International Journal of Approximate Reasoning 56 (2015) 224–248

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

www.elsevier.com/locate/ijar

Statistical modeling under partial identification: Distinguishing three types of identification regions in regression analysis with interval data

Georg Schollmeyer *, Thomas Augustin

Department of Statistics, LMU, Munich, Germany

ARTICLE INFO

Article history:
Received 19 November 2013
Received in revised form 1 July 2014
Accepted 3 July 2014
Available online 16 July 2014

Keywords:
Partial identification
Imprecise probabilities
Interval data
Interval censoring
Coarse data
Linear regression model

ABSTRACT

One of the most promising applications of the methodology of imprecise probabilities in statistics is the reliable analysis of interval data (or more generally coarsened data). As soon as one refrains from making strong, often unjustified assumptions on the coarsening process, statistical models are naturally only partially identified and set-valued parameter estimators (identification regions) have to be derived.

In this paper we consider linear regression analysis under interval data in the dependent variable. While in the traditional case of neglected imprecision different understandings of regression *modeling* lead to the same parameter estimators, we now have to distinguish between two different types of identification regions, called *(Sharp) Marrow Region (SMR)* and *(Sharp) Collection Region (SCR)* here. In addition, we propose the *Set-loss Region (SR)* as a compromise between *SMR* and *SCR* based on a set-domained loss function. We elaborate and discuss some fundamental properties of these regions and then illustrate the methodology in detail by an example, where the influence of different covariates on wine quality, measured by a coarse rating scale, is investigated. We also compare the different identification regions to classical estimates from a naive analysis and from common interval censorship modeling.

1. Introduction

The paper considers reliable regression analysis under interval data on the dependent variable. That is, we study the influence of certain covariates X (also called independent variables, explanatory variables, regressors or stimuli) on the response variable Y (dependent variable, outcome) under the additional difficulty that for Y only upper and lower bounds $\underline{Y}, \overline{Y}$ are observed, such that the interval $[\underline{Y}, \overline{Y}]$ contains Y.[1] This is a special case of so-called *coarse(ned) data*, i.e. data that are not observed in the resolution intended in the subject matter context (see, in particular, [26], for a discussion from a classical viewpoint, and, e.g., [1, Section 7.8] from an imprecise probability perspective).

Coarse data arise naturally in a big variety of situations.[2] If we consider, as a particular area, survey data, then at least four different types of instances should be mentioned. Firstly, in the case of sensitive survey questions (like questions on

* Corresponding author.
E-mail addresses: georg.schollmeyer@stat.uni-muenchen.de (G. Schollmeyer), thomas.augustin@stat.uni-muenchen.de (T. Augustin).

[1] This type of data imprecision is called epistemic data imprecision, opposed to the other type called ontic data imprecision, cf. [16] and the beginning of Section 4.

[2] For an overview of different practical situations where interval data occur very naturally, see [23].

http://dx.doi.org/10.1016/j.ijar.2014.07.003
0888-613X/

personal income) it is often recommended to use an a priori categorized scale instead of asking respondents for a concrete value, in order to avoid many refusals. Secondly, on the other hand, in the case of open questions, respondents often de facto provide interval-valued information only, by their tendency to heap concrete values towards certain attractive numbers (for the case of unemployment durations, see, e.g., [66,71]). Thirdly, coarsening information is quite a common technique for data protection/disclosure control, making the variables in the publicly available data sets set-valued. Fourthly, often rating scales (consisting of integer numbers, e.g. ranging from -3 to 3 or from 0 to 10) are used to judge the quality of certain goods or political actions, or to express the extent of agreement to a certain statement. Handling such scales in the statistical analysis as ordinal measurement procedures only may be unsatisfactory, while, on the other hand, handling each number as a precise metric measurement may be a sort of overshooting. A compromise could be to underly still a metric continuum in the background but to understand every number m on the scale as providing the interval-valued information $[m-0.5, m+0.5]$. Indeed, our data set in Section 6 is of that type when we study the influence of certain covariates on the rated quality of red wine.

