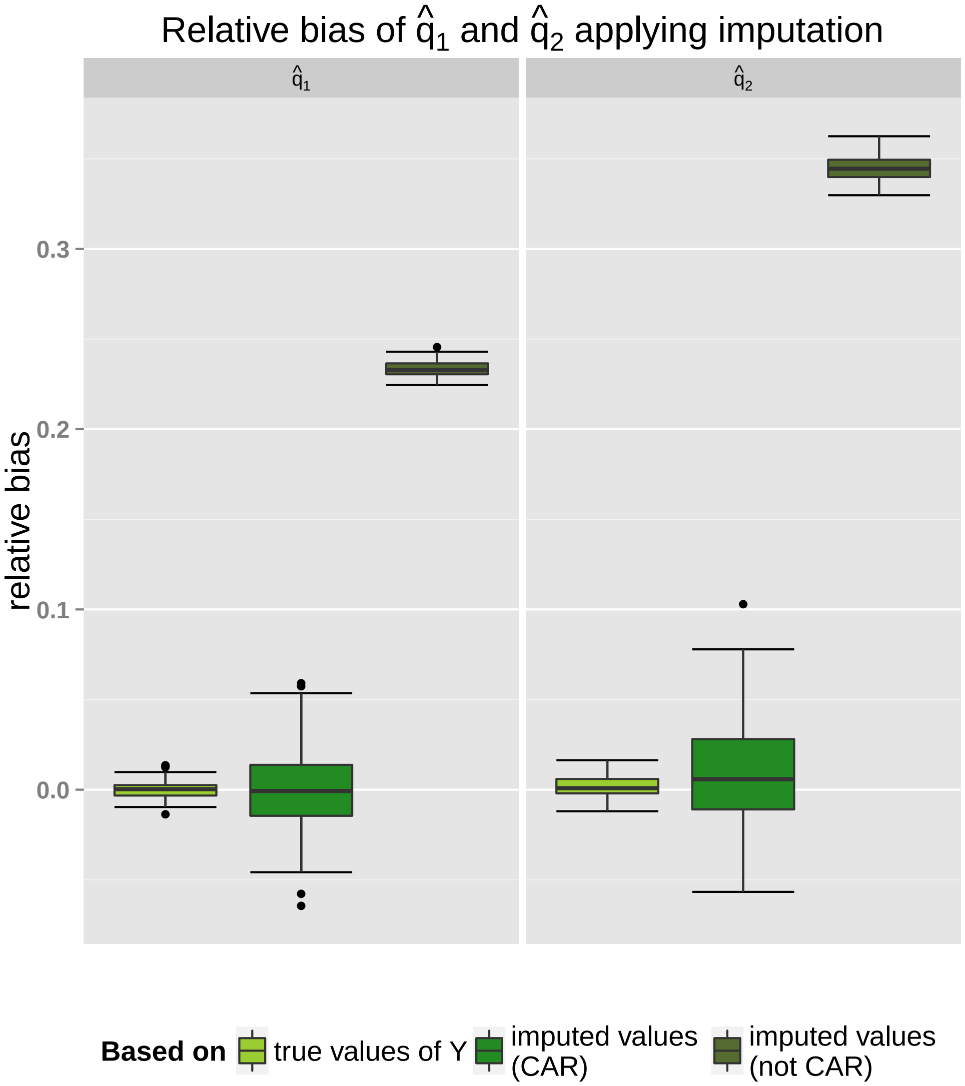


Figure 4.17.: Evaluation of $\hat{q}_1$ and $\hat{q}_2$ if multiple imputation based on five imputations for $Ycoarse11$ (CAR) and $Ycoarse13$ (not CAR) are conducted.

the relative bias of $\bar{\hat{q}}_1$ ($\bar{\hat{q}}_2$). Again, a similar relative bias of `0.0002419` and `0.003908` (`0.0007964` and `0.003464`) follows, if $\hat{q}_1$ ($\hat{q}_2$) has been calculated based on all true values of Y and on the imputed values in case of CAR, respectively. Nevertheless, at this time overestimation in the case in which CAR is not valid is substantially higher and thus a median relative bias of `0.2330` (`0.3446`) results. Here, the corresponding standard deviation is largest if the estimator is based on imputed values in the presence of CAR.

In summary, by means of the boxplots it has been illustratively shown that pretty good estimators can be concluded if imputation is applied for variables that have been coarsened at random. In order to investigate this finding more exactly, one should analyse in more detail how deviation of CAR can influence the resulting relative bias. Additionally, further research could contain a study of the consequences that are connected with an increasement of the proportion of coarsened values, i.e. an increasement of q_1 and q_2.

Until now, a situation of coarsened values that is equivalent to the missing data case has been regarded. But furthermore, the applicability of the described approach in more general situations of coarsened data is of interest, where here some first considerations in this respect will be mentioned only.

For this purpose a situation with seven observed categories is considered, namely “A”, “B”, “C”, “A XOR B”, “A XOR C”, “B XOR C” and “A - XOR B XOR C”. This situation does not correspond to the special case of missing data, as for instance by observing “A XOR C” additional information is available in the sense that either “A” or “C”, but not “B” represents the true category. Against this, in the case of missing data one would only know that the true category is an element of the sample space, namely either “A”, “B” or “C”. Hence, one can try to use this additional information implied by coarsened data within an imputation. One possible procedure of involving this information can be described by the following steps:

1. Determine $\hat{\pi}_{iA}^{imp}$, $\hat{\pi}_{iB}^{imp}$ and $\hat{\pi}_{iC}^{imp} = 1 - \hat{\pi}_{iA}^{imp} - \hat{\pi}_{iB}^{imp}$ in an analogue way as in equation (4.6).

2. Set those estimators equal to zero that can be excluded by means of the information included by the coarsened nature of the observation. For instance, if “A XOR C” has been observed, $\hat{pi}_{iB}^{imp}$ has to be equal to zero. Thus, for instance a normalization can be conducted, such that one obtains

$$\begin{aligned} \hat{\pi}_{iA}^{imp,n} &= \frac{\hat{\pi}_{iA}^{imp}}{\hat{\pi}_{iA}^{imp} + \hat{\pi}_{iC}^{imp}} \\ \hat{\pi}_{iC}^{imp,n} &= \frac{\hat{\pi}_{iC}^{imp}}{\hat{\pi}_{iA}^{imp} + \hat{\pi}_{iC}^{imp}}, \end{aligned}$$

where $\hat{\pi}_{iA}^{imp,n}$ and $\hat{\pi}_{iC}^{imp,n}$ represent the resulting normalized estimators.

3. These estimators are applied as sampling probabilities in the sense that for instance the value of estimator $\hat{\pi}_{iA}^{imp,n}$ is chosen as probability of sampling category "A" in case of a corresponding observation "A XOR C". Afterwards apply the combining rule from above.

As additional information is implied if imputation is used in case of general coarsened data, one expects even better results compared to the case which is equivalent to the missing data problem. In order to investigate the applicability of this explained procedure, one should validate it by means of simulated data and base further research on it.

Here, only first ideas concerning imputation as a method in case of coarsened data have been suggested, where some interesting aspects for further research have been proposed.

Nevertheless, partly different kinds of results are generated, as by imputation different estimators of π have been derived, while methods from the previous subsections have been based on estimations of coefficients β_0, β_1, β_2, which is not possible by proceding imputation in the way as it has been described here. Additionaly one generally should think about the basic goal of the analysis before performing imputation, as imputation should not be used in order to replace coarsened observations by the underlying true categories, but only for derivation of appropriate estimates.

5. A multinomial logit model based approach under ontologic uncertainty

The extension of the multinomial logit model that accounts for ontologic uncertainty is completely different compared to the one that incorporates epistemic uncertainty. Thus, first the multinomial logit model that accounts for ontologic uncertainty is proposed by addressing its peculiarity in a general way in Section 5.1. For reasons of consistency the *iid*-model as well as a multinomial logit model based approach with two covariates will be considered again in Section 5.2, where the corresponding data generating process, the loglikelihood as well as some results will be shown respectively. Moreover findings of the Dempster-Shafer theory will be applied in the context of prediction and the meaning of additional assumptions will be discussed in Section 5.3.

5.1. Idea and particularity of the model

Having the main idea of ontologic uncertainty in mind, namely that coarse categories as "$\{A, B\}$" reflect the truth in the sense that the corresponding individuals have not decided between category "A" and "B" yet, it could be reasonable to treat these coarse categories as own categories that require own estimators.

This idea corresponds to the one of Chapter 3 in which the *-notation has been introduced that suggests an analysis on the power set by means of

$$\begin{aligned} P^\star : \Omega^\star &\to [0,1] \\ A^\star &\to P^\star(A^\star) \end{aligned}$$

with $\Omega^\star = \mathcal{P}(\Omega) \setminus \emptyset$.

