Three Essays on the Econometric Analysis of Discrete Dependent Variables

Dissertation
der Wirtschaftswissenschaftlichen Fakultät
der Universität Zürich

zur Erlangung der Würde
eines Doktors der Ökonomie

vorgelegt von
Stefan Boes
von Deutschland

genehmigt auf Antrag von

Prof. Dr. Rainer Winkelmann
Prof. Michael Wolf, Ph.D.

Die Wirtschaftswissenschaftliche Fakultät der Universität Zürich gestattet hierdurch die Drucklegung der vorliegenden Dissertation, ohne damit zu den darin ausgesprochenen Anschauungen Stellung zu nehmen.

Zürich, den 03.10.2007

Der Dekan: Prof. Dr. H.P. Wehrli

Acknowledgments

On the completion of this thesis I benefited much from the work and interaction with a lot of people, colleagues and friends.

First and foremost, I want to thank Rainer Winkelmann, whose encouragement, scholarship, and support during the whole time of my dissertation made this a great experience and a successful project. I greatly appreciate all the stimulating and inspiring discussions that we had (and will have) about econometric problems, and what he taught me about applied econometric analyses.

Many thanks go to Michael Wolf for co-advising my thesis. Further, I am particularly grateful to Joao Santos Silva, Richard Smith, Josh Angrist, Guido Imbens, Richard Williams, and Martin Kukuk for all the constructive conversations, and to Philippe Mahler, Kevin Staub, Oliver Bachmann, Adrian Bruhin, and all other colleagues in Zurich for creating a very productive and friendly working atmosphere.

I owe inexpressively much to my beloved wife Alejandra, for her unconditional patience with my fascination for econometrics, with the books and papers spread all over our apartment, and with the amount of time I spent working on my projects. Finally, I am heavily indebted to my parents, who supported and motivated me my entire life, and to whom I dedicate the thesis.

Contents

Chapter 1

Introduction

Random variables that can only take a finite or countably infinite number of values are called *discrete* random variables. All qualitative variables are discrete, such as binary and multinomial variables, but also quantitative variables can be discrete, such as durations and counts. In this thesis, I discuss econometric models for discrete dependent variables. Examples for discrete variables are numerous and include the choice of holiday destination, the highest educational achievement, an individual's employment or health status, the retirement age, or information like a firm's number of patents and investment strategies, just to mention a few.

Discrete random variables need to be distinguished from *continuous* random variables that can take an infinite number of possible values. Examples include personal income, consumer good expenditures, distance to the nearest college, or the hours of work per year. In practice, the separation between discrete and continuous is mostly a gradual one since all variables can only be measured with a finite precision, and one may consider variables that take enough discrete values as continuous — which in this case is referred to as quasi-continuous. An example for a quasi-continuous variable is time to a particular event, measured for example in days, or in milliseconds.

The examples above are mainly drawn from economics, reflecting the subject matter of the thesis, but the list can be arbitrarily extended to examples from other disciplines in the social, natural, and applied sciences. In fact, even though the applications below mostly belong to the economics' domain, the models and methods developed here are

equally relevant to all other quantitative sciences.

Statistical methods to analyze discrete dependent variables have evolved rapidly over the last decades, on the one hand because mostly all large databases include this type of variable and modeling the relationship between discrete and/or continuous covariates and discrete responses has become a major concern, and on the other hand because increased computing power and availability of statistical software packages has stimulated the development and implementation of ever more sophisticated models as canned procedures, and, as a consequence, many models may nowadays be easily estimated.

Quite specialized tools for discrete dependent variables are necessary because a large number of models suitable for continuous dependent variables, e.g., the linear regression model, are simply inapplicable or at least inappropriate when applied to a regressand measured on a discrete scale. Among the reasons are:

- The conditional expectation function — which is the modeling subject of the linear regression model — may not exist, if, for example, the dependent variable has the character of a multinomial response.

- Discrete dependent variables often do not have a linear conditional expectation function (given that it exists) because of range restrictions (non-negative, 0/1, ...), and a presumed linearity may result in non-sensible predictions.

- With discrete dependent variables, each outcome has a positive probability and modeling these probabilities may be of independent interest.

This list is by no means exhaustive, but it identifies the main rationales for introducing econometric models for discrete dependent variables.

I will assume throughout the thesis that the data are observational (rather than experimental), i.e., I will place myself in the position of a data consumer that cannot control the data collection procedure, but I will assume that individual units have been randomly drawn from the population of interest. This implies that each member of the population has equal probability of appearing in the sample, but this does not imply that all the information relevant for the analysis is available in the sample at hand. In fact, more often than not such a lack of information will be present. To clarify this point, let Y

denote the discrete dependent variable of interest, and let W denote the vector of covariates. The most general framework in which one can analyze the relationship between Y and W is the probabilistic framework, more specifically the probability distribution of Y conditional on W. Let y denote a realization of Y, and let w denote a realization of W.[1] The conditional distribution of Y given W can be written as

$$P(Y = y|W = w) \tag{1.1}$$

The core of all econometric analyses is to learn something about $P(Y = y|W = w)$ from the available data. If for each individual all combinations (y, w) were observed and the number of observations were unlimited, then this would display an ideal situation where all conclusions of interest (positive or negative) can be drawn. Unfortunately, such a situation is not realistic as either the number of observations is limited — which is well in the domain of statistics and statistical inference —, or not all the relevant information is observed so that $P(Y = y|W = w)$ is not identified for all (y, w) (Manski 1995). The identification problem is well in the domain of econometrics, and empirical researchers usually proceed with imposing assumptions on the sampling process that, when combined with data, allow to draw conclusions with respect to economic hypotheses. The empirical strategies may roughly be distinguished in parametric, semiparametric, and nonparametric.

In parametric approaches, a conditional probability model for Y given W is specified up to a finite number of parameters and assumed to fully characterize the sampling process. The parametric approach has the advantage that estimation can be carried out in a familiar maximum likelihood framework, and obtaining point estimates, testing hypotheses, predicting outcomes, and the like is fairly straightforward. However, it is known as the common trade-off in econometrics that being able to draw strong conclusions generally requires strong assumptions. If these assumptions are not met, that is, if the conditional probability model is misspecified, then the maximum likelihood principle runs into serious problems and all the well-known asymptotic properties (consistency, normality, efficiency) no longer need to hold. There are a number of reasons why the presumed model can be misspecified, and I will give a brief overview of four of them.

[1] Throughout, I will use capital letters to denote random variables, and lower case letters to denote sample observations or realizations of random variables.

Model assumptions induce results

A conditional probability model imposes structure on the data that may or may not be reflected in the population of interest. For example, consider the change in the conditional probabilities $\Delta P_y = P(Y = y|W = w_1) - P(Y = y|W = w_0)$ for two different values w_1, w_0, and assume that Y takes three values $1, 2, 3$. Suppose the model either predicts

$$\Delta P_1 > 0 \quad \Delta P_2 < 0 \quad \Delta P_3 < 0 \qquad \text{or}$$
$$\Delta P_1 > 0 \quad \Delta P_2 > 0 \quad \Delta P_3 < 0$$

but the model can never predict

$$\Delta P_1 > 0 \quad \Delta P_2 < 0 \quad \Delta P_3 > 0$$

Thus, the probability changes can only switch from positive to negative (when moving from the smallest outcome $Y = 1$ to the largest $Y = 3$), but can never switch back to positive. If the population is indeed characterized by multiple switching, then the results above are more model driven rather than empirically determined. Examples where such a pattern can arise are the ordered probit and logit models, and the Poisson regression model; see for example Winkelmann and Boes (2006: Ch. 6/8).

Unobserved heterogeneity

Suppose that W can be separated into X and U, where X denotes the vector of observed covariates and U denotes unobservables summarizing everything that affects Y except X. Rewrite the conditional distribution in (1.1) as $P(Y = y|X = x, U = u)$. Since U is unobserved, the analysis must be restricted to what is observable, i.e.,

$$P(Y = y|X = x) \;=\; \sum_u P(Y = y|X = x, U = u)P(U = u|X = x)$$

where for simplicity it is assumed that U conditional on X has a discrete probability distribution $P(U = u|X = x)$. Specification of $P(Y = y|X = x, U = u)$ up to a finite number of parameters is not sufficient to perform a valid analysis, additional assumptions on $P(U = u|X = x)$ are necessary to link the model with the data. These additional assumptions may be hard to justify and likely be ill-conditioned as they need to impose structure on the relationship between observables X and unobservables U.

Endogeneity

Replace Y by Y_1 in (1.1) and suppose that W can be separated into Y_2 and X, where both Y_2 and X are observed. In order to access the problem of endogeneity it is helpful to first consider the joint distribution of (Y_1, Y_2), conditioned only on X, and then to discuss the implications for the conditional model $Y_1|Y_2$ and X. It must hold that

$$P(Y_1 = y_1, Y_2 = y_2|X = x) \;=\; P(Y_1 = y_1|Y_2 = y_2, X = x)P(Y_2 = y_2|X = x)$$

If the parameters of interest only appear in the model for $P(Y_1 = y_1|Y_2 = y_2, X = x)$, and the parameters in the model for $P(Y_2 = y_2|X = x)$ are merely nuisance parameters, then Y_2 is exogenous and inference based on the model for $P(Y_1 = y_1|Y_2 = y_2, X = x)$ alone is meaningful (Engle *et al.* 1983). Conversely, if the parameters of interest appear in both parts of the model, for $P(Y_1 = y_1|Y_2 = y_2, X = x)$ and for $P(Y_2 = y_2|X = x)$, then Y_2 is said to be endogenous and inference based on the conditional probability model for $P(Y_1 = y_1|Y_2 = y_2, X = x)$ alone is invalid.

Partial observability

Suppose that W can be separated into D and X, where D is a dummy variable indicating whether Y is observed ($D = 1$) or not ($D = 0$), and let X denote the vector of observed covariates. Two different cases need to be distinguished

$$P(Y = y|D = 1, X = x) \qquad \text{and} \qquad P(Y = y|D = 0, X = x)$$

Partial observability implies that the sampling process reveals information on $Y|X$ conditional on Y is observed, which is with probability $P(D = 1|X = x)$, but the sampling process is uninformative with respect to the conditional distribution of $Y|X$ given $D = 0$, which happens with probability $P(D = 0|X = x)$; see for example Manski (2003). Thus, if one is interested in the conditional distribution of $Y|X$, then

$$\begin{aligned} P(Y = y|X = x) \;=\; & P(Y = y|D = 1, X = x)P(D = 1|X = x) \\ & + P(Y = y|D = 0, X = x)P(D = 0|X = x) \end{aligned}$$

is not identified by the sampling process alone, given that $P(D = 0|X = x) > 0$. Additional assumptions on the conditional probability model for $P(Y = y|D = 0, X = x)$ are

required for identification. Examples for partial observability of Y are item non-response and censoring (or selection). Partial observability also occurs when evaluating programs and policies, and when analyzing treatment effects.

These specification issues are not to be meant mutually exclusive, i.e., for example partial observability can be related to endogeneity if the selection status depends on the outcome variable. There are a number of ways to deal with these problems, and often the solutions are specific to the type of data available and the question(s) one wants to answer. In the most general sense, the modeling options are either to impose less structure on the data and look for the conclusions one can draw with the smaller set of assumptions, or, when remaining in a parametric world, to look for a class of very flexible models that account for the particular features of the data.

Chapter 2 of this thesis (joint work with Rainer Winkelmann) follows the latter idea. The modeling subject is the response to a survey question "How satisfied are you with your life at present, all things considered?", and answers range from "completely dissatisfied" (coded as 0) to "completely satisfied" (coded as 10) on an eleven-point scale. Y is thus measured on a discrete ordinal scale, and we use models that account for this property. The background for our research is the increasing evidence from the empirical economic and psychological literature suggesting that positive and negative well-being are more than opposite ends of the same phenomenon. We investigate such asymmetries in the effect of income on subjective well-being using a single-item measure of general life satisfaction, as opposed to multi-item analyses in the previous literature. We pinpoint the shortcomings of standard ordered response models — like the ordered probit and the ordered logit model — in analyzing such questions and offer a flexible parametric solution: a multiple-index ordered probit panel data model with varying thresholds. In the above terminology, our model prevents the effects of income on well-being to be model-induced, and we account for individual specific unobserved heterogeneity by exploiting the panel structure of the dataset (drawn from the German Socio-Economic Panel 1984-2004). Our results suggest that income has only a minor effect on high satisfaction but significantly reduces dissatisfaction.

In order to perform the analysis in an easy-to-access fashion, I have written two new estimation commands — called `goprobit` and `regoprob` — that estimate the models of Chapter 2 in Stata.[2] The details of these modules (download, syntax, output, etc.) are described in Appendices A and B.

Parametric approaches in general have often been criticized for their ambiguity in the assumptions made, and for the lack of justification by the data. As the argument goes, empirical researchers willing to take this approach will always have to defend their results, in particular to what extent they are model- rather than data-determined. An alternative to flexible parametric models are semiparametric or nonparametric approaches that impose less structure on the data, i.e., the modeling strategy is to not specify all aspects of the model but only those parts that are relevant for the analysis. In practice, these approaches include relaxing functional form assumptions and/or distributional assumptions.

Chapter 3 of the thesis can be classified as semiparametric. The modeling subject is a count data valued dependent variable, i.e., Y takes on the values $0, 1, 2, \ldots$ with or without explicit upper limit. The analysis does not impose any assumptions on the distribution of Y, but rather focuses on the conditional expectation function of Y given a vector of explanatory variables X. Due to the fact that Y must be non-negative, an exponential model with observed heterogeneity (X) and unobserved heterogeneity (U) is imposed. As previously argued, the correlation between included and omitted regressors generally causes inconsistency of standard estimators for count data models. Using a specific residual function and suitable instruments, a consistent generalized method of moments (GMM) estimator can be obtained under conditional moment restrictions. I extend this approach by fully exploiting the model assumptions and thereby improving efficiency of the resulting estimator. Empirical likelihood (EL) estimation in particular has favorable properties in this setting compared to the two-step GMM procedure, which is demonstrated in a Monte Carlo experiment. The proposed method is applied to the estimation of a cigarette demand function, and the results show that bias and precision

[2] Stata is a registered trademark of StataCorp, College Station TX, USA.

of estimators can be significantly improved using the EL approach.

The model in Chapter 4 imposes even less assumptions on the sampling process than that in Chapter 3 — though the modeling subject and the question of interest are of a very different kind. Chapter 4 deals with identification of treatment effects when the outcome variable is ordered. If outcomes are measured ordinally, previously developed methods to investigate the impact of an endogenous binary regressor (or treatment) on average outcomes cannot be applied as the expectation of an ordered variable, in its strict sense, does not exist, and a shift in focus to distributional effects is indispensable. In an ordered potential outcomes framework, the chapter discusses several kinds of treatment effects, such as versions of the average treatment effect and the average treatment effect on the treated. Without imposing a fully fledged parametric model the treatment effects are generally not point-identified because of the partial observability problem. Assuming a threshold crossing model on both the ordered potential outcomes and the binary treatment variable leaving the distribution of error terms and functional forms unspecified, it is discussed how the treatment effects can be bounded and inference on the bounds can be conducted.

Chapter 4 probably reflects most the recent trend in microeconometric analyses of "estimating features of a model rather than estimating the full model" (Heckman and Vytlacil 2001). The analysis starts with as little assumptions on the sampling process as possible and then investigates the insights for a predefined set of parameters by adding more and more assumptions. In that sense, the results are robust against misspecification and one can easily take one step back if a certain set of assumptions seems implausible in a given data situation. On the downside, the weaker the conditions imposed, the smaller the set of inferences one can draw. The results may therefore not be conclusive with respect to the research question, but in this case the researcher can decide whether to believe in additional assumptions, or to try to get better data.

References

Engle, R.F., D.F. Hendry, and J.-F. Richard (1983): "Exogeneity," *Econometrica*, 51, 277-304.

Heckman, J.J., and E.J. Vytlacil (2001): "Local Instrumental Variables," in: C. Hsiao, K. Morimune, and J. Powell (eds.) *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, Cambridge: Cambridge University Press.

Manski, C.F. (1995): *Identification Problems in the Social Sciences*, Cambridge: Harvard University Press.

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

Winkelmann, R., and S. Boes (2006): *Analysis of Microdata*, Berlin: Springer.

Chapter 2

Random Effects Generalized Ordered Probit Models with an Application to Subjective Data

Joint work with Rainer Winkelmann.
Another version of this chapter has been published as *SOI Working Paper 0605*.

2.1 Introduction

Pinning down the income elasticity of subjective well-being is one of the great challenges in the emerging field of the economics of happiness (Layard 2005, Frey and Stutzer 2002, Bruni and Porta 2006). If this line of research is to have a lasting impact on economic policy making, a reliable estimate, and understanding, of the causal effect of income on well-being (the extent to which "money can buy happiness") will be a litmus test. The recent survey by Clark *et al.* (2006) bears witness to the intensive empirical economic research undertaken in this area.

The emotional model theory of subjective well-being, developed in the early 1980s by psychologist Ed Diener, posits that individuals' appraisals of their own lives (i.e., a person's individual judgment about his current status in the world) capture the essence of well-being (Diener 1984, Diener *et al.* 1985, Diener *et al.* 1999). The literature has

identified three core components of subjective well-being: positive affect, (the lack of) negative affect, and general life satisfaction (i.e., subjective appreciation of life's rewards), separable constructs that can be independently examined. Together these three capture a broad range of hedonic and eudemonic experience.

An important early result, sometimes referred to as "well-being paradox", is that average satisfaction in a country does not increase as countries grow wealthier (Easterlin 1974, 1995). At the individual level, there is a weak positive cross-sectional association between income and satisfaction. If one follows an individual over the life-cycle however, as income first increases and then levels off, subjective well-being remains unchanged. Income expectations and aspirations matter, which means that the effect is subject to habituation and comparison (Diener and Biswas-Diener 2002, Clark and Oswald 1996, Luttmer 2005). As expected, the estimated effects differ somewhat depending on whether long-term or short-term income fluctuations are considered, whether truly exogenous variation in income is available, how exactly subjective well-being is measured, and what other controls are included in the model.

The contribution of our paper is to explore, for general life satisfaction (GLS), whether the effect of income is different in different parts of the satisfaction distribution. Is it perhaps the case that the effect of income differs for persons who are relatively dissatisfied, relative to those who report a high life satisfaction, regardless of income? Such a finding would not only improve our understanding of the mechanism underlying the GLS responses, but also add another explanation to why the overall effect is rather small although income may have a substantial effect for parts of the population. Any evaluation of the well-being consequences of economic policies would need to account for such asymmetries.

We should briefly elaborate on what we mean by "response asymmetries". In the traditional interpretation of the single item GLS scale, satisfaction is just the absence of dissatisfaction. In this view, the effect of income on satisfaction is equal to minus the effect of income on dissatisfaction. We avoid such a cardinal interpretation and rather focus on the ordering. For simplicity, consider the case where the GLS scale has only three

categories: "satisfied", "neutral" and "dissatisfied".[1] The model we consider does not impose *a priori* that factors increasing the probability of satisfaction must also reduce the probability of dissatisfaction, and vice versa. This is new, as far as we can tell, although there have been a number of related approaches.

Huppert and Whittington (2003) use the General Health Questionnaire (GHQ-30) to identify positive items. The score on these positive items is then labeled "positive well-being", whereas a standard symptom measure of psychological distress, also from the GHQ-30, is used for "negative well-being". Similarly, Headey and Wooden (2004) compare well-being from a GLS question (as used in our paper) with *ill-being* obtained from a five-item scale on mental health (i.e., capturing anxiety, depression, and the like). These studies therefore do not investigate differences in the effects of a variable, such as income, at different poles of the same scale. Our approach also differs from the large literature on positive and negative affect, spurred by Bradburn (1969), since we focus on global life satisfaction, a person's conscious evaluative judgment of life, rather than affect.

With data from the German Socio-Economic Panel 1984-2004, we find that income significantly reduces the incidence of low satisfaction but it does not increase the incidence of high satisfaction in a subsample of men living in one-person households. This finding corroborates previous evidence of asymmetric effects from multi-item analyses of subjective well-being, this time with a single-item measure of general life satisfaction.

2.2 Happiness and Income in Economics

For economists, empirical evidence on the relationship between income and subjective well-being (SWB) is important for (at least) two reasons. First, the design and evaluation of economic policies often takes income as the target quantity of interest. The idea is, of course, that income is a good proxy for well-being, and that it is easy to measure. If the link between income and well-being is less strong than suspected, then economic policies based on income (or GDP) maximization alone may turn out to be inferior from an overall

[1] The question we actually use is a response to "How satisfied are you with your life, all things considered?" on an 11-point numerical scale, where "0" is labeled "completely dissatisfied" and "10" is labeled "completely satisfied".

well-being perspective.

Second, the relationship between income and well-being may be used to put a monetary value — or shadow price — on non-traded goods, usually in the context of cost-benefit analyses. The basic idea is one of compensation: in case of a "bad", how much of an increase in income is required to offset the negative effect of the bad, while keeping the person at the same level of SWB as in the absence of the bad? Similarly, in case of a good, one can implicitly determine the shadow price by asking how much income a person would be willing to give up in order to obtain the good, keeping SWB fixed.

Examples for this line of research are Blanchflower and Oswald (2004), who estimate the pecuniary value of a lasting marriage (relative to widowhood) to be $100,000 per year. Other examples include Winkelmann and Winkelmann (1998) who estimate the money-equivalent value of the psychological cost of unemployment, a trade-off that we will come back to below, and Schwarze (2003) who uses the principle to determine an income equivalence scale, i.e., the income compensation required to keep the same level of an individual's well-being with one additional household member present. Frey *et al.* (2004) estimate the value of public safety, or the absence of terrorism. Van Praag and Baarsma (2005) measure the external cost of air traffic noise for people living near the Amsterdam Airport.

Unfortunately, the implied compensation may be sensitive to the chosen model, and too restrictive assumptions may lead to spurious estimates. An obvious concern is that the same income change has a different meaning for poor than for rich people. This concern resonates throughout the literature. Typically, it is found that the correlation between income and subjective well-being is much stronger among the poor. While the absence of poverty does not guarantee happiness, the presence often prevents it (Diener and Biswas-Diener 2002). Such non-linearities can be addressed, for instance, by studying the correlation between GLS and logarithmic income. In this case, a proportionate effect is assumed: To achieve the same increase in satisfaction, larger and larger absolute changes in income are necessary. Semiparametric estimators have provided some support for a log-linear functional form.

The topic of our paper is different. Not all poor people are dissatisfied with their

lives, nor are all wealthy people satisfied. The general life satisfaction scale integrates the subjects' reflected valuation of various domains of their lives, weighting them in whatever way they choose (van Praag *et al.* 2003). In the broadest sense, one can distinguish two domains, a pecuniary domain and a non-pecuniary domain (that includes, perhaps most importantly, health and social relationships). Our working hypothesis is that the non-pecuniary domain moderates the effect of the pecuniary domain on GLS. Specifically, if the valuation of the non-pecuniary domain contributes to a low GLS, then the effect of the pecuniary domain becomes stronger, i.e., an income increase will have a more favorable effect on GLS, compared to the case where the non-pecuniary domain leads to a high GLS score. Such a framework will lead to the aforementioned response asymmetries: income will lower dissatisfaction more than it will increase satisfaction.

To test this hypothesis, we cannot use conventional regression or ordered response models, because in these models the effect of income at various satisfaction levels cannot be estimated freely but rather is dictated by functional form, essentially a single parameter. A naive approach would be to split the scale, for example by defining the outcomes "dissatisfied" for scores below an arbitrary cut-off, and "satisfied" for values above an arbitrary cut-off, and analyzing their response patterns separately. Slightly more sophisticated approaches can be based either on a latent class framework, or on generalized ordered probit models as proposed here.

In latent class models, one can define any number of latent groups and estimate the effect of income conditional on group membership. A recent example for such an approach is the study by Clark *et al.* (2005) who used GLS data from the European Community Household Panel. They found that the effect of income changes were larger in the "latent satisfied" than in the "latent dissatisfied" classes. Here we address the issue from a different angle: Rather than inferring response asymmetries from unobservable class membership, we model them directly using an alternative approach with outcome-specific parameters, a generalized ordered probit model for panel data. The technical details of the model are discussed in the next section.

2.3 Econometric Modeling

Most empirical work on the determinants of subjective well-being uses either linear regression or single-index ordered probit and logit models. While the latter account for the discreteness and ordering of the dependent variable, they impose an implicit cardinalization such that, for example, the trade-off ratios between income and other determinants of well-being must be constant across the distribution of outcomes (Boes and Winkelmann 2006). Since we want to estimate unrestricted income effects for low and high levels of well-being, we need to use more flexible models, and the multinomial logit with its multi-index structure is certainly one option. However, this model does not make any use of the ordering information and therefore cannot be efficient. We propose a generalization of Maddala's (1983) and Terza's (1985) model to panel data instead that is comparably flexible as the multinomial logit and in addition accounts for the ordinality.

Model and Assumptions

Let $Y_{it} \in \{1, \ldots, J\}$ denote the survey response to the GLS question of individual $i = 1, \ldots, n$ at time $t = 1, \ldots, T_i$, and let X_{it} denote the vector of covariates (including logarithmic income). The relationship between Y_{it} and X_{it} is specified in terms of cumulative conditional probabilities:

$$P(Y_{it} \leq y | X_{it}; \theta_y) = \Phi(-X'_{it}\theta_y) \qquad y = 1, \ldots, J-1 \tag{2.1}$$

where $\Phi(\cdot)$ denotes the cumulative density of the standard normal distribution, and θ_y denotes a vector of category-specific parameters, including a constant.[2] The function $\Phi(\cdot)$ maps the linear index onto the unit interval, and we require $\theta = (\theta_1 \ \ldots \ \theta_{J-1})$ to fulfill the strict inequalities $X'_{it}\theta_1 > \ldots > X'_{it}\theta_{J-1}$ such that the cumulative probabilities increase with each increment in y. Due to adding up $P(Y_{it} \leq J | X_{it}) = 1$, so that we can only identify $J-1$ category-specific parameter vectors. The model reduces to the standard ordered probit model if only the constant term in θ_y is category-specific.

[2] For the ease of exposition, we set up the model in terms of cumulative conditional probabilities. Like the standard ordered probit, the generalized model may also be motivated in terms of a latent variable and a threshold crossing mechanism generating the ordinal response variable. We refer to Winkelmann and Boes (2006: Ch. 6) for a detailed outline of the underlying assumptions and identification issues in this framework. See also the appendix to this chapter for more details.

In order to exploit the advantages of panel data more fully, the model can be augmented by individual specific time invariant effects. Conditioning on such effects avoids bias if, for example, unobserved personality traits affect well-being as well as observable characteristics (Ferrer-i-Carbonell and Frijters 2004). Let η_i denote such individual effects, and rewrite the cumulative probabilities (2.1) conditional on η_i as

$$P(Y_{it} \leq y | X_{it}, \eta_i; \theta_y) = \Phi(-X'_{it}\theta_y - \eta_i) \qquad y = 1, \ldots, J-1 \tag{2.2}$$

We assume that X_{it} is strictly exogenous conditional on η_i and that outcomes are independent conditional on (X_i, η_i), where X_i contains X_{it} for all t. The first assumption rules out lagged dependent variables in X_{it}, the second assumption allows for dependencies in Y_{it} across t if conditioned only on X_i. Note that the independence assumption restricts the covariance matrix of individual effects to be diagonal, i.e., $Cov(\eta_i, \eta_{i'}) = 0 \; \forall i \neq i'$.

Without specifying the relationship between X_{it} and η_i, i.e., treating η_i as fixed parameters to be estimated along with θ, a model based on (2.2) will suffer from the incidental parameters problem. For fixed time and large cross-sectional dimension, the number of parameters η_i is unbounded, with available information on η_i being fixed, which in general yields inconsistent estimators of η_i and θ. We solve this problem by treating η_i as random variable drawn along with (X_i, Y_i). Following the idea of Chamberlain (1980) and Mundlak (1978) we allow for possible correlation between η_i and X_i:

$$\eta_i = \bar{X}'_i\gamma + \alpha_i \tag{2.3}$$

where $\bar{X}_i$ is the vector of averages of X_{it} over time, γ is a conformable parameter vector, and α_i is an orthogonal error with $\alpha_i | X_i \sim Normal(0, \sigma^2_\alpha)$.[3] The distributional assumption and the independence ensure that the correlation matrix of the random effects is the identity matrix. If we replace η_i in (2.2) by (2.3), then we obtain

$$P(Y_{it} \leq y | X_{it}, \bar{X}_i, \alpha_i; \theta_y, \gamma) = \Phi(-X'_{it}\theta_y - \bar{X}'_i\gamma - \alpha_i) \qquad y = 1, \ldots, J-1 \tag{2.4}$$

or in terms of a conditional probability model for all $y = 1, \ldots, J$

$$P(Y_{it} = y | X_{it}, \bar{X}_i, \alpha_i; \theta, \gamma) = \Phi(-X'_{it}\theta_y - \bar{X}'_i\gamma - \alpha_i) - \Phi(-X'_{it}\theta_{y-1} - \bar{X}'_i\gamma - \alpha_i) \tag{2.5}$$

[3] A straightforward generalization of (2.3) would be to let γ vary by the satisfaction levels, i.e., replace γ by γ_y. Computationally somewhat more involved would be to let α_i vary by the satisfaction level. Note that only time-varying covariates are included in $\bar{X}_i$ because otherwise θ_y and γ would not be separately identified.

where α_i is the individual specific time invariant random effect, and it is understood that $\Phi(-X_{it}'\theta_0 - \bar{X}_i'\gamma - \alpha_i) = 0$ and $\Phi(-X_{it}'\theta_J - \bar{X}_i'\gamma - \alpha_i) = 1$ due to identification and adding up to one. The joint distribution of $Y_i = (Y_{i1}, \ldots, Y_{iT_i})$ conditional on observables but unconditional on α_i is obtained by integrating the joint distribution of Y_i and α_i over α_i,

$$
\begin{aligned}
&f(y_{i1}, \ldots, y_{iT_i} | x_i; \theta, \gamma, \sigma_\alpha) \\
&\quad = \int_{-\infty}^{\infty} \prod_{t=1}^{T_i} \prod_{y=1}^{J} P(Y_{it} = y | X_{it}, \bar{X}_i, \alpha_i; \theta, \gamma)^{\mathbf{1}(Y_{it}=y)} \frac{1}{\sigma_\alpha} \phi\left(\frac{\alpha_i}{\sigma_\alpha}\right) d\alpha_i
\end{aligned} \tag{2.6}
$$

where $\mathbf{1}(\cdot)$ is the logical indicator function. The inner product over all J categories selects the appropriate likelihood contribution for each observation (individual i at time t) according to the observed category, and the independence of Y_{it} conditional on (X_i, α_i) ensures that the joint probability of $(Y_{i1}, \ldots, Y_{iJ})|(X_i, \alpha_i)$ can be written as the product of single probabilities over all periods T_i. The integral in (2.6) does not have a closed form solution, but it can be rewritten in a form amenable to Gauss-Hermite quadrature for numerical approximation.[4]

Estimation of parameters by maximum likelihood is straightforward once the integral has been evaluated, and the resulting estimator is consistent, efficient, and approximately

[4] The normality assumption implies that the density function of α_i is given by

$$
f(\alpha_i) = \frac{1}{\sqrt{2\pi}\sigma_\alpha} \exp\left(-\frac{\alpha_i^2}{2\sigma_\alpha^2}\right) = \frac{1}{\sigma_\alpha} \phi\left(\frac{\alpha_i}{\sigma_\alpha}\right)
$$

Now consider a change of variables from α_i to $\xi_i = \alpha_i/(\sqrt{2}\sigma_\alpha)$. The inverse of ξ_i is $\alpha_i = \xi_i\sqrt{2}\sigma_\alpha$ with Jacobian $\sqrt{2}\sigma_\alpha$, so that the density of ξ_i can be derived as

$$
g(\xi_i) = \frac{1}{\sqrt{\pi}} \exp\left(-\xi_i^2\right)
$$

Define a function $h(\xi_i)$ as

$$
h(\xi_i) = \prod_{t=1}^{T_i} \prod_{y=1}^{J} P(Y_{it} = y | X_{it}, \bar{X}_i, \xi_i; \theta, \gamma, \sigma_\alpha)^{\mathbf{1}(Y_{it}=y)} / \sqrt{2}
$$

where the probabilities are given by (2.5) with α_i replaced by $\xi_i\sqrt{2}\sigma_\alpha$. The Gauss-Hermite approximation of the integral in (2.6) with M points is then given by

$$
\int_{-\infty}^{\infty} h(\xi_i) \exp\left(-\xi_i^2\right) d\xi_i \approx \sum_{m=1}^{M} w_m h(a_m)
$$

where w_m and a_m denote quadrature weights and abscissas, respectively. The approximation is the better the larger the number of points M. Abramowitz and Stegun (1964: 924) provide tables of (w_m, a_m) for different M.

normally distributed. The generalized ordered probit model with random effects specification has been implemented in a new Stata module called `regoprob` available via the `ssc` commands in Stata.[5]

Interpretation of the Model

There are a number of ways to interpret the estimated parameters, but we focus here on two quantities that offer a very intuitive interpretation when dealing with conditional probability models. First, we may ask the question "How does a *ceteris paribus* change in income affect the distribution of GLS responses?" which is answered by marginal probability effects (MPE's). Such effects are of particular interest for the asymmetry hypothesis since we are able to identify whether income effects on GLS differ for low and high GLS. Second, we may look at asymmetric effects from a different (probability) angle insofar as we do not investigate the change in the GLS distribution at different poles, but instead we keep the GLS distribution fixed and analyze income changes required to compensate for a change in another covariate, thereby distinguishing between trade-offs for low and and high GLS.

MPE's are defined as first derivatives of (2.5) with respect to the variable(s) of interest. Since α_i is an unobserved random variable, we cannot directly calculate the MPE's without further assumptions. One possibility would be to take advantage of the probit form and the normality of α_i and rewrite the conditional probabilities marginal on α_i as

$$\begin{aligned} P(Y_{it} = y|X_{it}, \bar{X}_i; \theta, \gamma, \sigma_\alpha) &= \Phi\left(\frac{-X_{it}'\theta_y - \bar{X}_i'\gamma}{\sqrt{1+\sigma_\alpha^2}}\right) - \Phi\left(\frac{-X_{it}'\theta_{y-1} - \bar{X}_i'\gamma}{\sqrt{1+\sigma_\alpha^2}}\right) \\ &= \Phi\left(-X_{it}'\vartheta_y - \bar{X}_i'\psi\right) - \Phi\left(-X_{it}'\vartheta_{y-1} - \bar{X}_i'\psi\right) \end{aligned} \tag{2.7}$$

where $\vartheta_y = \theta_y(1+\sigma_\alpha^2)^{-1/2}$ and $\psi = \gamma(1+\sigma_\alpha^2)^{-1/2}$ denote the population-averaged coefficient vectors. The coefficients are called population-averaged since they are obtained as the expectation of (2.5) over α_i. Taking derivatives of (2.7) yields

$$MPE_y^{(l)} = \frac{\partial P(Y_{it} = y|X_{it}, \bar{X}_i; \vartheta, \psi)}{\partial X_{it}^{(l)}}$$

[5] Stata is a registered trademark of StataCorp, College Station TX, USA. Type `net search regoprob` or `ssc install regoprob` in the command line of Stata to find out more about `regoprob`. See also Appendices A and B for details on the command syntax and the output generated by Stata.

$$= \phi(-X_{it}'\vartheta_{y-1} - \bar{X}_i'\psi)\vartheta_{y-1}^{(l)} - \phi(-X_{it}'\vartheta_y - \bar{X}_i'\psi)\vartheta_y^{(l)} \tag{2.8}$$

where $\phi(\cdot)$ denotes the density function of the standard normal distribution, and $X_{it}^{(l)}$ denotes the l-th element in X_{it} (here assumed to be logarithmic income) and $\vartheta_y^{(l)}$ the corresponding scaled (income) coefficient. The MPE's are functions of the covariates and therefore depend on the values of X_{it} and $\bar{X}_i$. We estimate the MPE's replacing the unknown coefficients by the maximum likelihood estimates and evaluating at the sample averages of the regressors.

