1 INTRODUCTION

1.1 PURPOSE

Regression analysis is the workhorse for the empirical analyses in the social and behavioral sciences, whenever the aim of the investigation is to find the effects of the independent variables on a dependent variable. When the dependent variable is a continuous variable, the standard additive-linear regression model is generally the preferred model. But when the dependent variable is a categorical variable, consisting of two or more discrete categories, researchers mostly turn to the logistic or the probit regression model. The outcomes of these logistic or probit regression analyses are often interpreted in essentially the same way as the results of standard linear regression. Logistic or probit regression is then regarded as if it were a normal standard linear regression, albeit not for a continuous dependent variable but for a categorical, often dichotomous one.

For example, when it is found that the logistic regression coefficient for the effect of Education on Voting ($0 = no, 1 = yes$) has the same positive sign, but a much higher value in the subgroup Men than in the subgroup Women, it seems perfectly logical to conclude that for men, Voting depends more heavily on their education than for women (and to start looking for an explanation of this finding). Or when in the simple logistic regression equation with Voting as the dependent and Education as the only independent variable, Age is added as a second independent variable, and it turns out that the strength of the original logistic effect of Education on Voting is reduced, it seems natural to explain this change of the educational effect just in terms of the (spurious) effects due to Age. However, such practices and conclusions, though straightforward in standard linear regression, are more problematic in logistic and probit regression.

The main reason for these potentially problematic interpretations is the possible (in)comparability of the pertinent logistic or probit effect coefficients. Regarding the above examples, the logistic regression coefficient for the effect of Education on Voting among Men may not be completely comparable with the corresponding logistic effect among Women and, similarly,

the logistic effect of Education on Voting in the simple logistic regression equation may not be completely comparable with the corresponding effect in the multiple logistic regression equation (after adding Age).

The main source of the possible incomparabilities can be explained in two different, but strongly related ways, which follow directly from the fact that the logistic/probit model can be derived from two different perspectives. As is shown in Chapter 2, the logistic/probit model can be seen as a DRM (discrete response model) or as an LVM (latent variable model). In logistic/probit DRM, the dependent variable is the ordinary, observed categorical variable Y (e.g., Voting or not and the logistic/probit regression equation is used to find the direct effects of the independent variables on the observed Y). In logistic/probit LVM, on the other hand, the dependent variable is a latent, not directly observed continuous variable Y^* (e.g., the underlying Propensity to vote), which is connected in a specific way to the observed categorical variable Y. The effects of the independent variables on Y^* are represented by a standard additive-linear regression equation.

From the LVM perspective, the logistic/probit regression coefficients for the effects on Y turn out to be scaled versions of the underlying unstandardized regression effects on latent variable Y^*. The scaling concerns the unexplained (error) variance in Y^*: The unexplained variance in Y^* is arbitrarily fixed to a certain value.

Now, if the unknown, unfixed unexplained variance in latent variable Y^* is different in one subgroup (Men) compared to another (Women), but the fixed unexplained variances are made the same for each subgroup, the error-scaled effects of Education on Voting will differ from each other, even when the unscaled effects on Y^* are the same. In this sense, the logistic regression effects among Men and Women may not be completely comparable and may provide misleading information about the underlying effects on Y^*.

Similarly, the unexplained variance in Y^* in the simple regression equation (with only Education as the independent variable) will generally be smaller than in the multiple regression equation (with Age added) and, therefore, the scaled logistic effects of Education on Voting in these two equations will no longer be strictly comparable, because the error variances in both equations will be fixed to the same value.

Looking at the comparability issue from the DRM perspective, there is the often overlooked fact that when an independent variable is added to a logistic (or probit) regression equation and this additional variable has a direct effect on the dependent variable but is statistically independent of the other independent variables in the equation, all logistic (or probit) effects from the original equation will get different values. This is unlike what happens to the regression coefficients in additive-linear regression analysis, which are not affected by the introduction of such an extra orthogonal variable (also called a *maverick*; see Section 2.4).

Regarding our examples, the logistic (probit) effect of Education on Voting in the multiple regression equation would be different from the corresponding effect in the simple logistic regression equation, even if the added variable Age would be statistically independent of Education and therefore no spurious Age effects could be present. Similarly, adding such an orthogonal variable to the subgroup analyses might affect the existing effects differently in different subgroups, especially when the effects of the orthogonal variable on *Y* is much stronger in one particular subgroup than in another. It looks as if the corresponding subgroup effects are only comparable if all variables that affect the dependent variable have been explicitly included.

The relationship between the *scaling problem* in logistic/logit/probit LVM and the *maverick problem* in logistic/logit/probit DRM (later called the *collapsing problem*) will become clear in the remaining chapters of this volume.

The principal purpose of this volume is to provide insight into the precise nature of the *comparability issues* and some related problems. The consequences these issues may have for the substantive conclusions are evaluated in such a way that readers can make optimal decisions on when and how to use logistic/probit regression for answering their own research questions.