In this way one is not interested in deriving precise values for coarse observations as not even precise values exist. Consequently, the nature of the model and the underlying problem is completely different to the one that includes epistemic uncertainty, as in this context one of the main tasks has been determining the underlying coarsening mechanism and thus estimating the parameters q_1 and q_2. As in the presence of ontologic uncertainty own estimators for these coarse categories are needed instead, an extended multinomial logit model based approach is introduced here as one can see from Figure 5.1.

Multinomial logit model based approach that accounts for a coarse dependent variable under ontologic uncertainty

Data under ontologic uncertainty:
Let Y_i be a categorical random variable with nominal scale of measurement that includes precise as well as coarse categories. Moreover let

- $\boldsymbol{x_i}$ be some covariates
- Ω be the sample space of all precise categories of Y_i
- $\Omega^\star = \mathcal{P}(\Omega) \setminus \emptyset$ be the sample space of all possible categories in the sense that coarse categories are included additionaly
- $m = |\Omega^\star|$ be the number of categories of Y_i.

Model under ontologic uncertainty:
The probability of occurence for category $r = 1,\ 2,\ 3, ...,\ m-1$ can be calculated by

$$P(Y_i = r|\boldsymbol{x_i}) = \frac{\exp(\boldsymbol{x_i^T}\boldsymbol{\beta}_r)}{1 + \sum_{s=1}^{m-1} \exp(\boldsymbol{x_i^T}\boldsymbol{\beta}_s)}$$

and for category m by

$$P(Y_i = m|\boldsymbol{x_i}) = \frac{1}{1 + \sum_{s=1}^{m-1} \exp(\boldsymbol{x_i^T}\boldsymbol{\beta}_s)}$$

Figure 5.1.: Extended multinomial logit model that accounts for ontologic uncertainty within the dependent variable

This model can be regarded as an extension of the precise multinomial logit model explained in Section 4.1. The only difference of these two models can be described by the fact that the model that accounts for coarse data under ontologic uncertainty is characterized by an increasement of categories in the sense that coarse categories are included as well. Therefore own parameter estimators have to be determined for those coarse categories, so that estimators $\hat{\boldsymbol{\beta}}_{\mathbf{1}}$,..., $\hat{\boldsymbol{\beta}}_{\boldsymbol{m-1}}$ result.

5.2. Illustration of the resulting model

In order to illustrate the model introduced in Section 5.1, this model is considered for simple situations with a sparse number of categories only. After having explained the underlying data generating process in Subsection 5.2.1, the optimization problem as well as some results will be shown in Subsection 5.2.2. Thereby, for reasons of consistency with Chapter 4 again the corresponding *iid*-model without covariates is addressed first, where considerations concerning the multinomial logit model with two covariates will follow afterwards. Both models will be called model $1^\star$ and model $2^\star$ in context of ontologic uncertainty respectively, in order to distinguish them from the *iid*-model and the model with two covariates in the presence of epistemic uncertainty and to establish a connection to the *-notation of Chapter 3.

5.2.1. The data generating process

As the nature of data under ontologic uncertainty is completely different from the one that exhibits epistemic uncertainty, a data generating process that differs from the one explained in Section 4.2 is needed. As the *iid*-model as well as the model with covariates will be of interest again, both corresponding data generating processes will be described here, where the parameters that have been involved within these both data generating processes can be inferred from Table 5.1. First the data generating process for the *iid* model (model $1^\star$) will be addressed.

General parameters (all 3 models)	number of observations per dataset: n=10000 number of datasets : $M = 100$
Parameters for DGP of *iid*- model	$\pi_A = 0.48$, $\pi_B = 0.44$ $\Rightarrow \pi_{AB} = 0.08$ number of categories of Y: $c = 3$
General parameters for models with covariates	number of covariates: $p = 2$ $X_1 \sim Po(\texttt{lambda} = \texttt{3})$ $X_2 \sim \mathcal{N}(\texttt{mean} = \texttt{0},\ \texttt{sd} = \texttt{2})$
Parameters for model with covariates and 3 categories of Y ($c = 3$)	$\beta_{A0} = 1.5$, $\beta_{A1} = -1$, $\beta_{A2} = 1.4$ $\beta_{B0} = -0.9$, $\beta_{B1} = 0.8$, $\beta_{B2} = -0.1$ reference category: $\boldsymbol{\beta_{AB}}$
Parameters for model with covariates and 7 categories of Y ($c = 7$)	$\beta_{A0} = 1.5$, $\beta_{A1} = 0.2$, $\beta_{A2} = 0.9$ $\beta_{B0} = 1.4$, $\beta_{B1} = 0.8$, $\beta_{B2} = 0.1$ $\beta_{C0} = 0.2$, $\beta_{C1} = 1.2$, $\beta_{C2} = 0.8$ $\beta_{AB0} = 0.8$, $\beta_{AB1} = -0.4$, $\beta_{AB2} = 0.1$ $\beta_{AC0} = 0.7$, $\beta_{AC1} = -0.1$, $\beta_{AC2} = -0.5$ $\beta_{BC0} = 1.4$, $\beta_{BC1} = 0.9$, $\beta_{BC2} = -0.2$ reference category: $\boldsymbol{\beta_{ABC}}$

Table 5.1.: Parameters of data generating processes under ontologic uncertainty.

Model 1*:

Compared to the data used in context of epistemic uncertainty (see Table 4.2), there is no variable Y_{coarse}, as under ontologic uncertainty true values of variable Y are not coarsened, but coarse values are the true values itself. As additionally in the *iid* model no covariates are included, only variable Y has to be generated that includes some precise and some coarse values that all correspond to the truth at the time of data collection. Again, a situation of two potential precise categories, namely "$\{A\}$" and "$\{B\}$", and one possible coarse observation, namely "$\{A, B\}$", is addressed, where these categories have been sampled with probabilities 0.48, 0.44, and $1 - 0.48 - 0.44 = 0.08$ respectively, as these probabilities will result from the multinomial logit model based approach and in this way the goal is pursued to consider models that are comparable. Because of the *iid* assumptions these probabilities have been used for every individual. An extraction of the resulting dataset with 10000 observations consisting of this coarse variable Y only can be seen in Table 5.2. In this way 100 datasets have been generated, which all differ because of the randomness of the corresponding sampling process, in order to be able to consider the distribution of evaluating measures as the relative bias later on.

Y
$\{A\}$
$\{B\}$
$\{A\}$
$\vdots$
$\{A,B\}$
$\{B\}$
$\{A\}$

Table 5.2.: Structure of simulated datasets that are used for the *iid* model under ontologic uncertainty.

Model 2*:

Against this, in the model with covariates the probabilities of occurence of particular categories are dependent on the values of the covariates of the underlying individuals and thus these covariates have to be generated first. As in Subsection 4.2 one discrete covariate, namely a Poisson distributed covariate $X_1 \sim Po(3)$, and one metric covariate, namely a normal distributed covariate $X_2 \sim \mathcal{N}(0,4)$, have been involved. But instead of sampling “A” and “B” with probabilities of Equation 4.4 and coarsening these categories afterwards as in Subsection 4.2, now three categories “$\{A\}$”, “$\{B\}$” and “$\{A,B\}$” are sampled with probabilities

$$\begin{aligned}
\pi_{iA} &= \frac{\exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2}) + \exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})} \\
\pi_{iB} &= \frac{\exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2}) + \exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})} \\
\pi_{iAB} &= \frac{1}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2}) + \exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})},
\end{aligned}$$

so that coarsening is not needed any more.

In Table 5.3 an extraction of the dataset with 10000 observations is depicted that will form the foundation of the analysis by means of a multinomial logit model that accounts for ontologic uncertainty. Again, 100 datasets of that kind are simulated.

Y	X1	X2
$\{A\}$	4	5.064233
$\{A,B\}$	4	2.431019
$\{A\}$	1	1.969333
$\vdots$	$\vdots$	$\vdots$
$\{B\}$	5	1.90167023

Table 5.3.: Structure of simulated datasets for the model under ontologic uncertainty with covariates.

In the framework of including some aspects of the Dempster-Shafer theory the case of more than one coarse category can be interesting. Therefore a third kind of the latter kind of dataset with two covariates representing the situation of seven true categories of Y, namely "$\{A\}$", "$\{B\}$", "$\{C\}$", "$\{A,B\}$", "$\{A,C\}$", "$\{B,C\}$", "$\{A,B,C\}$" (i.e. the power set of $\Omega = \{A,B,C\}$ without the empty set), has been generated. The underlying data generating process is analogous to the one that incorporates one coarse category only and the resulting kind of data is generated 100 times again.

These datasets will be used in the framework of analyses by means of model $1^\star$ and model $2^\star$ in the next subsection.