The second quantity of interest, the trade-off ratio, assesses the importance of income *relative* to other determinants. It follows from totally differentiating (2.7) that

$$dP(Y_{it} = y|X_{it}, \bar{X}_i; \vartheta, \psi) = MPE_y^{(l)} dX_{it}^{(l)} + MPE_y^{(m)} dX_{it}^{(m)} \tag{2.9}$$

where $X_{it}^{(l)}$ denotes logarithmic income, $X_{it}^{(m)}$ denotes any other covariate in X_{it}, and the MPE's are given by (2.8). The approximation in (2.9) directly leads to the concept of compensating variation: How much of a variation in one regressor (here income) is needed to offset the given change in another regressor such that $dP(Y_{it} = y|X_{it}, \bar{X}_i; \vartheta, \psi) = 0 \; \forall y$, i.e., all probabilities remain unchanged. Rearranging terms yields

$$\frac{dX_{it}^{(l)}}{dX_{it}^{(m)}} = -\frac{MPE_y^{(m)}}{MPE_y^{(l)}} \tag{2.10}$$

In the standard model, this trade-off ratio reduces to the ratio of coefficients which does not vary across outcomes, whereas in the generalized model such an restriction is not imposed. Rather, we can let the data speak and determine empirically how these trade-off ratios look like.

2.4 Data

The German Socio-Economic Panel (GSOEP) is a large annual panel survey of randomly selected households in Germany (see Burkhauser *et al.* 2001 for more details). Personal information is available for all household members aged 16 and above. Our data are drawn from the West German (A) subsample 1984-2004, yielding a maximum of 21 observations per individual (on average about five observations per individual). We apply a number of

standard selection criteria: included individuals are between 25 and 65 years old at the time of the survey, and we require non-missing information on all the included variables.[6]

In addition, we employ a novel restriction by considering single person households only. The rationale for this selection is that the match between reported household income and individual material well-being is much better in single-person households than we could possibly hope for in a multi-person household. General household surveys such as the GSOEP typically include two types of income measures, one being total household income (from all sources), the other being personal labor earnings. Clearly, personal labor earnings are not a very good indicator of material well-being, in particular, but not only, for persons who do not work, as it does not include any government transfers (e.g., child benefit, government grants, or rent subsidies). Household income (net of taxes and social security contributions) is in general a more appropriate measure. However, in multi-person households, there remain two types of ambiguities. First, there is an ongoing debate on the right equivalence scale in order to reflect economies of scale in household production and consumption. Secondly, we do not know whether resources are shared evenly within the household, but such an (arbitrary) assumption is required when assigning one income to several household members.

For these reasons, we find it instructive to study the relationship between income and SWB in the (reference) population of single person households. We do not claim that such a sample is representative for the whole population, and of course, this raises the question of external validity: To what extent can results for single person households be extrapolated to the population of all households? While single person households are non-representative with respect to a number of factors (such as age, and possibly also income), we controlled for this in our analysis, and it is *a priori* unclear why the well-being function (after including these factors) should be different for such persons.

All in all, this approach leaves us with 5008 person-year observations for men, and with 4727 person-year observations for women. The dependent variable is, as mentioned before, the response to the survey question "How satisfied are you with your life, all things

[6] The variables we include in the model generally have very high response rates with missing information for only a few respondents, in particular for the GLS variable. We therefore do not expect significant bias in the results from dropping these observations.

considered?". There are relatively few responses in the 0-2 range. For this reason, and to preserve some degrees of freedom (a full set of regression parameters is added for each additional category), we use a modified scale where the original 0-2 responses have been grouped into the lowest "dissatisfied" category.

Figure 2.1 depicts the frequency distribution of GLS responses in our sample, separately for men and women. People are mostly satisfied with their life: about two thirds report a GLS level of seven or higher, and women have a slightly higher average GLS level than men. The distribution in Figure 2.1 is characteristic of most SWB distributions in the sense that the majority of people reports a relatively high level of GLS, although the highest response category is chosen relatively infrequently.

— Insert Figure 2.1 about here —

In the regression analysis, control variables include — apart from logarithmic income — a second order polynomial in age and dummy variables for unemployment and health status. We use a relatively simple specification with only a few variables, which has two main advantages. First, since eight regression parameters are estimated for each variable, fewer regressors keep the model manageable. Second, many of the additional variables used in the previous literature are arguably endogenous choice variables, obstructing the interpretation of the results. Finally, all analyses are performed separately by gender.

Table 2.1 summarizes the sample means of the explanatory variables. Among one-person households, men have a significantly higher monthly income than women (about 260 Euros) and are on average more than five years younger. The unemployment rate is about 2.5 percentage points higher for men than for women, and 58.2 percent of the women are relatively satisfied with their health status (compared to 65.6 percent of the men). These variations can largely be explained by the different age distributions of single male and single female households. Men are mostly living alone when they are young and at the beginning of their career path. Women are more likely to live alone when they are older, contributing factors being a higher incidence of widowhood due to greater life-expectancy.

— Insert Table 2.1 about here —

Table 2.2 cross-tabulates the sample means of the dependent variable conditional on the GLS response, again separately for men and women. The income variable shows a lot of variation along the GLS dimension. For men (panel A), the lowest average monthly income (1124 Euro) is observed for individuals with very low GLS, the highest income (1519 Euro) for those with response "8". When moving from the utmost left part of the GLS distribution to the right, average income is first increasing then decreasing. A similar pattern can be observed for women (panel B), although on a lower level. Concerning unemployment and health, we find that among less satisfied people the unemployment rate is relatively high and that reported health status and GLS are positively correlated.

— Insert Table 2.2 about here —

2.5 Estimation Results

In this section, we report on the estimation results of the relationship between income and subjective well-being, the latter measured by general life satisfaction. We first present the estimated income parameters under several model assumptions, then turn our attention to the implications with respect to the asymmetry hypothesis, and finally discuss the robustness of our results.

We estimated two different models: A random effects ordered probit model (OProbit) including group means as additional regressors, and a generalized random effects ordered probit model (GOProbit), also including group means, where all parameters are outcome-specific. In both cases, the pooled models were clearly rejected against the panel models, which is reflected in Table 2.3 where we report the estimated variances (and standard errors) of the random effects, $\hat{\sigma}_\alpha^2$, separately for men and women. Furthermore, a joint significance test of the group means as additional regressors rejected the null hypothesis of zero correlation, and thus a simple random effects specification without $\bar{X}_i$ is rejected by the data as well.

— Insert Table 2.3 about here —

Table 2.4 displays the estimated coefficients on logarithmic income and unemployment separately for men (panel A) and women (panel B). Although the raw parameters are not very interesting *per se*, the comparison is useful for understanding our later results. For men, we find a positive and significant income parameter in the standard model (0.362 with z-value 6.67). In the generalized model, eight different parameters are estimated. The income coefficients are slightly higher for the parameter vectors θ_1 to θ_6 than the overall estimate in the standard model. The point estimate decreases but is still significant for θ_7, and finally turns negative and insignificant for θ_8. The estimated coefficients in the sample of women are smaller (in absolute value) and less significant than those for men indicating a weaker relative impact. For example, in the standard model we obtain an income point estimate of 0.131, which is only about a third of that for men, and the z-value decreases to 1.97. In the generalized model the income coefficients are significant on the 5%-level only for θ_4 and θ_5, while all other income coefficients are insignificant. For the unemployment coefficient we obtain point estimates for low/high satisfaction that are smaller/higher (in absolute terms) than the overall estimate in the standard model, for women we observe the opposite pattern.

— Insert Table 2.4 about here —

If we formally test the generalized ordered probit model against the standard model, we can reject the null hypothesis of equal slope parameters for men ($LR_{203} = 548.9$) and for women ($LR_{203} = 430.1$). The null hypothesis of equal income coefficients is also rejected for both, men and women, and equal unemployment coefficients is only rejected for women.

In order to interpret the estimated parameters and evaluate the effects of income on low and high GLS we now turn to the quantities introduced in Section 2.3 and the marginal probabilities first. Table 2.5, Figures 2.2 and 2.3 summarize the MPE's of income and unemployment by gender. Consider, for example, the results for men and take the *ceteris paribus* effect of an increase in logarithmic household income by a small amount on the probability of responding a GLS level of "8" (analogous interpretation applies to the effects at all other GLS levels). Table 2.5 shows a value of 0.059 for the standard

model. This means that the probability of a response of "8" increases by 0.059 percentage points if we increase logarithmic income by 0.01, which corresponds approximately to a one-percent increase in level income. A doubling of income, i.e., a change in logarithmic income by 0.693, increases the probability of response "8" by about $0.059 \times 0.693 \times 100$, or about 4.09 percentage points, ceteris paribus.

— Insert Table 2.5, Figures 2.2 and 2.3 about here —

Comparing the MPE's among the standard and the generalized models and over all possible outcomes, we obtain the following pattern. For men all models suggest that more income significantly reduces the probability of low GLS (0-5), and significantly increases the probability of response "8". For high GLS responses (9-10), the standard model predicts a significant positive effect, whereas the generalized model does not predict an effect significantly different from zero. Thus, based on the generalized ordered probit model, there is no evidence for income to have an effect on high satisfaction. Moreover, the effect of income is asymmetric: higher income decreases the probability of dissatisfaction, but it does not affect the probability of high satisfaction. Figure 2.2 illustrates the asymmetric effects and shows the differences between the MPE's in the standard ordererd probit model and the generalized ordered probit model.

For women the relationship between income and GLS is relatively weak. While the standard model finds small but significant effects for low and high GLS, the generalized model predicts a significant negative effect only for responses "5" and "6". Concerning unemployment, we find evidence for men that an increased unemployment probability reduces the probability of response "8", or higher, and increases the probability of low responses, but for women the relationship is less clear. For example, an increase in the probability of being unemployed by one percentage point reduces the probability of response "8" by about 0.096 percentage points for men, and raises the probability of the same outcome by about 0.051 percentage points for women. The gender difference might be explained by social norms that assign the role of primary income earner to men and therefore make income a relatively more important determinant of male well-being (e.g., Lalive and Stutzer 2004). Such a gender difference can also be observed when considering

unemployment.

The relationship between GLS, income, and unemployment, for men and women, at various parts of the GLS distribution can alternatively be illustrated by the trade-off ratios. Table 2.6, Figures 2.4 and 2.5 show the required changes in logarithmic income if the unemployment probability increases by one percentage point, given the GLS distribution is fixed. If we want to interpret the reported numbers, we need to be careful with respect to the significance of MPE's. The trade-off ratio does only make sense for significant income effects. In this case, the required change in income is either zero if the MPE of unemployment is statistically not different from zero, or the change is positive (or negative) for significant unemployment effects. We marked the four cases (non-sensible/zero/positive/negative) with $\times / \circ / + / -$.

— Insert Table 2.6, Figures 2.4 and 2.5 about here —

The numbers in Table 2.6 (multiplied by 100) approximate the percentage change in income, e.g., for men in the standard model a 0.019 means that income must increase by 1.9 percent to offset the increase in the unemployment probability by one percentage point. By construction, the trade-off ratios in the ordered probit model are constant for all levels of GLS, and interpretation therefore is not particularly interesting. In the generalized model, required income changes vary between 0.6 and 4.2 percent. An important observation is that income compensations are entirely ineffective for men with high GLS, and effective for medium to low satisfied men, though in an unsystematic way. For women, a compensation for unemployment in terms of income is rather unpromising, and other factors determining GLS need to be identified when looking for effective compensation schemes. Figures 2.4 and 2.5 provide a graphical illustration of the results.

While these results are obtained for a specific sample and a specific parametric model with its set of assumptions, we found a remarkable robustness of the main conclusions with respect to alternative specifications and samples. Possible alternatives include the use of different link functions (rather than the probit ones), including the logit, the log-logistic, and the complementary log-log; we estimated a series of binary models, where the dependent variables result from dichotomization of GLS responses, i.e., $Y_{it} > 2$ against $Y_{it} \leq 2$,

$Y_{it} > 3$ against $Y_{it} \leq 3$, and so on; conditioning on fixed effects using Chamberlain's (1982) conditional logit model; and possible endogeneity of income in the GLS equation. We could not find evidence for endogeneity. Neither provided alternative link functions a better fit, nor did the response asymmetry for men disappear under the alternative model assumptions.

2.6 Conclusion

The distinction between positive and negative well-being has been made for some time now. Huppert and Whittington (2003) point out that the determinants of positive and negative well-being are not necessarily the same. For example, in their study of participants in the British Health and Lifestyle Survey, paid employment was found to be an important determinant of positive well-being but to have less influence on psychological symptoms. Headey and Wooden (2004) also use two separate measures of well-being and ill-being. In their case, the pecuniary situation, captured through income and wealth, was found to affect both aspects equally.

Our paper takes a different approach. We also study the determinants of well-being, in particular the effect of income. However, we use a single item scale of general life satisfaction, where low scores are interpreted as a state of "dissatisfaction" and high scores signify "satisfaction". There are a number of advantages of such a single measure. It is widely available, and it allows for a straightforward computation of compensating income variations, an important application of this type of modeling in economics. We therefore propose a new and very flexible panel data model in which we can analyze whether income effects depend on the level of satisfaction. The model allows for individual specific effects and outcome-specific parameters, i.e., the effect of income on GLS may be non-monotonic.

In a sample of men in single-person households drawn from the German Socio-Economic Panel waves 1984 to 2004, we find support for the existence of asymmetric income effects. Based on our results, income has a large effect among men with low GLS responses, but no effect on men with high GLS responses. For women in single-person households income plays a minor role in the formation of GLS, and support of the asymmetry hypothesis is

rather weak.

Clearly, more research is needed in this area. We think that our methodological focus on flexible estimation of marginal probability effects and trade-off ratios with a single measure of well-being, namely general life satisfaction, should prove useful in further investigations. If one wants to estimate marginal probability effects and compensating variations in a meaningful way, then one should use the generalized ordered probit model rather than the simpler models prevailing in earlier research.

Tables and Figures

Figure 2.1: Marginal Distribution of Satisfaction Responses

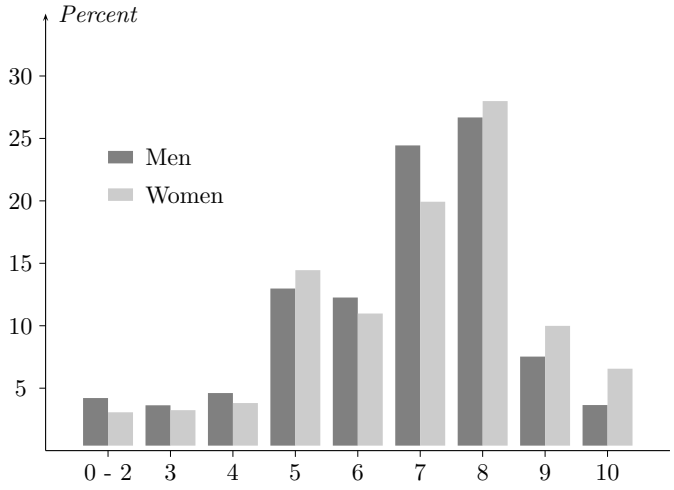

Figure 2.2: Marginal Probability Effects of Logarithmic Income — Men

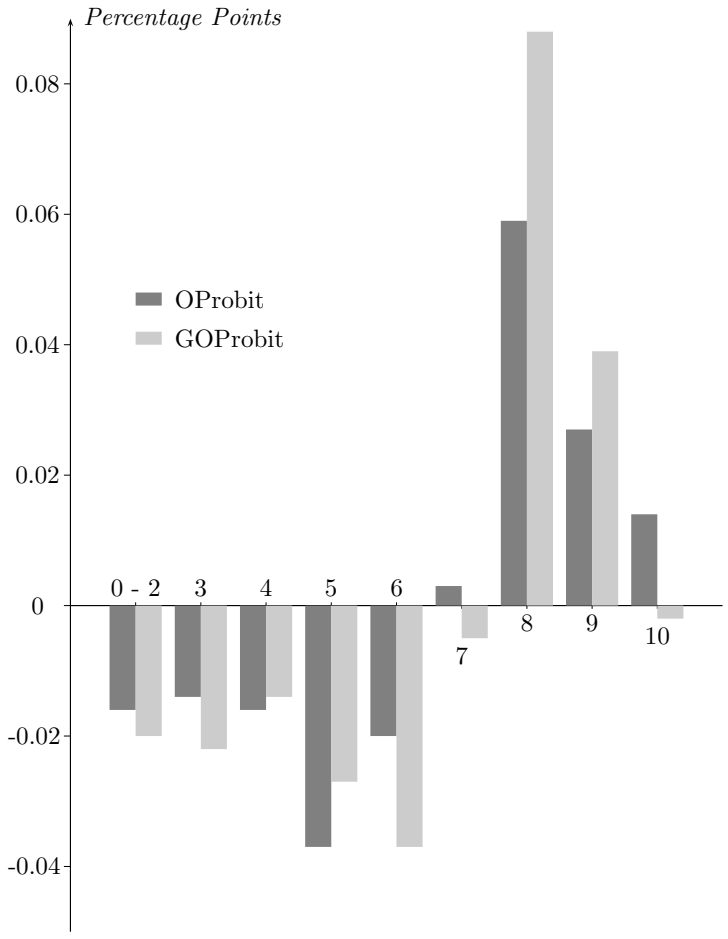

Figure 2.3: Marginal Probability Effects of Logarithmic Income — Women

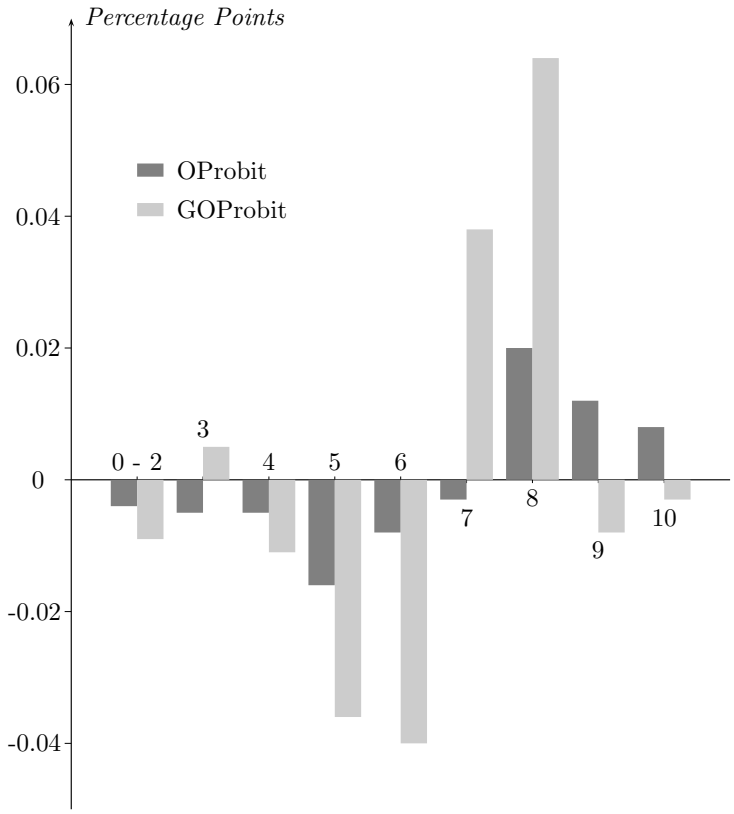

Figure 2.4: Trade-Off Ratios between Income and Unemployment — Men

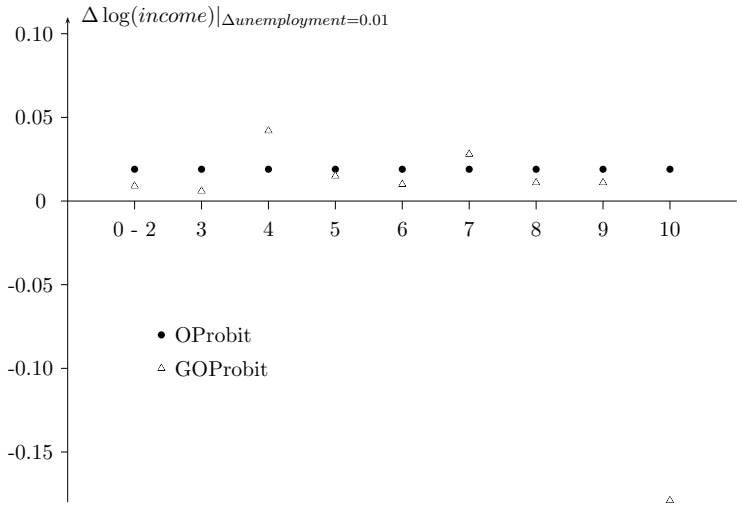

Figure 2.5: Trade-Off Ratios between Income and Unemployment — Women

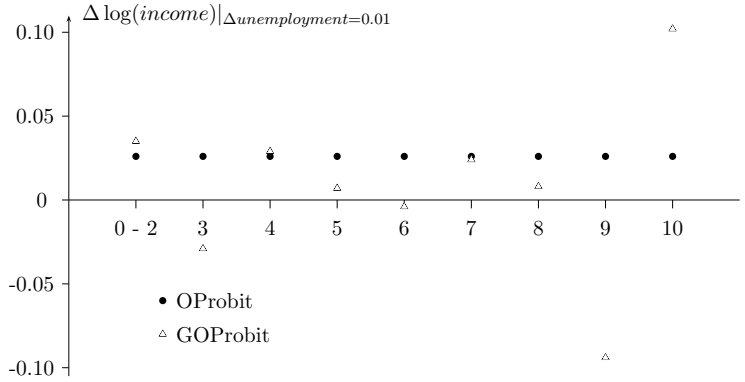

Table 2.1: Descriptive Statistics by Gender

	Men		Women	
Variable	Mean	Std.Err.	Mean	Std.Err.
Monthly income in EUR	1403.5	12.0	1140.9	10.3
Age in years	40.24	0.16	45.80	0.20
Unemployment (0/1)	0.083	0.004	0.058	0.003
Good health (0/1)	0.656	0.007	0.582	0.007
Number of Obs.	5008		4727	

Table 2.2: Sample Means by Gender and Satisfaction Level

	GLS Level								
Variable	0 - 2	3	4	5	6	7	8	9	10
A. Men									
Relative Freq.	4.21%	3.63%	4.61%	12.98%	12.26%	24.44%	26.68%	7.53%	3.65%
Income	1123.9	1152.0	1414.3	1255.8	1324.0	1477.9	1519.3	1473.6	1267.8
Age	43.86	41.83	41.07	43.55	40.32	38.99	38.91	38.19	43.79
Unemployment	0.336	0.176	0.182	0.126	0.103	0.056	0.030	0.029	0.022
Good Health	0.336	0.319	0.338	0.340	0.549	0.732	0.841	0.897	0.896
B. Women									
Relative Freq.	3.07%	3.24%	3.81%	14.45%	10.98%	19.93%	27.99%	9.99%	6.56%
Income	930.5	935.4	1047.4	978.1	1055.7	1196.7	1238.6	1290.7	1082.1
Age	47.50	45.75	45.59	49.28	46.25	43.37	44.62	45.34	49.84
Unemployment	0.234	0.124	0.172	0.089	0.052	0.036	0.039	0.013	0.035
Good Health	0.159	0.196	0.267	0.274	0.420	0.601	0.767	0.847	0.845

Table 2.3: Estimated Variances of the Random Effects by Gender and Model

	OProbit	GOProbit
Men	0.785	0.833
	(0.184)	(0.212)
Women	0.666	0.708
	(0.150)	(0.164)

Notes: The models are the ordered probit (OProbit) and the generalized ordered probit (GOProbit). Standard errors in parentheses.

Table 2.4: Estimated Income and Unemployment Coefficients by Gender and Model

	OProbit	GOProbit						
	overall	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6	θ_7	θ_8
A. Men									
Logarithmic Income	0.362	0.417	0.470	0.423	0.372	0.429	0.456	0.314	-0.066
	(0.054)	(0.141)	(0.109)	(0.093)	(0.082)	(0.079)	(0.082)	(0.115)	(0.152)
Unemployment	-0.696	-0.396	-0.351	-0.681	-0.587	-0.597	-0.661	-0.724	-1.189
	(0.077)	(0.173)	(0.147)	(0.125)	(0.111)	(0.111)	(0.124)	(0.189)	(0.297)
B. Women									
Logarithmic Income	0.131	0.256	0.060	0.144	0.234	0.329	0.175	-0.070	-0.045
	(0.067)	(0.225)	(0.144)	(0.131)	(0.108)	(0.100)	(0.098)	(0.117)	(0.146)
Unemployment	-0.347	-0.893	-0.635	-0.720	-0.463	-0.300	0.027	-0.256	0.458
	(0.096)	(0.218)	(0.185)	(0.167)	(0.147)	(0.139)	(0.140)	(0.198)	(0.244)

Notes: The models are the ordered probit (OProbit) and the generalized ordered probit (GOProbit). Each model controls for a quadratic form in age, good health (0/1), and time fixed effects. Individual effects are assumed to be decomposable into a linear function of individual group means and orthogonal error, and the likelihood for each individual is approximated using Gauss-Hermite quadrature. Standard errors in parentheses.

Table 2.5: Marginal Probability Effects of Income and Unemployment by Gender and Satisfaction Level

		Satisfaction Level								
		0 - 2	3	4	5	6	7	8	9	10
A. Men										
Logarithmic Income	OProbit	-0.016	-0.014	-0.016	-0.037	-0.020	0.003	0.059	0.027	0.014
		(0.003)	(0.001)	(0.001)	(0.003)	(0.009)	(0.003)	(0.009)	(0.005)	(0.005)
	GOProbit	-0.020	-0.022	-0.014	-0.027	-0.037	-0.005	0.088	0.039	-0.002
		(0.007)	(0.006)	(0.004)	(0.005)	(0.006)	(0.007)	(0.033)	(0.109)	(0.089)
Unemployment	OProbit	0.031	0.026	0.031	0.070	0.039	-0.005	-0.114	-0.051	-0.028
		(0.005)	(0.002)	(0.001)	(0.005)	(0.012)	(0.005)	(0.014)	(0.008)	(0.009)
	GOProbit	0.019	0.012	0.058	0.040	0.037	0.014	-0.096	-0.044	-0.041
		(0.009)	(0.007)	(0.009)	(0.010)	(0.007)	(0.006)	(0.027)	(0.023)	(0.013)
B. Women										
Logarithmic Income	OProbit	-0.004	-0.005	-0.005	-0.016	-0.008	-0.003	0.020	0.012	0.008
		(0.002)	(0.001)	(0.001)	(0.005)	(0.012)	(0.003)	(0.011)	(0.004)	(0.006)
	GOProbit	-0.009	0.005	-0.011	-0.036	-0.040	0.038	0.064	-0.008	-0.003
		(0.008)	(0.010)	(0.008)	(0.021)	(0.022)	(0.029)	(0.030)	(0.016)	(0.011)
Unemployment	OProbit	0.011	0.012	0.014	0.042	0.020	0.007	-0.052	-0.032	-0.022
		(0.003)	(0.001)	(0.001)	(0.008)	(0.017)	(0.004)	(0.017)	(0.006)	(0.009)
	GOProbit	0.030	0.013	0.031	0.026	-0.018	-0.091	0.051	-0.077	0.034
		(0.008)	(0.010)	(0.014)	(0.029)	(0.031)	(0.035)	(0.046)	(0.026)	(0.018)

Notes: See notes Table 2.4. The marginal probability effects have been calculated for logarithmic income and unemployment, evaluated at the sample means of the explanatory variables and marginal on the individual effect. An increase in income by one percent corresponds to an increase in logarithmic income by 0.01, i.e., reported numbers can be interpreted directly as percentage point changes. Similarly, if changes in the unemployment probability by 0.01 are considered, then the reported numbers directly give percentage point changes. Approximate standard errors (delta method) in parentheses.

Table 2.6: Trade-Off Ratios Between Income and Unemployment

	Satisfaction Level								
	0 - 2	3	4	5	6	7	8	9	10
A. Men									
OProbit	0.019^{+}	0.019^{+}	0.019^{+}	0.019^{+}	0.019^{+}	$0.019^{\times}$	0.019^{+}	0.019^{+}	0.019^{+}
	(0.004)	(0.001)	(0.001)	(0.002)	(0.011)	(0.036)	(0.004)	(0.005)	(0.007)
GOProbit	0.009^{+}	0.006^{+}	0.042^{+}	0.015^{+}	0.010^{+}	$0.028^{\times}$	0.011^{+}	$0.011^{\times}$	$-0.179^{\times}$
	(0.006)	(0.004)	(0.014)	(0.005)	(0.009)	(0.041)	(0.006)	(0.032)	(6.976)
B. Women									
OProbit	0.026^{+}	0.026^{+}	0.026^{+}	0.026^{+}	$0.026^{\times}$	$0.026^{\times}$	0.026^{+}	0.026^{+}	$0.026^{\times}$
	(0.017)	(0.001)	(0.002)	(0.011)	(0.052)	(0.033)	(0.019)	(0.011)	(0.024)
GOProbit	$0.035^{\times}$	$-0.029^{\times}$	$0.029^{\times}$	0.007°	-0.004°	$0.024^{\times}$	-0.008°	$-0.094^{\times}$	$0.102^{\times}$
	(0.034)	(0.102)	(0.052)	(0.004)	(0.004)	(0.122)	(0.073)	(1.420)	(0.778)

Notes: See notes Table 2.4. The trade-off ratios show the required change in logarithmic income to compensate for an increase in the unemployment probability by one percentage point, fixing the probability of a GLS response. The ratio of significant (at the 10% level) marginal income and unemployment effects is marked $+/-$ (if positive/negative). If the marginal income effect is insignificant, the ratio is marked $\times$. If the income effect is significant but the unemployment effect is not, the ratio is marked $\circ$. Approximate standard errors (delta method) in parentheses.

Technical Appendix: Generalized Ordered Response Models

This appendix characterizes ordered response models in greater detail than does the main part of Chapter 2, where attention is paid to general model formulas and not to their derivation. The first part motivates the standard ordered probit and logit models in a latent variable framework — the common approach in the economic literature — and considers interpretation of model parameters. In the second part, extensions of the standard model are presented, and their advantages and disadvantages are critically evaluated.

Standard Model

Let Y denote an ordinal dependent variable taking J different outcomes coded, without loss of generality, in a rank preserving manner with $1, \ldots, J$. The values y of Y are determined by a latent variable Y^* and a partition of the real line:

$$Y = y \quad \text{if and only if} \quad \kappa_{y-1} < Y^* \leq \kappa_y, \qquad y = 1, \ldots, J$$

where $\kappa_0, \ldots, \kappa_J$ denote threshold parameters, and it is understood that $\kappa_0 = -\infty$ and $\kappa_J = \infty$. The threshold parameters are assumed to fulfill an order restriction, formally $\kappa_1 < \ldots < \kappa_{J-1}$, so that higher values of Y are associated with a higher latent Y^* (given the thresholds). Alternatively, the threshold mechanism determining Y from Y^* may be written compactly as

$$Y = \sum_{y=1}^{J} y \mathbf{1}(\kappa_{y-1} < Y^* \leq \kappa_y)$$

where $\mathbf{1}(a)$ denotes the logical indicator function that equals one if a is true and zero otherwise. The model is completed by assuming that

$$Y^* = X'\beta + U$$

where X is a $k \times 1$ vector of covariates (excluding a constant), β is a conformable vector of parameters, and U is an error term.

In standard ordered response models it is assumed that the thresholds are parameters to be estimated along with β. Estimation of the $J - 1 + k$ parameters by maximum likelihood is straightforward once a distribution function $F_{U|X}$ has been specified. For notational simplicity, I will suppress the subscription $U|X$ in the following. The likelihood contributions are of the form

$$\begin{aligned} P(Y = y|X) &= P(\kappa_{y-1} < Y^* \leq \kappa_y|X) \\ &= P(\kappa_{y-1} - X'\beta < U \leq \kappa_y - X'\beta|X) \\ &= P(U \leq \kappa_y - X'\beta|X) - P(U \leq \kappa_{y-1} - X'\beta|X) \\ &= F(\kappa_y - X'\beta) - F(\kappa_{y-1} - X'\beta) \end{aligned} \tag{2.11}$$

The ordered logit model and the ordered probit model are obtained by substituting $F(\cdot)$ with the cumulative density function of the standard logistic and the standard normal distribution, respectively. Note that this specification includes a location and a scale normalization in order to identify the parameters of the model. The location is fixed by excluding the constant from X, the scale is fixed by using a normalized error term variance. Alternative normalizations are possible, but the ones imposed here are most common. For a sample of n independent observations (y_i, x_i), $i = 1, \ldots, n$, the log-likelihood function is given by

$$\ln L(\beta, \kappa_1, \ldots, \kappa_{J-1}) = \sum_{i=1}^{n} \sum_{y=1}^{J} \mathbf{1}(y_i = y) \ln[F(\kappa_y - x_i'\beta) - F(\kappa_{y-1} - x_i'\beta)]$$

and maximization of this function over the parameters yields the maximum likelihood estimators $\hat{\beta}, \hat{\kappa}_1, \ldots, \hat{\kappa}_{J-1}$. For more details about the standard model and estimation see McKelvey and Zavoina (1975) and McCullagh (1980).

Interpretation

There are a number of ways to interpret the parameters of this model. What does it mean for an element of β to be "large" or "small"? First, one might be tempted to interpret the coefficients in terms of the latent model for Y^*, since this part of the model has a simple linear form. However, the β's are only identified up to scale. Moreover, Y^*, being an artificial construct, is not of much interest *per se*. Potentially more interesting is a comparison based on a "compensating variation". Let $X^{(l)}$ denote the l-th element of the vector of covariates and $\beta^{(l)}$ the corresponding parameter. Now consider changing two covariates $X^{(l)}$ and $X^{(m)}$ at the same time such that $\Delta Y^* = 0$ (and therefore all probabilities are unchanged). This requires

$$\beta^{(l)}\Delta X^{(l)} = -\beta^{(m)}\Delta X^{(m)} \qquad \text{or} \qquad \frac{\Delta X^{(l)}}{\Delta X^{(m)}} = -\frac{\beta^{(m)}}{\beta^{(l)}}$$

If, like in the example of this chapter, $X^{(l)}$ is logarithmic income and $X^{(m)}$ is unemployment, then the above fraction gives the relative increase in income required to compensate for the negative effect of unemployment (assuming that $\beta^{(l)} > 0$ and $\beta^{(m)} < 0$).