There is an impressive amount of older and more recent literature that discusses the possible pitfalls of equating too easily standard linear and logistic regression and offers guidelines for the appropriate uses and interpretations of logistic regression (see, e.g., Allison, 1999; Breen et al., 2018; Guo & Geng, 1995; Hauck et al., 1991; Karlson et al., 2012; Kuha & Mills, 2017; Long, 1997; McKelvey & Zavoina, 1975; Mood, 2010; Winship & Mare, 1983, 1984; Yatchew & Griliches, 1985; and the many references mentioned in these articles).

On the basis of this literature, researchers have drawn all kinds of conclusions about the usefulness of logistic regression and, for that matter, of closely related techniques like probit, logit, and log-linear analysis. At one extreme, it is advocated to abandon logistic regression altogether; at the other, it is concluded that it is essentially *much ado about nothing*.

Obviously, the problematic aspects of logistic/probit regression cannot be simply ignored. At the same time, the types of problems are precisely known. So a researcher may know exactly under what circumstances the interpretation of the outcomes may become problematic. Moreover, in many cases, sound solutions for the problems are readily available. Finally, the distortions due to the comparability and other issues may be largely irrelevant for the answers to the research questions. Despite the difficulties a researcher may encounter when using logistic regression, the advantages of the logistic regression model for the analysis of a categorical dependent variable may far outweigh the disadvantages.

1.2 CONTENT

In this volume, three general comparative situations are dealt with that cover most of the substantive research questions researchers may try to answer by means of logistic/probit regression:

- How to interpret and compare the logistic/probit regression effects within one single regression equation (Chapter 3)
- How to interpret and compare the logistic/probit effects from different subgroups or time points (Chapter 4)
- How to interpret and compare the logistic/probit effects from different equations to find estimates for the total, direct, indirect, and spurious effects (Chapter 5)

Chapters 3 through 5 can be regarded as the core chapters, dealing successively with the three comparison situations. Because the issues involved in each of these comparison situations are somewhat different for LVM and DRM logistic regression, these issues are discussed separately for LVM and DRM. Moreover, an important part of the controversies surrounding the use of logistic/probit regression concerns the causal status of the logistic regression coefficients. Therefore, separate attention is paid to causality in each of the core chapters. A summary IN SUM is provided at the end of each core chapter.

In Chapter 2, the necessary background material is presented. Mainly those elements emphasized in Chapter 2 are needed in later chapters. Having some elementary knowledge of logistic regression analysis may be advantageous for the reader's understanding of this chapter.

In Chapter 6, some extensions of the logistic regression models are briefly discussed, especially models for dealing with a polytomous dependent variable. Further, a few summary remarks are made about on how to measure effects, more specifically about the use of logistic regression effects and odds ratios compared to standard regression effects and percentage differences. A few general concluding remarks close this chapter (and volume).

Throughout this volume, the discussions are extensively illustrated by means of several real-world and one simulated data example (see the book's webpage).

Experimental data from social psychology will be used in Chapter 3 in the discussions on the causal status of logistic regression effects. The general question underlying this experiment is whether the perceived moral standing of a 'creator' affects the acceptance of their creative products (Stavrova et al., 2016).

For the discussions on interaction effects and subgroup comparisons in Chapter 4, a data set from sociology is borrowed that is also used in Allison's

(1999) influential article on the use of logistic regression for subgroup comparisons (Long et al., 1993). The data set is about the chances assistant professors have of being promoted to associate professor and whether these chances are different for men and women.

The separation of a total logistic effect into its direct and indirect logistic components, as discussed in Chapter 5, is illustrated by means of a data set from economics regarding the acceptance of mortgage loan applications and how the race of the applicant affects the acceptance decision by the white loan officer (Hunter & Walker, 1996).

A simulated data set is used for the introduction of the basic logistic/probit regression model in Chapter 2. This data set is constructed with *university enrollment* in mind as the dependent variable. The simulated data set has the advantage here that several of the peculiar characteristics of logistic/probit regression can be more clearly seen and illustrated because the true state of affairs, the true model, and the true values of the logistic regression effects are known. The same simulated data set is also used in the first parts of Chapter 3.

Finally, simulated data appears in Section 2.3.3 to show the consequences of heteroskedasticity and in the book's webpage, mainly to illustrate the discussions around Figures 3.1 and 5.3.

More extensive descriptions of these data sets can be found in the pertinent chapters.

The main focus in this volume is on the logistic regression model, often in comparison with the standard linear-additive regression model (or its variant LPM [linear probability model]). However, the probit regression model is also introduced, and it is shown how the comparison difficulties and solutions for the logistic regression model apply similarly to the probit regression model. The choice to focus more on the logistic than on the probit regression model is mainly based on the more elegant interpretation that can be given to logistic regression coefficients compared to the probit regression effects.

Moreover, as follows from the close correspondence between the logistic regression model and the categorical logit or loglinear model, the findings for the logistic regression effects turn out to be similarly relevant for the effects in logit and loglinear models.

To explain the basic comparison issues involved, unnecessary complications are avoided. For one thing, this means that the discussions and examples are restricted to regression models for a dichotomous dependent variable. However, the basic insights into the comparison problems obtained in this way can be readily extended to models for a polytomous dependent variable. (In the last chapter, several logistic regression models for polytomous dependent variables are introduced.)