5.2.2. Illustrating analyses by means of model $1^\star$ and model $2^\star$

The log-likelihood that has to be optimized as well as some results concerning the evaluation of the estimation will be shown for the *iid*-model without covariates (model $1^\star$) first and model $2^\star$ afterwards.

Model 1*:

By regarding “A or B” as an own category within the *iid* model, the corresponding log-likelihood

$$l(q, \pi_{iA}) = n_A \cdot \ln(\pi_A) + n_B \cdot \ln(\pi_B) + n_{AB} \cdot \ln(1 - \pi_A - \pi_B) \tag{5.1}$$

results, where n_A represents the number of cases with observed category “$\{A\}$”, n_B with observed category “$\{B\}$” and n_{AB} with observed category “$\{A, B\}$” and the probability of observing category “$\{A, B\}$” π_{AB} can be represented as $1-\pi_A-\pi_B$. Equation (5.1) can also be interpreted in terms of the corresponding likelihood under epistemic uncertainty by setting q_1 and q_2 of equation (4.5) to zero because of the absence of coarsening precise categories and extending it by an additional part for the coarse category.
Optimizing this likelihood one obtains estimators for the probability of occurence for categories “A” and “B”, namely $\hat{\pi}_A$ and $\hat{\pi}_B$ (reference category $\{A,\ B\}$), which will be compared with the probabilities that have been involved within the simulation process, namely $\pi_A = 0.48$ and $\pi_B = 0.44$. Figure 5.2 shows the relative bias of estimator $\hat{\pi}_A$ and $\hat{\pi}_B$, where a relative median bias of 0.0005772373 for $\hat{\pi}_A$ and −0.0003761718 for $\hat{\pi}_B$ and a standard deviation of 0.009447005 for $\hat{\pi}_A$ and 0.01048505 for $\hat{\pi}_B$ classify these estimators as quite good. This is not surprising as because of the precise observation of the categories under ontologic uncertainty this model corresponds to the commonly used models that is extended by an additional category “$\{A, B\}$” only so that a precise estimation can be conducted.

Model 2*:

In a similar way the multinomial logit model is extended by a third coarse category “A or B”, where here the general findings from Section 5.1 can be easily applied for this simple situation of data (see Section 5.2.1). Thus in this case a loglikelihood results as follows

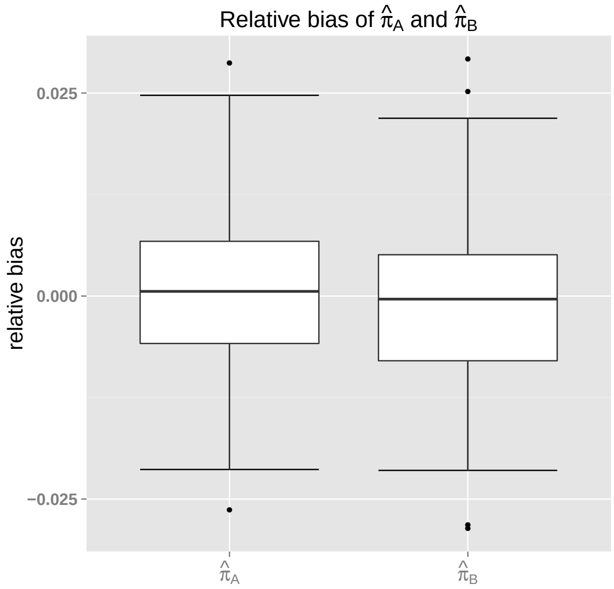


Figure 5.2.: Evaluation of $\hat{\pi}_A$ and $\hat{\pi}_B$ of model $1^{\star}$.

$$\begin{aligned} l(\boldsymbol{\beta_A}, \boldsymbol{\beta_B}) \quad &= \sum_{i=1}^{N_1} \ln\left(\frac{\exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2}) + \exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})}\right) + \\ &\sum_{i=N_1+1}^{N_2} \ln\left(\frac{\exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2}) + \exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})}\right) + \\ &\sum_{i=N_2+1}^{N} \ln\left(\frac{1}{1 + \exp(\beta_{A0} + x_{i1}\beta_{A1} + x_{i2}\beta_{A2}) + \exp(\beta_{B0} + x_{i1}\beta_{B1} + x_{i2}\beta_{B2})}\right), \end{aligned}$$

which refers to an ordered data structure, i.e. individuals $1, ..., N_1$ exhibit category "$\{A\}$", category "$\{B\}$" is observed for individuals $N_1 + 1, ..., N_2$ and individuals $N_2 + 1, ..., N$ are the indecisive ones, who show coarse category "A or B". Analogously the loglikelihood for the case with categories "$\{A\}$", "$\{B\}$", "$\{C\}$", "$\{A, B\}$", "$\{A, C\}$", "$\{B, C\}$" and "$\{A, B, C\}$" (see Section 5.2.1) can be determined.

In order to evaluate the estimators that result from optimizing the corresponding loglikelihoods for both situations of data, the boxplots of Figure 5.3 and Figure 5.4 can be considered, which show the relative bias. In the situation of three categories, a comparable large standard deviation of the relative bias

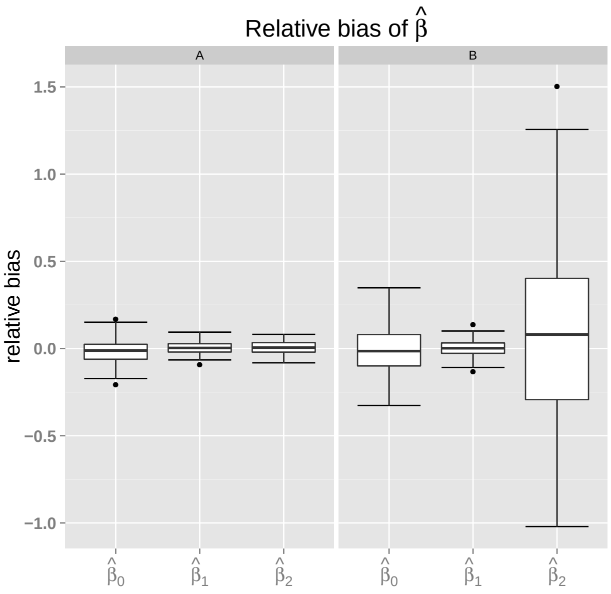


Figure 5.3.: Evaluation of the resulting $\boldsymbol{\beta}$ estimators of model $2^\star$ (3 categories).

of $\hat{\beta}_{B2}$, namely 0.4995586, is apparent, which could be acceptable because of a quite small median relative bias of 0.07985869. The other estimators show fairly small median and standard deviation of the relative bias. Against this, in the situation of seven categories there are some estimators that can be described by a quite large standard deviation as well as a comparably large median. Hence, the relative bias of estimator $\hat{\beta}_{AB2}$ exhibits a standard deviation of 1.80242 and a median of -0.3947519, the one of $\hat{\beta}_{B2}$ a standard deviation of 1.46592 and a median of -0.2831196, and the one of estimator $\hat{\beta}_{C0}$ a standard deviation of 1.702213 and a median of 0.3407546. While the relative bias of estimator $\hat{\beta}_{AC1}$ shows a quite large standard deviation (1.834586) as well, its median relative bias of -0.02087467 is comparably small.

Nevertheless, this kind of problem of comparably large standard deviations and median of the relative bias is not induced by the fact that the model is extended by implying coarse categories as well, as these coarse categories are regarded as normal categories representing the truth, which corresponds to the analysis on the power set.

Until now only the findings according to the general analysis on the power set under ontologic uncertainty have been applied in the context of this modelling approach. But in chapter 3 additionally a way of prediction by means of the DST has been presented that can be used if decisions have been made such that there are not any coarse categories left. Some considerations concerning this will be made in the next Section.