To move the interpretation closer to the observed outcomes of Y, the threshold mechanism needs to be taken into account. One way of doing so is to ask how much of a change in a covariate it takes to move over one response category. For this purpose, one can form the ratio of the interval length to the parameter $(\kappa_y - \kappa_{y-1})/\beta^{(l)}$. The smaller this ratio (in absolute terms), the smaller the maximum change in $X^{(l)}$ required to move the response from category y to category $y+1$.

These two measures, while certainly of interest in some applications, stop short of the most natural way of interpreting parameters in discrete probability models, that is in terms of marginal or discrete probability effects. The marginal probability effects are obtained directly from (2.11):

$$MPE_y^{(l)}(X) = \frac{\partial P(Y=y|X)}{\partial X^{(l)}} = [f(\kappa_{y-1} - X'\beta) - f(\kappa_y - X'\beta)]\beta^{(l)} \tag{2.12}$$

where $f(\cdot) = F'(\cdot)$ denotes the density of U. In general, the marginal probability effects are functions of X and therefore vary for each individual. Average marginal probability effects can be obtained by taking expectations:

$$AMPE_y^{(l)} = E_X\left[\frac{\partial P(Y=y|X)}{\partial X^{(l)}}\right] \tag{2.13}$$

A consistent estimator of $AMPE_y^{(l)}$ is obtained by replacing β in (2.12) with the maximum likelihood estimator $\hat{\beta}$ and averaging over the sample:

$$\widehat{AMPE}_y^{(l)} = \frac{1}{n}\sum_{i=1}^{n} \widehat{MPE}_y^{(l)}(x_i)$$

An alternative average effect can be obtained by evaluating the marginal probability effect at the expected value of X, i.e., to consider $MPE_y^{(l)}(E(X))$ which can be estimated by $\widehat{MPE}_y^{(l)}(\bar{x})$ where $\bar{x} = \sum_{i=1}^{n} x_i/n$. If the change in $X^{(l)}$ is discrete, then marginal probability effects can be used to approximate the discrete probability effect by

$$\Delta P_y^{(l)}(X) \approx MPE_y^{(l)}(X)\Delta X^{(l)}$$

Such an approximation can be poor, in particular if large changes in $X^{(l)}$ are considered and/or $P(Y = y|X)$ exhibits much kurtosis. In these cases it is advisable to calculate changes in probabilities by the exact formula

$$\Delta P_y^{(l)}(X) = P(Y = y|X + \Delta X^{(l)}) - P(Y = y|X)$$

It is interesting to note that despite their intuitive appeal, marginal or discrete probability effects are rarely reported in practice.

Once the focus is put on the full distribution of outcomes, and the marginal probability effects, it becomes immediately apparent that standard ordered response models are quite restrictive, and perhaps unnecessarily so. A first way to pinpoint the restrictive nature of the marginal effects is the observation that their relative magnitude is not allowed to vary over the outcomes

$$\frac{MPE_y^{(l)}(X)}{MPE_y^{(m)}(X)} = \frac{\beta^{(l)}}{\beta^{(m)}}$$

The relative marginal effects do not depend on y (nor do they depend on X); in other words, they are the same in every part of the outcome distribution. It is not possible, referring to the example above, that income is more important (relative to unemployment, say) in the left part of the outcome distribution than in the right part. This property holds regardless of the choice of $F(\cdot)$.

A second restrictive property is that the sign of the marginal and the discrete probability effects for increasing y is entirely determined by the distribution function $F(\cdot)$. For

example, with $F(\cdot)$ being either standard normal or logistic, $f(\cdot)$ is bell-shaped with a maximum at 0. It follows from (2.12) and the order restriction on the threshold parameters that

$$sgn[MPE_y^{(l)}(X)] = -sgn[\beta^{(l)}] \text{ if } \kappa_{y-1} < X'\beta \text{ and } \kappa_y \leq X'\beta$$

$$sgn[MPE_y^{(l)}(X)] = sgn[\beta^{(l)}] \text{ if } \kappa_{y-1} \geq X'\beta \text{ and } \kappa_y > X'\beta$$

where $sgn(a)$ is the sign function that takes 1 if $a > 0$, 0 if $a = 0$, and -1 if $a < 0$. The sign of $MPE_y^{(l)}(X)$ is indeterminate for $\kappa_{y-1} < X'\beta$ and $\kappa_y > X'\beta$. Such a pattern may be referred to as "single crossing" property in the effect of covariates on probabilities.

More specific results can be obtained once a specific distribution function $F(\cdot)$ is considered. The best known result is the proportional log-odds assumption of the ordered logit model. From (2.11)

$$P(Y \leq y|X) = \Lambda(\kappa_y - X'\beta) = \frac{\exp(\kappa_y - X'\beta)}{1 + \exp(\kappa_y - X'\beta)}$$

and therefore

$$\frac{P(Y > y|X)}{P(Y \leq y|X)} = \frac{1 - \Lambda(\kappa_y - X'\beta)}{\Lambda(\kappa_y - X'\beta)} = \exp(X'\beta - \kappa_y)$$

Hence, the logarithmic odds of an outcome greater than y relative to an outcome less or equal than y are a linear function of X and the slope does not depend on y.

The nature of these restrictive properties is the single index assumption, i.e., only one set of parameters β is contained in the model. More flexible response patterns can be obtained if index functions are allowed to vary across response categories. In the next two sections, such models will be discussed. For further discussion of some of the issues presented here, see Boes and Winkelmann (2006) and Williams (2006).

Generalized Threshold Models

When searching for more flexible parametric ordered response models, the multinomial logit model stands at one extreme in terms of high flexibility. The multinomial logit, however, does not make use of the ordering information, and therefore cannot be efficient. A very flexible model that uses the information is obtained by making the threshold parameters linear functions of the covariates (Maddala 1983, Terza 1985). Let

$$\kappa_y(X) = \kappa_y + X'\gamma_y$$

where X is the $k \times 1$ vector of covariates, excluding a constant, and γ_y is a conformable category-specific parameter vector. Substitution of $\kappa_y(X)$ for κ_y in (2.11) yields

$$\begin{aligned} P(Y = y|X) &= F(\kappa_y + X'\gamma_y - X'\beta) - F(\kappa_{y-1} + X'\gamma_{y-1} - X'\beta) \\ &= F(\kappa_y - X'\beta_y) - F(\kappa_{y-1} - X'\beta_{y-1}) \end{aligned} \tag{2.14}$$

where $\beta_y = \beta - \gamma_y$ and it is understood that $\kappa_0 = -\infty$ and $\kappa_J = \infty$, as before, so that $F(\kappa_0 - X'\beta_0) = 0$ and $F(\kappa_J - X'\beta_J) = 1$. The model as presented here presumes that the same set of covariates affects linearly both the latent variable Y^* and the thresholds $\kappa_y(X)$. In this case, the parameters β and γ_y cannot be identified separately, but only their difference β_y is identified.

Conversely, if one set of covariates, say X_1, affects the latent variable Y^*, and another set of covariates, say X_2, affects the threshold parameters, then both β and γ_y are separately identified; see Kerkhofs and Lindeboom (1995) for an example. Overlapping X_1 and X_2 is an intermediate case. An alternative identification strategy is to specify $\kappa_y(X)$ as a non-linear function of X, such as $\kappa_y(X) = \sum_{j=1}^{y} \exp(X'\gamma_j)$, and include a constant term in both the γ's and β, so that the main effects (from β) and the threshold effects (from the γ's) can be distinguished. In order to illustrate the main implications of a generalized threshold mechanism, I will confine myself to the simple linear case and (2.14), analogous arguments will apply to the alternative identification schemes.

The model now contains $J-1$ parameter vectors $\beta_1, \ldots, \beta_{J-1}$, plus $J-1$ constants $\kappa_1, \ldots, \kappa_{J-1}$ that can be estimated jointly by maximum likelihood. The generalized model nests the standard ordered model under the restriction

$$\beta_1 = \ldots = \beta_{J-1}$$

Hence, the restricted model has $(J-2) \times k$ additional degrees of freedom. Clearly, the proliferation of parameters, in particular when J is large, is a potential disadvantage. However, a test can be easily conducted, and one can economize on degrees of freedom by imposing partial restrictions in subsets of outcomes, such as $\beta_1 = \beta_2$, while allowing parameters to differ in other parts of the distribution.

The model has substantially more flexible marginal probability effects, since

$$MPE_y^{(l)}(X) = f(\kappa_{y-1} - X'\beta_{y-1})\beta_{y-1}^{(l)} - f(\kappa_y - X'\beta_y)\beta_y^{(l)} \tag{2.15}$$

All the statements in the previous subsection on constant relative effects and single crossing no longer need to hold. Rather, these effects can be determined empirically. Furthermore, in the logit case one obtains

$$P(Y \leq y|X) = \Lambda(\kappa_y - X'\beta_y)$$

and therefore

$$\frac{P(Y > y|X)}{P(Y \leq y|X)} = \frac{1 - \Lambda(\kappa_y - X'\beta_y)}{\Lambda(\kappa_y - X'\beta_y)} = \exp(X'\beta_y - \kappa_y)$$

so that the effects of covariates on the log-odds are category specific.

The greater flexibility in modeling ordered responses with generalized thresholds does not come without costs. First, the constraint of ascending constants in the standard model to ensure a well-defined likelihood now extends to

$$\kappa_1 - X'\beta_1 < \ldots < \kappa_{J-1} - X'\beta_{J-1} \tag{2.16}$$

In a model with generalized thresholds, it is necessary that the multiple indices satisfy the order restrictions for all observations. As a practical consequence, the large number of parameters in conjunction with the order restriction (2.16) may increase computation time considerably, as attempts of unproductive likelihood steps are routinely made. See also Appendices A and B for more on this.

Sequential Model

For an alternative approach of modeling an ordinal response variable, I now consider the class of sequential models. This kind of model has previously been discussed in the statistics literature (e.g., Tutz 1991). As before, I assume that Y is coded as $1, \ldots, J$ where "1" designates the smallest outcome and "J" the largest. The basic idea here is to cast the model in terms of conditional transition probabilities $P(Y = y|Y \geq y)$. These conditionals fully characterize the probability function of Y. For example,

$$\begin{aligned}
P(Y = 1) &= P(Y = 1|Y \geq 1)P(Y \geq 1) = P(Y = 1|Y \geq 1) \\
P(Y = 2) &= P(Y = 2|Y \geq 2)P(Y \geq 2) \\
&= P(Y = 2|Y \geq 2)[1 - P(Y = 1|Y \geq 1)] \\
P(Y = 3) &= P(Y = 3|Y \geq 3)P(Y \geq 3) \\
&= P(Y = 3|Y \geq 3)[1 - P(Y = 1) - P(Y = 2)] \\
&= P(Y = 3|Y \geq 3)[1 - P(Y = 1|Y \geq 1)][1 - P(Y = 2|Y \geq 2)]
\end{aligned}$$

and in general

$$P(Y = y) = P(Y = y|Y \geq y) \prod_{j=0}^{y-1} (1 - P(Y = j|Y \geq j)) \qquad y = 1, \ldots, J$$

where it is understood that $P(Y = 0|Y \geq 0) = 0$ and $P(Y = J|Y \geq J) = 1$. Note that this approach is in close analogy to discrete time hazard models in duration analysis. Let t_j, $j = 1, \ldots, J$ denote the possible exit points, ordered by time. Then $P(T = t_j|T \geq t_j)$ gives the probability of exit at time t_j, conditional on survival until t_j.

The sequential model naturally accounts for the ordering of responses without imposing any arbitrary cardinality assumption. Due to its analogy to hazard rate models, the sequential model may be interpreted in terms of an underlying response mechanism, starting in the lowest category, and the individual then sequentially chooses between one or at least two, two or at least three, and so on. The observed category is equivalent to the category where these conditional transitions stop.

While such an interpretation may be suitable in some cases, the model does not necessarily need such a literal representation of cognitive processes that are at work when

the respondent answers the question of interest. It even appears rather unlikely in many instances that individuals actually think that way. In this case, one may rather see the sequential model as a flexible tool for obtaining a model for ordered responses with unrestricted marginal probability effects.

In order to model the effects of covariates, the model is parameterized as follows:

$$P(Y = y|Y \geq y, X) = F(\alpha_y + X'\beta_y)$$

where α_y is a category specific constant, β_y is a vector of category specific slopes, and $F(\cdot)$ is any function mapping real numbers onto the unit interval, such as the cumulative density function of the standard normal or the standard logistic distribution, respectively. The corresponding probability function is

$$P(Y = y|X) = F(\alpha_y + X'\beta_y) \prod_{j=0}^{y-1} [1 - F(\alpha_j + X'\beta_j)] \tag{2.17}$$

where it is understood that $F(\alpha_0 + X'\beta_0) = 0$ and $F(\alpha_J + X'\beta_J) = 1$. A model with category specific constants and single slope parameter β is obtained as a special case.

An important advantage of this model over the generalized threshold model is that no restrictions on the parameter space are required to ensure the existence of a proper probability function. This simplifies estimation considerably. On the downside, calculation of marginal probability effects is more difficult:

$$\begin{aligned} MPE_1^{(l)}(X) &= f(\alpha_1 + X'\beta_1)\beta_1^{(l)} \\ MPE_y^{(l)}(X) &= f(\alpha_y + X'\beta_y)\beta_y^{(l)} \prod_{j=1}^{y-1} [1 - F(\alpha_j + X'\beta_j)] \\ &\quad -F(\alpha_y + X'\beta_y) \sum_{j=1}^{y-1} MPE_j^{(l)}(X) \end{aligned}$$

Like in the generalized threshold model the effect of changing one covariate are local, i.e., vary by each category. In order to estimate the parameters of the model one needs to perform $J - 1$ consecutive binary regressions. The dependent variable Y_y of these binary models is equal to one if $Y = y$ and equal to zero if $Y > y$. In each step, only observations "at risk" are included, i.e., those for which it is the case that $Y \geq y$.

References

Abramowitz, M., and I.A. Stegun (1964): *Handbook of Mathematical Functions*, National Bureau of Standards, Applied Mathematical Series – 55, Washington D.C.

Blanchflower, D., and A. Oswald (2004): "Well-Being Over Time in Britain and the USA," *Journal of Public Economics*, 88, 1359-1386.

Boes, S. and R. Winkelmann (2006): "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90, 165-179.

Bradburn, N.M. (1969): *The Structure of Psychological Well-Being*, Aldine Publishing Co.

Bruni, L., and P.L. Porta (2006): *Economics and Happiness*, Oxford University Press.

Burkhauser, R.V., B.A. Butrica, M.C. Daly and D.R. Lillard (2001): "The Cross-National Equivalent File: A product of cross-national research," in I. Becker, N. Ott, and G. Rolf (eds.) *Soziale Sicherung in einer dynamischen Gesellschaft. Festschrift für Richard Hauser zum 65. Geburtstag*, Frankfurt/New York: Campus, 354-376.

Chamberlain, G. (1982): "Multivariate Regression Models for Panel Data," *Journal of Econometrics*, 18, 5-45.

Clark, A.E., F. Etilé, F. Postel-Vinay, C. Senik, and K. van der Straeten (2005): "Heterogeneity in Reported Well-Being: Evidence from Twelve European Countries," *The Economic Journal*, 115, C118-C132.

Clark, A.E., P. Frijters, and M.A. Shields (2006): "Income and Happiness: Evidence, Explanations and Economic Implications," *Working paper #5, NCER Working Paper Series.*

Clark, A., and A. Oswald (1996): "Satisfaction and Comparison Income," *Journal of Public Economics*, 61, 359-381.

Diener, E. (1984): "Subjective Well-Being," *Psychological Bulletin*, 95, 542-575.

Diener, E., and R. Biswas-Diener (2002): "Will Money Increase Subjective Well-Being: A Literature Review and Guide and Needed Research," *Social Indicators Research*, 57, 119-169.

Diener, E., R.A. Emmons, R.J. Larsen, S. Griffin (1985): "The Satisfaction with Life Scale," *Journal of Personality Assessment*, 49, 71-75.

Diener, E., E.M. Suh, R.E. Lucas, and H.L. Smith (1999): "Subjective Well-Being: Three Decades of Progress," *Psychological Bulletin*, 125, 276-302.

Easterlin, R. (1974): "Does Economic Growth Improve the Human Lot? Some Empirical Evidence," in P. David and M. Reder (eds.) *Nations and Households in Economic Growth: Essays in Honor of Moses Abramowitz*, 89-125, New York: Academic Press.

Easterlin, R. (1995): "Will Raising the Incomes of All Increase the Happiness of All?" *Journal of Economic Behaviour and Organization*, 27, 35-48.

Ferrer-i-Carbonell, A., and P. Frijters (2004): "How Important is Methodology for the Estimates of the Determinants of Happiness," *The Economic Journal* 114, 641-659.

Frey, B.S., S. Luechinger, and A. Stutzer (2004): "Valuing Public Goods: The Life Satisfaction Approach," *CESifo Working Paper No. 1158.*

Frey, B.S., and A. Stutzer (2002): *Happiness and Economics*, Princeton: Princeton University Press.

Headey, B.W., J. Kelly, and A.J. Wearing (1993): "Dimensions of Mental Health: Life-Satisfaction, Positive Effect, Anxiety and Depression," *Social Indicators Research*, 29, 63-82.

Headey, B.W., and M. Wooden (2004): "The Effects of Wealth and Income on Subjective Well-Being and Ill-Being," *The Economic Record*, 80, S24-S33.

Huppert, F.A., and J.E. Whittington (2003): "Evidence for the independence of positive and negative well-being: Implications for quality of life assessment," *British Journal of Health Psychology*, 8, 107-122.

Kerkhofs, M., and M. Lindeboom (1995): "Subjective Health Measures and State Dependent Reporting Errors," *Health Economics*, 4, 221-235.

Lalive, R. and A. Stutzer (2004): "Approval of Equal Rights and Gender Differences in Well-Being," *IZA Discussion Paper No. 1202.*

Layard, R. (2005): *Happiness: Lessons from a New Science*, New York: Penguin Press.

Luttmer E.F.P. (2005): "Neighbors as Negatives: Relative Earnings and Well-Being," *The Quarterly Journal of Economics*, 20, 963-1002.

Maddala, G. (1983): *Limited-dependent and qualitative variables in econometrics*, Cambridge: Cambridge University Press.

McCullagh, P. (1980): "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, 109-142.

McKelvey, R., and W. Zavoina (1975): "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology*, 4, pp. 103–120.

Mundlak, Y. (1978): "On the Pooling of Time Series and Cross Section Data," *Econometrica*, 46, 69-85.

Schwarze, J. (2003): "Using panel data on income satisfaction to estimate the equivalence scale elasticity," *Review of Income and Wealth*, 49, 359-372.

Terza, J. (1985): "Ordinal Probit: A Generalization," *Communications in Statistics – Theory and Methods*, 14, 1-11.

Tutz, G. (1991): "Sequential Models in Categorical Regression," *Computational Statistics and Data Analysis*, 11, 275-295.

Van Praag, B.M.S., and B.E. Baarsma (2005): "Using Happiness Surveys to Value Intangibles: The Case of Airport Noise," *The Economic Journal*, 115, 224-246.

Van Praag, B.M.S., Frijters, P. and Ferrer-i-Carbonell, A. (2003). The Anatomy of Subjective Well-being. Journal of Economic Behavior and Organization, 51 , (1), 29-49.

Williams, R. (2006): "Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables," *The Stata Journal*, 6(1), 58-82. A pre-publication version is available at http://www.nd.edu/ rwilliam/gologit2/gologit2.pdf.

Winkelmann, L. and R. Winkelmann (1998): "Why Are the Unemployed So Unhappy? Evidence from Panel Data," *Economica*, 65, 1-15.

Winkelmann, R., and S. Boes (2006): *Analysis of Microdata*, Berlin: Springer.

Chapter 3

Count Data Models with Correlated Unobserved Heterogeneity: An Empirical Likelihood Approach

Another version of this chapter has been published as *SOI Working Paper 0704*.

3.1 Introduction

Regression models for count data have become a standard tool in empirical analyses with applications in all fields of economics. Examples include the number of patents applied for by a firm (Hausman *et al.* 1984), the number of doctor visits (Pohlmeier and Ulrich 1995), the number of children borne to a woman (Winkelmann and Zimmermann 1995), and the number of days a worker is absent from his job (Delgado and Kniesner 1997).

If the regressand Y is measured as a non-negative integer, then the applied model should somehow account for this characteristic. One option is to use a specific conditional probability model for Y, given a vector of observed explanatory variables X, such as the Poisson regression model. The Poisson model, however, presumes that the researcher is able to account for the full amount of individual heterogeneity just by including X. Additional unobserved heterogeneity is not allowed for and ruled out by the model assumptions. Various generalizations of the Poisson model have been proposed that ac-

count for unobserved heterogeneity. Standard approaches employ mixture distributions, either parametrically by introducing, for example, Gamma distributed unobservables (the negative binomial models), or semiparametrically by leaving the mixing distribution unspecified (Gurmu *et al.* 1998). Winkelmann (2003: Ch. 4) gives an overview.

Mullahy (1997) extends the discussion to the important case when independence between observed and unobserved heterogeneity fails. He considers the conditional expectation function, formally $E(Y|X,\nu)$, specified as the exponential of a linear predictor $X'\beta$, with multiplicative unobserved heterogeneity ν. Mullahy (1997) points out that, given nonzero correlation between X and ν, standard estimators like Poisson pseudo maximum likelihood or non-linear least squares will generally be inconsistent for β because the usual residual function is not orthogonal to X. Also, a non-linear instrumental variables (IV) strategy based on this residual function will be inconsistent due to the non-separability of the observable and the unobservable factors.

Fortunately, a simple transformation of the model yields a residual function, say $\rho(Y,X;\beta)$, that is additively separable in X and ν, and the assumption of mean independence between the latter and instruments Z can be used to construct conditional moment restrictions of the form $E[\rho(Y,X;\beta)|Z] = 0$. As proposed by Mullahy (1997), estimation can be based on the generalized method of moments (GMM) using moment functions $g(Y,X,Z;\beta) = a(Z)\rho(Y,X;\beta)$ for some function $a(Z)$. The GMM estimator will be consistent for β and asymptotically normally distributed. The estimator is not necessarily efficient, though, because the asymptotic variance depends on the choice of $a(Z)$.

The aim of this paper is to extend Mullahy's (1997) approach using optimal instruments $a^*(Z)$ that fully utilize the information given by the conditional moment restrictions. In this, I follow Donald *et al.* (2003) who approximate the conditional moment restrictions by a series of unconditional moments using a general vector of approximating functions. From a theoretical point of view, semiparametric efficiency is achieved if linear combinations of these functions may well approximate the optimal instrument matrix of Chamberlain (1987) and if the dimension of the vector is increased with the sample size. As a practical matter, I will select the number of unconditional moments according to the

mean squared error criteria in Donald *et al.* (2005).

Clearly, the idea of using functions of the conditioning variables as additional instruments is not new; see for example Wooldridge (2001). In fact, one motivation of GMM is that all possible information — as given by the conditional moment restrictions — can be used in an efficient manner by choosing the "right" weighting matrix. A general vector of approximating functions like the one employed here has the advantage of systematically using the information at hand. This will in general improve the efficiency of the resulting estimator compared to a baseline where $a(Z) = Z$, or compared to any other vague choice of $a(Z)$. On the downside, many approximating functions, and thus unconditional moment conditions, may be needed to obtain the optimal estimator.

This requirement can be a serious matter, in particular in light of recent work concerning the finite sample properties of GMM. These studies emphasize the poor performance of the two-step procedure with increasing number of moment conditions, and alternatives are proposed, for example the empirical likelihood (EL) estimator of Owen (1988), Qin and Lawless (1994) and Imbens (1997). Other moment estimators exist as well (Hansen *et al.* 1996, Kitamura and Stutzer 1997, Imbens *et al.* 1998). Smith (1997) introduces the class of generalized empirical likelihood (GEL) estimators that include the aforementioned estimators as special cases, and asymptotic equality of GEL and GMM was shown. Further studies by Newey and Smith (2004) and Imbens and Spady (2006) examine the higher order properties of GEL and GMM and evidence the relative advantage of EL compared to two-step GMM in terms of higher order asymptotic bias and higher order efficiency (after bias correction) with increasing degree of overidentification.

The novelty of this paper is the application of the approximating functions approach to an inherently non-linear IV model, first in a generated data experiment and then with real data in a model for cigarette demand. The model and moment conditions will be laid out in the next section. Section 3 briefly discusses EL and GMM estimation, and the moment selection criteria. Section 4 compares the properties of the estimators in a simulated data environment. The results indicate that the EL estimator has indeed favorable properties in terms of bias and efficiency, as it was to be expected from earlier theoretical results. Section 5 applies the method to estimate a cigarette demand function. Fully exploiting the

model assumptions considerably improves the efficiency of the estimators. For example, just by including the optimal vector of approximating functions for one instrument, the t-statistic for the parameter of interest is more than doubled compared to the baseline IV estimator.

3.2 Exponential Model, Heterogeneity, and Moment Conditions

Let Y denote a random variable with support being the non-negative integers, let X denote a $k \times 1$ vector of explanatory variables (including a constant), and let Z denote a $q \times 1$ vector of instruments ($q \geq k$) with properties to be defined below. Assume that n observations of (Y, X, Z) form a random sample of the population, and suppose that the main objective is to estimate the effect of elements of X on Y.

The paper focuses on the relationship between Y and X as summarized in the conditional expectation function (CEF). Specifically, assume that the data-generating process is consistent with the CEF

$$E(Y|X, \nu; \beta) = \exp(X'\beta)\nu \tag{3.1}$$

where β is the $k \times 1$ vector of unknown parameters, and $\nu = \exp(\omega) > 0$ is unobservable to the researcher. Without loss of generality the normalization $E(\nu) = 1$ can be invoked as a constant term is included in X. Note that observable and unobservable characteristics are treated symmetrically in (3.1) because the CEF is log-linear in both X and ω. The specific functional form of the CEF might appear restrictive at first, but there is no *a priori* reason for X and ω to enter the CEF asymmetrically. Moreover, the linear index $X'\beta$ is sufficiently flexible to approximate any non-linear function in the regressors arbitrarily close, and the exponential function ensures (3.1) to be positive, as required for a count dependent variable. Strictly speaking, it is not necessary for (3.1) to be fulfilled that Y is a count. What follows is equally relevant to any other data-generating process consistent with such an exponential CEF.

The specification of the CEF in (3.1) implies the non-linear regression model

$$Y = \exp(X'\beta)\nu + \varepsilon \tag{3.2}$$

where the regression error ε has the property $E(\varepsilon|X, \nu) = 0$, by construction. Windmeijer and Santos Silva (1997) consider estimation of models like (3.2) in situations where some of the regressors may be simultaneously determined with the dependent count. In this case, there is a crucial distinction between additive and multiplicative (for that matter structural) errors, the two otherwise being observationally equivalent (Wooldridge 1992). Grogger (1990) discusses the additive approach and testing for exogeneity of the regressors using a Hausman-type test.

In the given context, it is natural to maintain the notation in (3.2) to distinguish between regression error and unobservable characteristics, the latter not being accounted for in the regression and potentially correlated with X. Mullahy (1997) gives conditions for consistent estimation of β in such a model. In a nutshell, if ν and X are mean independent, then pseudo maximum likelihood (PML) estimation of the Poisson model is consistent for β (see Gourieroux *et al.* 1984, Wooldridge 1997). Contrary to that, if mean independence fails, then PML will generally be inconsistent, and estimation with instrumental variables based on appropriately defined residuals is suggested alternatively. Mullahy (1997) imposes two key assumptions on the instrument vector Z. The first assumption is an independence condition that ν and Z must be mean independent, formally $E(\nu|Z) = E(\nu)$. The second assumption imposes the restriction $E(Y|X, \nu, Z) = E(Y|X, \nu)$ which implies for the regression error that $E(\varepsilon|X, Z, \nu) = 0$.

With the assumptions on Z, a conditional moment restriction can be constructed via the residual function $\rho(Y, X; \beta) = Y \exp(-X'\beta) - 1$ since

$$E[\rho(Y, X; \beta)|Z] = E[Y \exp(-X'\beta) - 1|Z] = 0 \tag{3.3}$$

by iterated expectations. As noted by Mullahy (1997), the crucial step in deriving such a residual function is that ν needs to be additively separable from X which can be achieved by dividing both sides of (3.2) by $\exp(X'\beta)$. The conditional moment restriction is assumed to uniquely identify the true parameter value β. Now let $a(Z)$ denote a matrix-valued function of Z with dimension $s \geq k$, which in the simplest case is the identity

function $a(Z) = Z$. It is common practice to derive unconditional (population) moment restrictions from (3.3) as

$$E\left[a(Z)\rho(Y, X; \beta)\right] = 0 \tag{3.4}$$

and the estimator of β is obtained as the solution to sample analogues $\sum_i a(z_i)\rho(y_i, x_i; \hat{\beta}) = 0$, with estimation operationalized, for example, in a GMM or non-linear IV framework. Such a procedure, however, is suboptimal for at least two reasons. First, the conditional moment restriction is stronger than the unconditional ones implying that an estimator based on the latter does not necessarily exploit all the available information. Second, the procedure is only valid under the presumption that $a(Z)$ (or in the simplest case Z) identifies β, which must not necessarily be so; see Dominguez and Lobato (2004).

Let $\mathcal{D}(Z) = E\left[\partial\rho(Y, X; \beta)/\partial\beta' | Z\right]$ denote the Jacobian and $\mathcal{V}(Z) = E\left[\rho(Y, X; \beta)^2 | Z\right]$ the variance obtained from the conditional moment restriction in (3.3). Chamberlain (1987) shows that the asymptotic efficiency bound for any $\sqrt{n}$-consistent semiparametric estimator based on (3.3) is given by $\mathcal{I}^{-1} = E_Z\left[\mathcal{D}(Z)'\mathcal{V}(Z)^{-1}\mathcal{D}(Z)\right]^{-1}$. This efficiency lower bound is derived under the initial assumption of independently and identically distributed data following a multinomial distribution, in which case the usual parametric efficiency bound applies, under the restriction (3.3). Since any distribution can be approximated arbitrarily close by the multinomial distribution, and the efficiency bound does not depend on the support of the distribution, the bound derived under the multinomial distribution also applies in the general semiparametric case.

An optimal GMM estimator based on the unconditional moment restrictions in (3.4) that attains the semiparametric efficiency bound requires instruments

$$a^*(Z) = \mathcal{D}(Z)'\mathcal{V}(Z)^{-1}$$

In general, such an estimator is not feasible as both expectations forming $a^*(Z)$ are unknown. It is shown already in Chamberlain (1987) that a GMM estimator based on a particular sequence of unconditional moment restrictions may come arbitrarily close to the semiparametric efficiency bound. Related to this idea, Donald *et al.* (2003) use a series of functions of Z to form unconditional moment restrictions, and let the dimension K of the vector of approximating functions grow with the sample size. Let $q^K(Z)$ denote

such a vector. Under certain regularity conditions, the sequence of unconditional moment restrictions

$$E[q^K(Z)\rho(Y, X; \beta)] = 0 \tag{3.5}$$

is equivalent to the conditional moment restriction in (3.3). Efficiency is established if linear combinations of $q^K(Z)$ can approximate $a^*(Z)$, with approximation error diminishing as K grows, since the asymptotic variance of the optimal GMM estimator with instruments $a^*(Z)$ reaches the semiparametric efficiency bound (see also Newey 1993).

Donald *et al.* (2003) suggest using splines as approximating functions. If Z is univariate, the s-th order spline with knots $t_1, \ldots, t_{K-s-1}$ is given by

$$q^K(Z) = (1, Z, \ldots, Z^s, [1(Z > t_1)Z]^s, \ldots, [1(Z > t_{K-s-1})Z]^s)' \tag{3.6}$$

with indicator function $1(\cdot)$. Common choice is $s = 3$ for cubic splines. For Z multivariate, the approximating functions may be generated by products of univariate splines for each element of Z. Under the assumption that Z is continuously distributed with compact support and density bounded away from zero, Donald *et al.* (2003) derive limits on the growth rate of K to obtain asymptotic efficiency. The method can be easily implemented in existing procedures that utilize unconditional moment restrictions, a potential advantage over alternative approaches such as Kitamura *et al.* (2004) and Dominguez and Lobato (2004).

3.3 Estimation Methods and Moment Selection

3.3.1 Generalized Method of Moments

The GMM principle has become a well-established estimation technique for moment conditions such as (3.5) since Hansen (1982); see also Hall (2005). To describe it, let $g_i(\beta) = q^K(z_i)\rho(y_i, x_i; \beta)$ and $\hat{g}_n(\beta) = \sum_{i=1}^n g_i(\beta)/n$. The GMM estimator $\hat{\beta}_{gmm}$ minimizes the weighted squared distance of sample and population moments, algebraically

$$\hat{\beta}_{gmm} = \arg\min_{\beta} \hat{g}_n(\beta)' \Upsilon \hat{g}_n(\beta) \tag{3.7}$$

where Υ is a $K \times K$ weighting matrix. For optimal GMM, the weighting matrix is chosen such that $\Upsilon = \hat{\Omega}_n(\tilde{\beta})^{-1}$ with $\hat{\Omega}_n(\beta) = \sum_{i=1}^n g_i(\beta)g_i(\beta)'/n$ and preliminary consistent estimator $\tilde{\beta}$. Under mild regularity conditions the resulting estimator $\hat{\beta}_{gmm}$ is consistent and the stabilizing transformation $\sqrt{n}(\hat{\beta}_{gmm} - \beta)$ is asymptotically normal with zero expectation and estimated covariance matrix

$$\hat{\Sigma}_{gmm} = \left[\hat{G}_n(\hat{\beta}_{gmm})'\hat{\Omega}_n(\hat{\beta}_{gmm})^{-1}\hat{G}_n(\hat{\beta}_{gmm})\right]^{-1}$$

where $G_i(\beta) = \partial g_i(\beta)/\partial\beta'$ and $\hat{G}_n(\beta) = \sum_{i=1}^n G_i(\beta)/n$.

Accumulating empirical evidence and recent theoretical work on the properties of two-step GMM, however, reveals that point estimates and inference based on the asymptotic normal distribution may be highly unreliable in finite samples (Hansen *et al.* 1996 and Hall 2005, among others). Newey and Smith (2004) discuss higher order asymptotic properties of GMM as possible explanation for the finite sample behavior. In particular, note that the optimization problem for two-step GMM implies first order conditions

$$\hat{G}_n(\hat{\beta}_{gmm})'\hat{\Omega}_n(\tilde{\beta})^{-1}\hat{g}_n(\hat{\beta}_{gmm}) = 0$$

and thus, in the optimum, a linear combination of sample equivalents to (3.5) must equal zero. It is shown, *inter alia*, that asymptotic (higher order) bias of the two-step GMM estimator arises from estimating the Jacobian matrix (left term) and the matrix of second moments (middle term) by sample averages, and the weighting matrix depending on a first step (inefficient) estimator.