Another consequence is that hardly any attention will be paid to statistical inference issues (but see Hosmer & Lemeshow, 1989; Long, 1997). In the

light of the ongoing *statistics war* (Mayo, 2018), this must certainly not be interpreted as a sign that the authors think that statistical inferences and statistical significance tests are unimportant and that they advocate to do away with *p*-values and statistical tests. On the contrary. The only reason not to deal extensively with statistical inference is that it is just another topic than the one we deal with here. And luckily, standard statistical packages mostly provide all that is needed in this respect; STATA is especially a good choice (Long & Freese, 2014). Complex sampling schemes can often be accommodated and then there is always resampling, bootstrapping, and Bayesian approaches to be used in nonstandard situations, as Feinberg (2012) reminds us.

1.3 CAUSALITY

As indicated earlier, in Chapters 3 through 5, explicit attention is paid to the causal interpretation of the logistic regression outcomes. This explicit focus on causality is needed because part of the controversies surrounding logistic regression has to do with the possibility to interpret logistic/probit regression equations and their effect coefficients in a causal sense. The final verdict on whether this is possible or not and in what way, depends to a large extent on one's view on causality and on what is regarded as a proper causal analysis. Hence, some brief remarks on our views.

Even from a very cursory glance at the philosophical literature it is clear that the concept of causality is not a simple one and, more importantly, far from unequivocal. There are several different conceptualizations and accounts of causality each with their own advantages, problems, and different emphases (see Beebee et al., 2009; Cartwright, 2007, 2014; Elster, 1983, 2007; Kern, 2004; Kistler, 2018; Losee, 2011; Mumford & Anjum, 2013; among many others). Not surprisingly then, also in social and behavioral research many different approaches to the investigation of causality can be found, using different causal accounts and ranging from more qualitatively oriented methods to a variety of quantitative methods in (quasi-)experimental and observational studies (see, e.g., Berzuini et al., 2012; Cox & Wermuth, 2004; Diamond & Robinson, 2010; Freedman, 2008a; George & Bennett, 2005; Gerring, 2007, Chapter 7; Hill, 1965; Illari & Russo, 2014; Imbens & Rubin, 2015; Mahoney, 2003; Morgan, 2013; Morgan & Winship, 2007; Pearl, 2009, 2010; Pearl & Mackenzie, 2018; Ragin, 2000; Rubin, 1974, 2005; Shadish et al., 2002; Vayda & Walters, 2011).

Looking at the various approaches of causality in the social and behavioral sciences, it is our view that, in the end, it is all about telling a convincing causal story, where *convincing* is not meant here in a rhetorical or empathic way, but convincing because of the basic methodological rules that have been followed. Among many other requirements, the story must be evidence

based. This presupposes that the theoretical statements in the story are logically consistent with each other and are empirically testable. Further, for the story to be convincing, the predictions following from crucial theoretical causal statements must be empirically supported, and perhaps most difficult, plausible alternative explanations for the empirical relationships must be excluded.

Formal statistical procedures may be very enlightening and helpful to make the causal story methodologically convincing. The two most elaborated and influential statistical procedures in the social and behavioral sciences are probably the closely related causality accounts by Rubin and by Pearl (Imbens & Rubin, 2015; Pearl, 2009, 2010; Pearl & Mackenzie, 2018; Rubin, 1974, 2005). Pearl's structural theory of causation makes a particular use of parametric and nonparametric structural equation models (SEMs), path diagrams, and (Directed) Graphs (especially DAGs [directed acyclic graphs]). It is closely related to the kind of causal modeling that is common in observational studies with an emphasis on the elimination of the confounding effects of observed and unobserved variables causing *spuriousness*.

Rubin's PO (potential outcome) model is more in line with the tradition of causal analysis by means of (quasi-)experimentation with an explicit focus on the design of an investigation and the possibly selective way subjects have been "assigned" to the (quasi-)experimental conditions. Both approaches can be largely translated into each other's languages.

Underlying these two approaches, there may be also a difference in methodological orientation regarding causality. Researchers following Rubin's PO approach are more inclined to equate causal effect with the difference in the dependent variable found after manipulation of the experimental factor in a pure, randomized experiment. This difference *is* the causal effect. Within the Pearl/SEM/graphical approach, causality is more seen as a theoretical concept. Regression coefficients as such, for example, in SEMs (or for that matter, in experiments) are not causal effects. A causal interpretation of a regression coefficient must be based on extra theoretical considerations. The authors of this volume are more inclined to subscribe to the latter position.

More on these two approaches and especially on how logistic regression fits into these two causal accounts is discussed in Chapters 3 through 5.

As a final point, it should not be forgotten that despite the fact that causal analysis and causal explanation may be seen as the ultimate goals of social and behavioral science research, a large part of this research can probably best be characterized as descriptive or as somewhere in between the extremes purely causal and purely descriptive. And descriptive research, learning the facts, the true state of the world, how it looks like, is often important in itself and very often difficult enough to achieve. Even if the use of logistic regression analysis would be confined to description—a position we do not accept—it would still serve an important role in social and behavioral science research.