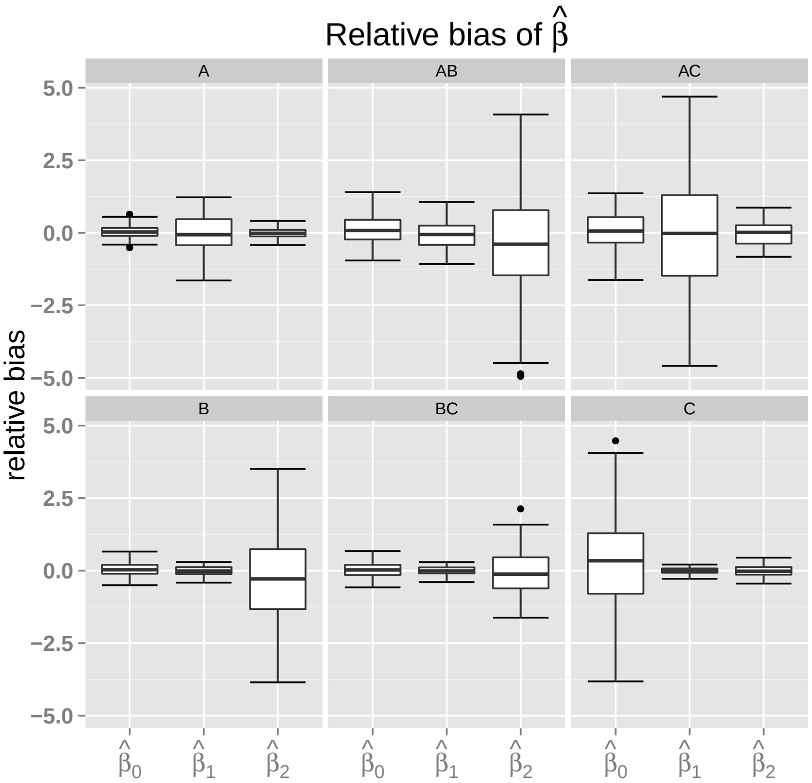


Figure 5.4.: Evaluation of the resulting $\boldsymbol{\beta}$ estimators of model $2^{\star}$ (7 categories).

5.3. Including aspects of the DST and implication of additional information

One can be concerned with situations under ontologic uncertainty that require a decision at some time. Thereby it is either imaginable that the respondent has found out rationally and by own motivation which of the imaginable options that are consistent with his coarse answer fits best to his preferences or that decisions have been enforced because of external circumstances as for

instance the election day.
Here in Subsection 5.3.1 it will be illustrated how predicition intervals can be obtained by means of DST. Although there are reasons why this interval should not be shrunk, in Subsection 5.3.2 an approach will be suggested that leads to point identified predictions.

5.3.1. Obtaining prediction intervals by means of DST

In Chapter 3 it has been explained how one is able to make predicitons by means of the belief function as well as the plausibility function from the DST, which have been called $\underline{F}^\star$ and $\overline{F}^\star$ in the context of the presence of ontolgic uncertainty. In the following these notions will be applied for the dataset that exhibits categories "A", "B", "C", "$\{A, B\}$", "$\{A, C\}$", "$\{B, C\}$" and "$\{A, B, C\}$", where in Table 5.4 the rounded mean (ratio scale) number of cases based on the 100 datasets is depicted that show these categories respectively. For illustration one can imagine that there are three parties "A", "B" and "C" that can be elected and Table 5.4 shows the answers of 10000 respondents that have been interviewed before election day. Now one is interested in the confidence at the day of election that can be attributed to the fraction of party "B" (questionary set $Q^\star = B$), which results from calculating the lower bound $\underline{F}^\star(B)$

party P	A	B	C	AB	AC	BC	ABC
no. of cases	1934	1358	5341	93	54	1146	58

Table 5.4.: Illustration of prediction in case of ontologic uncertainty.

$$\underline{F}^\star(B) = \sum_{P \subseteq B} m^\star(P) = \frac{n_B}{n} = \frac{1358}{10000} = 0.1358$$

and upper bound $\overline{F}^\star(B)$

$$\begin{aligned}\overline{F}^\star(B) = \sum_{P \cap B \neq \emptyset} m^\star(P) &= \frac{n_B + n_{AB} + n_{BC} + n_{ABC}}{n} \\ &= \frac{1358 + 93 + 1146 + 58}{10000} = 0.2655\end{aligned}$$

Hence interval

$$F^{\star}(B) = [0.1358, 0.2655]$$

can be obtained that represents the confidence of the fraction of party “B”. The length of this interval is fairly large compared to the one that results if one is interested in the confidence of the fraction of party “A” ($F^{\star}(A) = [0.1958, 0.2146]$), as in the latter case the cases that reveal coarse categories, which have to be accounted within the calculation of $\overline{F}^{\star}(A)$, is smaller. Nevertheless, the question whether these intervals and thus the underlying extent of ontologic uncertainty has to be evaluated as rather large or not has to be evaluated in the corresponding contentual context.

There might be cases in which the analyst assesses the resulting interval as too inexact, wherefore it could be reasonable to discuss whether and in which way further assumptions can be implied.

At the first glance imposing additional restrictions seems to be meaningless as this contradicts the inherent idea of data that are coarse induced by ontologic uncertainty. Under ontologic uncertainty data are coarse as not even the respondent himself knows which answer to prefer. In this way all answers that are consistent with the coarse answer are classified as imaginable options by him, which all seem to be potential answers that have to be almost equiprobable. Otherwise, if one category was available that outperforms the others, the respondent would have chosen it and he would not be indecisive. This means that all available information concerning the variable of interest has already been involved by accounting for the coarse answers of the respondents.

Thereby, one should not underestimate the benefit that is induced by implying the coarse data, as in comparison to a situation in which coarse answers are not possible and respondents have to express their indecision by choosing category “Don’t know”, substantially less information can be generated. Thus, applying the example from above concerning party “A” to this latter situation an essentially larger interval $[0.1958, 0.3292]$ is obtained, where it is formed by the same lower bound and the upper bound has been calculated by involving the fraction of category “A” as well as of all coarse categories. The improvement of implying coarse categories instead of simply having one global category for indecision increases, the larger the occurence of coarse values and the larger the number of coarse categories.

5.3.2. Suggestion of a point identifying approach in case of prediction

Although it has been pointed out in Subsection 5.3 that all the available information that can be revealed by the answer of the respondent is implied within the intervals that result by applying the idea of DST, here a way will be shown how analysts who require more precise estimations can involve additional information implied by the covariates. Thereby, it is important to emphasize that results of these procedures have to be treated with caution for reasons as mentioned above.

In order to get an idea about the influence of the covariates on clear decisions concerning the variable of interest, one can estimate the coefficients based on the subset of respondents that has already made a decision. In this way one obtains estimators $\hat{\beta}^d_{A0}$, $\hat{\beta}^d_{A1}$, $\hat{\beta}^d_{A2}$, $\hat{\beta}^d_{B0}$, $\hat{\beta}^d_{B1}$ and $\hat{\beta}^d_{B2}$, where category C has been chosen to be the reference category and d denotes "decisive". These estimators can be used to calculate estimators $\hat{\pi}^{ind}_{iA}$, $\hat{\pi}^{ind}_{iB}$ and $\hat{\pi}^{ind}_{iC}$ for the subset of respondents with covariates x^{ind}_{i1} and x^{ind}_{i2} that is indecisive:

$$\begin{aligned}
\hat{\pi}^{ind}_{iA} &= \frac{\exp(\hat{\beta}_{A0} + x^{ind}_{i1}\hat{\beta}_{A1} + x^{ind}_{i2}\hat{\beta}_{A2})}{1 + \exp(\hat{\beta}_{A0} + x^{ind}_{i1}\hat{\beta}_{A1} + x^{ind}_{i2}\hat{\beta}_{A2}) + \exp(\hat{\beta}_{B0} + x^{ind}_{i1}\hat{\beta}_{B1} + x^{ind}_{i2}\hat{\beta}_{B2})} \\
\hat{\pi}^{ind}_{iB} &= \frac{\exp(\hat{\beta}_{B0} + x^{ind}_{i1}\hat{\beta}_{B1} + x^{ind}_{i2}\hat{\beta}_{B2})}{1 + \exp(\hat{\beta}_{A0} + x^{ind}_{i1}\hat{\beta}_{A1} + x^{ind}_{i2}\hat{\beta}_{A2}) + \exp(\hat{\beta}_{B0} + x^{ind}_{i1}\hat{\beta}_{B1} + x^{ind}_{i2}\hat{\beta}_{B2})} \\
\hat{\pi}^{ind}_{iC} &= 1 - \hat{\pi}^{ind}_{iA} - \hat{\pi}^{ind}_{iB}
\end{aligned} \tag{5.2}$$

These estimators $\hat{\pi}^{ind}_{iA}$, $\hat{\pi}^{ind}_{iB}$ and $\hat{\pi}^{ind}_{C}$ will be used in order to determine sampling probabilities of precise categories "A", "B" and "C" for the indecisive respondents.