As the asymptotic bias formulae are known, an analytical bias correction of $\hat{\beta}_{gmm}$ becomes available. The bias arising from estimation of the Jacobian matrix is particularly important, and a bias corrected GMM estimator can be obtained as

$$\hat{\beta}_{bcgmm} = \hat{\beta}_{gmm} + \hat{\Sigma}_{gmm}\sum_{i=1}^{n}\hat{G}_i\hat{P}\hat{g}_i/n \tag{3.8}$$

where $\hat{g}_i = g_i(\hat{\beta}_{gmm})$, $\hat{G}_i = G_i(\hat{\beta}_{gmm})$, and $\hat{P} = \hat{\Omega}^{-1} - \hat{\Omega}^{-1}\hat{G}\hat{\Sigma}_{gmm}\hat{G}'\hat{\Omega}^{-1}$ with $\hat{G} = \hat{G}_n(\hat{\beta}_{gmm})$, $\hat{\Omega} = \hat{\Omega}_n(\hat{\beta}_{gmm})$; see Newey and Smith (2004) and Donald *et al.* (2005) for details.

In comparison to two-step GMM, other moment estimators imply first order conditions in which the Jacobian and second moment matrix are estimated more efficiently. Among

the alternatives, the empirical likelihood estimator received considerable attention and was found to possess some desirable higher order properties. In particular, it was shown that the asymptotic bias of GMM grows with the number of overidentifying restrictions, whereas the bias of EL is bounded. I will therefore discuss EL estimation of β next.

3.3.2 Empirical Likelihood

Empirical likelihood estimation was first introduced in the biostatistics literature, see Owen (1988, 1991) and Qin and Lawless (1994, 1995) for details on EL and its application to moment condition models; see also Owen (2001) for a monograph on empirical likelihood. More recent surveys by Imbens (2002) and Kitamura (2006) point out the richness of the EL approach, in particular as an alternative to the two-step GMM procedure.

Let p_i denote an unknown probability weight assigned to the sample outcome (y_i, x_i, z_i) of one observation i with $0 < p_i < 1 \ \forall i$, impose the normalization $\sum_i p_i = 1$, and let $p = (p_1, \ldots, p_n)'$. A nonparametric likelihood estimator of p is obtained by maximizing the nonparametric log-likelihood function, algebraically

$$\hat{p} = \arg\max_p \sum_{i=1}^n \ln p_i \quad \text{s.t.} \quad \sum_{i=1}^n p_i = 1 \tag{3.9}$$

Without further restrictions, optimal probability weights are given by $\hat{p}_i = 1/n$. In order to incorporate special features of the data-generating process, one may impose empirical moments as additional restrictions, which can be specified from (3.5) as $\sum_i p_i g_i(\beta) = 0$. Following Kitamura (2006), the optimization problem yields the Lagrangian function

$$\mathcal{L} = \sum_{i=1}^n \ln p_i + \kappa \left(1 - \sum_{i=1}^n p_i\right) - n\lambda' \sum_{i=1}^n p_i g_i(\beta) \tag{3.10}$$

where λ and κ denote Lagrangian multipliers. It can be shown that the first order conditions are solved by $\hat{\kappa} = n$,

$$\begin{aligned} \hat{p}_i(\beta) &= \frac{1}{n\left[1 + \hat{\lambda}(\beta)' g_i(\beta)\right]} \\ \hat{\lambda}(\beta) &= \arg\min_\lambda -\sum_{i=1}^n \ln\left[1 + \lambda' g_i(\beta)\right] \end{aligned}$$

Optimal probability weights $\hat{p}_i$ and optimal Langrangian multipliers $\hat{\lambda}$ both depend on the unknown parameter vector β. Plugging the optimality conditions into the objective function in (3.9) yields the empirical log-likelihood function for β

$$\ln L_{el}(\beta) = \min_{\lambda} \; -\sum_{i=1}^{n} \ln\left[1 + \lambda' g_i(\beta)\right] - n \ln n$$

and the EL estimator is defined as

$$\hat{\beta}_{el} = \arg\max_{\beta} \; \ln L_{el}(\beta) = \arg\max_{\beta} \min_{\lambda} \; -\sum_{i=1}^{n} \ln\left[1 + \lambda' g_i(\beta)\right] \tag{3.11}$$

Since maximization of (3.11) does not have a simple closed form solution, numerical methods have to be applied to obtain the value of $\hat{\beta}_{el}$. Owen (2001) and Kitamura (2006) provide details on computational algorithms that have stable convergence properties in the above problem.

Under similar regularity conditions as in the GMM framework, Qin and Lawless (1994) show consistency of the empirical likelihood estimator and prove asymptotic normality of the stabilizing transformation $\sqrt{n}(\hat{\beta}_{el} - \beta)$ with zero expectation and estimated covariance matrix

$$\hat{\Sigma}_{el} = [\hat{G}_p(\hat{\beta}_{el})' \hat{\Omega}_p(\hat{\beta}_{el})^{-1} \hat{G}_p(\hat{\beta}_{el})]^{-1}$$

where $\hat{G}_p(\beta) = \sum_{i=1}^{n} \hat{p}_i(\beta) \partial g_i(\beta)/\partial \beta'$ and $\hat{\Omega}_p(\beta) = \sum_{i=1}^{n} \hat{p}_i(\beta) g_i(\beta) g_i(\beta)'$. Note that the terms in the EL covariance matrix are estimated using probability weights $\hat{p}_i(\hat{\beta}_{el})$ obtained from an empirical likelihood optimization, whereas the terms in the GMM variance are estimated using sample weights $1/n$.

It can be shown that optimal probability weights $\hat{p}_i$ and Langrangian multipliers $\hat{\lambda}$, both evaluated at the EL estimator, imply first order conditions

$$\hat{G}_p(\hat{\beta}_{el})' \hat{\Omega}_p(\hat{\beta}_{el})^{-1} \hat{g}_n(\hat{\beta}_{el}) = 0$$

As with two-step GMM, a linear combination of sample moments must equal zero. EL uses empirical moments for the Jacobian term and the matrix of second moments, and probability weights p_i are chosen efficiently. Moreover, the EL estimator does not depend on a preliminary, possibly inefficient estimator $\tilde{\beta}$. Based on these properties, Newey and Smith (2004) show that the EL estimator is preferable to the GMM estimator in terms of higher order asymptotic bias, and higher order efficiency after bias correction.

3.3.3 Moment Selection Criteria

To describe the moment selection criteria of Donald *et al.* (2005), some further notation needs to be introduced. Let $\hat{\beta}_K$ denote any of the three estimators — GMM, bias corrected GMM, or EL — given that the vector of approximating functions has dimension K. Let $t'\hat{\beta}_K$ denote a linear combination of $\hat{\beta}_K$ for some linear combination coefficients t. Let

$$\hat{\rho}_i = \rho(w_i; \hat{\beta}_K), \quad \hat{G} = \hat{G}_n(\hat{\beta}_K), \quad \hat{\Omega} = \hat{\Omega}_n(\hat{\beta}_K), \quad \hat{\Sigma} = [\hat{G}'\hat{\Omega}^{-1}\hat{G}]^{-1}, \quad \hat{\tau} = \hat{\Sigma}t$$

$$\hat{d}_i = \hat{G}' \left[\sum_{j=1}^{n} q^K(z_j) q^K(z_j)'/n \right]^{-1} q^K(z_i), \qquad \hat{\eta}_i = \partial\hat{\rho}_i/\partial\beta - \hat{d}_i$$

$$\hat{\xi}_i = q^K(z_i)'\hat{\Omega}^{-1} q^K(z_i)/n, \qquad \hat{\Lambda}(K) = \sum_{i=1}^{n} (\hat{\tau}'\hat{\eta}_i)^2 \hat{\xi}_i, \qquad \hat{\Pi}(K) = \sum_{i=1}^{n} (\hat{\tau}'\hat{\eta}_i)\hat{\xi}_i\hat{\rho}$$

$$\hat{\Phi}(K) = \hat{\Lambda}(K) - \hat{\tau}'\hat{\Sigma}^{-1}\hat{\tau}, \qquad \hat{Q} = \sum_{i=1}^{n} q^K(z_i)\hat{\rho}(\hat{\tau}'\hat{\eta}_i) q^K(z_i)'$$

$$\hat{\Pi}_b(K) = tr(\hat{\Omega}^{-1/2}\hat{Q}\hat{\Omega}^{-1}\hat{Q}\hat{\Omega}^{-1/2}), \qquad \hat{D}_i = \hat{G}'\hat{\Omega}^{-1} q^K(z_i)$$

$$\hat{\Xi}(K) = \sum_{i=1}^{n} \{5(\hat{\tau}'\hat{d}_i)^2 - \hat{\rho}^4(\hat{\tau}'\hat{D}_i)^2\}\hat{\xi}_i$$

$$\hat{\Xi}_{el}(K) = \sum_{i=1}^{n} \{3(\hat{\tau}'\hat{d}_i)^2 - \hat{\rho}^4(\hat{\tau}'\hat{D}_i)^2\}\hat{\xi}_i$$

The selection criteria are

$$\begin{aligned} S_{gmm}(K) &= \hat{\Pi}(K)^2/n + \hat{\Phi}(K) \\ S_{bcgmm}(K) &= [\hat{\Lambda}(K) + \hat{\Pi}_b(K) + \hat{\Xi}(K)]/n + \hat{\Phi}(K) \\ S_{el}(K) &= [\hat{\Lambda}(K) - \hat{\Pi}_b(K) + \hat{\Xi}(K) - 2\,\hat{\Xi}_{el}(K)]/n + \hat{\Phi}(K) \end{aligned} \tag{3.12}$$

The optimal dimension K^* of the vector of approximating functions is chosen such that $S(K)$ is minimal, i.e., $K^* = \arg\min_K S(K)$, which is shown to minimize the higher-order mean squared error (MSE) of each estimator. The terms in each criterion contain second and higher order moments, for details on the interpretation see Newey and Smith (2004) and Donald *et al.* (2005).

3.4 Monte Carlo Evidence

In this section, I compare the finite sample behavior of EL and GMM in a generated count data experiment with correlated unobserved heterogeneity. The model imposes a conditional moment restriction as the one introduced in the discussion above, and I investigate the performance of the proposed estimators with increasing dimension of the vector of approximating functions.

The sampling process is based on the Poisson model with Gamma distributed heterogeneity. The model is non-standard compared to the well-known negative binomial models in that the heterogeneity term is correlated with the single observed regressor X. Specifically, consider the following data-generating process

$$(\varphi, \psi) \sim BVN(0, 0, 1, 1, 0), \qquad \zeta = \varphi + \gamma\psi - (1 + \gamma^2)/2$$

$$Z \sim N(0, 1) \quad \text{or} \quad Z \sim LN(0, 1)$$

$$X = (1,\, \alpha Z + \psi)', \qquad \mu = \exp(X'\beta), \qquad \nu|\zeta \sim Gamma[1, \exp(\zeta)]$$

$$Y|X, \nu \sim Poisson(\mu\nu)$$

where $BVN(\cdot)$ stands for the bivariate normal distribution with zero means, unit variances, and zero correlation, $N(0, 1)$ stands for the standard normal, and $LN(0, 1)$ for the standard log-normal distribution. It is assumed that only (Y, X, Z) are observed. The conditional distribution of $\nu|\zeta$ is normalized such that $E(\nu|\zeta) = \exp(\zeta)$ and $Var(\nu|\zeta) = \exp(2\zeta)$. The location normalization of ζ implies that $E(\nu) = E[E(\nu|\zeta)] = E[\exp(\zeta)] = 1$. For α fixed, the parameter γ determines the correlation between X and ζ. If γ equals zero, the unobserved heterogeneity is independent of the regressor and PML consistently estimates β. For nonzero γ, the conditional expectation $E(\nu|X)$ is non-constant in X, and PML estimation will generally be inconsistent. Given that ν and Z are statistically independent, the assumption imposed here is somewhat stronger than required, and $\alpha \neq 0$, then moment estimation as outlined above using the instrument Z can be applied.

The parameter vector β is fixed at $(0,\, 1)'$, and γ is set to 0.5. In order to vary the correlation between instrument and regressor, two different values of α are chosen —

0.3 and 0.7. Two different sample sizes are considered — $n = 500$ and $n = 2000$ — and samples are drawn for all variables in each of 1000 Monte Carlo replications. Since $\gamma \neq 0$, PML estimation will be inconsistent for β in each of the settings. The experiment shows that, depending on the variation in X, the median bias in the estimated slope $\hat{\beta}_{1,pml}$ varies between 0.264 and 0.381 in the normal case, and between 0.377 and 0.446 in the log-normal case. These numbers need to be compared with the results for the other estimators, displayed in Tables 3.1 – 3.4.

— Insert Tables 3.1 and 3.2 about here —

Consider Tables 3.1 and 3.2 with $n = 500$ observations first. The columns in Table 3.1 correspond to the median of the estimated standard error of $\hat{\beta}_1$ (Med.SE) and the rejection rate for an overidentifying test (in the case of $K > 2$) with 5% significance level. Table 3.2 shows the median bias (Med.Bias) and the median absolute deviation (MAD) from the true value, and the probabilities of $\hat{\beta}_1$ deviating from 1 by more than 0.1 and 0.2, respectively. Robust measures of central tendency and dispersion are presented as the existence of (finite-sample) moments might be an issue (Kunitomo and Matsushita 2003, Guggenberger 2005, 2007, Guggenberger and Hahn 2005, Davidson and MacKinnon 2006). Five different specifications of $q^K(Z)$ are presented. The first, as a benchmark, is basic IV with instrument Z, i.e., the vector of approximating functions is simply $q^2(Z) = (1, Z)'$. The next three rows give the results with augmented instrument vector according to (3.6) and having dimensions $K = 4, 8, 16$, and optimal K^*. The approximating functions are chosen such that they form a basis for the set of cubic splines, i.e., $s = 3$, and the knots $t_1, \ldots, t_{K-4}$ are set equal to the quantiles of the empirical Z-distribution. The first-step weighting matrix for the two-step GMM estimator is chosen to be the $K \times K$ identity matrix. For the selection criteria, the linear combination coefficients pick the slope as parameter of interest.

The results in Table 3.1 indicate that there are considerable efficiency gains by increasing the dimension of the vector of approximating functions. These gains are higher with a low value of α and for the EL estimator more than for the GMM estimators. If Z is normally distributed, EL seems to perform better than GMM, if Z follows a log-normal

distribution there is no clear advantage for one of the three estimators. In all cases, the optimal K^* yields the lowest median standard error. Due to the variation in K^*, it is suggestive to choose the dimension of $q^K(Z)$ according to the MSE criteria, as opposed to a rule-of-thumb fixed choice of K. The rejection rate for the overidentifying restrictions test is always close to the nominal level.

Despite the efficiency gains, it is important to note that the estimators behave quite differently when looking at the summary statistics of $\hat{\beta}_1$ in Table 3.2. In all cases, the basic IV estimator produces consistent results, which is reflected in almost zero median bias. As it was expected from previous theoretical results, the GMM estimator exhibits significant bias if K and thus the number of overidentifying restrictions grows, and even under the optimal choice K^* the bias remains. Bias correction helps to improve upon the standard two-step GMM procedure, but in all settings the EL estimator has lowest bias. With respect to the median absolute deviation and the deviation probabilities, there are only minor differences between the three estimators.

— Insert Tables 3.3 and 3.4 about here —

Tables 3.3 and 3.4 report the simulation results for $n = 2000$ observations. In this case, GMM and EL perform similarly, which was to be expected as they are all first order asymptotically equivalent. It is noteworthy that even with 2000 observations, the two-step GMM estimator with large degree of overidentification exhibits bias that does not occur with bias corrected GMM and EL. The efficiency gains from augmenting the vector of approximating functions, however, are much smaller in the large sample than they are in the small sample experiment.

3.5 Cigarette Demand and Smoking Habits

As a final exercise, I apply the proposed methods to the estimation of a cigarette demand function. Cigarette demand is measured as the number of cigarettes smoked per day, and thus Y has the character of a count dependent variable. Mullahy (1985) studies the dynamic link between today's demand for cigarettes and an individual's smoking habits

amassed over lifetime. If included in a regression model, such habits can be interpreted as a lagged dependent variable, and there is good reason to believe that unobserved smoking determinants are also dynamically linked. One would thus suspect, given a positive correlation between unobservables over time, that the smoking habit dynamics may be overestimated in a simple Poisson regression model. With suitable instruments, a non-linear IV strategy as outlined above may be applied that does not suffer from such bias.

The analysis is based on a sample of $n = 1140$ male observations taken from the data in Mullahy (1997); see also Mullahy (1985) for a description. The data stem from the Smoking Supplement of the 1979 US National Health Interview Survey and contain information on the respondent's socioeconomic characteristics as well as information on various health topics and smoking behavior. For the regressions, the dependent variable has been scaled to the number of cigarette packs smoked per day (number of cigarettes divided by 20). Mullahy (1985) constructs the smoking habit measure from the total time smoked and the number of cigarettes consumed. This measure is zero for non-smokers, and positive for smokers, the exact value depending on the discount rate (here 10 percent) and not having direct unit interpretation. Apart from the smoking habit measure as the key variable of interest, the estimated models control for age (in years), the years of schooling, a dummy variable indicating race, family income (in thousand US Dollars), household size, average state-level cigarette price (in US Dollars per pack in 1979), and an indicator whether smoking in restaurants had been restricted (in 1979).

The excluded instruments are the cigarette price in 1978 and the total number of years smoking in restaurants had been restricted (before and with 1979). The rationale for the instruments is that both should affect smoking habits, i.e., smoking behavior in 1978 and before, but they should not have a direct effect on current cigarette demand. The latter exclusion restriction is plausible, since cigarette prices and indicators of smoking restrictions in 1979, i.e., at the time current cigarette demand is recorded, are explicitly controlled for, and thus there is no reason to believe why the instruments should have an effect on Y other than the habits channel. Compared to the data in Mullahy (1997), I restrict the sample to individuals aged younger than 25, as those are the most responsive

to changes in the instruments.

— Insert Table 3.5 about here —

Table 3.5 displays the results for the smoking habit coefficient. The columns correspond to the Poisson pseudo maximum likelihood (PML) estimator, the two-step and bias-corrected GMM estimators, and the EL estimator. For the ease of exposition, the estimated parameters and standard errors have been multiplied by 1000. The PML estimate shows a value of 12.53 with estimated standard error 0.81. This value indicates that the expected number of packages smoked per day increases by $100[\exp(12.53/1000)-1] = 1.26$ percent for an unit increase in the smoking habit measure. Multiplied by the average value of the smoking habits (35.65), this gives an elasticity of 0.45, i.e., if the smoking habit measure increases by 1 percent, then the expected number of cigarettes smoked per day (measured in packs) increases by 0.45 percent. The elasticity may of course be evaluated at other values than the average smoking habits.

Using the basic IV setting with instruments all regressors except the smoking habits plus the cigarette price in 1978 and the number of years the smoking restrictions had been in place, the estimated parameters drop by around 5 to 10 percent with much larger standard error. The IV point estimates confirm the expectation that PML might overestimate the true smoking habit effect. On the downside, from a statistical point of view, smoking habits do not significantly affect current smoking behavior, which contradicts the perspective of smoking habits entering cigarette demand as a psychological and/or physiological addiction. Note that the overidentifying test statistic is sufficiently small as to not reject the null hypothesis of valid instruments. Note too that the basic setting does not fully exploit the model assumptions and, provided that the instruments fulfill mean independence, an improvement over these results might be possible.

The remaining of Table 3.5 shows the estimation results for various specifications of the vector of approximating functions. Among the many options to specify this vector, a reasonable working guess is to first find the optimal dimension, say K_l^* for the l-th element of the instrument vector, given basic specification for all other instruments, and then gradually combine the optimal K_l^* including interactions if suitable. The table first

reports the results for the optimal specification of the excluded instruments, i.e., the number of years smoking restrictions had been in place and the cigarette price in 1978, respectively. In curly brackets is the number of additional approximating functions, e.g., for the cigarette price in 1978 one additional element (its square) has been included. This number plus one are the degrees of freedom for the overidentifying restrictions test with test statistic reported in square brackets.

The point estimates of the smoking habit coefficient drop compared to PML and basic IV. Using the square of cigarette prices in 1978 as additional instrument, for example, even turns the sign of the coefficient negative for bias-corrected GMM and EL. Although the overidentifying restrictions are not rejected for this specification, there is only a minor gain in the value of the moment selection criteria. For the restaurant smoking restrictions, the overidentifying restrictions are not rejected either, but in this case there is a considerable drop in the value of the selection criteria indicating higher potential efficiency gains by augmenting the instrument vector. Note that in both cases the null hypothesis of a zero smoking habit coefficient cannot be rejected. Clearly, an element-wise optimization may also be done for the included instruments.

Next, I combine the optimal approximating functions for each excluded instrument to further explore the model assumptions. It turns out that the optimal number of approximating functions K_l^* for each instrument can be combined to obtain the optimal number of approximating functions when both instruments are considered simultaneously. Presumably, this result is specific to the data and does not hold in general, but such a strategy might in any case be a good starting point to explore the validity of mean independence. Using the additional approximating functions and including interactions does not change the point estimates by much, but the standard errors become smaller due to the additional information that is used.

Finally, combining the optimal dimension K_l^* for excluded and included instruments and adding interactions if indicated, the optimal vector of approximating functions for the GMM estimators has five additional terms for restaurant smoking restrictions, household size and the cigarette price in 1979, two additional terms for family income, and the square of cigarette price in 1978. For the EL estimator, the interaction between smoking

restrictions and cigarette prices in 1978 and an additional term for family income is included. The results show point estimates of 8.86 for two-step GMM, 7.61 for bias corrected GMM, and 7.08 for EL. In terms of elasticities, a one percent increase in the smoking habit measure leads to an increase in the expected number of cigarette packs consumed per day by about 0.32 percent for the GMM estimator, 0.27 percent for the bias corrected GMM estimator, and 0.25 percent for the EL estimator, respectively. In all cases, the estimated coefficients are statistically different from zero at the 5 percent level, and therefore there is a much higher precision in the IV estimates than has previously been obtained.

3.6 Concluding Remarks

This paper extends Mullahy's (1997) IV approach for the estimation of count data models with correlated unobserved heterogeneity. Based on transformed residuals and a mean independence assumption, the model implies conditional moment restrictions that can be estimated by common moment estimators, such as the generalized method of moments (GMM) and empirical likelihood (EL). As the asymptotic variance typically depends on the choice of instruments, the paper proposes the use of a general vector of approximating functions, opting ideas of Donald *et al.* (2003), to improve efficiency of the resulting estimator.

A small Monte Carlo experiment points out the benefits of the method and outlines the relative advantage of EL compared to two-step GMM. Finally, the approach is applied to estimate the effect of smoking habits on cigarette demand. Compared to the standard Poisson PML estimator, the estimated elasticities of cigarette demand with respect to smoking habits change from 0.45 to 0.32 (GMM) and 0.25 (EL), respectively, a drop that is conformable to previous findings. Importantly, since the methods applied here fully exploit the model assumptions, the parameters have been estimated with much higher precision than before.

Tables

Table 3.1: Simulation Results for $se(\hat{\beta}_1)$ and χ^2-test; $n = 500$

	GMM		BCGMM		EL	
	Med.SE	Overid.	Med.SE	Overid.	Med.SE	Overid.
$z \sim N(0,1)$, $\alpha = .3$						
$K = 2$	.359	—	.359	—	.359	—
$K = 4$	.287	.058	.285	.051	.289	.063
$K = 8$	.246	.051	.246	.048	.243	.046
$K = 16$	.206	.063	.201	.052	.198	.058
$K = K^*$	.187	.053	.186	.061	.179	.059
$z \sim N(0,1)$, $\alpha = .7$						
$K = 2$	.158	—	.158	—	.158	—
$K = 4$	.141	.062	.141	.051	.140	.064
$K = 8$	.146	.065	.145	.049	.143	.060
$K = 16$	.146	.048	.139	.058	.137	.052
$K = K^*$	.125	.052	.125	.052	.125	.053
$z \sim LN(0,1)$, $\alpha = .3$						
$K = 2$	.141	—	.141	—	.141	—
$K = 4$	.107	.056	.106	.049	.104	.058
$K = 8$	.117	.054	.118	.051	.114	.062
$K = 16$	.116	.043	.114	.044	.108	.047
$K = K^*$	.069	.052	.070	.051	.073	.051
$z \sim LN(0,1)$, $\alpha = .7$						
$K = 2$	.061	—	.061	—	.061	—
$K = 4$	.045	.052	.045	.052	.043	.055
$K = 8$	.054	.053	.054	.046	.049	.054
$K = 16$	.053	.049	.052	.053	.048	.046
$K = K^*$	.032	.052	.032	.049	.031	.053

Notes: Med.SE is the median of the estimated standard error of $\hat{\beta}_1$, Overid. is the rejection rate for an overidentifying restrictions test with 5% nominal level. $K = 2$ is the basic IV setting with only Z included. $K \geq 2$ specifies a fixed number of elements in the $q^K(Z)$ vector, K^* is the optimal number.

Table 3.2: Simulation Results for $\hat{\beta}_1$; $n = 500$

	GMM				BCGMM				EL			
	Med.Bias	MAD	$P(\text{AD} > d)$		Med.Bias	MAD	$P(\text{AD} > d)$		Med.Bias	MAD	$P(\text{AD} > d)$	
			d=.1	d=.2			d=.1	d=.2			d=.1	d=.2
$z \sim N(0,1)$, $\alpha = .3$												
$K = 2$	.005	.247	.819	.609	.005	.246	.819	.609	.005	.246	.819	.609
$K = 4$	-.021	.207	.717	.510	-.005	.216	.741	.537	-.006	.220	.740	.545
$K = 8$	-.024	.180	.692	.452	.006	.225	.756	.543	.001	.241	.778	.574
$K = 16$	-.043	.135	.617	.321	.006	.180	.683	.465	.002	.177	.707	.442
$K = K^*$	-.020	.127	.589	.303	-.001	.143	.621	.335	.001	.145	.655	.339
$z \sim N(0,1)$, $\alpha = .7$												
$K = 2$	.005	.111	.538	.244	.005	.111	.538	.244	.005	.111	.538	.244
$K = 4$	-.012	.110	.540	.253	-.001	.113	.547	.250	.003	.121	.593	.249
$K = 8$	-.033	.120	.573	.269	.001	.130	.609	.308	.005	.119	.561	.276
$K = 16$	-.088	.120	.569	.268	-.008	.132	.602	.314	-.007	.116	.564	.279
$K = K^*$	-.028	.092	.474	.164	-.010	.094	.470	.168	-.004	.096	.477	.171
$z \sim LN(0,1)$, $\alpha = .3$												
$K = 2$	-.002	.113	.542	.236	-.002	.113	.542	.236	-.002	.112	.542	.236
$K = 4$	-.026	.116	.534	.283	.001	.115	.557	.294	.001	.105	.509	.234
$K = 8$	-.032	.107	.514	.204	-.009	.116	.552	.232	-.009	.116	.558	.256
$K = 16$	-.050	.095	.460	.153	-.013	.099	.496	.200	-.007	.110	.535	.232
$K = K^*$	-.018	.091	.458	.195	-.009	.091	.464	.189	-.006	.088	.441	.154
$z \sim LN(0,1)$, $\alpha = .7$												
$K = 2$	-.005	.049	.192	.030	-.005	.049	.192	.030	-.005	.049	.192	.030
$K = 4$	.004	.048	.196	.031	-.001	.048	.192	.029	.002	.043	.157	.018
$K = 8$	-.010	.052	.208	.026	-.007	.051	.209	.026	-.004	.050	.211	.022
$K = 16$	-.040	.044	.164	.018	-.015	.041	.140	.010	-.009	.048	.176	.024
$K = K^*$	-.019	.048	.188	.028	-.009	.048	.170	.028	-.003	.043	.177	.028

Notes: See the notes of Table 3.1. Med.Bias is the median bias of $\hat{\beta}_1$ from the true value $\beta_1 = 1$, MAD is the median absolute deviation from the true value, and $P(AD > d)$ is the probability of $\hat{\beta}_1$ deviating from 1 by more than d.

Table 3.3: Simulation Results for $se(\hat{\beta}_1)$ and χ^2-test; $n = 2000$

	GMM		BCGMM		EL	
	Med.SE	Overid.	Med.SE	Overid.	Med.SE	Overid.
$z \sim N(0,1)$, $\alpha = .3$						
$K = 2$	.172	—	.172	—	.171	—
$K = 4$	.167	.051	.168	.052	.169	.051
$K = 8$	.166	.050	.166	.055	.163	.047
$K = 16$	.157	.049	.157	.048	.155	.052
$K = K^*$	.143	.051	.143	.049	.145	.048
$z \sim N(0,1)$, $\alpha = .7$						
$K = 2$	.092	—	.091	—	.091	—
$K = 4$	.079	.050	.079	.049	.078	.052
$K = 8$	.084	.053	.083	.048	.083	.051
$K = 16$	.091	.051	.091	.051	.089	.051
$K = K^*$	.074	.048	.074	.047	.074	.053

Notes: See the notes of Table 3.1.

Table 3.4: Simulation Results for $\hat{\beta}_1$; $n = 2000$

	GMM				BCGMM				EL			
	Med.Bias	MAD	$P(\text{AD} > d)$		Med.Bias	MAD	$P(\text{AD} > d)$		Med.Bias	MAD	$P(\text{AD} > d)$	
			d=.1	d=.2			d=.1	d=.2			d=.1	d=.2
$z \sim N(0,1)$, $\alpha = .3$												
$K = 2$	.005	.131	.572	.326	.005	.131	.572	.326	.005	.131	.572	.326
$K = 4$	.010	.116	.547	.293	.001	.122	.558	.312	-.002	.123	.575	.307
$K = 8$	.014	.118	.584	.268	.008	.140	.620	.325	.001	.144	.594	.355
$K = 16$	.027	.108	.536	.231	.002	.143	.658	.351	.002	.156	.684	.386
$K = K^*$	.023	.091	.470	.144	.003	.094	.473	.150	.001	.092	.485	.155
$z \sim N(0,1)$, $\alpha = .7$												
$K = 2$	.003	.092	.431	.087	.003	.092	.431	.087	.003	.092	.431	.087
$K = 4$	.004	.060	.237	.027	.007	.062	.247	.028	.007	.059	.250	.022
$K = 8$	-.016	.067	.284	.042	-.007	.067	.308	.044	-.002	.070	.332	.055
$K = 16$	-.034	.072	.375	.068	-.011	.082	.411	.093	.005	.077	.357	.070
$K = K^*$	-.008	.052	.205	.009	-.007	.053	.203	.011	-.005	.052	.209	.011

Notes: See the notes of Tables 3.1 and 3.2.

Table 3.5: The Effect of Smoking Habits on Cigarette Demand

	Poisson ML	GMM	BCGMM	EL
	12.53			
	(0.81)			
Basic instruments		11.33	12.04	11.63
		(14.35)	(14.98)	(14.97)
		[0.59]	[0.58]	[0.58]
Optimized over				
(a) rest. smoking restrictions {5}		8.05	7.23	7.17
		(6.13)	(5.89)	(5.90)
		[2.04]	[2.05]	[1.95]
(b) cigarette price in 1978 {1}		1.42	-0.57	-0.98
		(7.27)	(6.87)	(6.90)
		[2.28]	[2.04]	[2.04]
(a) and (b) {6}		7.04	5.82	5.58
		(5.80)	(5.49)	(5.35)
		[7.70]	[7.52]	[7.92]
(a) and (b) plus interaction {7}		6.34	4.47	7.09
		(5.50)	(5.11)	(5.40)
		[7.69]	[7.73]	[8.18]
all variables		8.86	7.61	7.08
{GMM, BCGMM: 19; EL: 21}		(4.04)	(3.85)	(3.51)
		[26.91]	[24.25]	[24.88]

Notes: All models control for age, years of schooling, dummy variables indicating race and smoking restrictions in 1979, cigarette price in 1979, household income, and household size. The first value is the estimated coefficient; the second value (in round brackets) is the estimated asymptotic standard error; the third value (in square brackets) is the overidentifying test statistic with degrees of freedom the number in curly brackets +1.

Excluded instruments: Cigarette price in 1978; number of years smoking restrictions in place. In curly brackets is the number of additional elements, compared to the basic set of instruments, according to the specification of the $q^K(Z)$ vector. Optimization over all variables adds functions of the included instruments and interactions.

References

Cameron, A.C. and P.K. Trivedi (1998): *Regression Analysis of Count Data*, Econometric Society Monograph No. 30, Cambridge University Press.

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305-334.

Davidson, R. and J.G. MacKinnon (2006): "Moments of IV and JIVE Estimators," *unpublished manuscript*.

Delgado, M.A. and T.J. Kniesner (1997): "Count Data Models with Variance of Unknown Form: An Application to a Hedonic Model of Worker Absenteeism," *Review of Economics and Statistics*, 79, 41-49.

Dominguez, M. and I. Lobato (2004): "Consistent estimation of models defined by conditional moment restrictions," *Econometrica*, 72, 1601-1615.

Donald, S.G., G.W. Imbens and W.K. Newey (2003): "Empirical Likelihood Estimation and Consistent Tests with Conditional Moment Restrictions," *Journal of Econometrics*, 117, 55-93.

Donald, S.G., G.W. Imbens and W.K. Newey (2005): "Choosing the Number of Moments in Conditional Moment Restriction Models," *unpublished manuscript*.

Gourieroux, C., A. Monfort and A. Trognon (1984): "Pseudo Maximum Likelihood Methods: Applications to Poisson Models," *Econometrica*, 52, 701-720.

Grogger, J.T. (1990): "A Simple Test for Exogeneity in Probit, Logit, and Poisson Regression Models," *Economics Letters*, 33, 329-332.

Guggenberger, P. (2005): "Monte-carlo evidence suggesting a no moment problem of the continuous updating estimator," *Economics Bulletin*, 3, 1-6.

Guggenberger, P. (2007): "Finite Sample Evidence Suggesting a Heavy Tail Problem of the Generalized Empirical Likelihood Estimator", *Econometric Reviews*, forthcoming.

Guggenberger, P. J. Hahn (2005): "Finite Sample Properties of the 2step Empirical Likelihood Estimator" *Econometric Reviews*, 24, 247-263.

Gurmu, S., P. Rilstone and S. Stern (1998): "Semiparametric Estimation of Count Regression Models," *Journal of Econometrics*, 88, 123-150.

Hall, A.R. (2005): *Generalized Method of Moments*, Oxford University Press.

Hansen L.P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029-1054.

Hansen, L.P., J. Heaton and A. Yaron (1996): "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics*, 14, 262-280.