When handling coarse data, classical statistical approaches, by their inherent confinement to precise models and results, are damned to escape the data imprecision somewhere in the analysis. A direct, first way is to represent every interval by a certain precise value, typically the mid-point of the corresponding interval, and then to perform a standard analysis based on these fictitious values (cf. also Section 6.3 below). More sophisticated classical approaches try to model the coarsening process, either explicitly or by presupposing situations where the coarsening can be incorporated directly into the likelihood[3] (see also the discussion of the (non-informative) interval censorship model in Section 6.4).

However, such a way to proceed requires strong additional assumptions that quite often are not supported by substantive arguments. In the last decade in statistics and econometrics awareness has been growing that this price to pay for the seemingly precise results is too high. Adding unjustified assumptions undermines the credibility of the conclusions, and therefore nothing less than the practical relevance of the statistical analysis.[4] Thus, a paradigmatic shift is taking place, appreciating the extent of imprecision as a constitutive element of statistical analysis, and stressing the importance to develop approaches that, by reflecting the underlying imprecision in the data properly, grant reliability of the conclusions. Quite often the imprecise results are still unambiguous enough to give a – now profound – answer to the underlying substantive science questions, and, moreover, in the case of an ambiguous result the scientist has learned that without further external information any stronger conclusion drawn from the data may be a mere artefact.

Against this background, in different areas of application, related approaches, basically considering all possible precise data compatible with the observed set of values, have emerged almost independently. They include the approach to reliable descriptive statistics in social sciences proposed in [47, Chapter 17f] (see also [46]), the analysis of fuzzy data (see, e.g., [20]) and concepts of reliable computing and interval analysis in the engineering sciences (see, e.g., [43]). This cautious processing of sets of precise data points is also of importance in generalized Bayesian inference (e.g., [18,73]) and is closely related to the (profile-)likelihood approach for set-valued data developed by [74] and [10]. Moreover, it can be embedded into the methodologies of partial identification and systematic sensitivity analysis (see, in particular, [36,64] and [70], respectively), which provide a general framework for observationally equivalent statistical models arising for instance from coarse data. Instead of single-valued parameters one obtains so-called **identification regions**, i.e. sets of all parameters compatible with the data and the modeling assumptions maintained. Further developments in that area include work on corresponding confidence region procedures (see, e.g., [7,42,8]), improvements with respect to computational feasibility (e.g., [11,2,62,53]), and extensions to generalized correction methods for misclassification [40,32]. Selected applications include [44]'s study of income poverty measures based on coarsened survey data, [33]'s use of register data to evaluate the German unemployment compensation reform and [60,41,58]'s investigation of treatment effects in observational studies. In another direction, approaches aiming at "most committed regions" containing at least a prespecified percentage of the data can be found in e.g., [49] or [54]. Proper handling of coarse data is also a prominent topic on ISIPTA conferences, see, e.g., the contributions by [27,72,68,61,69,10].

The present paper tries to contribute to the vivid discussion how to deal with interval data in the response variable of linear regression models. It is organized as follows. After some basic definitions in Section 2 detailing the framework we are working in, we illustrate in Section 3 the importance to be very careful which modeling assumptions exactly one is ready to maintain when working in the generalized framework of partial identification. In contrast to traditional linear regression analysis it makes a substantial difference whether or not one takes the model relationship as the indeed truly underlying structure (the "marrow" in the parlance below). One approach would indeed rely on the assumption that the conditional expectation of the response variable is exactly a linear function of the covariates. Opposed to this, another approach would be to understand the linear model merely as an auxiliary means to guide the predictions of responses given covariates. Then it is only assumed that a prediction that is linear in the covariates could approximate the conditional expectations of the response variable well enough. In the traditional context the estimators arising from both ways of modeling directly coincide in the least squares estimator. In the extended setting, however, two different types of identification regions have to be distinguished (cf. also [45]), called *Sharp Marrow Region (SMR)* and *Sharp Collection Region (SCR)*, respectively here. The latter, arising from collecting all predictions based on all possible values compatible with the interval information, naturally

[3] For the background, see, in particular, [26], extending Little and Rubin's [35] classification of missing data mechanisms to coarsening.
[4] Compare also Manski's Law of Decreasing Credibility [36, p. 1].