As the estimators $\hat{\pi}^{ind}_{iA}$, $\hat{\pi}^{ind}_{iB}$ and $\hat{\pi}^{ind}_{iC}$ do not reflect the information that is generated by the coarse answer of the respondent, namely that respondents who are indecisive between "B" and "C" will not report the answer "A", weighted sampling probabilities $\pi^{ind}_{iA,w}$, $\pi^{ind}_{iA,w}$ and $\pi^{ind}_{iA,w}$ will be applied respectively, where index “w” denotes “weighted”. Hence, for respondents with answer “$\{B, C\}$” weighted estimators

$$\begin{aligned}
\hat{\pi}_{iA,w}^{ind} &= 0 \\
\hat{\pi}_{iB,w}^{ind} &= \frac{\hat{\pi}_{iB}^{ind}}{\hat{\pi}_{iB}^{ind} + \hat{\pi}_{iC}^{ind}} \\
\hat{\pi}_{iC,w}^{ind} &= \frac{\hat{\pi}_{iC}^{ind}}{\hat{\pi}_{iB}^{ind} + \hat{\pi}_{iC}^{ind}}
\end{aligned}$$

are calculated which are then used as these sampling probabilities $\pi_{iA,w}^{ind}$, $\pi_{iA,w}^{ind}$ and $\pi_{iA,w}^{ind}$. Sampling probabilities $\pi_{iA,w}^{ind}$, $\pi_{iB,w}^{ind}$ and $\pi_{iC,w}^{ind}$ can be derived analogously for answers “$\{A, B\}$”, “$\{A, C\}$” and “$\{A, B, C\}$”. In this way unique categories for indecisive respondents can be obtained, which are used in order to calculate the fraction of A-, B- and C-values if decisions have been made. In this case if again the mean fractions based on the 100 datasets are considered, precise fractions of 0.2045 and 0.1776 result for “A” and “B” respectively, which both correspond approximately to the center of the intervals determined by DST. This is in accordance with the fact that information has been generated from the decisive respondents only as the coefficients have been calculated based on this subset. Therefore, the assumption has been required that the decisive and indecisive respondents do not differ concerning making their decision.

Dependent from the underlying situation, this can be regarded as problematic, as for instance indecisive respondents potentially decide either more rationally as they take more time for their decision or more arbitrarily as they cannot definitely choose a particular category. This argumentation reminds of the discussion in the framework of imputation in case of missing data, where it is required that the missingness does not depend on the true value. The described procedure in order to determine point identified answers is indeed similar to the one of classical regression imputation which calculates coefficients based on the observed observations and uses them to impute the values of the missing values. As imputation should only be applied under missing (completely) at random in the context of missing data, it is required for the described procedure that the probability of answering in a coarse way is not dependent on the value which is actually given if a decision has to be made. Please note, that

this requirement is different to the one that has been made in context of imputation under epistemic uncertainty in Subsection 4.6, as here no coarsening mechanism is available and instead of true underlying categories the answer that is ultimately given forms the reference.
Nevertheless, the interval obtained by DST already contains valuable information and if the available covariates do not have any effect on the final decision, one should rely on these intervals instead of insisting on precise results.
Although there seem to be ways how coarse data under ontologic uncertainty can be incorporated within the model, in practice one often deals with this problem by either forcing indecisive respondents to answer precisely or one omits their answer. The latter case does not only lead to a loss of efficiency because of a reduction of cases, but can also induce a substantial bias of the required estimators as the answer of these omitted indecisive respondents might differ from the other respondents for reasons already mentioned. This is even worse than in the framework of the imputation-like procedure proposed in this subsection, as within this procedure at least the information of the coarse answers as well as the covariates of the indecisive respondents have been involved.

In summary, accounting for ontologic uncertainty in the framework of a multionomial logit model is comparably simple since coarse values already represent the truth and thus can be involved within the model as own categories. Hence, the only modification that results compared to the precise commonly known multinomial logit model consists of adding further categories that represent the coarse answers. Moreover, findings from Chapter 3 could be applied, as the idea of including these coarse categories corresponds to the analysis on the power set that has been suggested and additionally prediction intervals could be obtained by means of DST.
After having made some considerations how one can account for epistemic as well as ontologic uncertainty within a multinomial logit model in Chapter 4 and this chapter respectively, it could be interesting to compare their underlying idea and to investigate the consequences if the wrong type of uncertainty is assumed. These problems will be addressed in the next chapter.

6. Comparison of the proposed modelling approaches under epistemic and ontologic uncertainty

6.1. Comparison concerning different aspects

By reason of the inherently different conceptions of epistemic and ontologic uncertainty (see Chapter 1), both resulting multinomial logit models that account for one of these two types respectively are completely different. In this way both models pursue different goals, the formulations of the resulting models differ and one is concerned with different problems. Therefore, a conclusive comparison of those two proposed modelling approaches under epistemic and ontologic uncertainty concerning these three aspects could be insightful.

While under epistemic uncertainty it is of main interest to detect the underlying coarsening structure in order to be able to infer the true values of the coarsened observations so that parameters of interest can be estimated, under ontologic uncertainty coarse values already represent the truth so that the corresponding goal consists of finding a way how these coarse values can be incorporated within the model.

These different goals indicate that the precise multinomial logit model that has been presented in Section 4.1 has to be modified differently for those two underlying models. Consequently, the likelihood of the multinomial logit model that accounts for epistemic uncertainty is described in dependence of the parameters that characterize the coarsening q that in most cases has to be estimated as well. Against this, the multinomial logit model under ontologic uncertainty extends the precise multinomial logit model by involving coarse categories as well and thus by an enlargement of the number of categories, so that own parameters estimators for coarse categories result. In this way, if the easy situation of observing categories “A”, “B” and “A XOR B” is considered, parameter $\boldsymbol{\beta_A}$ with reference category $\boldsymbol{\beta_B}$ as well as parameters q_1 and q_2 have to be estimated within the multinomial logit model that accounts for epistemic uncertainty, where for the case of ontologic uncertainty estimators $\hat{\boldsymbol{\beta}}_{\boldsymbol{A}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ with reference category $\hat{\boldsymbol{\beta}}_{\boldsymbol{AB}}$ are derived.

While in the latter model no big changes compared to the precise multinomial logit model result, so that there do not occur any problems in the course of parameter estimation, an identification problem induced by the relation of q_1 and q_2 results in case of implying epistemic uncertainty. Therefore, in connection with parameter estimation under epistemic uncertainty considerations have to be made whether assumptions as CAR or more general assumptions about a particular value of $R = \frac{q_2}{q_1}$ are justified, so that point identification can still be ensured or whether partial identification has to be conducted otherwise. It has been illustrated in the framework of partial identification that even weak assumptions concerning the relation of q_1 and q_2 for instance in the sense that more true “A” values are coarsened to “A or B” than true “B” values ($\Rightarrow R < 1$) can generate valuable information such that the resulting partial identified interval can be substantially shrunk. Although under ontologic uncertainty there are no problems of that kind so that precise parameter estimators result, in many situations decisions have to be made at some point and hence one is interested in the precise values that are going to result then. For this purpose, intervals showing the probability of occurence of special precise categories can be derived from DST, where it has been discussed whether these intervals should be further restricted or not. One way in order to obtain precise predictions has been suggested, but the associated problem of requiring a situation that is described by coarse categories that are “coarse at random” should not be forgotten in this context.

The most important differences between the multinomal logit model under epistemic and ontologic uncertainty are summarized in Table 6.1. Thereby the structure from above is seized again.

	epistemic uncertainty	**ontologic uncertainty**
the goal	detect coarsening structure	incorporate coarse data into the model
the model	extension of precise likelihood by additional terms implying q $\Rightarrow$ estimators: $\hat{\boldsymbol{\beta}}_{\boldsymbol{A}}$, $\hat{q}_1$, $\hat{q}_2$ (est. of reference cat.: $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$)	increase of number of categories compared to precise model by involving coarse categories additionaly $\Rightarrow$ estimators: $\hat{\boldsymbol{\beta}}_{\boldsymbol{A}}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{B}}$ (est. of reference cat.: $\hat{\boldsymbol{\beta}}_{\boldsymbol{AB}}$)
problems and solutions	identification problem $\Rightarrow$ implying assumptions as CAR or PI	restriction of prediction interval from DST? $\Rightarrow$ be careful with suggested imputation-similar procedure

Table 6.1.: Comparison of the multinomial logit model under epistemic and ontologic uncertainty.

6.2. The importance of dinstinguishing between these types of uncertainty

Although by means of Section 6.1 a first valuable insight could be gained into the importance of of distinguishing between epistemic and ontologic uncertainty, as completely different models result, here some further considerations and first illustrative analyses concerning this aspect will be shown.

In practice problems can arise, as analysts sometimes do not accept that coarse observations that express indecision represent the truth in the sense that they want to get an idea about the true values. In this way they analyse data that are actually coarse because of underlying ontologic uncertainty by epistemic methods. Therefore, it is interesting to investigate the resulting effect of the parameter estimators that is associated by this procedure in order to have an idea about the extent of the problem that occurs from wrongly assuming epistemic uncertainty.

As in many cases analysts require point identified results and hence apply simplified assumptions, here parameter estimation under assumed epistemic uncertainty has been conducted by implying the CAR assumption. But as in this case the error that is induced by wrongly assuming epistemic uncertainty can not be separated from the one that is caused by wrongly assuming CAR, further research should address the general case as well, in which intervals that result from partial identification should be compared with the true values of β.