Hausman, J., B.H. Hall and Z. Griliches (1984): "Econometric Models for Count Data with an Application to the Patents - R&D Relationship," *Econometrica*, 52, 909-938.

Imbens, G.W. (1997): "One-Step Estimators for Over-Identified Generalized Method of Moment Models," *Review of Economic Studies*, 64, 359-383.

Imbens, G.W. (2002): " Generalized Method of Moments and Empirical Likelihood," *Journal of Business & Economic Statistics*, 20, 493-506.

Imbens, G.W. and R.H. Spady (2006): "The Performance of Empirical Likelihood and its Generalizations," in: D.W.K. Andrews (ed.) *Identification and Inference for Econometric Models: Essays in Honour of Thomas Rothenberg.*

Imbens, G.W., R.H. Spady and P. Johnson (1998): "Information Theoretic Approaches to Inference in Moment Condition Models," *Econometrica*, 66, 333-357.

Kitamura, Y. (2006): "Empirical Likelihood Methods in Econometrics: Theory and Practice", *Cowles Foundation Discussion Paper No. 1569.*

Kitamura, Y. and M. Stutzer (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65, 861-874.

Kitamura, Y., G. Tripathi, and H. Ahn (2004): "Empirical likelihood based inference in conditional moment restriction models," Econometrica, 72, 1667-1714.

Kunitomo, N. and Y. Matsushita (2003): "Finite Sample Distributions of the Empirical Likelihood Estimator and the GMM Estimator," *CIRJE Discussion paper F-200.*

Mullahy, J. (1985): "Cigarette Smoking: Habits, Health Concern, and Heterogeneous Unobservables in a Microeconometric Analysis of Consumer Demand," Ph.D. Dissertation, University of Virginia.

Mullahy, J. (1997): "Instrumental-Variable Estimation of Count Data Models: Applications to Models of Cigarette Smoking Behavior," *Review of Economics and Statistics*, 79, 586-593.

Newey, W. (1993): "Efficient Estimation of Models with Conditional Moment Restrictions," in: G. Maddala, C. Rao and H. Vinod (eds.) *Handbook of Statistics Vol. 11*, Elsevier Science, North Holland.

Newey, W.K. and R.J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219-255.

Owen, A.B. (1988): "Empirical Likelihood Ratio Confidence Regions for a Single Functional," *Biometrika*, 75, 237-249.

Owen, A.B. (1991): "Empirical Likelihood for Linear Models," *Annals of Statistics*, 19, 1725-1747.

Owen, A.B. (2001): *Empirical Likelihood*, Chapman & Hall/CRC, Boca Raton.

Pohlmeier, W. and V. Ulrich (1995): "An Econometric Model of the Two-Part Decisionmaking Process in the Demand for Health Care," *Journal of Human Resources*, 30, 339-361.

Qin, J. and J. Lawless (1994): "Empirical Likelihood and General Estimating Equations," *Annals of Statistics*, 22, 300-325.

Qin, J. and J. Lawless (1995): "Estimating Equations, Empirical Likelihood, and Constraints on Parameters," *Canadian Journal of Statistics*, 23, 145-159.

Smith, R.J. (1997): "Alternative Semi-Parametric Likelihood Approaches to Generalised Method of Moments Estimation," *Economic Journal*, 107, 503-519.

Windmeijer, F.A.G. and J.M.C. Santos Silva (1997): "Endogeneity in Count Data Models: An Application to Demand for Health Care," *Journal of Applied Econometrics*, 12, 281-294.

Winkelmann, R. (2003): *Econometric Analysis of Count Data*, Springer Verlag, Berlin.

Winkelmann, R. and K.F. Zimmermann (1994): "Count Data Models for Demographic Data," *Mathematical Population Studies*, 4, 205-221.

Wooldridge, J.M. (1992): "Some Alternatives to the Box-Cox Regression," *International Economic Review*, 33, 935-955.

Wooldridge, J.M. (1997): "Quasi-Likelihood Methods for Count Data," in: M.H. Pesaran and P. Schmidt (eds.) *Handbook of Applied Econometrics Vol. 2 - Microeconomics*, Blackwell.

Wooldridge, J.M. (2001): "Applications of Generalized Method of Moments Estimation," *Journal of Economic Perspectives*, 15, 87-100.

Chapter 4

Nonparametric Analysis of Treatment Effects in Ordered Response Models

Another version of this chapter has been published as *SOI Working Paper 0709*.

4.1 Introduction

Suppose one is interested in the effect of a binary treatment D on an ordered response Y. The treatment variable is such that $D = 1$ whenever the treatment is received, and $D = 0$ otherwise. It is often useful to think of D as a dummy endogenous variable in the model for Y, provided that the treatment status is determined by self-selected individuals rather than randomly assigned treatment groups. In terms of potential outcomes (Neyman 1923, Rubin 1974), let Y_1 denote the potential outcome with treatment, and let Y_0 denote the potential outcome without treatment. The measured outcome Y is related to potential outcomes (Y_1, Y_0) so that

$$Y = DY_1 + (1 - D)Y_0 \tag{4.1}$$

Assume that the total number of categories is independent of the treatment status, i.e., irrespective of being treated or not the individual will face the same set of mutually exclusive and exhaustive ordered categories. Without loss of generality, let $\mathcal{Y} = \{1, 2, \ldots, J\}$

denote the set of possible outcomes of Y, where "1" $<$ "2" $<$ ... $<$ "J". The assigned values in $\mathcal{Y}$ are entirely meaningless, as long as they keep the ordering, and are just for notational convenience.

One can think of a number of applications with a binary treatment and an ordered response. For example, in medical research the effectiveness of a new drug may be evaluated regarding the patient's health status, in educational economics one may be interested in the effect of out-of-school training programs on exam grades, and in labor economics the sorting of workers into public and private sector jobs may be analyzed with respect to their economic performance, the latter measured as promotion, lateral move, or demotion.

The ordinal nature of Y needs to be taken into account when defining treatment effects. With quantitative and binary outcomes, the individual treatment effect $Y_1 - Y_0$ has potential interest. For example, if Y measures wages and D is participation in job training, then $Y_1 - Y_0$ gives the wage difference with and without the training program. If Y indicates, for example, one-year survival after cardiac surgery and D indicates medical treatment, then $Y_1 - Y_0$ shows whether the individual would survive with medication and die without (1), is not affected by medication (0), or would survive without and die with medication (-1). For ordinal variables such an interpretation does not exist because the distance between outcomes is not defined.

In practice, only one of two potential outcomes Y_1 or Y_0 can be observed because each individual either receives the treatment, or does not. Thus, it is impossible to recover the individual treatment effect and the literature typically focuses on averages of $Y_1 - Y_0$, such as the average treatment effect, $E(Y_1 - Y_0)$, or the average treatment effect on the treated, $E(Y_1 - Y_0|D = 1)$. Under certain assumptions, these parameters can at least partly be recovered from observed data (Heckman and Robb 1985, 1986, Manski 1990, 1994, 1995, Imbens and Angrist 1994, Angrist *et al.* 1996, Heckman *et al.* 1999, among many others). With ordinal data, again, the case is different: Any rank preserving recoding of the elements in $\mathcal{Y}$ should not affect the parameters of interest. $E(Y_1)$ and $E(Y_0)$, however, will be affected by such a value conversion, so that the concept of averages needs to be replaced by a concept insensitive to the definition of $\mathcal{Y}$.

For these reasons I propose to analyze treatment effects for ordinal outcomes in terms of probabilities rather than expectations. Investigating treatment effects in terms of probabilities is particularly attractive for discrete responses as each outcome occurs with a positive probability, and analyzing probability effects is thus of interest on its own. Let the "average" treatment effect (ATE) be defined as the probability difference of observing a particular outcome with and without the treatment, formally

$$\Delta_y^{ATE} \equiv P(Y_1 = y) - P(Y_0 = y) \qquad y = 1, \ldots, J \tag{4.2}$$

Note that there are indeed J effects, one for each outcome of Y. If the treatment affects responses positively — adopting the convention that higher outcomes of Y are in some way "better" than smaller outcomes —, then one would expect Δ_y^{ATE} negative for low y and positive for high y. In practice, there may not exist such a clear systematic indicating whether the treatment has a positive or a negative effect, but the shift in focus to probability effects allows for a detailed analysis of the effects of the treatment in all parts of the outcome distribution.

Analogously, the effect on outcome probabilities for individuals who actually received the treatment can be defined as treatment on the treated parameter (TT)

$$\Delta_y^{TT} \equiv P(Y_1 = y|D = 1) - P(Y_0 = y|D = 1) \qquad y = 1, \ldots, J \tag{4.3}$$

Both treatment parameters are robust against the particular values assigned to outcomes, but rely on the "same scale" assumption. Yet this assumption is not overly restrictive, as otherwise it would be difficult to compare the Y_1 and the Y_0 distribution. One may also define other treatment effect parameters in terms of probabilities rather than expectations, such as the local average treatment effect (LATE) of Imbens and Angrist (1994), or the marginal treatment effect (MTE) of Heckman (1997). In this paper, I will confine myself on the parameters in (4.2) and (4.3), but some remarks on other parameters will be given below.

Δ_y^{ATE} and Δ_y^{TT} are not immediately identified from the population distribution of (Y, D). To see why, consider the average treatment effect and $P(Y_1)$.[1] By the law of total

[1] From now on, I will drop the y argument in the probability statements if possible to save some notation, e.g., $P(Y_1)$ will be shorthand notation for $P(Y_1 = y)$, or $P(Y_1|D = 1)$ will be shorthand for $P(Y_1 = y|D = 1)$. If not mentioned otherwise, the equations will hold for all $y = 1, \ldots, J$.

probability,

$$P(Y_1) = P(Y_1|D=1)P(D=1) + P(Y_1|D=0)P(D=0)$$

The sampling process identifies the probability of treatment selection, $P(D=1)$, and the outcome probability with treatment given treatment has been received, $P(Y_1|D=1) = P(Y|D=1)$. The sampling process is uninformative, however, regarding $P(Y_1|D=0)$, which is the outcome probability with treatment, given the treatment has not been received. In the common terminology such a term is referred to as counterfactual probability. $P(Y_0)$ is not identified either, because the sampling process does not reveal $P(Y_0|D=1)$, and therefore the average treatment effect is not identified. Lack of observability of the counterfactual $P(Y_0|D=1)$ also makes identification of the treatment on the treated parameter fail.

The aim of this paper is to find reasonable bounds on counterfactual probabilities in a setting with ordinal outcomes and binary treatment, and thus to bound the treatment effect parameters. As a starting point and without imposing any assumptions on the data-generating process, it must certainly hold that both counterfactuals, $P(Y_1|D=0)$ and $P(Y_0|D=1)$, are bounded by zero and one. The average treatment effect is thus bounded by

$$\Delta_y^{ATE} \in \left[L\mathcal{B}1_y^{ATE}, U\mathcal{B}1_y^{ATE}\right] \qquad \text{with} \tag{4.4}$$

$$\begin{aligned} L\mathcal{B}1_y^{ATE} &= P(D=1,Y=y) - P(D=1) - P(D=0,Y=y) \\ U\mathcal{B}1_y^{ATE} &= P(D=1,Y=y) + P(D=0) - P(D=0,Y=y) \end{aligned}$$

Analogously, the average treatment effect on the treated is restricted to the interval

$$\Delta_y^{TT} \in \left[L\mathcal{B}1_y^{TT}, U\mathcal{B}1_y^{TT}\right] \qquad \text{with} \tag{4.5}$$

$$L\mathcal{B}1_y^{TT} = P(Y=y|D=1) - 1, \qquad U\mathcal{B}1_y^{TT} = P(Y=y|D=1)$$

The intervals in (4.4) and (4.5) define identification regions for the treatment parameters since all valid probability distributions $P(Y_1|D=0)$ and $P(Y_0|D=1)$ necessarily yield treatment effects within the stated bounds (Manski 2000, 2003). Note that the width of

the regions is one, which is the logical maximum for a probability effect. Note too that the bounds are not informative regarding the sign of both treatment parameters as zero is included in the range of possible values. The question to be investigated in the following sections is how further assumptions on the sampling process do narrow these bounds.

More specifically, the paper will explore a nonparametric threshold crossing model on both the ordered potential outcomes and the treatment selection. Ordinal data modeling is traditionally based on latent variables and threshold crossing mechanisms. For example, parametric models like the ordered probit and the ordered logit model follow this structure (McKelvey and Zavoina 1975, McCullagh 1980), but also semiparametric approaches like Klein and Sherman (2002), Bellemare *et al.* (2002), Coppejans (2007), Lewbel (1997, 2003), and Stewart (2004) impose a threshold crossing model to generate ordinality in the response variable. It therefore seems natural to analyze the implications of such a model structure in a nonparametric bounding analysis. The model is nonparametric in the sense that no distributional assumptions, and no functional form assumptions will be imposed other than the threshold mechanism.

Three recent papers are related to mine. First, Shaikh and Vytlacil (2005) discuss treatment effect bounds with a binary response variable and a binary treatment. They impose nonparametric threshold crossing models on both the treatment selection and the binary potential outcomes, whereas the model here assumes ordinal potential outcomes. As it will be worked out below, this requires a slightly different bounding strategy, and supplemental interpretations can be given in the extended setting. Second, Scharfstein *et al.* (2004) analyze bounds on the distribution of ordinal outcomes, but their model setup is different from mine because they consider two outcome variables where the first is always observed and the second (sequentially following the first) is potentially missing so that the joint distribution of the two outcomes is not identified. Third, Li and Tobias (2007) describe Bayesian estimation of treatment effects for ordinal outcomes. They impose more structure on the model than it is imposed here and focus on mean treatment effect parameters (and therefore require additional implicit assumptions on the kind of ordinal response variable that is analyzed).

4.2 Model and Assumptions

The model for the treatment status and the potential outcomes is a version of the model in Shaikh and Vytlacil (2005) generalized to the case of ordinal outcomes and defined as

$$\begin{aligned} D^* &= s(Z) - \nu \qquad D = \mathbf{1}(D^* \geq 0) \\ Y_0^* &= r_0(X) + \varepsilon_0 \qquad Y_0 = \sum_{y=1}^{J} y\mathbf{1}(\kappa_{0y-1} < Y_0^* \leq \kappa_{0y}) \\ Y_1^* &= r_1(X) + \varepsilon_1 \qquad Y_1 = \sum_{y=1}^{J} y\mathbf{1}(\kappa_{1y-1} < Y_1^* \leq \kappa_{1y}) \end{aligned} \tag{4.6}$$

where (X, Z) is a random vector of observed covariates, ν, ε_0, and ε_1 are unobserved random variables, and $\mathbf{1}(\cdot)$ is the logical indicator function. The model is a latent index model with latent variables D^*, Y_0^*, and Y_1^*, and a threshold crossing mechanism that generates the treatment status D and the potential outcomes Y_0 and Y_1. The model is nonparametric in the sense that the functional forms of $s(Z)$, $r_0(X)$, and $r_1(X)$ are left unspecified and no parametric assumption on the distribution of $(\varepsilon_0, \varepsilon_1, \nu)$ is made. The model presumes that the error terms and the functions of observable factors are additively separable; see Vytlacil (2002, 2006) for a discussion of this property in latent index threshold crossing models. Finally, the observed outcome Y is generated according to (4.1), completing the model.

The definition of treatment parameters and the identification regions stated in the introduction still hold conditional on the vector of observed covariates X. In this case, the average treatment effect and the average treatment effect on the treated are local (conditional on X), and unconditional treatment effects may be obtained as weighted averages. The model as presented above also includes a vector Z that affects the treatment selection. Z may contain all elements of X, and additional elements in Z will generally be referred to as instrumental variables. X may or may not contain an element that is not included in Z. If such an element exists, then this information can be gainfully employed in the bounding analysis. Let $\mathcal{X}$ denote the support of the random vector X, and let $\mathcal{Z}$ denote the support of the random vector Z.

The assumptions imposed on the model (extending Shaikh and Vytlacil 2005) are:

(A1) The threshold parameters κ_{0j}, κ_{1j}, $j = 0, \ldots, J$ are fixed and fulfill the order requirement $-\infty = \kappa_{00} < \kappa_{01} < \ldots < \kappa_{0J} = \infty$, and $-\infty = \kappa_{10} < \kappa_{11} < \ldots < \kappa_{1J} = \infty$.

(A2) For some $x_0 \in \mathcal{X}$ let $r_0(x_0) = 0$, and for some $x_1 \in \mathcal{X}$ let $r_1(x_1) = 0$.

(A3) The distribution of ν is absolutely continuous with respect to Lebesgue measure.

(A4) $(\varepsilon_0, \varepsilon_1, \nu) \perp\!\!\!\perp (Z, X)$.

(A5) $\varepsilon_j|\nu \sim \varepsilon|\nu$, $j = 0, 1$.

(A6) The distribution of $\varepsilon_j|\nu$ has strictly positive density with respect to Lebesgue measure on $\mathbf{R}$, $j = 0, 1$.

(A7) $s(Z)$ is non-degenerate conditional on X.

(A8) The support of the distribution of (X, Z) is compact, and $r_0(\cdot)$, $r_1(\cdot)$, $s(\cdot)$ are continuous.

For a detailed discussion of assumptions (A3) to (A8) see Shaikh and Vytlacil (2005). Crucial in the following analysis are the independence assumption (A4) and the restriction to equal distributions of ε_1 and ε_0 conditional on ν (A5). The additional assumptions (A1) and (A2) are imposed due to the ordinal nature of Y. (A1) in combination with the model equation explicitly accounts for the order information. The threshold parameters are assumed to be unknown, although the extension to known thresholds (interval data) is possible. In the latter case, knowledge of thresholds in both treatment statuses is required, unless they are independent of treatment and thus equal. Knowledge of κ_0 and κ_1 will considerably simplify the analysis, and remarks will be given at the appropriate places when the additional information can be used.

The model allows for much flexibility in the threshold mechanism since no distributional or functional form assumptions are imposed and the (unknown) threshold parameters are allowed to vary by the treatment status. In particular, the model does not restrict the shape of treatment effects in a way similar to the single crossing property of probability effects in standard parametric ordered probit and logit models (Boes and

Winkelmann 2006), nor does it require a specific model for the threshold parameters in order to relax this property.

Assumption (A2) is an identifying assumption that simplifies exposition and is standard in parametric models. If (A2) is not met, then parametric ordinal response models may only identify location-normalized instead of absolute threshold parameters, i.e., κ_0, κ_1 will be replaced by $\kappa_0 - r_0$ and $\kappa_1 - r_1$, respectively, where r_0, r_1 denote the constant terms in $r_0(X), r_1(X)$. As it is irrelevant for the following analysis if all thresholds are shifted equally to the right or to the left, (A2) is purely simplifying and does not restrict the analysis in any way.

4.3 Bounds on Treatment Effects

For the ease of exposition, I will first consider bounds on the treatment effect parameters when no X covariates are available (Sections 4.3.1 and 4.3.2). In this case, the latent potential outcome equations of the model simplify to $Y_1^* = \varepsilon_1$ and $Y_0^* = \varepsilon_0$. The extension to the case when X covariates are present will be separately discussed below (Section 4.3.3).

4.3.1 Bounds under the Independence Assumption

The first bounding strategy follows Manski (1990, 1994). Assume that potential outcomes (Y_0, Y_1) are independent of Z, but that treatment selection D varies with Z. One may interpret such a condition as exclusion restriction, and Z is an instrumental variable. It is easy to verify that the model assumptions in Section 4.2 imply this condition but not vice versa, i.e., the assumptions imposed by the model are stronger than the exclusion restriction alone. Given independence, it must hold that $P(Y_1|Z = z) = P(Y_1)$ for all $z \in \mathcal{Z}$.[2] Moreover, write

$$P(Y_1|Z) = P(Y_1|D = 1, Z)P(D = 1|Z) + P(Y_1|D = 0, Z)P(D = 0|Z)$$

[2] In order to save some notation, I will drop the particular value z (or later on x) that is conditioned on if it is not critical in the given context. It will be implicitly assumed that all expressions are only evaluated over the appropriate support, i.e., at all evaluation points the conditional probabilities exist and are well-defined.

In this expression, all probabilities but the counterfactual $P(Y_1|D=0,Z)$ are identified from the population distribution (Y,D,Z). The unidentified probability is bounded by zero and one which in turn imposes upper and lower bounds on $P(Y_1|Z)$. Due to the independence assumption, the smallest of $P(D=1,Y|Z=z)+P(D=0|Z=z)$ — which is the upper bound of $P(Y_1|Z=z)$ — over all $z \in \mathcal{Z}$ may be used as a new upper bound for $P(Y_1)$, and the largest of $P(D=1,Y|Z=z)$ — which is the lower bound of $P(Y_1|Z=z)$ — over all $z \in \mathcal{Z}$ may be used as a new lower bound for $P(Y_1)$. Analogously, new upper and lower bounds for $P(Y_0)$ may be obtained and the average treatment parameter can be bounded by

$$\Delta_y^{ATE} \in \left[L\mathcal{B}\mathcal{2}_y^{ATE}, U\mathcal{B}\mathcal{2}_y^{ATE}\right] \qquad \text{with} \tag{4.7}$$

$$\begin{aligned} L\mathcal{B}\mathcal{2}_y^{ATE} &= \sup_{z\in\mathcal{Z}}\{P(D=1,Y=y|Z=z)\} \\ &\quad - \inf_{z\in\mathcal{Z}}\{P(D=1|Z=z)+P(D=0,Y=y|Z=z)\} \\ U\mathcal{B}\mathcal{2}_y^{ATE} &= \inf_{z\in\mathcal{Z}}\{P(D=1,Y=y|Z=z)+P(D=0|Z=z)\} \\ &\quad - \sup_{z\in\mathcal{Z}}\{P(D=0,Y=y|Z=z)\} \end{aligned}$$

where $\sup\{\cdot\}$ denotes the supremum and $\inf\{\cdot\}$ the infimum of the argument in curly brackets over the values indicated in the subscript.

For the treatment on the treated effect note that in general $P(Y_0|D=1,Z) \neq P(Y_0|D=1)$, i.e., $Y_0|D=1$ is not independent of Z, as the instrument does affect the treatment status. One option to proceed would be to re-define the treatment on the treated parameter conditional on Z, or conditional on $P(D=1|Z)$, and then obtain the unconditional parameter by integration. I follow an alternative strategy and rewrite the counterfactual $P(Y_0|D=1)$ in terms of an identified probability and a probability that can be bounded under independence. It must hold that

$$\begin{aligned} P(Y_0=y|D=1) &= P(D=1,Y_0=y)/P(D=1) \\ &= [P(Y_0=y)-P(D=0,Y_0=y)]/P(D=1) \end{aligned}$$

by Bayes' theorem and the law of total probability. The sampling process identifies $P(D=1)$ and $P(D=0,Y_0)=P(D=0,Y)$, but only partially identifies $P(Y_0)$. Given

$Y_0 \perp\!\!\!\perp Z$, one may construct upper and lower bounds on $P(Y_0)$ in the same manner as above. Rewrite the treatment on the treated parameter as

$$\begin{aligned}
\Delta_y^{TT} &= [P(D=1,Y_1=y) - P(D=1,Y_0=y)]/P(D=1) \\
&= [P(D=1,Y=y) - P(Y_0=y) + P(D=0,Y=y)]/P(D=1) \\
&= [P(Y=y) - P(Y_0=y)]/P(D=1) \qquad (4.8)
\end{aligned}$$

so that

$$\Delta_y^{TT} \in \left[L\mathcal{B}\mathcal{2}_y^{TT}, U\mathcal{B}\mathcal{2}_y^{TT}\right] \qquad \text{with} \qquad (4.9)$$

$$\begin{aligned}
L\mathcal{B}\mathcal{2}_y^{TT} &= \Big[P(Y=y) - \inf_{z\in\mathcal{Z}}\{P(D=1|Z=z) \\
&\qquad + P(D=0,Y=y|Z=z)\}\Big]/P(D=1) \\
U\mathcal{B}\mathcal{2}_y^{TT} &= \Big[P(Y=y) - \sup_{z\in\mathcal{Z}}\{P(D=0,Y=y|Z=z)\}\Big]/P(D=1)
\end{aligned}$$

Note that the bounds in (4.7) and (4.9) do not exploit the ordinal nature of the response variable, nor do they exploit the threshold crossing structure of the model. The analysis may therefore be applied to any nominal response Y and binary treatment D. The question to be investigated in the following is how such additional assumptions on the structure of the data can be used to improve upon (4.7) and (4.9).

4.3.2 Bounds Under the Threshold Crossing Model Structure

The bounding strategy of this section generalizes Heckman and Vytlacil (2001) and Shaikh and Vytlacil (2005) to the case of ordinal potential outcomes. Given the threshold crossing structure of the treatment selection equation and the independence assumption, it follows that for any two evaluation points $z_1, z_0 \in \mathcal{Z}$

$$\begin{aligned}
P(D=1|Z=z_1) > P(D=1|Z=z_0) &\Leftrightarrow P(s(z_1) \geq \nu) > P(s(z_0) \geq \nu) \\
&\Leftrightarrow s(z_1) > s(z_0)
\end{aligned}$$

Furthermore, let

$$\begin{aligned}
z^u &= \arg\sup_{z\in\mathcal{Z}} P(D=1|Z=z) \\
z^l &= \arg\inf_{z\in\mathcal{Z}} P(D=1|Z=z)
\end{aligned}$$

This information can be used in two ways. First, by definition of z^u and z^l it must hold that $s(z^u) \geq s(z)$ and $s(z^l) \leq s(z)$ for all $z \in \mathcal{Z}$. The following lemma then simplifies the supremum and infimum expressions in the bounds on the average treatment and the treatment on the treated parameters as stated in (4.7) and (4.9):

Lemma 1 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A4) and (A7)-(A8) are fulfilled. Then,*

$$(a) \quad \sup_{z \in \mathcal{Z}} \{P(D=1, Y=y|Z=z)\} = P(D=1, Y=y|Z=z^u)$$

$$(b) \quad \sup_{z \in \mathcal{Z}} \{P(D=0, Y=y|Z=z)\} = P(D=0, Y=y|Z=z^l)$$

$$(c) \quad \inf_{z \in \mathcal{Z}} \{P(D=1, Y=y|Z=z) + P(D=0|Z=z)\}$$
$$= P(D=1, Y=y|Z=z^u) + P(D=0|Z=z^u)$$

$$(d) \quad \inf_{z \in \mathcal{Z}} \{P(D=1|Z=z) + P(D=0, Y=y|Z=z)\}$$
$$= P(D=1|Z=z^l) + P(D=0, Y=y|Z=z^l)$$

Proof. First consider part (a) of the lemma. Recall that at all evaluation points the conditional probabilities exist and are well-defined. The assumptions of the lemma ensure that

$$\begin{aligned} & P(D=1, Y|Z=z^u) - P(D=1, Y|Z=z) \qquad (4.10) \\ &= P(\nu \leq s(z^u), Y_1) - P(\nu \leq s(z), Y_1) \\ &= P(s(z) < \nu \leq s(z^u), Y_1) \geq 0 \end{aligned}$$

where the weak inequality follows by definition of z^u. The supremum of $P(D=1, Y=y|Z=z)$ over z is equivalent to the infimum of (4.10) over z. As (4.10) must be non-negative, necessary and sufficient condition for an infimum of (4.10) is that $z = z^u$. Analogously,

$$\begin{aligned} & P(D=0, Y|Z=z^l) - P(D=0, Y|Z=z) \qquad (4.11) \\ &= P(\nu > s(z^l), Y_0) - P(\nu > s(z), Y_0) \\ &= P(\nu \leq s(z), Y_0) - P(\nu \leq s(z^l), Y_0) \\ &= P(s(z^l) < \nu \leq s(z), Y_0) \geq 0 \end{aligned}$$

where the weak inequality follows by definition of z^l. The supremum in part (b) of the lemma is equivalent to the infimum of (4.11) over z, and, by the assumptions of the model and given the weak inequality, $z = z^l$ is necessary and sufficient for a supremum of $P(D = 0, Y = y|Z = z)$. In order to show part (c) of the lemma, write

$$\begin{aligned} & P(D = 1, Y|Z = z^u) + P(D = 0|Z = z^u) \\ & \quad -P(D = 1, Y|Z = z) - P(D = 0|Z = z) \qquad (4.12) \\ = \; & P(D = 1, Y|Z = z^u) - P(D = 1, Y|Z = z) \\ & \quad -[P(D = 1|Z = z^u) - P(D = 1|Z = z)] \\ = \; & P(s(z) < \nu \leq s(z^u), Y_1) - P(s(z) < \nu \leq s(z^u)) \leq 0 \end{aligned}$$

where the weak inequality follows by definition of z^u and the law of total probability. The infimum of $P(D = 1, Y = y|Z = z) + P(D = 0|Z = z)$ is equivalent to the supremum of (4.12) both over z. As (4.12) must be non-positive, necessary and sufficient condition for a supremum of (4.12) is that $z = z^u$. Analogous arguments prove part (d) of the lemma. □

A direct implication of Lemma 1 is that the bounds on the average treatment and the treatment on the treated parameters as stated in Section 4.3.1 simplify to

$$\Delta_y^{ATE} \in \left[L\mathcal{B}\beta_y^{ATE}, U\mathcal{B}\beta_y^{ATE}\right] \qquad \text{with} \qquad (4.13)$$

$$\begin{aligned} L\mathcal{B}\beta_y^{ATE} &= P(D = 1, Y = y|Z = z^u) - P(D = 1|Z = z^l) \\ &\quad -P(D = 0, Y = y|Z = z^l) \\ U\mathcal{B}\beta_y^{ATE} &= P(D = 1, Y = y|Z = z^u) + P(D = 0|Z = z^u) \\ &\quad -P(D = 0, Y = y|Z = z^l) \end{aligned}$$

and

$$\Delta_y^{TT} \in \left[L\mathcal{B}\beta_y^{TT}, U\mathcal{B}\beta_y^{TT}\right] \qquad \text{with} \qquad (4.14)$$

$$\begin{aligned} L\mathcal{B}\beta_y^{TT} &= [P(Y = y) - P(D = 1|Z = z^l) - P(D = 0, Y = y|Z = z^l)]/P(D = 1) \\ U\mathcal{B}\beta_y^{TT} &= [P(Y = y) - P(D = 0, Y = y|Z = z^l)]/P(D = 1) \end{aligned}$$

Compared to the bounds in (4.7) and (4.9), the bounds in (4.13) and (4.14) can be readily evaluated once z^u and z^l are determined. It is also possible to calculate their width; for the average treatment effect the width is given by $P(D=0|Z=z^u)+P(D=1|Z=z^l)$, and for the treatment on the treated parameter the width is given by $P(D=1|Z=z^l)/P(D=1)$. Both are smaller than one given that treatment selection varies with Z, i.e., for both treatment parameters the independence assumption together with the threshold crossing treatment selection is informative and yields narrower bounds than the identification regions stated in the introduction. Note however that the bounds in (4.13) and (4.14) do not yield tighter bounds than those in (4.7) and (4.9), because the former are simply a special case of the latter, but the imposed model structure considerably simplifies the form and the calculation of the bounds.

The second implication of the threshold crossing treatment selection can be derived in combination with the threshold model for the potential outcomes. Let

$$sgn(a) = \begin{cases} -1 & \text{if } a<0 \\ 0 & \text{if } a=0 \\ 1 & \text{if } a>0 \end{cases}$$

denote the sign function, and consider the following lemma:

Lemma 2 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then for any two evaluation points z_1, z_0 with $P(D=1|Z=z_1) > P(D=1|Z=z_0)$,*

$$sgn[P(Y\leq y|Z=z_1) - P(Y\leq y|Z=z_0)] = sgn(\kappa_{1y}-\kappa_{0y}) \equiv \delta_y$$

so that δ_y can take three values -1,0,1 depending on whether the difference $\kappa_{1y}-\kappa_{0y}$ is negative, zero, or positive, respectively.