Here the analysis has been based on the dataset that implies ontologic uncertainty (see Subsection 5.2.1) and that shows three observed categories, namely "A", "B" and "$\{A,\ B\}$". By means of the boxplots in Figure 6.1 one can note a substantially larger relative bias of $\hat{\beta}_{A0}$, $\hat{\beta}_{A1}$, $\hat{\beta}_{A2}$, $\hat{\beta}_{B0}$, $\hat{\beta}_{B1}$ and $\hat{\beta}_{B2}$ and comparably large standard deviations of the relative bias in the case that parameters have been derived by optimizing the loglikelihood under epistemic uncertainty and assuming CAR compared to the case that ontologic uncertainty rightly is assumed. Thus, parameter estimators are comparably strongly biased if the approach that accounts for epistemic uncertainty is applied, even if in

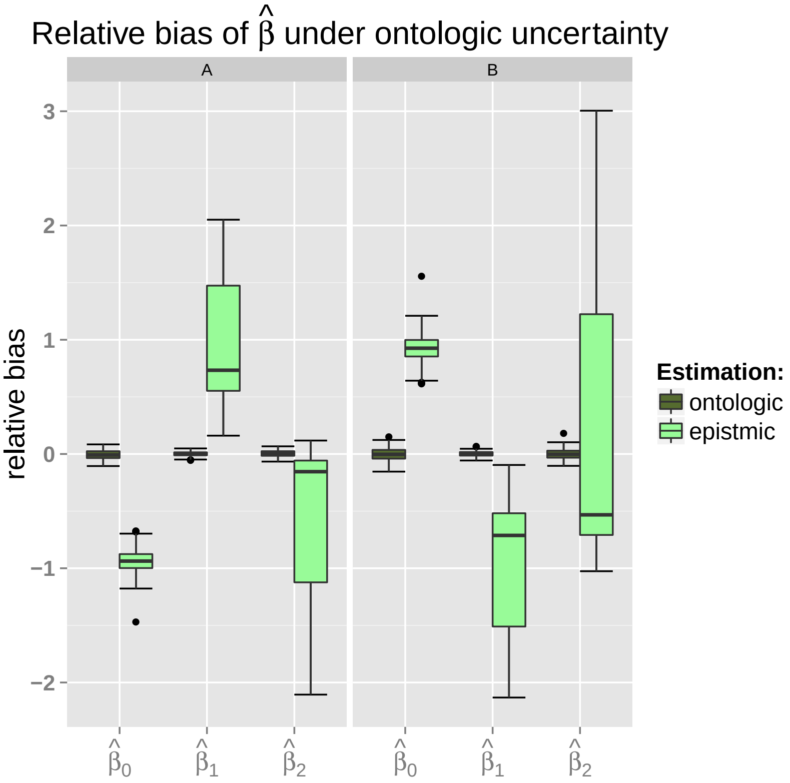


Figure 6.1.: Relative bias of $\hat{\beta}$ in the presence of ontologic uncertainty if estimation is conducted based on 1.) an ontologic and 2.) an epistemic approach.

the addressed case the number of coarse values is quite small (mean numbers: $n_A = 4844.94$, $n_B = 4364.35$, $n_{AB} = 790.71$). If the relative amount of coarse observations is increased, these estimators are expected to be even worse, which could be investigated in the framework of further research as well.
While in the model under ontologic uncertainty three categories are implied ("A", "B" and "$\{A,\ B\}$") and category "$\{A,\ B\}$" has been chosen as reference category, the model under epistemic uncertainty involves two categories only ("A" and "B"), where category "B" has been the corresponding reference. For this reason, effect coding has been applied, so that all estimated coefficients can be interpreted in regard to the corresponding mean values and hence parameter estimators from both procedures can be comparable.

In summary, it has been illustrated that there are essential differences between multinomial logit model based approaches that are designed to account for epistemic and ontologic uncertainty. For this reason it is very important to distinguish between those two types of uncertainty as otherwise substantially biased results can be obtained.

7. Summary and outlook

Although coarse data are widely present, there are still no commonly used methods that prescribe how to analyse data of that kind. Therefore, this thesis attended to the investigation of some approaches that address coarse categorical data under epistemic and ontologic uncertainty. In the course of this, in a first step considerations concerning general approaches have been made by applying already existing approaches of other areas in this context first, where their main ideas have been involved within a multinomial logit model based approach afterwards.

Under epistemic uncertainty gaining information about the unknown coarsening mechanism is central. While the assumption of "coarsening at random" can simplify a lot in this context, approaches as partial identification and sensitivity analysis can be applied more generally as they involve justified assumptions only. Even if both latter methods are commonly known from the missing data problem, by means of the relation between missing and coarse data their basic conceptions could be applied in the context of coarsened data. Among the derivation of an upper bound for the coarsening parameters, further includable contentual assumptions as information about the fraction of the coarsening parameters (e.g. $R = \frac{q_1}{q_2}$) have been suggested. Although partial identification as well as sensitivity analysis pursue the same goal and end up with similar results, their direction of analysis differs essentially.

In the presence of ontologic uncertainty the development of a framework that allows for coarse values is of peculiar interest, wherefore the *-notation has been introduced. While the idea of considering the power set has been inspired by some foundations of finite random set theory, some conceptions from the Dempster-Shafer theory have been adapted in order to be able to make predictions if decisions have been made.

After having considered these general approaches, it has been addressed how a categorical dependent variable that is coarse either induced by epistemic or by ontologic uncertainty can be involved within a multinomial logit model.

In the framework of optimizing the loglikelihood that accounts for epistemic uncertainty, an identification problem arises in case of an unknown coarsening mechanism. Although point identified results can be obtained if the assumption of CAR is involved in the estimation, analyses have shown the necessity of its actual validity, so that possibly methods such as sensitivity analysis and partial identification should be preferred. Here, it has been found that even the incorporation of some weak knowledge about the relation of the coarsening parameters can essentially shrink the resulting interval. Moreover, it is interesting that regression parameter estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ have been nearly unbiased for all conducted analyses, even in case of wrongly assuming CAR. Alternatively, an imputation based approach can be enlightening, as coarse data even allows to involve more information than in the case of missing data.

Against this, a basically different model could be introduced for the case of ontologic uncertainty. This model solely differs from the commonly known precise multinomial logit model by the fact that coarse values can be involved in terms of own categories. By means of some conceptions of the Dempster-Shafer theory predictions are achievable, where additional assumptions that shrink the prediction interval should be treated with caution.

The main difference concerning these two models consists of the fact that under epistemic uncertainty the coarsening structure is analysed, whereas in the presence of ontologic uncertainty a general model has been of peculiar interest that is able to involve the actual coarse categories.

Several worthwhile ideas for further research have been mentioned at some points of this thesis, where some especially considerable questions will be shown here:

Generally, one could analyse to what extent the proposed methods are adaptable in case that the underlying coarse variables of interest are metric instead of categorical and apply them as far as possible for this case.

It could also be worth to concentrate on some particular extensions of the general approaches of Chapter 2 and 3. In Subsection 2.2.4 the relation of coarse data to the problem of misclassification has already been worked out, so that it could be promising to investigate the applicability of some methods that are common in this area in the context of coarse data. In the case of ontologic uncertainty some ideas from random set theory and Dempster-Shafer theory have motivated the framework for the analysis, but the theory of hints by Kohlas and Monney [1995] could be insightful as well.

Furthermore, some generalizations concerning the modelling approaches of Chapter 4 and 5 are conceivable. Generally one should regard the case of coarse explanatory variables as well. Concerning the investigated case of a coarse categorical dependent variable, one could for instance consider models with dependent variables of a higher scale of measurement, as for instance the cumulative or sequential threshold model.

With regard to adjustments of the proposed model under epistemic uncertainty the case of more than two true categories and a coarsening mechanism that is dependent on the covariates could be considered. Moreover, it could be interesting to investigate how a variation of the coarsening parameters influences the relative empirical bias, where it is expected that the bias increases with increased values of the coarsening parameters induced by the associated additional uncertainty. In the same way, the impact of the proportion of particular true values could be analysed in more detail as its importance with respect to the general assymmetry of the underlying problem has been noted. In the framework of the inclusion of the upper bounds it already has been addressed that one should think about a way to involve full information generated by the estimation problem. Concerning the proposed imputation based approach one should consider the general case of more than two true categories in more detail.

In the framework of the model that accounts for ontologic uncertainty one could compare the resulting estimators with the ones that are obtained if coarse categories are summarized within one category of “Don’t know”.

Although there are several starting points for further research, the importance of the distinction between coarse data under epistemic and ontologic uncertainty is obvious. As some approaches from other areas could be applied in the context of coarse data such that there are already some methods that are able to deal with data of that kind, the analysis of coarse data seems to be a promising topic.