Proof. Consider the cumulative outcome probability conditional on the instrument

$$\begin{aligned} P(Y\leq y|Z) &= P(D=1, Y\leq y|Z) + P(D=0, Y\leq y|Z) \\ &= P(D=1, Y_1\leq y|Z) + P(D=0, Y_0\leq y|Z) \\ &= P(\nu\leq s(z), \varepsilon_1\leq \kappa_{1y}) + P(\nu > s(z), \varepsilon_0\leq \kappa_{0y}) \\ &= P(\nu\leq s(z), \varepsilon\leq \kappa_{1y}) + P(\nu > s(z), \varepsilon\leq \kappa_{0y}) \end{aligned}$$

where the first equality follows by the law of total probability, the second equality follows by (4.1), the third equality follows by the model and the independence assumption, and the last equality follows by assumption (A5). Now take the difference of the cumulative outcome probabilities evaluated at any two evaluation points z_1, z_0 with $P(D = 1|Z = z_1) > P(D = 1|Z = z_0)$ such that $s(z_1) > s(z_0)$. Then,

$$\begin{aligned} & P(Y \leq y|Z = z_1) - P(Y \leq y|Z = z_0) \\ = \; & P(s(z_0) < \nu \leq s(z_1), \varepsilon \leq \kappa_{1y}) - P(s(z_0) < \nu \leq s(z_1), \varepsilon \leq \kappa_{0y}) \\ = \; & \begin{cases} P(s(z_0) < \nu \leq s(z_1), \kappa_{0y} < \varepsilon \leq \kappa_{1y}) & \text{iff} \quad \kappa_{1y} > \kappa_{0y} \\ 0 & \text{iff} \quad \kappa_{1y} = \kappa_{0y} \\ -P(s(z_0) < \nu \leq s(z_1), \kappa_{1y} < \varepsilon \leq \kappa_{0y}) & \text{iff} \quad \kappa_{1y} < \kappa_{0y} \end{cases} \end{aligned}$$

Thus, the sign of the difference in the cumulative probabilities can be used to identify the relative magnitude of threshold parameters. More precisely, the difference will be positive if and only if the difference between upper treated and upper nontreated threshold parameters is positive. The difference will be zero if and only if the upper thresholds are equal, and negative if and only if the difference between upper treated and upper non-treated thresholds is negative. □

Lemma 2 is analogous to Lemma 4.2 of Shaikh and Vytlacil (2005), but now with respect to the properties of ordinal potential outcomes. Information on the relative magnitude of threshold parameters can be used to tighten the bounds on the unidentified probabilities $P(Y_0|D = 1, Z)$ and $P(Y_1|D = 0, Z)$. Consider $P(Y_1|D = 0, Z)$ and recall that so far it was assumed that this probability was bounded by zero and one. Now write

$$P(Y_1 = y|D = 0, Z) = P(Y_1 \leq y|D = 0, Z) - P(Y_1 \leq y - 1|D = 0, Z)$$

which follows from the ordinal nature of Y. Furthermore, the difference

$$\begin{aligned} & P(Y_1 \leq y|D = 0, Z) - P(Y_0 \leq y|D = 0, Z) \qquad (4.15) \\ = \; & P(\varepsilon_1 \leq \kappa_{1y}|\nu > s(z)) - P(\varepsilon_0 \leq \kappa_{0y}|\nu > s(z)) \\ = \; & P(\varepsilon \leq \kappa_{1y}|\nu > s(z)) - P(\varepsilon \leq \kappa_{0y}|\nu > s(z)) \end{aligned}$$

has the same sign as $\kappa_{1y} - \kappa_{0y}$, and $\delta_y \equiv sgn(\kappa_{1y} - \kappa_{0y})$ is identified by Lemma 2. This must hold for all possible outcomes y, so that by the model assumptions, the sign of the difference

$$\begin{aligned} & P(Y_1 \leq y-1|D=0,Z) - P(Y_0 \leq y-1|D=0,Z) \qquad (4.16) \\ = \;& P(\varepsilon_1 \leq \kappa_{1y-1}|\nu > s(z)) - P(\varepsilon_0 \leq \kappa_{0y-1}|\nu > s(z)) \\ = \;& P(\varepsilon \leq \kappa_{1y-1}|\nu > s(z)) - P(\varepsilon \leq \kappa_{0y-1}|\nu > s(z)) \end{aligned}$$

equals $\delta_{y-1} \equiv sgn(\kappa_{1y-1} - \kappa_{0y-1})$. The strategy to bound the unidentified probabilities is a pairwise comparison of terms in the difference

$$\begin{aligned} & P(Y_1 = y|D=0,Z) - P(Y_0 = y|D=0,Z) \qquad (4.17) \\ = \;& [P(Y_1 \leq y|D=0,Z) - P(Y_1 \leq y-1|D=0,Z)] \\ & -[P(Y_0 \leq y|D=0,Z) - P(Y_0 \leq y-1|D=0,Z)] \\ = \;& [P(Y_1 \leq y|D=0,Z) - P(Y_0 \leq y|D=0,Z)] \\ & -[P(Y_1 \leq y-1|D=0,Z) - P(Y_0 \leq y-1|D=0,Z)] \end{aligned}$$

With three different outcomes of both δ_y and δ_{y-1} there are in total nine possibilities to consider. The following lemma states and summarizes the results for both unidentified probabilities:

Lemma 3 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then,*

$$\begin{aligned} \delta_y \;&>\; \delta_{y-1} \\ &\Leftrightarrow\; P(Y_1 = y|D=0,Z) > P(Y_0 = y|D=0,Z) = P(Y = y|D=0,Z) \\ &\qquad P(Y_0 = y|D=1,Z) < P(Y_1 = y|D=1,Z) = P(Y = y|D=1,Z) \\ \delta_y \;&=\; \delta_{y-1} \;=\; 0 \\ &\Leftrightarrow\; P(Y_1 = y|D=0,Z) = P(Y_0 = y|D=0,Z) = P(Y = y|D=0,Z) \\ &\qquad P(Y_0 = y|D=1,Z) = P(Y_1 = y|D=1,Z) = P(Y = y|D=1,Z) \\ \delta_y \;&<\; \delta_{y-1} \\ &\Leftrightarrow\; P(Y_1 = y|D=0,Z) < P(Y_0 = y|D=0,Z) = P(Y = y|D=0,Z) \\ &\qquad P(Y_0 = y|D=1,Z) > P(Y_1 = y|D=1,Z) = P(Y = y|D=1,Z) \end{aligned}$$

If $\delta_y = \delta_{y-1} = \pm 1$*, then the sign of the difference* $P(Y_1 = y|D = 0, Z) - P(Y_0 = y|D = 0, Z)$ *and the sign of the difference* $P(Y_0 = y|D = 1, Z) - P(Y_1 = y|D = 1, Z)$ *are indeterminate.*

Proof. Immediately follows by application of Lemma 2, (4.15), (4.16) and (4.17). Note that the case $\delta_y > \delta_{y-1}$ includes possibilities $(1, 0)$, $(1, -1)$, and $(0, -1)$ for pairs (δ_y, δ_{y-1}), and $\delta_y < \delta_{y-1}$ includes possibilities $(0, 1)$, $(-1, 1)$, and $(-1, 0)$. □

Lemma 2 identifies $\delta_y \equiv sgn(\kappa_{1y} - \kappa_{0y})$ for all $y \in \mathcal{Y}$. Lemma 3 then uses the information to impose bounds on counterfactual probabilities tighter than the logical unit range. Without loss of generality, take the two evaluation points z^l and z^u with $s(z^u) > s(z^l)$, and apply Lemma 2 to identify the relative magnitude of threshold parameters. Suppose, the information is revealed that $\delta_y > \delta_{y-1}$. Then $P(Y = y|D = 0, Z)$ can be used as a lower bound for $P(Y_1 = y|D = 0, Z)$ instead of zero, and $P(Y = y|D = 1, Z)$ can be used as an upper bound for $P(Y_0 = y|D = 1, Z)$ instead of one. Bounds on $P(Y_1|Z)$ and $P(Y_0|Z)$ are thus given by

$$\begin{aligned} P(Y = y|Z) &\leq P(Y_1 = y|Z) \leq P(D = 1, Y = y|Z) + P(D = 0|Z) \\ P(D = 0, Y = y|Z) &\leq P(Y_0 = y|Z) \leq P(Y = y|Z) \end{aligned}$$

If alternatively the information is revealed that $\delta_y < \delta_{y-1}$, then the bounds on $P(Y_1|Z)$, $P(Y_0|Z)$ can be derived as

$$\begin{aligned} P(D = 1, Y = y|Z) &\leq P(Y_1 = y|Z) \leq P(Y = y|Z) \\ P(Y = y|Z) &\leq P(Y_0 = y|Z) \leq P(D = 1|Z) + P(D = 1, Y = y|Z) \end{aligned}$$

If upper and lower treated and non-treated thresholds are equal, then the outcome of Y does not vary with the treatment status because the cumulative probabilities are unchanged, and the unidentified probabilities become identified, i.e., $P(Y_1|Z) = P(Y|Z) = P(Y_0|Z)$. The bounds imposed by Lemma 3 thus depend on the category under consideration, i.e., one may have $\delta_y > \delta_{y-1}$, but $\delta_{y+1} < \delta_y$, such that the restrictions on counterfactual probabilities in category y are different from the restrictions in category

$y+1$. If Lemmas 2 and 3 do not reveal further information on the counterfactual probabilities, then the lower bound zero and the upper bound one on $P(Y_1|D=0,Z)$ and $P(Y_0|D=1,Z)$ still apply.

As argued above in the derivation of bounds under independence, the model assumptions imply that $P(Y_1|Z)=P(Y_1)$ and $P(Y_0|Z)=P(Y_0)$. $P(Y_1)$ and $P(Y_0)$ must therefore necessarily lie within the intersection over all possible z so that lower bounds can be replaced by supremum expressions, and upper bounds can be replaced by infimum expressions. With the exception of $\sup_{z\in\mathcal{Z}}\{P(Y=y|Z=z)\}$ and $\inf_{z\in\mathcal{Z}}\{P(Y=y|Z=z)\}$, all terms reduce according to Lemma 1. Simplification of the former is possible as well:

Lemma 4 *Assume that (Y_0,Y_1,D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then,*

$$\begin{aligned}
&(a1) && \sup_{z\in\mathcal{Z}}\{P(Y=y|Z=z)\} &&= P(Y=y|Z=z^u) && \text{if } \delta_y>\delta_{y-1}\\
&(a2) && \inf_{z\in\mathcal{Z}}\{P(Y=y|Z=z)\} &&= P(Y=y|Z=z^u) && \text{if } \delta_y<\delta_{y-1}\\
&(b1) && \inf_{z\in\mathcal{Z}}\{P(Y=y|Z=z)\} &&= P(Y=y|Z=z^l) && \text{if } \delta_y>\delta_{y-1}\\
&(b2) && \sup_{z\in\mathcal{Z}}\{P(Y=y|Z=z)\} &&= P(Y=y|Z=z^l) && \text{if } \delta_y<\delta_{y-1}
\end{aligned}$$

Proof. Consider part (a1) and recall that $s(z^u)\geq s(z)$ for all z. The assumptions ensure that

$$\begin{aligned}
&P(Y=y|Z=z^u)-P(Y=y|Z=z) \qquad (4.18)\\
&= P(D=0,Y=y|Z=z^u)+P(D=1,Y=y|Z=z^u)\\
&\quad -P(D=0,Y=y|Z=z)-P(D=1,Y=y|Z=z)\\
&= P(\nu>s(z^u),\kappa_{0y-1}<\varepsilon\leq\kappa_{0y})+P(\nu\leq s(z^u),\kappa_{1y-1}<\varepsilon\leq\kappa_{1y})\\
&\quad -P(\nu>s(z),\kappa_{0y-1}<\varepsilon\leq\kappa_{0y})-P(\nu\leq s(z),\kappa_{1y-1}<\varepsilon\leq\kappa_{1y})\\
&= P(s(z)<\nu\leq s(z^u),\kappa_{1y-1}<\varepsilon\leq\kappa_{1y})\\
&\quad -P(s(z)<\nu\leq s(z^u),\kappa_{0y-1}<\varepsilon\leq\kappa_{0y})\geq 0
\end{aligned}$$

where the last inequality follows by definition of $s(z^u)$ and $\delta_y>\delta_{y-1}$. Since the supremum in part (a1) of the lemma is equivalent to the infimum of (4.18) over z and (4.18) must be non-negative, necessary and sufficient condition for an infimum of (4.18) is that $z=z^u$.

If $\delta_y < \delta_{y-1}$, then (4.18) holds under the weak inequality ≤ 0, and the infimum in part (a2) of the lemma is equivalent to the supremum of (4.18) over z. As (4.18) must be non-positive, necessary and sufficient condition for a supremum is that $z = z^u$. Following analogous arguments for the infimum in the case $\delta_y > \delta_{y-1}$ and the supremum in the case $\delta_y < \delta_{y-1}$ proves parts (b1) and (b2) of the lemma. □

The following proposition uses the bounds on $P(Y_0)$ and $P(Y_1)$ under the threshold crossing model structure of treatment selection and potential outcomes to bound the average treatment and the treatment on the treated parameters:

Proposition 1 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then,*

$$\Delta_y^{ATE} \in \left[L\mathcal{B}4_y^{ATE}, U\mathcal{B}4_y^{ATE}\right] \qquad with \tag{4.19}$$

$$L\mathcal{B}4_y^{ATE} = \begin{cases} P(Y=y|Z=z^u) - P(Y=y|Z=z^l) & if \ \ \delta_y > \delta_{y-1} \\ 0 & if \ \ \delta_y = \delta_{y-1} = 0 \\ L\mathcal{B}3_y^{ATE} & if \ \ \delta_y < \delta_{y-1} \\ L\mathcal{B}3_y^{ATE} & if \ \ \delta_y = \delta_{y-1} = \pm 1 \end{cases}$$

$$U\mathcal{B}4_y^{ATE} = \begin{cases} U\mathcal{B}3_y^{ATE} & if \ \ \delta_y > \delta_{y-1} \\ 0 & if \ \ \delta_y = \delta_{y-1} = 0 \\ P(Y=y|Z=z^u) - P(Y=y|Z=z^l) & if \ \ \delta_y < \delta_{y-1} \\ U\mathcal{B}3_y^{ATE} & if \ \ \delta_y = \delta_{y-1} = \pm 1 \end{cases}$$

and

$$\Delta_y^{TT} \in \left[L\mathcal{B}4_y^{TT}, U\mathcal{B}4_y^{TT}\right] \qquad with \tag{4.20}$$

$$L\mathcal{B}4_y^{TT} = \begin{cases} [P(Y=y) - P(Y=y|Z=z^l)]/P(D=1) & if \ \ \delta_y > \delta_{y-1} \\ 0 & if \ \ \delta_y = \delta_{y-1} = 0 \\ L\mathcal{B}3_y^{TT} & if \ \ \delta_y < \delta_{y-1} \\ L\mathcal{B}3_y^{TT} & if \ \ \delta_y = \delta_{y-1} = \pm 1 \end{cases}$$

$$
UB4_y^{TT} = \begin{cases} UB3_y^{TT} & \text{if } \delta_y > \delta_{y-1} \\ 0 & \text{if } \delta_y = \delta_{y-1} = 0 \\ [P(Y=y) - P(Y=y|Z=z^l)]/P(D=1) & \text{if } \delta_y < \delta_{y-1} \\ UB3_y^{TT} & \text{if } \delta_y = \delta_{y-1} = \pm 1 \end{cases}
$$

For known threshold parameters (interval data), (4.19) and (4.20) still hold, but δ_y and δ_{y-1} can a-priori be determined and there is no uncertainty about the four cases.

Proof. Follows directly by Lemmas 1, 2, 3, and 4, and the discussion preceding Lemma 4. For known threshold parameters the identification strategy of Lemma 2 becomes redundant. Given the additional information, bounds on the unidentified counterfactuals $P(Y_0|D=1,Z)$ and $P(Y_1|D=0,Z)$ can be directly imposed as described in Lemma 3 with δ_y, δ_{y-1} known. □

Note that the width of the bounds in (4.19) and (4.20) is at maximum the same and in many cases smaller than the width of the bounds in (4.13) and (4.14). If $\delta_y > \delta_{y-1}$, then the upper bound in (4.19) corresponds to the upper bound in (4.13), but the lower bound in (4.19) is larger than the lower bound in (4.13), since

$$
\begin{aligned} LB4_y^{ATE} - LB3_y^{ATE} &= P(D=0, Y=y|Z=z^u) \\ &\quad -P(D=1, Y=y|Z=z^l) + P(D=1|Z=z^l) > 0 \end{aligned}
$$

With the same argument, if $\delta_y < \delta_{y-1}$, then the lower bounds in (4.19) and (4.13) are the same, but the upper bound in (4.19) is lower than the upper bound in (4.13), i.e., $UB4_y^{ATE} - UB3_y^{ATE} < 0$.

Analogously, for the treatment on the treated parameter and a positive sign of the difference $\delta_y - \delta_{y-1}$, the lower bound in (4.20) is larger than the lower bound in (4.14), i.e., $LB4_y^{TT} - LB3_y^{TT} > 0$, with the upper bounds unchanged, and if $\delta_y - \delta_{y-1}$ is negative, then the upper bound in (4.20) is lower than the upper bound in (4.14), i.e., $UB4_y^{TT} - UB3_y^{TT} < 0$, with the lower bounds unchanged. If $\delta_y = \delta_{y-1} = 0$, then both treatment parameters become point-identified to be zero. Only if $\delta_y = \delta_{y-1} = \pm 1$, then the width of the bounds does not change and the threshold mechanism is uninformative on the treatment parameters.

Note that unlike for the bounds constructed before, the sign of Δ_y^{ATE} and Δ_y^{TT} as bounded by Proposition 1 can be identified if $\delta_y \lesseqgtr \delta_{y-1}$ or $\delta_y = \delta_{y-1} = 0$. This follows because the lower bounds $L\mathcal{B}4_y^{ATE}$ and $L\mathcal{B}4_y^{TT}$ of both treatment parameters are positive in the case $\delta_y > \delta_{y-1}$, and in the case $\delta_y < \delta_{y-1}$ the upper bounds $U\mathcal{B}4_y^{ATE}$ and $U\mathcal{B}4_y^{TT}$ are negative. Finally, if $\delta_y = \delta_{y-1} = 0$, then the sign of the treatment effects is point-identified to be zero.

The final remark on (4.19) and (4.20) is related to the case of known thresholds. Given the assumptions of the model and provided that no X covariates are available, the only way that treated and non-treated individuals may differ are the threshold parameters. If the thresholds do not vary by the treatment status, and are thus equal, then $\delta_y = \delta_{y-1} = 0$ in all cases and the treatment parameters are point-identified to be zero, as predicted by Proposition 1.

4.3.3 Including Covariates

I now turn to the case when X covariates are available and to the full model (4.6). The treatment parameters conditional on X are defined as

$$\Delta_y^{ATE}(x) = P(Y_1 = y|X = x) - P(Y_0 = y|X = x) \tag{4.21}$$

$$\begin{aligned} \Delta_y^{TT}(x) &= P(Y_1 = y|D = 1, X = x) - P(Y_0 = y|D = 1, X = x) \\ &= [P(Y = y|X = x) - P(Y_0 = y|X = x)]/P(D = 1|X = x) \end{aligned} \tag{4.22}$$

By the preceding discussion, it is straightforward to show that $P(Y_1|X)$ and $P(Y_0|X)$ are only partially identified, and so are the treatment parameters. The offending terms are, as before, the counterfactuals $P(Y_1|D = 0, X)$ and $P(Y_0|D = 1, X)$, respectively. All the results derived before in (4.7) and (4.9), Lemma 1, and (4.13) and (4.14) are trivially extended to X conditioned on.

In principle, the same holds true for the whole discussion in the preceding section, i.e., Lemmas 2, 3, 4, and Proposition 1 may easily be extended to hold conditional on X. There is, however, a potential source of narrowing the bounds, given that X varies conditional on Z, i.e., there exists at least one element in X that is not included in Z.

This extra variation can be explored as follows. Consider a modified version of Lemma 2:

Lemma 5 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then for any evaluation points x_0, x_1, z_0, z_1 with $P(D = 1|X = x_j, Z = z_1) > P(D = 1|X = x_j, Z = z_0)$, $j = 0, 1$,*

$$\begin{aligned} sgn\Big\{&[P(D=1, Y \le y|X = x_1, Z = z_1) - P(D = 1, Y \le y|X = x_1, Z = z_0)] \\ &-[P(D=0, Y \le y|X = x_0, Z = z_0) - P(D = 0, Y \le y|X = x_0, Z = z_1)]\Big\} \\ &= sgn(\kappa_{1y}(x_1) - \kappa_{0y}(x_0)) \equiv \delta_y(x_1, x_0) \end{aligned}$$

so that $\delta_y(x_1, x_0)$ can take three values $-1, 0, 1$ depending on whether the difference between $\kappa_{1y}(x_1) \equiv \kappa_{1y} - r_1(x_1)$ and $\kappa_{0y}(x_0) \equiv \kappa_{0y} - r_0(x_0)$ is negative, zero, or positive, respectively.

Proof. Consider the probability differences in the sign function separately:

$$\begin{aligned} &P(D=1, Y \le y|X = x_1, Z = z_1) - P(D = 1, Y \le y|X = x_1, Z = z_0) \qquad (4.23) \\ &= P(D=1, Y_1 \le y|X = x_1, Z = z_1) - P(D = 1, Y_1 \le y|X = x_1, Z = z_0) \\ &= P(\nu \le s(z_1), \varepsilon_1 \le \kappa_{1y} - r_1(x_1)) - P(\nu \le s(z_0), \varepsilon_1 \le \kappa_{1y} - r_1(x_1)) \\ &= P(s(z_0) < \nu \le s(z_1), \varepsilon_1 \le \kappa_{1y}(x_1)) = P(s(z_0) < \nu \le s(z_1), \varepsilon \le \kappa_{1y}(x_1)) \end{aligned}$$

and

$$\begin{aligned} &P(D=0, Y \le y|X = x_0, Z = z_0) - P(D = 0, Y \le y|X = x_0, Z = z_1) \qquad (4.24) \\ &= P(D=0, Y_0 \le y|X = x_0, Z = z_0) - P(D = 0, Y_0 \le y|X = x_0, Z = z_1) \\ &= P(\nu > s(z_0), \varepsilon_0 \le \kappa_{0y} - r_0(x_0)) - P(\nu > s(z_1), \varepsilon_0 \le \kappa_{0y} - r_0(x_0)) \\ &= P(s(z_0) < \nu \le s(z_1), \varepsilon_0 \le \kappa_{0y}(x_0)) = P(s(z_0) < \nu \le s(z_1), \varepsilon \le \kappa_{0y}(x_0)) \end{aligned}$$

by the assumptions of the lemma, and $\kappa_{1y}(x_1) \equiv \kappa_{1y} - r_1(x_1)$ and $\kappa_{0y}(x_0) \equiv \kappa_{0y} - r_0(x_0)$. Taking the difference between (4.23) and (4.24) yields

$$\begin{aligned} &P(D=1, Y \le y|X = x_1, Z = z_1) - P(D = 1, Y \le y|X = x_1, Z = z_0) \\ &\quad -[P(D=0, Y \le y|X = x_0, Z = z_0) - P(D = 0, Y \le y|X = x_0, Z = z_1)] \\ &= \begin{cases} P(s(z_0) < \nu \le s(z_1), \kappa_{0y}(x_0) < \varepsilon \le \kappa_{1y}(x_1)) & \text{iff} \quad \kappa_{1y}(x_1) > \kappa_{0y}(x_0) \\ 0 & \text{iff} \quad \kappa_{1y}(x_1) = \kappa_{0y}(x_0) \\ -P(s(z_0) < \nu \le s(z_1), \kappa_{1y}(x_1) < \varepsilon \le \kappa_{0y}(x_0)) & \text{iff} \quad \kappa_{1y}(x_1) < \kappa_{0y}(x_0) \end{cases} \end{aligned}$$

Thus, the sign of the double difference in the cumulative probabilities can be used to identify the relative magnitude of $\kappa_{1y}(x_1)$ and $\kappa_{0y}(x_0)$. More precisely, the double difference will be positive if and only if the difference between $\kappa_{1y}(x_1) \equiv \kappa_{1y} - r_1(x_1)$ and $\kappa_{0y}(x_0) \equiv \kappa_{0y} - r_0(x_0)$ is positive. It will be zero if and only if the indices, accounting for the upper bound of the threshold mechanism, are equal, and negative if and only if $\kappa_{1y}(x_1) - \kappa_{0y}(x_0)$ is negative. □

Lemma 5 can be used to obtain bounds on the counterfactuals $P(Y_1 = y|D = 0, X, Z)$ and $P(Y_0 = y|D = 1, X, Z)$ tighter than the logical unit range. Consider the former counterfactual probability, and recall that

$$P(Y_1 = y|D = 0, X, Z) = P(Y_1 \leq y|D = 0, X, Z) - P(Y_1 \leq y - 1|D = 0, X, Z)$$

by the ordinal nature of Y. Take the first cumulative probability, evaluated at x_1, and subtract the identified probability $P(Y \leq y|D = 0, X, Z)$ evaluated at x_0 to obtain

$$\begin{aligned} & P(Y_1 \leq y|D = 0, X = x_1, Z) - P(Y_0 \leq y|D = 0, X = x_0, Z) \\ &= P(\varepsilon \leq \kappa_{1y}(x_1)|\nu > s(z)) - P(\varepsilon \leq \kappa_{0y}(x_0)|\nu > s(z)) \end{aligned}$$

The sign of the (unidentified) difference only depends on the sign of the difference $\kappa_{1y}(x_1) - \kappa_{0y}(x_0)$, which is identified by Lemma 5. Thus, if $\delta_y(x_1, x_0) > 0$, and hence $\kappa_{1y}(x_1) > \kappa_{0y}(x_0)$, then the above difference will be positive. If $\delta_y(x_1, x_0) < 0$, then the above difference will be negative, and if $\delta_y(x_1, x_0) = 0$, then $P(Y_1 \leq y|D = 0, X, Z) = P(Y_0 \leq y|D = 0, X, Z)$ becomes point-identified. Since Lemma 5 holds for all $y \in \mathcal{Y}$, analogous arguments prove that the difference

$$\begin{aligned} & P(Y_1 \leq y - 1|D = 0, X = x_1, Z) - P(Y_0 \leq y - 1|D = 0, X = x_0, Z) \\ &= P(\varepsilon \leq \kappa_{1y-1}(x_1)|\nu > s(z)) - P(\varepsilon \leq \kappa_{0y-1}(x_0)|\nu > s(z)) \end{aligned}$$

has the same sign as $\delta_{y-1}(x_1, x_0)$. A pairwise comparison of terms in the difference

$$\begin{aligned} & P(Y_1 = y|D = 0, X = x_1, Z) - P(Y_0 = y|D = 0, X = x_0, Z) \\ &= P(Y_1 \leq y|D = 0, X = x_1, Z) - P(Y_1 \leq y - 1|D = 0, X = x_1, Z) \\ &\quad -[P(Y_0 \leq y|D = 0, X = x_0, Z) - P(Y_0 \leq y - 1|D = 0, X = x_0, Z)] \end{aligned} \tag{4.25}$$

$$
\begin{aligned}
&= P(Y_1 \leq y|D=0, X=x_1, Z) - P(Y_0 \leq y|D=0, X=x_0, Z) \\
&\quad -[P(Y_1 \leq y-1|D=0, X=x_1, Z) - P(Y_0 \leq y-1|D=0, X=x_0, Z)] \\
&= P(\varepsilon \leq \kappa_{1y}(x_1)|\nu > s(z)) - P(\varepsilon \leq \kappa_{0y}(x_0)|\nu > s(z)) \\
&\quad -[P(\varepsilon \leq \kappa_{1y-1}(x_1)|\nu > s(z)) - P(\varepsilon \leq \kappa_{0y-1}(x_0)|\nu > s(z))]
\end{aligned}
$$

may thus be used to obtain bounds on the unidentified counterfactual probabilities. For example, if Lemma 5 reveals the information that $\delta_y(x_1, x_0) > \delta_{y-1}(x_1, x_0)$, then the difference between the former two probabilities after the last equality in (4.25) must be larger than the difference between the latter two, so that the overall sign is positive, and $P(Y_0 = y|D=0, X=x_0, Z)$ can be used as lower bound for $P(Y_1 = y|D=0, X=x_1, Z)$ instead of zero. By the same arguments, bounds on the counterfactual probability $P(Y_0 = y|D=1, X, Z)$ can be obtained. The following lemma summarizes and states the results:

Lemma 6 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then,*

$$
\begin{aligned}
(a) \quad \delta_y(x, \tilde{x}) &> \delta_{y-1}(x, \tilde{x}) \\
&\Leftrightarrow P(Y_1 = y|D=0, X=x, Z) > P(Y = y|D=0, X=\tilde{x}, Z) \\
\delta_y(x, \tilde{x}) &= \delta_{y-1}(x, \tilde{x}) = 0 \\
&\Leftrightarrow P(Y_1 = y|D=0, X=x, Z) = P(Y = y|D=0, X=\tilde{x}, Z) \\
\delta_y(x, \tilde{x}) &< \delta_{y-1}(x, \tilde{x}) \\
&\Leftrightarrow P(Y_1 = y|D=0, X=x, Z) < P(Y = y|D=0, X=\tilde{x}, Z)
\end{aligned}
$$

If $\delta_y(x, \tilde{x}) = \delta_{y-1}(x, \tilde{x}) = \pm 1$, then the sign of the difference $P(Y_1 = y|D=0, X=x, Z) - P(Y_0 = y|D=0, X=\tilde{x}, Z)$ is indeterminate. And,

$$
\begin{aligned}
(b) \quad \delta_y(\tilde{x}, x) &> \delta_{y-1}(\tilde{x}, x) \\
&\Leftrightarrow P(Y_0 = y|D=1, X=x, Z) < P(Y = y|D=1, X=\tilde{x}, Z) \\
\delta_y(\tilde{x}, x) &= \delta_{y-1}(\tilde{x}, x) = 0 \\
&\Leftrightarrow P(Y_0 = y|D=1, X=x, Z) = P(Y = y|D=1, X=\tilde{x}, Z) \\
\delta_y(\tilde{x}, x) &< \delta_{y-1}(\tilde{x}, x)
\end{aligned}
$$

$$\Leftrightarrow \quad P(Y_0 = y|D = 1, X = x, Z) > P(Y = y|D = 1, X = \tilde{x}, Z)$$

If $\delta_y(\tilde{x}, x) = \delta_{y-1}(\tilde{x}, x) = \pm 1$*, then the sign of the difference* $P(Y_0 = y|D = 1, X = x, Z) - P(Y_1 = y|D = 1, X = \tilde{x}, Z)$ *is indeterminate.*

Proof. Part (a) follows directly by application of Lemma 5 and (4.25). Part (b) follows by analogous arguments applying Lemma 5 and $P(Y_0 = y|D = 1, X = x_0, Z) - P(Y_1 = y|D = 1, X = x_1, Z)$ replacing the probability difference in (4.25). □

Lemma 6 holds for all evaluation points $\tilde{x}$ in the support of X. Clearly, there might be some evaluation points $\tilde{x}$ for that $\delta_y(x, \tilde{x}) > \delta_{y-1}(x, \tilde{x})$, and other evaluation points $\tilde{x}$ for that $\delta_y(x, \tilde{x}) < \delta_{y-1}(x, \tilde{x})$, or $\delta_y(x, \tilde{x}) = \delta_{y-1}(x, \tilde{x}) = 1$, for example. In order to use the full information, let

$$\begin{aligned}
\mathcal{X}_0^l(x_1) &= \{x_0 : \ \delta_y(x_1, x_0) > \delta_{y-1}(x_1, x_0)\} \\
\mathcal{X}_0^u(x_1) &= \{x_0 : \ \delta_y(x_1, x_0) < \delta_{y-1}(x_1, x_0)\}
\end{aligned}$$

and

$$\begin{aligned}
\mathcal{X}_1^l(x_0) &= \{x_1 : \ \delta_y(x_1, x_0) < \delta_{y-1}(x_1, x_0)\} \\
\mathcal{X}_1^u(x_0) &= \{x_1 : \ \delta_y(x_1, x_0) > \delta_{y-1}(x_1, x_0)\}
\end{aligned}$$

It is made explicit in the definition of sets that these are either over x_0 for x_1 fixed (and thus are a function of x_1), or over x_1 for x_0 fixed (and thus are a function of x_0). Bounds on the counterfactual probability $P(Y_1 = y|D = 0, X, Z)$, conditional on all values z in the support of Z can then be derived as

$$\begin{aligned}
&\sup_{\tilde{x} \in \mathcal{X}_0^l(x)} \{P(Y = y|D = 0, X = \tilde{x}, Z)\} \\
&\qquad \leq P(Y_1 = y|D = 0, X = x, Z) \leq \inf_{\tilde{x} \in \mathcal{X}_0^u(x)} \{P(Y = y|D = 0, X = \tilde{x}, Z)\}
\end{aligned}$$

If there exists $\tilde{x}$ such that $\delta_y(x, \tilde{x}) = \delta_{y-1}(x, \tilde{x}) = 0$ (for x fixed), then point-identification of the counterfactual probability follows, i.e., $P(Y_1 = y|D = 0, X = x, Z) = P(Y = y|D = 0, X = \tilde{x}, Z)$. If no such $\tilde{x}$ exists, and no $\tilde{x}$ for that Lemma 6 yields tighter bounds than the unit range, then $\mathcal{X}_0^l$ and $\mathcal{X}_0^u$ are empty and it is understood that the bounds zero

and one still apply. Analogously, for $P(Y_0 = y|D = 1, X, Z)$ the bounds can be derived as

$$\begin{aligned}
&\sup_{\tilde{x}\in\mathcal{X}_1^l(x)} \{P(Y = y|D = 1, X = \tilde{x}, Z)\} \\
&\quad \leq P(Y_0 = y|D = 1, X = x, Z) \leq \inf_{\tilde{x}\in\mathcal{X}_1^u(x)} \{P(Y = y|D = 1, X = \tilde{x}, Z)\}
\end{aligned}$$

with point-identification $P(Y_0 = y|D = 1, X = x, Z) = P(Y = y|D = 1, X = \tilde{x}, Z)$ if there exists $\tilde{x}$ such that $\delta_y(\tilde{x}, x) = \delta_{y-1}(\tilde{x}, x) = 0$, and bounds zero and one if $\mathcal{X}_1^l$ and $\mathcal{X}_1^u$ are empty.

Replacing the bounds for the counterfactual probabilities in $P(Y_1|X, Z)$ and $P(Y_0|X, Z)$ and following the same arguments as under the independence assumption, the bounds on $P(Y_1|X)$ and $P(Y_0|X)$ are given by

$$\begin{aligned}
L\mathcal{B}_y^1(x) \equiv \sup_{z\in\mathcal{Z}} \Big\{ & P(D = 1, Y = y|X = x, Z = z) \\
& + \sup_{\tilde{x}\in\mathcal{X}_0^l(x)} \{P(Y = y|D = 0, X = \tilde{x}, Z = z)\} P(D = 0|X = x, Z = z) \Big\} \\
\leq P(Y_1 = y|X = x) \leq & \qquad (4.26) \\
U\mathcal{B}_y^1(x) \equiv \inf_{z\in\mathcal{Z}} \Big\{ & P(D = 1, Y = y|X = x, Z = z) \\
& + \inf_{\tilde{x}\in\mathcal{X}_0^u(x)} \{P(Y = y|D = 0, X = \tilde{x}, Z = z)\} P(D = 0|X = x, Z = z) \Big\}
\end{aligned}$$

and

$$\begin{aligned}
L\mathcal{B}_y^0(x) \equiv \sup_{z\in\mathcal{Z}} \Big\{ & \sup_{\tilde{x}\in\mathcal{X}_1^l(x)} \{P(Y = y|D = 1, X = \tilde{x}, Z = z)\} P(D = 1|X = x, Z = z) \\
& + P(D = 0, Y = y|X = x, Z = z) \Big\} \\
\leq P(Y_0 = y|X = x) \leq & \qquad (4.27) \\
U\mathcal{B}_y^0(x) \equiv \inf_{z\in\mathcal{Z}} \Big\{ & \inf_{\tilde{x}\in\mathcal{X}_1^u(x)} \{P(Y = y|D = 1, X = \tilde{x}, Z = z)\} P(D = 1|X = x, Z = z) \\
& + P(D = 0, Y = y|X = x, Z = z) \Big\}
\end{aligned}$$

The following proposition uses the bounds in (4.26) and (4.27) under the threshold crossing model structure and the full model to impose bounds on the average treatment effect and the average treatment effect on the treated conditional on X:

Proposition 2 *Assume that (Y_0, Y_1, D) are generated according to model (4.6), and assume that conditions (A1)-(A8) are fulfilled. Then,*

$$\Delta_y^{ATE}(x) \in \left[L\mathcal{B}5_y^{ATE}(x), U\mathcal{B}5_y^{ATE}(x)\right] \qquad \textit{with} \tag{4.28}$$

$$\begin{aligned} L\mathcal{B}5_y^{ATE}(x) &= L\mathcal{B}_y^1(x) - U\mathcal{B}_y^0(x) \\ U\mathcal{B}5_y^{ATE}(x) &= U\mathcal{B}_y^1(x) - L\mathcal{B}_y^0(x) \end{aligned}$$

and

$$\Delta_y^{TT}(x) \in \left[L\mathcal{B}5_y^{TT}(x), U\mathcal{B}5_y^{TT}(x)\right] \qquad \textit{with} \tag{4.29}$$

$$\begin{aligned} L\mathcal{B}5_y^{TT}(x) &= [P(Y=y|X=x) - U\mathcal{B}_y^0(x)]/P(D=1|X=x) \\ U\mathcal{B}5_y^{TT}(x) &= [P(Y=y|X=x) - L\mathcal{B}_y^0(x)]/P(D=1|X=x) \end{aligned}$$

Proof. Follows directly by Lemmas 5, 6, and the discussion preceding the proposition. □

The bounds imposed by Proposition 2 depend on the amount of variation in X conditional on Z, and therefore it is difficult to make a general statement about their properties. However, two important conclusions can be drawn. First, if X does not vary conditional on Z, then the bounds in (4.28) and (4.29) simplify to the bounds in (4.19) and (4.20) with X conditioned on, but there is no possibility to further narrow the bounds. The reason is that if X is degenerate conditional on Z, then there exists only one $\tilde{x} = x$ in Lemma 6, which then becomes equivalent to Lemma 3 conditional on X. Thus, the cases $\delta_y(x,x) \lesseqgtr \delta_{y-1}(x,x)$, $\delta_y(x,x) = \delta_{y-1}(x,x) = 0$ allow to impose new upper or / and lower bounds on the counterfactual probabilities, if $\delta_y(x,x) = \delta_{y-1}(x,x) \pm 1$, then the bounds zero and one still apply, and as a consequence, the bounds in Proposition 2 collapse to those in Proposition 1 (conditional on X).