8. List of Figures

9. List of Tables

10. Bibliography

Baker, S., W. Rosenberger, and R. Dersimonian. 1992. Closed–Form Estimates for Missing Counts in Two–Way Contingency Tables. Statistics in Medicine 11:643–657.

Beierle, C., and G. Kern-Isberner, 2005. Wissensbasierte Systeme im Überblick. Pages 6–19 *in* W. Bibel and R. Kruse, editors. Methoden wissensbasierter Systeme, volume 3. Vieweg–Verlag.

Bellenger, A., and S. Gatepaille. 2011. Uncertainty in Ontologies: Dempster–Shafer Theory for Data Fusion Applications. Computing Research Repository .

Bernoulli, J. 1713. Ars Conjectandi. Impensis Thurnisiorum, Fratrum.

Beynon, M., B. Curry, and P. Morgan. 2000. The Dempster–Shafer Theory of Evidence: An Alternative Approach to Multicriteria Decision Modelling. Omega 28:37–50.

Blumenthal, S. 1968. Multinomial Sampling with Partially Categorized Data. Journal of the American Statistical Association 63:542–551.

Casella, G., and R. Berger. 1990. Statistical Inference. Duxbury Press Belmont.

Cormode, G., and D. Srivastava, 2010. Anonymized Data: Generation, Models, Usage. Pages 1211–1212 *in* Li, F., editor. ICDE'10 : 26th International Conference on Data Engineering. IEEE.

Davidson, R., D. Kendall, and E. Harding. 1974. Stochastic Geometry: A Tribute to the Memory of Rollo Davidson. Wiley–Interscience.

De Finetti, B. 1977. Theory of Probability. Bulletin of the American Mathematical Society 83:94–97.

Dempster, A. 2008. The Dempster–Shafer Calculus for Statisticians. International Journal of Approximate Reasoning 48:365–377.

Dempster, A. 1967. Upper and Lower Probabilities Induced by a Multivalued Mapping. The Annals of Mathematical Statistics 38:325–339.

Fahrmeir, L., H. Brachinger, A. Hamerle, and G. Tutz. 1996. Multivariate Statistische Verfahren. Walter de Gruyter.

Fahrmeir, L., T. Kneib, and S. Lang. 2007. Regression: Modelle, Methoden und Anwendungen. Springer.

Fine, T. 1977. Review: A Mathematical Theory of Evidence. Bulletin of the American Mathematical Society 83:667–672.

Gill, R., M. Laan, and J. Robins, 1997. Coarsening at Random: Characterizations, Conjectures, Counter–Examples. Pages 255–294 *in* D. Lin and T. Fleming, editors. Proceedings of the First Seattle Symposium in Biostatistics, volume 123. Springer.

Heitjan, D. 1993. Ignorability and Coarse Data: Some Biomedical Examples. Biometrics 49:1099–1109.

Heitjan, D. 1994. Ignorability in General Incomplete–Data Models. Biometrika 81:701–708.

Heitjan, D., and S. Basu. 1996. Distinguishing ”Missing at Random” and ”Missing Completely at Random”. The American Statistician 50:207–213.

Heitjan, D., and D. Rubin. 1991. Ignorability and Coarse Data. The Annals of Statistics 19:2244–2253.

Hocking, R., and H. Oxspring. 1974. The Analysis of Partially Categorized Contingency Data. Biometrics 30:469–483.

Horowitz, J., and C. Manski. 2000. Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data. Journal of the American Statistical Association 95:77–84.

Imbens, G., and C. Manski. 2004. Confidence Intervals for Partially Identified Parameters. Econometrica 72:1845–1857.

Jaeger, M. 2005. Ignorability for Categorical Data. The Annals of Statistics 33:1964–1981.

Kalbfleisch, J., and R. Prentice. 2011. The Statistical Analysis of Failure Time Data. 2nd edition. Wiley–Interscience.

Kauermann, G. and Küchenhoff, H. 2011. Stichproben: Methoden und praktische Umsetzung mit R. Springer.

Kenward, M., E. Goetghebeur, and G. Molenberghs. 2001. Sensitivity Analysis for Incomplete Categorical Data. Statistical Modelling 1:31–48.

Kohlas, J., and P. Monney. 1995. A Mathematical Theory of Hints: An Approach to the Dempster–Shafer Theory of Evidence. Springer.

Kolmogorov, A. 1933. Grundbegriffe der Wahrscheinlichkeitsrechnung. Springer.

Koopmans, T., and O. Reiersol. 1950. The Identification of Structural Characteristics. The Annals of Mathematical Statistics 21:165–181.

Küchenhoff, H., T. Augustin, and A. Kunz. 2012. Partially Identified Prevalence Estimation under Misclassification Using the Kappa Coefficient. International Journal of Approximate Reasoning 53:1168–1182.

Lagakos, S. 1979. General Right Censoring and Its Impact on the Analysis of Survival Data. Biometrics 35:139–156.

Little, R. 1994. A Class of Pattern–Mixture Models for Normal Incomplete Data. Biometrika 81:471–483.

Little, R., and D. Rubin. 2002. Statistical Analysis with Missing Data. 2nd edition. Wiley–Interscience.

Manski, C. 2003. Partial Identification of Probability Distributions. Springer.

Manski, C. 2005. Partial Identification with Missing Data: Concepts and Findings. International Journal of Approximate Reasoning 39:151–165.

Manski, C., and E. Tamer. 2002. Inference on Regressions with Interval Data on a Regressor or Outcome. Econometrica 70:519–546.

Matheron, G. 1975. Random Sets and Integral Geometry. Wiley New York.

Meintrup, D., and S. Schäffler. 2007. Stochastik. Springer.

Moeschberger, M., and H. David. 1971. Life Tests Under Competing Causes of Failure and the Theory of Competing Risks. Biometrics 27:909–933.

Molchanov, I., 2005. Random Closed Sets. Pages 135–149 *in* Bilodeau, M. and Meyer, F. and Schmitt, M., editor. Space, Structure and Randomness, volume 183. Springer.

Molenberghs, G., E. Goetghebeur, S. Lipsitz, and M. Kenward. 1999. Non-random Missingness in Categorical Data: Strengths and Limitations. The American Statistician 53:110–118.

Molenberghs, G., M. Kenward, and E. Goetghebeur. 2001. Sensitivity Analysis for Incomplete Contingency Tables: The Slovenian Plebiscite Case. Journal of the Royal Statistical Society: Series C 50:15–29.

Molinari, F. 2008. Partial Identification of Probability Distributions with Misclassified Data. Journal of Econometrics 144:81–117.

Narens, L. 2007. Theories in Probability: An Examination of Logical and Qualitative Foundations. World Scientific Publishing Co., Inc.

Nauck, B. and Brüderl, J. and Huinink, J. and Walper, S. 2013. Beziehungs- und Familienpanel (pairfam). GESIS Datenarchiv. Köln. ZA5678 Datenfile Version 3.

Neyman, J. 1977. Frequentist Probability and Frequentist Statistics. Synthese 36:97–131.

Nguyen, H. 2006. An Introduction to Random Sets. CRC Press.

Nguyen, H. 2007. A Continuous Lattice Approach to Random Sets. Thai Journal of Mathematics 5:137–142.

Nguyen, H., 2012. On Belief Functions and Random Sets. Pages 1–19 *in* Denoeux, Thierry and Masson, Marie-Hélène, editor. Belief functions: Theory and applications, volume 164. Springer.

Nordheim, E. 1984. Inference from Nonrandomly Missing Categorical Data: An Example from a Genetic Study on Turner's Syndrome. Journal of the American Statistical Association 79:772–780.

Robins, J., et al. 1997. Non–Response Models for the Analysis of Non–Monotone Non–Ignorable Missing Data. Statistics in Medicine 16:21–37.

Rubin, D. 1976. Inference and Missing Data. Biometrika 63:581–592.

Rubin, D., H. Stern, and V. Vehovar. 1995. Handling "Don't know" Survey Responses: The Case of the Slovenian Plebiscite. Journal of the American Statistical Association 90:822–828.

Schreiber, T. 2000. Statistical Inference from Set–Valued Observations. Probability and Mathematical Statistics 20:223–235.

Shafer, G. 1976. A Mathematical Theory of Evidence. Princeton University Press Princeton.

Smets, P., and R. Kennes. 1994. The Transferable Belief Model. Artificial Intelligence 66:191–234.

Stoyan, D. 1998. Random Sets: Models and Statistics. International Statistical Review 66:1–27.

Stoye, J. 2009*a*. More on Confidence Intervals for Partially Identified Parameters. Econometrica 77:1299–1315.

Stoye, J. 2009*b*. Partial Identification and Robust Treatment Choice: An Application to Young Offenders. Journal of Statistical Theory and Practice 3:239–254.

Tamer, E. 2010. Partial Identification in Econometrics. Annual Review of Economics 2:167–195.