Second, the sign of the treatment effects is always identified by the bounds in Proposition 2. First consider the bounds in (4.28) and assume that the true average treatment effect is positive, i.e., $P(Y_1 = y|X = x) > P(Y_0 = y|X = x)$. Then $\delta_y(x,x) > \delta_{y-1}(x,x)$

so that $x \in \mathcal{X}_0^l(x)$ and $x \in \mathcal{X}_1^u(x)$ by Lemma 5. Thus, for the lower bound it must hold that

$$\begin{aligned} & L\mathcal{B}_y^1(x) - U\mathcal{B}_y^0(x) \\ = \; & \sup_{z \in Z} \{P(D=1, Y=y|X=x, Z=z) + P(D=0, Y=y|X=x, Z=z)\} \\ & - \inf_{z \in Z} \{P(D=1, Y=y|X=x, Z=z) + P(D=0, Y=y|X=x, Z=z)\} \\ = \; & P(Y=y|X=x, Z=z^u) - P(Y=y|X=x, Z=z^l) > 0 \end{aligned}$$

which follows by Lemma 4 conditional on X, Lemma 6, and the definition of z^u and z^l. The inequality holds for $\tilde{x} = x$, if other $\tilde{x} \in \mathcal{X}_0^l(x)$ and $\tilde{x} \in \mathcal{X}_1^u(x)$ exist, then $L\mathcal{B}_y^1(x)$ may get larger but never can get smaller by the supremum condition, and $U\mathcal{B}_y^0(x)$ may get smaller but never can get larger by the infimum condition, so that the inequality will still hold, and the lower bound in (4.28) will strictly be positive. By similar arguments, one can show that for the upper bound in the case of $P(Y_1 = y|X = x) < P(Y_0 = y|X = x)$, $U\mathcal{B}_y^1(x) - L\mathcal{B}_y^0(x)$ is negative for $\tilde{x} = x$, and will always be negative for all $\tilde{x} \in \mathcal{X}_0^u(x)$ and $\tilde{x} \in \mathcal{X}_1^l(x)$ other than x. If $P(Y_1 = y|X = x) = P(Y_0 = y|X = x)$, then $\delta_y(x, x) = \delta_{y-1}(x, x) = 0$, so that the counterfactual probabilities become identified by Lemma 6, and the average treatment effect is point-identified to be zero.

Next consider the treatment on the treated parameter and assume that the true parameter is positive. Then $\delta_y(x, x) > \delta_{y-1}(x, x)$ by Lemma 5 so that $x \in \mathcal{X}_1^u(x)$. The sign of $L\mathcal{B}5_y^{TT}(x)$ is determined by the sign of $P(Y = y|X = x) - U\mathcal{B}_y^0(x)$. Simplifying terms yields

$$P(Y=y|X=x) - P(Y=y|X=x, Z=z^l) > 0$$

by Lemma 4 conditional on X, Lemma 6, and the definition of z^l. Thus, the lower bound is positive for $\tilde{x} = x$, and will always be positive for all $\tilde{x} \in \mathcal{X}_1^u(x)$ other than x due to the infimum condition. By analogous steps, one can show that the upper bound of the treatment on treated parameter will always be negative if the true parameter is negative, and the bounds collapse to zero and thus provide point-identification if the true parameter is zero.

4.4 Inference

Shaikh and Vytlacil (2005) describe the construction of confidence sets given a discontinuity in the form of the bounds. Special attention to inference is necessary in this case because the usual approach of estimating probabilities by relative frequencies (or replacing population features by sample counterparts) will be inconsistent at the jump points. Their approach is based on the construction of a random set $\mathcal{CI}$ that will asymptotically cover, with probability at least $1-\alpha$ for fixed $\alpha \in (0,1)$, all treatment effects as identified by the population bounds.

The confidence set approach can also be implemented here. To simplify exposition, let the observed data be n independently and identically distributed drawings (Y_i, D_i, X_i, Z_i) from the population of interest, and let X and Z be discrete random variables. Furthermore, assume that the probability of treatment selection varies with each single outcome of Z, i.e., for all evaluation points $z_1 \neq z_0$ we have that $P(D=1|X, Z=z_1) \neq P(D=1|X, Z=z_0)$. In order to illustrate ideas, consider first the construction of confidence sets for the average treatment parameters defined by (4.13), i.e., in the case of no X covariates, threshold crossing treatment selection, but without further assumptions on the mechanism generating the outcome variable. Let

$$\hat{P}(z) = \frac{1}{|\{i : Z_i = z\}|} \sum_{i:Z_i=z} D_i$$

denote a consistent estimator of $P(D=1|Z=z)$. The evaluation points z^l and z^u may then be estimated by $\hat{z}^l = \min_z \hat{P}(z)$ and $\hat{z}^u = \max_z \hat{P}(z)$. Given the assumptions and with n large enough, one can show that $plim\, \hat{z}^l = z^l$ and $plim\, \hat{z}^u = z^u$. Consistent estimators of the bounds $L\mathcal{B}\mathcal{3}_y^{ATE}$ and $U\mathcal{B}\mathcal{3}_y^{ATE}$ can be obtained by

$$\begin{aligned}
\widehat{L\mathcal{B}\mathcal{3}}_y^{ATE} &= \frac{1}{|\{i : Z_i = \hat{z}^u\}|} \sum_{i:Z_i=\hat{z}^u} D_i Y_{iy} - \frac{1}{|\{i : Z_i = \hat{z}^l\}|} \sum_{i:Z_i=\hat{z}^l} D_i \\
&\quad - \frac{1}{|\{i : Z_i = \hat{z}^l\}|} \sum_{i:Z_i=\hat{z}^l} (1-D_i) Y_{iy} \\
\widehat{U\mathcal{B}\mathcal{3}}_y^{ATE} &= \frac{1}{|\{i : Z_i = \hat{z}^u\}|} \sum_{i:Z_i=\hat{z}^u} D_i Y_{iy} + \frac{1}{|\{i : Z_i = \hat{z}^u\}|} \sum_{i:Z_i=\hat{z}^u} (1-D_i) \\
&\quad - \frac{1}{|\{i : Z_i = \hat{z}^l\}|} \sum_{i:Z_i=\hat{z}^l} (1-D_i) Y_{iy}
\end{aligned}$$

where Y_{iy} is a dummy variable taking the value one if $Y_i = y$, and zero otherwise. Each of these estimators contains sums of means of binary variables, such that large sample theorems can be evoked to establish, for example,

$$\sqrt{n}\left(\widehat{LB\mathcal{Z}}_y^{ATE} - LB\mathcal{Z}_y^{ATE}\right) \xrightarrow{d} N(0, \sigma^2_{lb3,y})$$

with asymptotic variance $\sigma^2_{lb3,y}$. Analogously, asymptotic normality of the estimated upper bound can be established, and aymptotically valid confidence intervals for $LB\mathcal{Z}_y^{ATE}$ and $UB\mathcal{Z}_y^{ATE}$ can be found by the estimated lower and upper bounds plus/minus a measure of variation. The confidence intervals for the bounds in turn can be used to construct a random set that will asymptotically cover, with probability at least $1-\alpha$, the average treatment effects as defined by the population bounds in (4.13). Let $q_{1-\alpha}$ denote the $(1-\alpha)$-quantile of the standard normal distribution, and let $\hat{\sigma}^2_{lb3,y}$, $\hat{\sigma}^2_{ub3,y}$ denote consistent estimators of the variances in the asymptotic distributions of the estimated lower and upper bounds, respectively. Then,

$$P\left(LB\mathcal{Z}_y^{ATE} > \widehat{LB\mathcal{Z}}_y^{ATE} - \frac{\hat{\sigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}}\right)$$

and

$$P\left(UB\mathcal{Z}_y^{ATE} < \widehat{UB\mathcal{Z}}_y^{ATE} + \frac{\hat{\sigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right)$$

both converge in probability to $1-\alpha$. For each Δ_y^{ATE} in the interval $\left[LB\mathcal{Z}_y^{ATE}, UB\mathcal{Z}_y^{ATE}\right]$ it must therefore hold that in the limit the probability of $\widehat{LB\mathcal{Z}}_y^{ATE} - \hat{\sigma}_{lb3,y} q_{1-\alpha}/\sqrt{n}$ being smaller than the true average treatment effect, and the probability of $\widehat{UB\mathcal{Z}}_y^{ATE} + \hat{\sigma}_{ub3,y} q_{1-\alpha}/\sqrt{n}$ being larger than the true average treatment effect are at least $1-\alpha$, with equality if Δ_y^{ATE} is exactly at the lower (upper) boundary. Thus, with probability at least $1-\alpha$ and for large n, the interval

$$\mathcal{CI}\mathcal{Z}_y^{ATE} = \left[\widehat{LB\mathcal{Z}}_y^{ATE} - \frac{\hat{\sigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB\mathcal{Z}}_y^{ATE} + \frac{\hat{\sigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right] \tag{4.30}$$

will cover the true average treatment effects as defined by (4.13). For details on this approach see also Imbens and Manski (2004). Alternative approaches of obtaining asymptotically valid confidence sets exist, such as Horowitz and Manski (2000), or Chernozhukov *et al.* (2007), but I will restrict myself to the confidence set approach as outlined above.

The confidence set for the average treatment on the treated parameters, as bounded by (4.14), can be derived by parallel arguments. Consistent estimators of the lower and the upper bounds of the average treatment on the treated effect can be found by

$$\begin{aligned}\widehat{LB\mathcal{I}}_y^{TT} &= \left[\frac{1}{n}\sum_{i=1}^{n} Y_{iy} - \frac{1}{|\{i : Z_i = \hat{z}^l\}|}\sum_{i:Z_i=\hat{z}^l} D_i \right. \\ &\quad \left. - \frac{1}{|\{i : Z_i = \hat{z}^l\}|}\sum_{i:Z_i=\hat{z}^l}(1-D_i)Y_{iy}\right] \Big/ \left(\frac{1}{n}\sum_{i=1}^{n} D_i\right) \\ \widehat{UB\mathcal{I}}_y^{TT} &= \left[\frac{1}{n}\sum_{i=1}^{n} Y_{iy} - \frac{1}{|\{i : Z_i = \hat{z}^l\}|}\sum_{i:Z_i=\hat{z}^l}(1-D_i)Y_{iy}\right] \Big/ \left(\frac{1}{n}\sum_{i=1}^{n} D_i\right)\end{aligned}$$

Furthermore, let $\varsigma^2_{lb3,y}$, $\varsigma^2_{ub3,y}$ denote the asymptotic variances of the estimated lower and upper bounds of the average treatment on the treated parameter, respectively, and $\hat{\varsigma}^2_{lb3,y}$, $\hat{\varsigma}^2_{ub3,y}$ the corresponding consistent estimators. Then, the random set constructed as

$$\mathcal{CI}_y^{TT} = \left[\widehat{LB\mathcal{I}}_y^{TT} - \frac{\hat{\varsigma}_{lb3,y}q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB\mathcal{I}}_y^{TT} + \frac{\hat{\varsigma}_{ub3,y}q_{1-\alpha}}{\sqrt{n}}\right] \tag{4.31}$$

will cover asymptotically the true average treatment on the treated parameter, as defined by the bounds in (4.14), with probability at least $1-\alpha$.

The construction of confidence sets for the average treatment and average treatment on the treated parameters as bounded by Proposition 1 proceeds in a similar way. Consider first Δ_y^{ATE} and the bounds in (4.19), and let $A_y^{ATE} \equiv P(Y=y|Z=z^u) - P(Y=y|Z=z^l)$ which can be consistently estimated by

$$\hat{A}_y^{ATE} = \frac{1}{|\{i : Z_i = \hat{z}^u\}|}\sum_{i:Z_i=\hat{z}^u} Y_{iy} - \frac{1}{|\{i : Z_i = \hat{z}^l\}|}\sum_{i:Z_i=\hat{z}^l} Y_{iy}$$

Large sample results ensure that

$$\sqrt{n}\left(\hat{A}_y^{ATE} - A_y^{ATE}\right) \xrightarrow{d} N(0, \sigma^2_{a,y})$$

where $\sigma^2_{a,y}$ denotes the variance of the asymptotic normal distribution. For the average treatment on the treated parameter and bounds (4.20), let $A_y^{TT} \equiv [P(Y=y) - P(Y=y|Z=z^l)]/P(D=1)$ which can be consistently estimated by

$$\hat{A}_y^{TT} = \left[\frac{1}{n}\sum_{i=1}^{n} Y_{iy} - \frac{1}{|\{i : Z_i = \hat{z}^l\}|}\sum_{i:Z_i=\hat{z}^l} Y_{iy}\right] \Big/ \left(\frac{1}{n}\sum_{i} D_i\right)$$

and again, by large sample arguments

$$\sqrt{n}\left(\hat{A}_y^{TT} - A_y^{TT}\right) \xrightarrow{d} N(0, \varsigma_{a,y}^2)$$

with asymptotic variance $\varsigma_{a,y}^2$. Thus, for each of the terms in (4.19) and (4.20) a consistent estimator exists and an asymptotically valid confidence interval can be constructed.

An additional complication arises because the bounds in (4.19) and (4.20) are discontinuous functions of δ_y and δ_{y-1}. This discontinuity needs to be taken into account when constructing the random set that will asymptotically cover the true parameter with predefined probability. In order to do that, the uncertainty about δ_y should be considered as well. For an analogous argument in a nonparametric regression context see also Gijbels *et al.* (2004). Recall that δ_y was defined as the sign of the difference between two cumulative probabilities, specifically as the sign of $d_y \equiv P(Y \leq y|Z = z_1) - P(Y \leq y|Z = z_0)$ for any two evaluation points z_1, z_0 with $P(D = 1|Z = z_1) > P(D = 1|Z = z_0)$. A consistent estimator of d_y can be obtained as

$$\hat{d}_y(z_1, z_0) = \frac{1}{|\{i : Z_i = z_1\}|} \sum_{i:Z_i=z_1} \sum_{j=1}^{y} Y_{ij} - \frac{1}{|\{i : Z_i = z_0\}|} \sum_{i:Z_i=z_0} \sum_{j=1}^{y} Y_{ij}$$

with z_1, z_0 such that $\hat{P}(z_1) > \hat{P}(z_0)$. The estimator $\hat{d}_y(z_1, z_0)$ uses the information of only two evaluation points, but it is possible to account for the additional information of *all* combinations z_1, z_0 satisfying the condition $P(D = 1|Z = z_1) > P(D = 1|Z = z_0)$, which will generally improve the precision of the estimator. The modified version

$$\hat{d}_y = \frac{1}{|\{(z_1, z_0) : \hat{P}(z_1) > \hat{P}(z_0)\}|} \sum_{(z_1,z_0):\hat{P}(z_1)>\hat{P}(z_0)} \hat{d}_y(z_1, z_0) \tag{4.32}$$

will therefore be used in the following. The estimator in (4.32) can be constructed for each outcome $y \in \mathcal{Y}$, and pairs $\hat{d}_{y,y-1} = (\hat{d}_y, \hat{d}_{y-1})$ will asymptotically be bivariate normally distributed with

$$\sqrt{n}\left(\hat{d}_{y,y-1} - d_{y,y-1}\right) \xrightarrow{d} N(0, \Sigma_{y,y-1})$$

The asymptotic covariance matrix $\Sigma_{y,y-1}$ has $Var(\hat{d}_y)$ and $Var(\hat{d}_{y-1})$ the main diagonal entries, and $Cov(\hat{d}_y, \hat{d}_{y-1})$ the off-diagonal entries. An asymptotic confidence ellipse for $d_{y,y-1}$ can be constructed as

$$(\hat{d}_{y,y-1} - d_{y,y-1})'\hat{\Sigma}_{y,y-1}^{-1}(\hat{d}_{y,y-1} - d_{y,y-1}) \leq \chi_{2,1-\alpha}^2 \tag{4.33}$$

where $\hat{\Sigma}_{y,y-1}$ is a consistent estimator of $\Sigma_{y,y-1}$, and $\chi^2_{2,1-\alpha}$ is the $1-\alpha$ quantile of the Chi-square distribution with two degrees of freedom. For n growing large, the ellipse defined by (4.33) will cover the true $d_{y,y-1}$ with probability $1-\alpha$.

The confidence sets for the average treatment effects and the average treatment on the treated effects as defined by Proposition 1 can then be constructed as follows. In the d_y, d_{y-1}-plane (where d_y is on the abscissa and d_{y-1} is on the ordinate), if the confidence ellipse defined by (4.33)

1. ... lies entirely in the fourth quadrant (d_y positive, d_{y-1} negative), or intersects with the abscissa ($d_{y-1}=0$) only in the first/fourth quadrant, or intersects with the ordinate ($d_y=0$) only in the third/fourth quadrant, then use the random set

$$\begin{aligned}\mathcal{CI}_4 a_y^{ATE} &= \left[\hat{A}_y^{ATE} - \frac{\hat{\sigma}_{a,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB3}_y^{ATE} + \frac{\hat{\sigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right] \\ \mathcal{CI}_4 a_y^{TT} &= \left[\hat{A}_y^{TT} - \frac{\hat{\varsigma}_{a,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB3}_y^{TT} + \frac{\hat{\varsigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right]\end{aligned}$$

2. ... intersects with both axes, then use the random set

$$\begin{aligned}\mathcal{CI}_4 b_y^{ATE} &= \left[\widehat{LB3}_y^{ATE} - \frac{\hat{\sigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB3}_y^{ATE} + \frac{\hat{\sigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right] \\ \mathcal{CI}_4 b_y^{TT} &= \left[\widehat{LB3}_y^{TT} - \frac{\hat{\varsigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB3}_y^{TT} + \frac{\hat{\varsigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right]\end{aligned}$$

3. ... lies entirely in the second quadrant (d_y negative, d_{y-1} positive), or intersects with the abscissa ($d_{y-1}=0$) only in the second/third quadrant, or intersects with the ordinate ($d_y=0$) only in the first/second quadrant, then use the random set

$$\begin{aligned}\mathcal{CI}_4 c_y^{ATE} &= \left[\widehat{LB3}_y^{ATE} - \frac{\hat{\sigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \hat{A}_y^{ATE} + \frac{\hat{\sigma}_{a,y} q_{1-\alpha}}{\sqrt{n}}\right] \\ \mathcal{CI}_4 c_y^{TT} &= \left[\widehat{LB3}_y^{TT} - \frac{\hat{\varsigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \hat{A}_y^{TT} + \frac{\hat{\varsigma}_{a,y} q_{1-\alpha}}{\sqrt{n}}\right]\end{aligned}$$

4. ... lies entirely in the first quadrant (both d_y and d_{y-1} are positive), or entirely in the third quadrant (both d_y and d_{y-1} are negative), then use the random set

$$\begin{aligned}\mathcal{CI}_4 d_y^{ATE} &= \left[\widehat{LB3}_y^{ATE} - \frac{\hat{\sigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB3}_y^{ATE} + \frac{\hat{\sigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right] \\ \mathcal{CI}_4 d_y^{TT} &= \left[\widehat{LB3}_y^{TT} - \frac{\hat{\varsigma}_{lb3,y} q_{1-\alpha}}{\sqrt{n}},\ \widehat{UB3}_y^{TT} + \frac{\hat{\varsigma}_{ub3,y} q_{1-\alpha}}{\sqrt{n}}\right]\end{aligned}$$

One can show that asymptotically the random sets $\mathcal{CI}4_y^{ATE}$, consisting of $\mathcal{CI}4\,a_y^{ATE}$ to $\mathcal{CI}4\,d_y^{ATE}$, and $\mathcal{CI}4_y^{TT}$, consisting of $\mathcal{CI}4\,a_y^{TT}$ to $\mathcal{CI}4\,d_y^{TT}$, cover the true average treatment effect and the average treatment effect on the treated, respectively, with probability at least $1-\alpha$. In order to see why the intervals constructed as such will cover the true parameter with probability at least $1-\alpha$, consider the average treatment effect and assume that $\Delta_y^{ATE} > 0$ such that $\delta_y > \delta_{y-1}$ and $\Delta_y^{ATE} \in \left[A_y^{ATE}, U\mathcal{B}\mathcal{Z}_y^{ATE}\right]$. With probability approaching one, the confidence interval constructed in (4.33) will fulfill the conditions to choose $\mathcal{CI}4\,a_y^{ATE}$, and $\mathcal{CI}4\,a_y^{ATE}$ covers all parameters $\Delta_y^{ATE} \in \left[A_y^{ATE}, U\mathcal{B}\mathcal{Z}_y^{ATE}\right]$ with probability at least $1-\alpha$, as desired. Analogous arguments show that in all other cases the desired coverage probability is obtained.

The confidence sets for the average treatment and the average treatment on the treated effects in the case of X covariates present, i.e., for the parameters as identified by Proposition 2, can be constructed following a similar strategy as in the case of no X covariates available. The steps involved are as follows. To begin with, let

$$\hat{P}(x,z) = \frac{1}{|\{i : X_i = x, Z_i = z\}|} \sum_{i:X_i=x,Z_i=z} D_i$$

denote a consistent estimator of $P(D=1|X=x,Z=z)$, and define

$$\begin{aligned} &d_y(x_1,x_0;z_1,z_0) \equiv \\ &\quad [P(D=1,Y\le y|X=x_1,Z=z_1) - P(D=1,Y\le y|X=x_1,Z=z_0)] \\ &\quad -[P(D=0,Y\le y|X=x_0,Z=z_0) - P(D=0,Y\le y|X=x_0,Z=z_1)] \end{aligned}$$

which can be consistently estimated by

$$\begin{aligned} \hat{d}_y(x_1,x_0;z_1,z_0) \;=\; & \frac{1}{|\{i : X_i = x_1, Z_i = z_1\}|} \sum_{i:X_i=x_1,Z_i=z_1} \sum_{j=1}^{y} D_i Y_{ij} \\ & - \frac{1}{|\{i : X_i = x_1, Z_i = z_0\}|} \sum_{i:X_i=x_1,Z_i=z_0} \sum_{j=1}^{y} D_i Y_{ij} \\ & - \left[\frac{1}{|\{i : X_i = x_0, Z_i = z_0\}|} \sum_{i:X_i=x_0,Z_i=z_0} \sum_{j=1}^{y} (1-D_i) Y_{ij} \right. \\ & \left. - \frac{1}{|\{i : X_i = x_0, Z_i = z_1\}|} \sum_{i:X_i=x_0,Z_i=z_1} \sum_{j=1}^{y} (1-D_i) Y_{ij} \right] \end{aligned}$$

with z_1, z_0 such that $\hat{P}(x_j, z_1) > \hat{P}(x_j, z_0)$, $j = 0, 1$. Accounting for the information of all such evaluation points z_1, z_0 yields the estimator

$$\hat{d}_y(x_1, x_0) = \frac{\sum_{(z_1,z_0):\hat{P}(x_j,z_1)>\hat{P}(x_j,z_0),j=0,1} \hat{d}_y(x_1, x_0; z_1, z_0)}{|\{(z_1, z_0) : \hat{P}(x_j, z_1) > \hat{P}(x_j, z_0), j = 0, 1\}|}$$

Then consider the estimator either as a function of x_1 keeping x_0 fixed, $\hat{d}_y(x_1|x_0)$, or as a function of x_0 keeping x_1 fixed, $\hat{d}_y(x_0|x_1)$, and note that the estimators hold for all $y \in \mathcal{Y}$ such that pairs $\left(\hat{d}_y(x_1|x_0), \hat{d}_{y-1}(x_1|x_0)\right)$ or $\left(\hat{d}_y(x_0|x_1), \hat{d}_{y-1}(x_0|x_1)\right)$ can be created. From these pairs, one may construct asymptotically valid confidence ellipses with regions as defined in the construction of $\mathcal{CI}4$ (entirely in each quadrant, and the 5 intersection possibilities with the axes).

The bounds $L\mathcal{B}_y^1(x)$, $U\mathcal{B}_y^1(x)$, $L\mathcal{B}_y^0(x)$, $U\mathcal{B}_y^0(x)$ depend on the sets $\mathcal{X}_0^l(x)$, $\mathcal{X}_0^u(x)$, $\mathcal{X}_1^l(x)$, $\mathcal{X}_1^u(x)$, the latter defined by $\delta_y(x_1, x_0)$ relative to $\delta_{y-1}(x_1, x_0)$. This dependence needs to be taken into account when constructing the confidence sets for the parameters. Let $\mathcal{AX}_0^l(x_1)$ denote an alternative set of all x_0 (given x_1) satisfying that in the $d_y(x_0|x_1), d_{y-1}(x_0|x_1)$-plane the confidence ellipse lies entirely in the fourth quadrant ($d_y(x_0|x_1)$ positive, $d_{y-1}(x_0|x_1)$ negative), or intersects with the abscissa ($d_{y-1}(x_0|x_1) = 0$) only in the first/fourth quadrant, or intersects with the ordinate ($d_y(x_0|x_1) = 0$) only in the third/fourth quadrant. Similarly, define $\mathcal{AX}_0^u(x_1)$ as an alternative set of all x_0 (given x_1) satisfying that in the $d_y(x_0|x_1), d_{y-1}(x_0|x_1)$-plane the confidence ellipse lies entirely in the second quadrant ($d_y(x_0|x_1)$ negative, $d_{y-1}(x_0|x_1)$ positive), or intersects with the abscissa ($d_{y-1}(x_0|x_1) = 0$) only in the second/third quadrant, or intersects with the ordinate ($d_y(x_0|x_1) = 0$) only in the first/second quadrant. Analogously, define sets $\mathcal{AX}_1^l(x_0)$ and $\mathcal{AX}_1^u(x_0)$ in the $d_y(x_1|x_0), d_{y-1}(x_1|x_0)$-plane. These alternative sets can be interpreted as estimators of the population sets $\mathcal{X}_j^k(x)$, $j = 0, 1$, $k = l, u$.

Empirical analogues of the upper and lower bounds on $P(Y_1 = y|X = x)$ and $P(Y_0 = y|X = x)$ can be derived from (4.26) and (4.27) replacing the population sets by the alternative sets defined above and the probabilities by the appropriate relative frequencies. From these estimators, one may construct estimators of the upper and lower bounds for the average treatment effect and the average treatment effect on the treated. Because of the dependence on the (estimated) alternative sets, obtaining an upper bound of an

one-sided $1-\alpha$ confidence interval for $U\mathcal{B}5_y^{ATE}(x)$, and a lower bound of an one-sided $1-\alpha$ confidence interval for $L\mathcal{B}5_y^{ATE}(x)$ is not straightforward. One option, also referred to by Shaikh and Vytlacil (2005), is subsampling; see Politis *et al.* (1999) for details, in particular Chapter 2. Let $\widehat{L\mathcal{B}5}^{ATE}_{y;1-\alpha^-}(x)$ and $\widehat{U\mathcal{B}5}^{ATE}_{y;1-\alpha^+}(x)$ denote such bounds of the confidence interval, then an asymptotically valid confidence interval for the average treatment effect can be obtained by

$$\mathcal{CI}5_y^{ATE} = \left[\widehat{L\mathcal{B}5}^{ATE}_{y;1-\alpha^-}(x), \widehat{U\mathcal{B}5}^{ATE}_{y;1-\alpha^+}(x)\right] \tag{4.34}$$

By analogous arguments, an asymptotically valid confidence interval

$$\mathcal{CI}5_y^{TT} = \left[\widehat{L\mathcal{B}5}^{TT}_{y;1-\alpha^-}(x), \widehat{U\mathcal{B}5}^{TT}_{y;1-\alpha^+}(x)\right] \tag{4.35}$$

for the average treatment on the treated effect can be constructed.

4.5 Moving Beyond ATE and TT

The previous sections have focused on two treatment parameters, namely the average treatment effect and the average treatment effect on the treated. Both parameters were defined in terms of probabilities rather than expectations to circumvent the problem of ordinal but arbitrary coding of the elements in $\mathcal{Y}$. The term "average" was introduced because the parameters reflect how an individual's probability of responding in each of the J ordinal categories will change with and without the receipt of treatment, and where probability was defined from a frequentist perspective as what would happen on average if the same individual was considered repeatedly.

The average treatment effect and the average treatment effect on the treated certainly are the treatment parameters that occur most often in the literature. The former is defined for an individual that is randomly drawn from the entire population of interest, the latter is defined for an individual randomly drawn from those that actually received the treatment. However, alternative parameters have been considered as well for different subgroups of the population. For example, the local average treatment effect (LATE) of Imbens and Angrist (1994) is defined as the average treatment effect for the subgroup of compliers, i.e., those individuals who would comply with the exogenous modification of

instruments. This concept can also be translated to probabilities. Let z_1, z_0 denote two evaluation points with $P(D = 1|Z = z_1) > P(D = 0|Z = z_0)$ such that, by the threshold crossing treatment selection, $s(z_1) > s(z_0)$. Then,

$$\begin{aligned} & P(Y = y|Z = z_1) - P(Y = y|Z = z_0) \qquad (4.36) \\ &= P(D = 1, Y = y|Z = z_1) + P(D = 0, Y = y|Z = z_1) \\ &\quad -P(D = 1, Y = y|Z = z_0) - P(D = 0, Y = y|Z = z_0) \\ &= P(\nu \leq s(z_1), Y_1 = y) + P(\nu > s(z_1), Y_0 = y) \\ &\quad -P(\nu \leq s(z_0), Y_1 = y) - P(\nu > s(z_0), Y_0 = y) \\ &= P(s(z_0) < \nu \leq s(z_1), Y_1 = y) - P(s(z_0) < \nu \leq s(z_1), Y_0 = y) \\ &= \Big[P(Y_1 = y|s(z_0) < \nu \leq s(z_1)) - P(Y_0 = y|s(z_0) < \nu \leq s(z_1))\Big] \\ &\quad P(s(z_0) < \nu \leq s(z_1)) \end{aligned}$$

where the first equality follows by the law of total probability, the second equality follows by the observation rule in (4.1), the threshold crossing treatment selection, and the independence assumption (A4), the third equality follows by $s(z_1) > s(z_0)$, and the last equality follows by Bayes' theorem.

From (4.36) define the local average treatment effect as

$$\begin{aligned} \Delta_y^{LATE}(z_1, z_0) &\equiv P(Y_1 = y|s(z_0) < \nu \leq s(z_1)) \\ &\quad -P(Y_0 = y|s(z_0) < \nu \leq s(z_1)) \qquad (4.37) \\ &= \frac{P(Y = y|Z = z_1) - P(Y = y|Z = z_0)}{P(s(z_0) < \nu \leq s(z_1))} \\ &= \frac{P(Y = y|Z = z_1) - P(Y = y|Z = z_0)}{P(\nu \leq s(z_1)) - P(\nu \leq s(z_0))} \\ &= \frac{P(Y = y|Z = z_1) - P(Y = y|Z = z_0)}{P(D = 1|Z = z_1) - P(D = 1|Z = z_0)} \end{aligned}$$

where the second equality follows by the derivation above, and the last equalities follow by the assumptions of the treatment selection model. Thus, the local average treatment effect gives the change in the probability distribution for those individuals who would not select into treatment if Z was externally set to z such that $s(z) \leq s(z_0)$, and who would select into treatment if Z was externally set to z such that $s(z) \geq s(z_1)$. An important aspect of

the local average treatment effect is that it is identified from the population distribution of (Y, D, Z) for all combinations z_1, z_0 with $P(D = 1|Z = z_1) > P(D = 1|Z = z_0)$, which is made explicit in the definition of $\Delta_y^{LATE}(z_1, z_0)$ including z_1 and z_0 in the argument.

A marginal version of the local average treatment effect has been introduced in Heckman (1997). Consider the limit $s(z_0) \to s(z_1)$ of (4.37) and define the marginal treatment effect as

$$\Delta_y^{MTE}(z_1) \equiv P(Y_1 = y|\nu = s(z_1)) - P(Y_0 = y|\nu = s(z_1)) \tag{4.38}$$

Thus, the marginal treatment effect gives the change in the probability distribution for those individuals that would just be indifferent between being selected into or out of the treatment if Z was externally set to z such that $s(z) = s(z_1)$. Starting from (4.38), one can show that the other treatment parameters, Δ_y^{ATE}, Δ_y^{TT}, and Δ_y^{LATE}, are integrated versions of Δ_y^{MTE} over different intervals and with different weighting functions (Heckman and Vytlacil 2001). An estimator of Δ_y^{MTE} can be obtained by $\partial P(Y = y|Z = z_1)/\partial P(D = 1|Z = z_1)$ given that the derivative exists and is finite in a small neighborhood of z_1. Since both Δ_y^{MTE} and Δ_y^{LATE} are identified, identification of Δ_y^{ATE} and Δ_y^{TT} in principle is possible. However, this requires observability of a sufficiently large support of $P(D = 1|Z = z)$, which must not necessarily hold in practice, and therefore the bounding analysis of Section 4.3 is more general by imposing identification regions for the treatment parameters.

While the previous treatment parameters were defined for different subgroups of the population, the ordinal nature of the response variable allows for a more thorough analysis of the effect on the outcome distribution, either in the entire population or in the subgroup of treated individuals. In particular, analyzing probabilities rather than expectations provides a much richer set of treatment parameters beyond the common mean effects. For example, consider the concept of stochastic order (SO) in two random variables (Mann and Whitney 1947). Let

$$\Delta_y^{SO} \equiv P(Y_1 \leq y) - P(Y_0 \leq y) \tag{4.39}$$

If $\Delta_y^{SO} \leq 0$ for all y, then Y_0 is said to be stochastically smaller than Y_1, i.e., Y_0 tends to have higher probability for low y, and smaller probability for high y compared to Y_1.

Analogously, if $\Delta_y^{SO} \geq 0$ for all y, then Y_0 is said to be stochastically larger than Y_1, and if $\Delta_y^{S0} = 0$ for all y, then Y_0 and Y_1 are said to be stochastically equivalent. Clearly, one may also analyze the stochastic order of Y_1 and Y_0 in the subgroup of the treated (SOT)

$$\Delta_y^{SOT} \equiv P(Y_1 \leq y|D = 1) - P(Y_0 \leq y|D = 1) \tag{4.40}$$

where, for example, Y_1 is said to be stochastically larger than Y_0, now conditional on $D = 1$, if $\Delta_y^{SOT} \leq 0$ for all y. If neither of the three cases is true for all y, i.e., Y_1 is not stochastically larger or smaller than, nor equivalent to Y_0, then one may at least analyze the degree of stochastic order starting from $y = 1$ moving to $y = J$, or the other way round.

Yet another way to look at the effect of treatment on the outcome distribution, related to the concept of stochastic ordering, is in terms of the relative odds, specifically,

$$\Omega_y \equiv \frac{P(Y_0 \leq y)/P(Y_0 > y)}{P(Y_1 \leq y)/P(Y_1 > y)} \tag{4.41}$$

and

$$\Omega_y^T \equiv \frac{P(Y_0 \leq y|D = 1)/P(Y_0 > y|D = 1)}{P(Y_1 \leq y|D = 1)/P(Y_1 > y|D = 1)} \tag{4.42}$$

These parameters show the factor by which the ratio of the odds $Y_0 \leq y$ relative to $Y_0 > y$ in the non-treatment group change compared to the odds $Y_1 \leq y$ relative to $Y_1 > y$ in the treatment group. With a positive treatment effect, i.e., the probability of higher outcomes increases with the receipt of treatment, this factor should be larger than one. If, on the other hand, the treatment effect is negative, then the odds ratio is smaller than one, and if the treatment effect is zero, then the odds ratio is one. Note that there exist $J-1$ odds ratios, one for each $y = 1, \ldots, J-1$.

Neither the stochastic order parameters, nor the odds ratios are immediately identified from the population distribution of (Y, D, Z), by the same argument as the average treatment and the average treatment on the treated are not identified. However, one may impose bounds on the unidentified probabilities and thus impose bounds on the parameters in (4.39)-(4.42).

4.6 Conclusion

The properties of ordinally measured variables, in a strict sense, require the shift in focus from mean treatment effects to probability treatment effects. Parametric ordered response models to estimate such effects already exist and are typically based on threshold crossing mechanisms. This is the first paper, to the best of my knowledge, that discovers the informational content of a threshold crossing mechanism in a nonparametric bounding analysis with ordinal potential outcomes; only Scharfstein *et al.* (2004) consider bounds on treatment effects with ordinal responses, but in a very particular prospective data situation.

The approach taken here is closely related to Shaikh and Vytlacil (2005), who consider a model with binary instead of ordinal outcomes, and the obtained results therefore complement their work. The extension to ordinal outcomes requires a slightly different identification and bounding strategy, where multiple thresholds need to be taken into account. As a central result, the imposed bounds always identify whether the treatment effect is positive, zero, or negative, although point-identification except for the zero treatment effect fails in the nonparametric setting. It is interesting to note that an additional set of parameters becomes available with ordinal outcomes that might be of interest in evaluating the effect of a treatment.

References

Angrist, J.D., G.W. Imbens, and D. Rubin (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444-455.

Bellemare, C., B. Melenberg, and A. van Soest (2002): "Semi-parametric models for satisfaction with income," Portuguese Economic Journal, 1, 181-203.

Boes, S., and R. Winkelmann (2006): "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90, 165-180.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Confidence Regions for Parameter Sets in Econometric Models," *Econometrica*, forthcoming.

Coppejans, M. (2007): "On efficient estimation of the ordered response model," *Journal of Econometrics*, 137, 577-614.

Gijbels, I., P. Hall, and A. Kneip (2004): "Interval and band estimation for curves with jumps," *Journal of Applied Probability*, 41A, 65-79.

Heckman, J.J. (1997): "Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations," *Journal of Human Resources*, 32, 441-462.

Heckman, J.J., R.J. LaLonde, and J.A. Smith (1999): "The Economics and Econometrics of Active Labor Market Programs," in O. Ashenfelter, and D. Card (eds.), *Handbook of Labor Economics Volume III*. Amsterdam: North Holland, 1865-2097

Heckman, J.J., and R. Robb (1985): "Alternative Methods for Estimating the Impact of Interventions," in J.J. Heckman, and B. Singer (eds.) *Longitudinal Analysis of Labor Market Data*, New York: Cambridge University Press, 156-245.

Heckman, J.J., and R. Robb (1986): "Alternative Methods for Solving the Problem of Selection Bias in Evaluating the Impact of Treatments on Outcomes," in J. Wainer

(ed.) *Drawing Inference From Self-Selected Samples*, New York: Springer-Verlag, 63-107.

Heckman, J.J., and E. Vytlacil (2001): "Local Instrumental Variables," in C. Hsiao, K. Morimune, and J. Powell (eds.) *Nonlinear Statistical Inference: Essays in Honor of Takeshi Amemiya*, Cambridge: Cambridge University Press.

Horowitz, J., and C.F. Manski (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95, 77-84.

Imbens, G.W., and J.D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-476.

Imbens, G.W., and C.F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845-1857.

Klein, R.W., and Sherman, R.P. (2002): "Shift restrictions and semiparametric estimation in ordered response models," *Econometrica*, 70, 663-692.

Lewbel, A. (1997): "Semiparametric estimation of location and other discrete choice moments," *Econometric Theory*, 13, 32-51.

Lewbel, A. (2003): "Ordered response threshold estimation," *unpublished working paper.*

Li, M., and J.L. Tobias (2007): "Bayesian Analysis of Treatment Effects in an Ordered Potential Outcomes Model in D. Millimet, J. Smith, and E. Vytlacil (eds.) *Advances in Econometrics, Volume 21: Estimating and Evaluating Treatment Effects in Econometrics*, forthcoming.

Mann, H.B., and D.R. Whitney (1947): "On a test of whether one of two random variables is stochastically larger than the other," *Annals of Mathematical Statistics*, 18, 50-60.

Manski, C.F. (1990): "Nonparametric Bounds on Treatment Effects," *American Economic Review, Papers and Proceedings*, 80, 319-323.

Manski, C.F. (1994): "The Selection Problem," in C. Manski, and D. McFadden (eds.) *Advances in Econometrics: Sixth World Congress*, Cambridge University Press, Cambridge.

Manski, C.F. (1995): *Identification Problems in the Social Sciences*, Harvard University Press, Cambridge and London.

Manski, C.F. (2000): "Identification Problems and Decisions Under Ambiguity: Empirical Analysis of Treatment Response and Normative Analysis of Treatment Choice," *Journal of Econometrics*, 95, 415-442.

Manski, C.F. (2003): *Partial Identification of Probabilities Distributions*, Springer: New York and Heidelberg.

McCullagh, P. (1980): "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society. Series B (Methodological)*, 42, 109-142.

McKelvey, R.D. and W. Zavoina (1975): "A statistical model for the analysis of ordinal level dependent variables," *Journal of Mathematical Sociology*, 4, 103-120.

Neyman, J. (1923): "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9," translated in *Statistical Science*, 5, 465-480.

Politis, D.N., J.P. Romano, and M. Wolf (1999): *Subsampling*, Springer, New York.

Rubin, D.B. (1974): "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Scharfstein, D.O., C.F. Manski, and J.C. Anthony (2004): "On the Construction of Bounds in Prospective Studies with Missing Ordinal Outcomes: Application to the Good Behavior Game Trial," *Biometrics*, 60, 154-164.

Shaikh, A.M., and E. Vytlacil (2005): "Threshold Crossing Models and Bounds on Treatment Effects: A Nonparametric Analysis," *unpublished working paper*.

Stewart, M.B (2004): "Semi-nonparametric Estimation of Extended Ordered Probit Models," *Stata Journal*, 4, 27-39.

Vytlacil, E. (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica*, 70, 331-341.

Vytlacil, E. (2006): "A Note on Additive Separability and Latent Index Models of Binary Choice: Representation Results," Oxford Bulletin of Economics and Statistics, 68, 515-518.

Appendix A

`goprobit` — A Stata Module to Estimate Generalized Ordered Probit Models

This appendix describes a new Stata module I wrote while working on my dissertation. Standard ordered response models have been implemented in all variants in almost all modern statistical software packages, but the availability of canned procedures to estimate generalized ordered response models is only moderate to non-existent. Exceptions are the estimators available in LIMDEP (written by W. Greene) and the Stata modules written by Williams (2006) and Fu (1998).

`goprobit` extends the "ready-to-use" procedures in Stata by estimating generalized ordered probit models. A short description of the model is enclosed in the documentation, for more details see the references at the end of the appendix. `goprobit` is not included in the basic distribution of Stata and needs to be downloaded and installed before use. The easiest way is via the `ssc` commands in Stata (`ssc install goprobit`). Alternatively, type `findit goprobit` in the command line and follow the on-screen instructions. If you experience any problems downloading `goprobit`, then try

http://ideas.repec.org/c/boc/bocode/s456603.html, or

http://econpapers.repec.org/software/bocbocode/s456603.htm

for manual download. A panel data version of `goprobit` with random effects can be estimated by `regoprob`; see Appendix B and the help file of `regoprob` (if installed).

Syntax

```
goprobit depvar [indepvars] [weight] [if exp] [in range] [, pl pl(varlist)
    npl npl(varlist) constraints(clist) robust cluster(varname) level(#)
    score(newvarlist|stub*) maximize_options ]
```

`goprobit` shares the features of all estimation commands; see [R] **estimates**.
`goprobit` typed without arguments redisplays previous results.
`fweights`, `iweights`, and `pweights` are allowed; see [U] **11.1.6 weights**.

The syntax of `predict` following `goprobit` is

```
predict [type] newvarname(s) [if exp] [in range] [, statistic
    outcome(outcome) ]
```

where *statistic* is

`p`	probability (specify one new variable and `outcome()` option, or specify J new variables, $J = \#$ of outcomes); the default
`xb`	linear prediction (`outcome()` option required)
`stdp`	S.E. of linear prediction (`outcome()` option required)
`stddp`	S.E. of difference in linear predictions (`outcome()` option is `outcome(`*outcome1,outcome2*`)`)

Note that you specify one new variable with `xb`, `stdp`, and `stddp` and specify either one or J new variables with `p`. These statistics are available both in and out of sample; type "`predict ... if e(sample) ...`" if wanted only for the estimation sample.

Description

`goprobit` is a user-written program that estimates generalized ordered probit models. The actual values taken on by the dependent variable are irrelevant except that larger values are assumed to correspond to "higher" outcomes. The model relaxes the parallel regression assumption of the standard ordered probit model; see [R] **oprobit** and below. `goprobit` supports linear constraints and allows the user to partially relax equal coefficients by specifying variables in `npl()` or `pl()`.

`goprobit` is a modified version of Vincent Kang Fu's `gologit` (Fu 1998) and Richard Williams' `gologit2` (Williams 2006) programs. The current version of `gologit2` allows to estimate the generalized ordered probit model using the `link(`*probit*`)` option and therefore produces results equivalent to `goprobit`. `goprobit` was written for Stata 8 and many of the references in this documentation are for Stata 8 manuals and commands.

Options

`pl, npl, npl(), pl()` provide alternative means for imposing or relaxing equal coefficients. Only one may be specified at a time.

`pl` specified without parameters constrains all independent variables to meet the parallel regression assumption. It will produce results that are equivalent to `oprobit`.

`npl` specified without parameters relaxes the parallel regression assumption for all explanatory variables. This is the default option.

`pl(`*varlist*`)` constrains the specified explanatory variables to meet the parallel regression assumption. All other variables do not need to meet the assumption. The variables specified must be a subset of the explanatory variables.

`npl(`*varlist*`)` frees the specified explanatory variables from meeting the parallel regression assumption. All other explanatory variables are constrained to meet the assumption. The variables specified must be a subset of the explanatory variables.

`constraints(`*clist*`)` specifies the linear constraints to be applied during estimation. The default is to perform unconstrained estimation. Constraints are defined with the constraint command; [R] **constraint**. `constraints(1)` specifies that the model is to be constrained according to constraint 1; `constraints(1-4)` specifies constraints 1 through 4; `constraints(1-4,8)` specifies 1 through 4 and 8. Keep in mind that the `pl` and `npl` options work by generating across-equation constraints, which may affect how any additional constraints should be specified. When using the constraint command, refer to equations by their equation `#`, e.g. `#1`, `#2`, etc.

`robust` specifies that the Huber/White/sandwich estimator of variance is to be used in place of the traditional calculation; see [U] **23.14 Obtaining robust variance estimates**. `robust` combined with `cluster()` allows observations which are not independent within cluster (although they must be independent between clusters). If you specify `pweights`, `robust` is implied.

`cluster(`*varname*`)` specifies that the observations are independent across groups (clusters) but not necessarily within groups. *varname* specifies to which group each observation belongs; e.g., `cluster(`*personid*`)` in data with repeated observations on individuals. `cluster()` affects the standard errors and variance-covariance matrix of the estimators (VCE), but not the estimated coefficients. `cluster()` can be used with `pweights` to produce estimates for unstratified cluster-sampled data.

`level(#)` specifies the confidence level in percent for the confidence intervals of the coefficients; see [R] **level**.

`score(`*newvarlist*`|`*stub*`*)` creates $J-1$ new variables, where J is the number of observed outcomes. Each new variable contains the contributions to the scores for an equation in the model; see [U] **23.15 Obtaining scores**.

If `score(`*newvarlist*`)` is specified, $J-1$ new variables must be provided.

If `score(`*stub*`*)` is specified, then variables *stub*1, ..., *stub*$J-1$ will be created.

The first variable contains $\mathrm{d}(\log L_i)/\mathrm{d}(x_i'\beta_1)$; the second variable contains $\mathrm{d}(\log L_i)/\mathrm{d}(x_i'\beta_2)$; and so on.

maximize_options control the maximization process; see help maximize. You should never have to specify them.

Options for `predict`

`p`, the default, calculates predicted probabilities.

If you do not specify the `outcome()` option, you must specify k new variables. For instance, say you fitted your model by typing "`goprobit happy income health`" and that `happy` takes on three values. Then you could type "`predict p1 p2 p3, p`" to obtain all three predicted probabilities.

If you also specify the `outcome()` option, then you specify one new variable. Say that `happy` took on values 1, 2, and 3. Then typing "`predict p1, p outcome(1)`" would produce the same `p1` as above, "`predict p2, p outcome(2)`" the same `p2` as above, etc. If `happy` took on values 7, 22, and 93, you would specify `outcome(7)`, `outcome(22)`, and `outcome(93)`. Alternatively, you could specify the outcomes by referring to the equation number (`outcome(#1)`, `outcome(#2)`, and `outcome(#3)`).

`xb` calculates the linear prediction. You must also specify the `outcome()` option.

`stdp` calculates the standard error of the linear prediction. You must specify option `outcome()`.

`stddp` calculates the standard error of the difference in two linear predictions. You must specify option `outcome()`, in this case with two particular outcomes of interest inside the parentheses; for example, "`predict sed, stdp outcome(1,3)`".

`outcome()` specifies for which outcome the statistic is to be calculated. `equation()` is a synonym for `outcome()`: it does not matter which one you use. `outcome()` and `equation()` can be specified using (1) `#1`, `#2`, ..., with `#1` meaning the first category of the dependent variable, `#2` the second category, etc.; or (2) values of the dependent variable.

Remarks

The `oprobit` command included with Stata imposes what is called the parallel regression assumption. By default, `goprobit` relaxes the parallel regression assumption and allows the effects of the explanatory variables to vary with the point at which the categories of the dependent variable are dichotomized. However, if the `pl` option is specified, `goprobit` estimates the standard ordered probit model, e.g. the commands `oprobit y x1 x2 x3` and `goprobit y x1 x2 x3, pl` will produce equivalent results.

In practice, the parallel regression assumption is often violated by the data. Standard advice in such situations is to go to a non-ordinal model, such as the multinomial logit model; see the help file to the Stata command `mlogit`. Unfortunately, such models do not take into account the ordinal nature of the dependent variable and therefore cannot be efficient. `goprobit` provides an alternative generalized ordered response model introduced by Maddala (1983: 46) and Terza (1985); see also Boes and Winkelmann (2006a). This model possibly relaxes the parallel regression assumption for some explanatory variables while being maintained for others. For example, the command `goprobit y x1 x2 x3, npl(x1)` would relax the parallel regression assumption for X_1 while maintaining it for X_2 and X_3. An equivalent command is `goprobit y x1 x2 x3, pl(x2 x3)` which forces X_2 and X_3 to meet the parallel regression assumption while not imposing it on X_1.

More formally, suppose we have an ordinal dependent variable Y which takes on the values $y = 1, 2, \ldots, J$. The generalized ordered probit model estimates a set of coefficients (including one for the constant) for each of the $J-1$ points at which the dependent variable can be dichotomized. The cell probabilities of Y are thus given by

$$\begin{aligned} P(Y = 1|X) &= F(-X'\beta_1) \\ P(Y = y|X) &= F(-X'\beta_y) - F(-X'\beta_{y-1}) \qquad y = 2, \ldots, J-1 \\ P(Y = J|X) &= 1 - F(-X'\beta_{J-1}) \end{aligned}$$

The generalized ordered probit model uses the normal distribution for $F(\cdot)$, although other distributions may also be used; see `gologit` and `gologit2`.

The standard ordered probit model (estimated by Stata's `oprobit` command and by `goprobit` with the `pl` option) restricts the β_y coefficients to be the same for every dividing point $y = 1, ..., J-1$. The generalized ordered probit model (estimated in `goprobit` via the `npl()` and `pl()` options) restricts some β_y coefficients to be the same for some dividing points while others are free to vary.

Note that the generalized ordered probit model imposes explicit restrictions on the parameters. Since probabilities are by definition constrained to be in the unit range, valid combinations must satisfy the following inequalities:

$$X'\beta_1 > X'\beta_2 > X'\beta_3 > \ldots > X'\beta_{J-1}$$

The current version of `goprobit` does not explicitly impose these restrictions during the maximization process. After fitting the model, the user should verify the validity of the model by calculating predicted probabilities. See the `gologit2` command and http://www.nd.edu/~rwilliam/gologit2/ for further discussion on this topic.

Saved Results

`goprobit` saves in `e()`:

Scalars:

`e(rc)`	return code from `maximize`	`e(k)`	# of parameters
`e(converged)`	1 if converged, 0 otherwise	`e(k_eq)`	# of equations
`e(rank)`	rank of `e(V)`	`e(ic)`	# of iterations
`e(df_m)`	model degrees of freedom	`e(N)`	# of observations
`e(ll)`	log-likelihood, full	`e(k_cat)`	# of categ. of Y
`e(ll_0)`	log-likelihood, constant-only	`e(p)`	p-value of χ^2 test
`e(chi2)`	χ^2 test, full model against constant-only		

Macros:

`e(cmd)`	estimation command	`e(title)`	title in output
`e(depvar)`	name of Y	`e(xvars)`	names of X
`e(nplvars)`	names of variables meeting the `npl` option		
`e(plvars)`	names of variables meeting the `pl` option		
`e(clustvar)`	name of variable identifying clusters		
`e(predict)`	program used to implement `predict`		
`e(opt)`	ml maximization	`e(ml_method)`	ml method
`e(crittype)`	optimization criterion	`e(technique)`	max technique
`e(user)`	name of user provided program that calculates the log-likelihood		
`e(chi2type)`	"Wald" or "LR"; type of model χ^2 test		
`e(properties)`	estimator properties; "`b V`"		

Matrices:

`e(cat)`	values of Y	`e(ilog)`	iteration log
`e(b)`	coefficient vector	`e(gradient)`	gradient vector
`e(V)`	variance-covariance matrix of the estimators		

Functions:

`e(sample)`	marks estimation sample

See [R] **maximize** for a complete list of all returned results specific to ml. See [P] **ereturn** for a description of results obtained post estimation.

Examples

- `goprobit happy linc unempl health if male == 1, robust`
- `goprobit happy linc unempl health if male == 1, robust npl(linc)`
- `goprobit, level(99)`
- `predict xb1, xb outcome(#1)`

Acknowledgements

Richard Williams of the Notre Dame Department of Sociology wrote `gologit2`. He kindly gave me permission to use parts of his code for the `goprobit` project. For a more detailed description of `gologit2` and its features, see the reference below or the help file of `gologit2`.

References

Boes, S. and R. Winkelmann (2006): "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90, 165-179.

Fu, V.K. (1998): "Estimating Generalized Ordered Logit Models," *Stata Technical Bulletin*, 8, 160-164.

Long, J.S and J. Freese (2003): *Regression Models for Categorical Dependent Variables Using Stata*, revised edition, Stata Press.

Maddala, G. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press: Cambridge.

Terza, J. (1985): "Ordered Probit: A Generalization," *Communications in Statistics A. Theory and Methods*, 14, 1-11.

Williams, R. (2006): "Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables," *The Stata Journal*, 6(1), 58-82. A pre-publication version is available at http://www.nd.edu/ rwilliam/gologit2/gologit2.pdf.

Winkelmann, R. and S. Boes (2006): *Analysis of Microdata*, Springer: Berlin.

Appendix B

`regoprob` — A Stata Module to Estimate Random Effects Generalized Ordered Probit Models

This appendix describes an extension of the `goprobit` module to panel data. In a panel dataset, each individual (or group) is followed over time such that observations within a group may not be independent. The generalized ordered probit model with random effects (estimated by `regoprob`) efficiently accounts for the panel information by specifying each outcome probability conditional on an individual specific random effect. Within-group observations are therefore explicitly allowed to be correlated.

Like the `goprobit` module, `regoprob` is not included in the basic distribution of Stata and needs to be downloaded and installed before use. Again, the easiest way in Stata is via `ssc install regoprob`. Alternatively, type `findit regoprob` in the command line and follow the on-screen instructions. If you experience any problems downloading `regoprob`, then try

http://ideas.repec.org/c/boc/bocode/s456604.html

http://econpapers.repec.org/software/bocbocode/s456604.htm

for manual download. `regoprob` needs commands `goprobit` and `ghquadm` for execution. When downloading `regoprob`, both commands should be automatically downloaded by Stata (if not already installed).

Syntax

```
regoprob depvar [indepvars] [if exp] [in range] [, i(varname) quadrat(#)
    pl pl(varlist) npl npl(varlist) constraints(clist) level(#)
    maximize_options ]
```

`regoprob` shares the features of all estimation commands; see [R] **estimates**.
`regoprob` typed without arguments redisplays previous results.

The syntax of `predict` following `regoprob` is

```
predict [type] newvarname(s) [if exp] [in range] [, statistic
    outcome(outcome) ]
```

where *statistic* is

`p`	probability marginal on the individual effect (specify one new variable and `outcome()` option, or specify J new variables, $J = \#$ of outcomes); default
`xb`	linear prediction (`outcome()` option required)
`stdp`	S.E. of linear prediction (`outcome()` option required)
`stddp`	S.E. of difference in linear predictions (`outcome()` option is `outcome(`*outcome1,outcome2*`)`)

Note that you specify one new variable with `xb`, `stdp`, and `stddp` and specify either one or k new variables with `p`. These statistics are available both in and out of sample; type "`predict ... if e(sample) ...`" if wanted only for the estimation sample.

Description

`regoprob` is a user-written program that estimates panel data generalized ordered probit models with random effects. The actual values taken on by the dependent variable are irrelevant except that larger values are assumed to correspond to "higher" outcomes. The model relaxes the parallel regression assumption of the standard ordered probit model; see [R] **oprobit** and the help file to its random effects counterpart **reoprob** (if installed). `regoprob` supports linear constraints and allows the user to partially relax equal coefficients by specifying variables in `npl()` or `pl()`. The likelihood contribution for each unit is approximated using Gauss-Hermite quadrature.

`regoprob` is a modified version of `goprobit` and requires installation of `goprobit` and `ghquadm` before using. `regoprob` was written for Stata 8 and many of the references in this documentation are for Stata 8 manuals and commands.

Options

`i()` specifies the variable corresponding to an independent unit (e.g., a subject id). `i(`*varname*`)` is not optional.

`quadrat()` specifies the number of points to use for the Gauss-Hermite quadrature. It is optional, and the default is 12. Increasing this value improves accuracy, but also increases computation time.

`pl, npl, npl(), pl()` provide alternative means for imposing or relaxing equal coefficients. Only one may be specified at a time.

`pl` specified without parameters constrains all independent variables to meet the parallel regression assumption. It will produce results equivalent to `reoprob`.

`npl` specified without parameters relaxes the parallel regression assumption for all explanatory variables. This is the default option.

`pl(`*varlist*`)` constrains the specified explanatory variables to meet the parallel regression assumption. All other variables do not need to meet the assumption. The variables specified must be a subset of the explanatory variables.

`npl(`*varlist*`)` frees the specified explanatory variables from meeting the parallel regression assumption. All other explanatory variables are constrained to meet the assumption. The variables specified must be a subset of the explanatory variables.

`constraints(`*clist*`)` specifies the linear constraints to be applied during estimation. The default is to perform unconstrained estimation. Constraints are defined with the constraint command; [R] **constraint**. `constraints(1)` specifies that the model is to be constrained according to constraint 1; `constraints(1-4)` specifies constraints 1 through 4; `constraints(1-4,8)` specifies 1 through 4 and 8. Keep in mind that the `pl` and `npl` options work by generating across-equation constraints, which may affect how any additional constraints should be specified. When using the constraint command, refer to equations by their equation `#`, e.g. `#1`, `#2`, etc.

`level(#)` specifies the confidence level in percent for the confidence intervals of the coefficients; see [R] **level**.

maximize_options control the maximization process; see help maximize. You should never have to specify them.

Options for `predict`

`p`, the default, calculates predicted probabilities *marginal* on the individual effect.

If you do not specify the `outcome()` option, you must specify k new variables. For instance, say you fitted your model by typing "`regoprob happy income health, i(persnr)`" and that `happy` takes on three values. Then you could type "`predict p1 p2 p3, p`" to obtain all three predicted probabilities.

If you also specify the `outcome()` option, then you specify one new variable. Say that `happy` took on values 1, 2, and 3. Then typing "`predict p1, p outcome(1)`" would produce the same `p1` as above, "`predict p2, p outcome(2)`" the same `p2` as above, etc. If `happy` took on values 7, 22, and 93, you would specify `outcome(7)`, `outcome(22)`, and `outcome(93)`. Alternatively, you could specify the outcomes by referring to the equation number (`outcome(#1)`, `outcome(#2)`, and `outcome(#3)`).

`xb` calculates the linear prediction. You must also specify the `outcome()` option.

`stdp` calculates the standard error of the linear prediction. You must specify option `outcome()`.

`stddp` calculates the standard error of the difference in two linear predictions. You must specify option `outcome()`, in this case with two particular outcomes of interest inside the parentheses; for example, "`predict sed, stdp outcome(1,3)`".

`outcome()` specifies for which outcome the statistic is to be calculated. `equation()` is a synonym for `outcome()`: it does not matter which one you use. `outcome()` and `equation()` can be specified using (1) `#1`, `#2`, ..., with `#1` meaning the first category of the dependent variable, `#2` the second category, etc.; or (2) values of the dependent variable.

Remarks and Methods

Standard ordered response models such as `oprobit` or `reoprob` impose what is called the parallel regression assumption. By default, `regoprob` relaxes this assumption and allows the effects of the explanatory variables to vary with the point at which the categories of the dependent variable are dichotomized. However, if the `pl` option is specified, `regoprob` estimates the standard random effects ordered probit model, e.g., the commands `reoprob y x1 x2, i(id)` and `regoprob y x1 x2, i(id) pl` will produce equivalent results.

The cross-sectional generalized ordered probit model has been introduced by Maddala (1983: 46) and Terza (1985), an extension to panel data has been proposed by Boes and Winkelmann (2006b). More formally, suppose Y_{it} is an ordinal dependent variable which takes on the values $y = 1, \ldots, J$, where i denotes cross-sectional units and t the time dimension of the (panel) dataset; see the Stata manual [XT] **xt** – Introduction to `xt` commands. The random effects generalized ordered probit model estimates a set of coefficients (including one for the constant) for each of the $J-1$ dichotomization points of Y_{it}. The outcome probabilities conditional on the individual effect α_i are equal to

$$\begin{aligned} P(Y_{it} = 1|X_{it}, \alpha_i) &= F(-X'_{it}\beta_1 - \alpha_i) \\ P(Y_{it} = y|X_{it}, \alpha_i) &= F(-X'_{it}\beta_y - \alpha_i) - F(-X'_{it}\beta_{y-1} - \alpha_i) \qquad y = 2, \ldots, J-1 \\ P(Y_{it} = J|X_{it}, \alpha_i) &= 1 - F(-X'_{it}\beta_{J-1} - \alpha_i) \end{aligned}$$

The random effects generalized ordered probit model uses the standard normal distribution as the cumulative distribution for $F(.)$, although other distributions may also be used. The individual effects are assumed to be normally distributed with zero mean and variance σ^2 which is parameterized as $\rho = \sigma^2/(1+\sigma^2)$.

The standard random effects ordered probit (estimated by `reoprob` and `regoprob` with the `pl` option) restricts the β_y coefficients to be the same for every dividing point $y = 1, \ldots, J-1$. The random effects generalized ordered probit model (estimated in `regoprob` via the `npl()` and `pl()` options) restricts some β_y coefficients to be the same for every dividing point while others are free to vary.

Note that the generalized ordered probit model imposes explicit restrictions on the parameters. Since probabilities are by definition constrained to be in the unit range, valid combinations must satisfy the following inequalities:

$$X'_{it}\beta_1 > X'_{it}\beta_2 > X'_{it}\beta_3 > \ldots > X'_{it}\beta_{J-1}$$

The current version of `regoprob` does not explicitly impose these restrictions during the maximization process. After fitting the model, the user should verify the validity of the model by calculating predicted probabilities. See the `gologit2` command and http://www.nd.edu/~rwilliam/gologit2/ for further discussion on this topic.

The likelihood contribution for each cross-sectional unit is approximated using a Gauss-Hermite quadrature. See Butler and Moffitt (1982) for details about using Gauss-Hermite quadrature to approximate such integrals and Footnote 4 (Page 17). Note that the results are in terms of $\rho = \sigma_\alpha^2/(1+\sigma_\alpha^2)$, rather than σ_α directly.

`regoprob` uses the d1 method (analytic first derivatives) of Stata's `ml` commands.

Saved Results

`regoprob` saves in `e()`:

Scalars:

`e(rc)`	return code from `maximize`	`e(k)`	# of parameters
`e(converged)`	1 if converged, 0 otherwise	`e(k_eq)`	# of equations
`e(rank)`	rank of `e(V)`	`e(ic)`	# of iterations
`e(df_m)`	model degrees of freedom	`e(N)`	# of observations
`e(ll)`	log-likelihood, full	`e(k_cat)`	# of categ. of Y
`e(ll_0)`	log-likelihood, constant-only	`e(p)`	p-value of χ^2 test
`e(chi2)`	χ^2 test, full model against constant-only		

Macros:

`e(cmd)`	estimation command	`e(title)`	title in output
`e(depvar)`	name of Y	`e(xvars)`	names of X
`e(nplvars)`	names of variables meeting the `npl` option		
`e(plvars)`	names of variables meeting the `pl` option		
`e(clustvar)`	name of variable identifying clusters		
`e(predict)`	program used to implement `predict`		
`e(opt)`	ml maximization	`e(ml_method)`	ml method
`e(crittype)`	optimization criterion	`e(technique)`	max technique
`e(user)`	name of user provided program that calculates the log-likelihood		
`e(chi2type)`	"Wald" or "LR"; type of model χ^2 test		
`e(properties)`	estimator properties; "`b V`"		

Matrices:

`e(cat)`	values of Y	`e(ilog)`	iteration log
`e(b)`	coefficient vector	`e(gradient)`	gradient vector
`e(V)`	variance-covariance matrix of the estimators		

Functions:

`e(sample)`	marks estimation sample

See [R] **maximize** for a complete list of all returned results specific to ml. See [P] **ereturn** for a description of results obtained post estimation.

Examples

- `regoprob happy linc unempl t2 t3 t4 if male == 1, i(id)`
- `regoprob happy linc unempl t2 t3 t4 if male == 1, i(id) npl(linc)`
- `regoprob, level(99)`
- `predict xb1, xb outcome(#1)`

Acknowledgements

Guillaume R. Frechette of the New York University wrote `reoprob`. He kindly gave me permission to use parts of his code for `regoprob`. See the help file of `reoprob` for a description of the `reoprob` command (if installed), and the references listed below.

Richard Williams of the Notre Dame Department of Sociology wrote `gologit2`. He kindly gave me permission to use parts of his code for the `goprobit` project. For a more detailed description of `gologit2` and its features, see also the references below or the help file of `gologit2` (if installed).

`regoprob` combines the features of `goprobit` and `reoprob`, i.e., estimates panel data generalized ordered probit models.

References

Boes, S. and R. Winkelmann (2006a): "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90, 165-179.

Boes, S. and R. Winkelmann (2006b): "The Effect of Income on Positive and Negative Subjective Well-Being," SOI Working Paper 0605.

Butler, J.S. and R. Moffitt (1982): "A computationally efficient quadrature procedure for the one-factor multinomial probit model," *Econometrica*, 50, 761-764.

Frechette, G.R. (2001): "sg158.1: Update to random-effects ordered probit," *Stata Technical Bulletin*, 61, 12. Reprinted in *Stata Technical Bulletin Reprints*, Vol. 10, 266-267.

Frechette, G.R. (2001): "sg158: Random-effects ordered probit," *Stata Technical Bulletin*, 59, 2001, 23-27. Reprinted in *Stata Technical Bulletin Reprints*, Vol. 10, 261-266.

Maddala, G. (1983): *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge University Press: Cambridge.

Terza, J. (1985): "Ordered Probit: A Generalization," *Communications in Statistics A. Theory and Methods*, 14, 1-11.

Williams, R. (2006): "Generalized Ordered Logit/ Partial Proportional Odds Models for Ordinal Dependent Variables," *The Stata Journal*, 6(1), 58-82. A pre-publication version is available at http://www.nd.edu/ rwilliam/gologit2/gologit2.pdf.

Winkelmann, R. and S. Boes (2006): *Analysis of Microdata*, Springer: Berlin.

Curriculum Vitae

Personal

Birthdate: April 18, 1978
Citizenship: German
Email: boes@sts.uzh.ch

Education

10/2003–09/2007 Doctoral Studies in Economics, University of Zurich, Switzerland

10/1998–09/2003 M.Sc. in Economics, University of Konstanz, Germany

Research Interests

Microeconometrics, in particular discrete response models, semi- and nonparametric methods; applied socio-economic modeling in general.

Publications and Work in Progress

Boes, S., M. Lipp, and R. Winkelmann (2007): "Money Illusion Under Test," *Economics Letters*, 94, 332-337.

Boes, S., and R. Winkelmann (2006): "Ordered Response Models," *Allgemeines Statistisches Archiv*, 90(1), 165-180.

Winkelmann, R., and S. Boes (2006): *Analysis of Microdata*, Springer: Berlin.

Boes, S. (2007): "Nonparametric Analysis of Treatment Effects in Ordered Response Models," *SOI Working Paper No. 0709.*

Boes, S. (2007): "Count Data Models with Unobserved Heterogeneity: An Empirical Likelihood Approach," *SOI Working Paper No. 0704.*

Boes, S., and R. Winkelmann (2006): "The Effect of Income on Positive and Negative Subjective Well-Being," *SOI Working Paper No. 0605.*

Boes, S. (2006): "regoprob – A Stata Module to Estimate Random Effects Generalized Ordered Probit Models."

Boes, S. (2006): "goprobit – A Stata Module to Estimate Generalized Ordered Probit Models."