Tutz, G. 2000. Die Analyse kategorialer Daten: Anwendungsorientierte Einführung in Logit–Modellierung und kategoriale Regression. Oldenbourg Verlag.

Vansteelandt, S., E. Goetghebeur, M. Kenward, and G. Molenberghs. 2006. Ignorance and Uncertainty Regions as Inferential Tools in a Sensitivity Analysis. Statistica Sinica 16:953–979.

Wu, W., and J. Mi, 2008. An interpretation of belief functions on infinite universes in the theory of rough sets. Pages 71–80 *in* C. Chan, J. GrzymalaBusse, and W. Ziarko, editors. Rough Sets and Current Trends in Computing, Proceedings, volume 5306. Springer.

Zadeh, L. 1984. Review of a Mathematical Theory of Evidence. AI Magazine 5:81–83.

Zadeh, L. 1986. A Simple View of the Dempster–Shafer Theory of Evidence and its Implication for the Rule of Combination. AI Magazine 7:85–90.

Zeng, D. 2004. Estimating Marginal Survival Function by Adjusting for Dependent Censoring Using Many Covariates. The Annals of Statistics 32:1533–1555.

A. Appendix

Alternative derivation of an upper bound for q_1 (analogous for q_2)

If one faces the basic equation (2.1) from Chapter 2, one can start finding an upper bound by assuming $P(\mathcal{Y} = (A\ XOR\ B)|Y = B)$ to be zero:

$$\begin{aligned} P(\mathcal{Y} = A\ XOR\ B) &\geq P(\mathcal{Y} = (A\ XOR\ B)|Y = A) \cdot P(Y = A) \Leftrightarrow \\ \frac{P(\mathcal{Y} = (A\ XOR\ B))}{P(Y = A)} &\geq P(\mathcal{Y} = (A\ XOR\ B)|Y = A) = q_1 \end{aligned}$$

The quantity on the left hand side is maximal if its denominator is as small as possible. Probability $P(Y = A)$ is smallest if all true "A"-vaules are precisely observed such that no coarsened observations "A XOR B" are produced by these true values. As this restriction is only useful, whenever $P(\mathcal{Y} = (A\ XOR\ B)) < P(\mathcal{Y} = A)$ only, the following upper bound $\overline{q}_1$ can be derived:

$$\overline{q}_1 = \begin{cases} \frac{P(\mathcal{Y}=(A\ XOR\ B))}{P(\mathcal{Y}=A)}, & \text{if } P(\mathcal{Y} = (A\ XOR\ B)) < P(\mathcal{Y} = A) \\ 1, & \text{else} \end{cases} \qquad \text{(A.1)}$$

If one is concerned with large sample sizes, probabilities $P(\mathcal{Y} = AB)$ and $P(\mathcal{Y} = A)$ can be approximated by their empirical estimators (if sampling variability is ignored), namely $\frac{n_{AB}}{n}$ and $\frac{n_A}{n}$. Thus, one obtains:

$$\overline{\hat{q}}_1 = \begin{cases} \frac{n_{AB}}{n_A}, & \text{if } n_{AB} < n_A \\ 1, & \text{else} \end{cases}$$

As the upper bound derived in Chapter 2 is more general in the sense that it is always smaller than 1, the upper bound of equation (A.1) has been shown in the appendix only.

Analysis of relative empirical bias of $\hat{q}$ in case of CAR ($q_1 = q_2 = q$, here: $q = 0.3$)

In this thesis the results concerning the relative empirical bias has been shown for the estimators of main interest only, namely $\hat{\pi}_A$ in model 1 and $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$ in model 2. In case of CAR parameter q has to be estimated, where the evaluation of the resulting estimator will be shown here.
For illustration observed data that has been generated by $q_1 = q_2 = q = 0.3$ has been used. Truely assuming CAR, the following result concerning the relative empirical bias of $\hat{q}$ can be obtained differentiated by model 1 and model 2:

Model 1

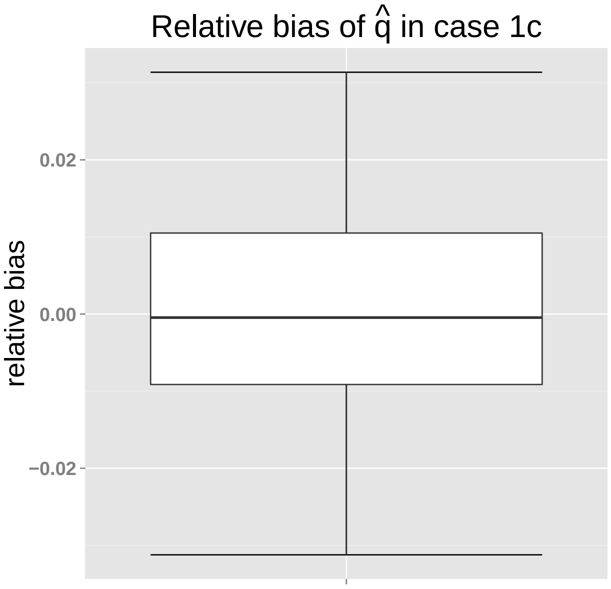


Figure A.1.: Boxplot showing the relative bias of $\hat{q}$ for the model without covariates

Minimum and maximum relative bias of `-0.03123` and `0.03137` respectively as well as a median of `-0.00046` show $\hat{q}$ as a (nearly) unbiased estimator (see Figure A.1). The standard deviation of `0.01366265`, which can also be ascribed to the large sample size of $n = 10000$. Thus, $\hat{q}$ can be considered as a quite good estimator.

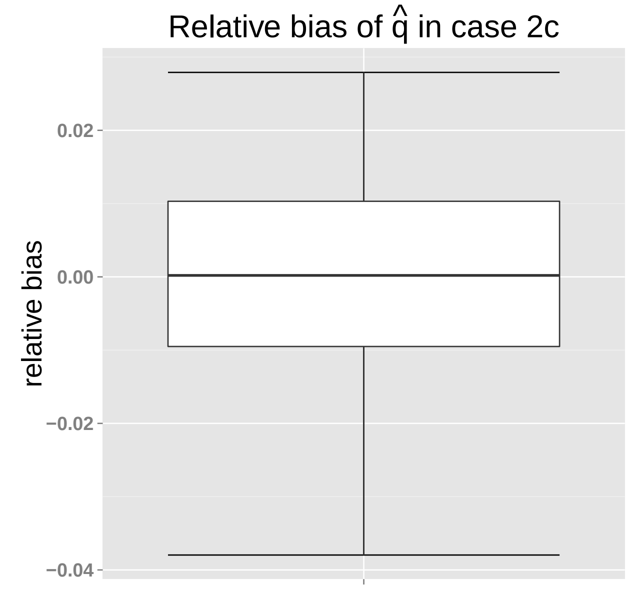


Figure A.2.: Boxplot showing the relative bias of $\hat{q}$ for the model with covariates

Model 2

The minimum and maximum of `-0.037960` and `0.0279100`, the median of `0.0001997` and the underlying standard deviation of `0.01525166` classify $\hat{q}$ as a quite good estimator if CAR is valid indeed (see Figure A.2).

B. Electronic appendix

The results of the analyses that have been shown in this thesis are based on the `R`-Code one can find on the attached CD-ROM. A brief summary of the files that are contained within the folder “`R`-Code”is given here:

- R-Code epistemic
  - `1_Creating Datasets.R`: data generating process (dgp) for model 1 and model 2
  - `2_A1_Analyse.R`: simple analyses for model 1 based on $M = 10$ datasets only
  - `3_A1_M=100_Analyse.R`: analyses for model 1 based on $M = 100$ datasets (results have been shown here)
  - `4_A2_M=100_Analyse.R`: analyses for model 2 based on $M = 100$ datasets
  - `5_Imputation.R`: Multiple imputation by means of observed variable `Ycoarse11` (=CAR) and `Ycoarse13` (= not CAR)
- R-Code ontologic
  - `1_Dataset_ont`: dgp for model $1^\star$ and model $2^\star$ (two situations of data: 3 and 7 categories)
  - `2_B_Analysen`: analyses (general analysis, prediction by means of DST, additional restrictions)
  - `3_B_Comparison_large_nAB`: Assuming ontologic uncertainty and estimation by means of epistemic approach (large n_{AB})
  - `4_B_Comparison_small_nAB`: Assuming ontologic uncertainty and estimation by means of epistemic approach (small n_{AB})

Additionaly, the CD-ROM shows a folder “graphics” containing all generated graphics, a folder “saved objects” containing all saved `.RDATA`-files and a digital version of this thesis.

Declaration of Authorship

I hereby confirm that I have authored this master's thesis independently and without use of others than the indicated resources.

Munich, 15th of July, 2013

(Julia Plaß)