NBER WORKING PAPER SERIES

ANALYZING BOUNDED COUNT DATA

John Mullahy

Working Paper 31814
http://www.nber.org/papers/w31814

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2023

Thanks are owed to Joao Santos Silva, Jon Skinner, and Jeff Wooldridge for helpful comments on an earlier version of this paper. Beyond this paper many colleagues and seminar participants have previously provided helpful comments on various aspects of the ideas presented here. Some of the work was supported by an Evidence for Action grant (#73336) from RWJF to UW-Madison. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Analyzing Bounded Count Data
John Mullahy
NBER Working Paper No. 31814
October 2023
JEL No. I1,I10

ABSTRACT

This paper presents and assesses analytical strategies that respect the bounded count structures of outcomes that are encountered often in health and other applications. The paper's main motivation is that the applied econometrics literature lacks a comprehensive discussion and critique of strategies for analyzing and understand such data. The paper's goal is to provide a treatment of prominent issues arising in such analyses, with particular focus on evaluations in which bounded count outcomes are of interest, and on econometric modeling of their probability and moment structures. Hopefully the paper will provide a toolkit for researchers so they may better appreciate the range of questions that might be asked of such data and the merits and limitations of the analytical methods they might contemplate to study them. It will be seen that the choice of analytical method is often consequential: questions of interest may be unanswerable when some familiar analytical methods are deployed in some circumstances.

John Mullahy
University of Wisconsin-Madison
Dept. of Population Health Sciences
787 WARF, 610 N. Walnut Street
Madison, WI 53726
and NBER
jmullahy@facstaff.wisc.edu

> It is never true in reality that two and two make four; for we cannot add unlike things and there are no two real things in the universe which are exactly alike. It is only to completely abstract units, entirely without content, that the most familiar laws of number and quantity apply. Yet no one questions the practical utility of such laws.
>
> Frank H. Knight, "The Limitations of Scientific Method in Economics" (1924)

1. Preliminaries, Definitions, and Assumptions

Figure 1 depicts the sample frequency distributions of four health-related outcomes:

Panel (a): the number of days on which adult respondents reported engaging in vigorous or moderate exercise in the prior week (VM7; 2015 Health Survey of England (HSE))

Panel (b): the number of chronic conditions from a list of eight, reported by adult respondents (CC8; 2021 Behavioral Risk Factors Surveillance System (BRFSS))

Panel (c): the number of physically healthy days in the previous 30 days, reported by adult respondents (HD30; 2021 BRFSS)

Panel (d): the PHQ-2 depression screener score for adult respondents (PHQ2; 2020 Medical Expenditure Panel Survey (MEPS))

While the health phenomena they describe are quite different these four outcomes share in common two features that in tandem define this paper's main focus: they are measured as nonnegative integers, and they are bounded from below and above. While many familiar health measures possess both measurement attributes it will turn out that some of these, like the PHQ-2 score in panel (d), have essential measurement properties that differentiate them from those shown in panels (a), (b), and (c).

This paper assesses analytical strategies that respect the bounded count structures of outcomes like those whose distributions are shown in panels (a)-(c) of figure 1. In some instances the study of outcomes having such structures can proceed with little special treatment, but in others their bounded count nature requires attention. The paper's goal is to serve as a comprehensive overview or toolkit for researchers working with such data so they might better appreciate the range of questions that might be asked and the merits and limitations of the empirical strategies they might deploy to analyze them.

The Anatomy of BCOs and the Nature of Their Measurement

While some of the subsequent discussion is not novel it should nonetheless prove useful to formalize ideas that sometimes lack a rigorous basis in the analysis of bounded count outcome (BCO)

data. Hopefully these formalities will serve to streamline discussion and provide it a sturdy foundation.[1]

Let s be the observed scalar BCO of interest, and let its probability distribution P have non-negative integer support $V = \langle 0, 1, \ldots, M \rangle$, where M is a finite positive integer.[2] It is in this sense that s is a BCO: the only values it can assume with positive probability are the nonnegative integers in V. The paper's setting is one with a known finite number, M, of "experiments" that result in a set of M binary outcomes. Experiments may be trials, games, survey items, time periods, geographical units, or any context where M outcomes are experienced for population or sample subjects. The binary-outcome experiments may be simple (coin toss Bernoulli trials where each outcome's probability is the same) or less simple (indications of matches in matching games, counts of chronic conditions, etc., where outcomes' marginal probabilities may vary over $m \in E = \{1, \ldots, M\}$). The M outcomes may or may not be independent or exchangeable across experiments and, while realized, may or may not be observed.[3]

Define the M-vector $\mathbf{y} = [y_1, \ldots, y_M]$ where each $y_m \in \{0, 1\}$ indicates the result of experiment $m \in E$. The count outcome s is then defined as the sum $s = \sum_{m \in E} y_m$. As such s counts the number of "successful" experiments. For some purposes s might usefully be interpreted as an aggregate or a dimension-reducing measure or score.[4] Indeed the symbol s is chosen to connote *sum*, *score*, or *successes*.

For the most part this paper assumes that s has ratio-scale measure. In his classic paper on measurement, Stevens, 1946, notes how true count measures naturally possess ratio-scale properties:

> Foremost among the ratio scales is the scale of number itself—cardinal number—the scale we use when we count such things as eggs, pennies, or apples. This scale of numerosity of aggregates is so basic and so common that it is ordinarily not even mentioned in discussions of measurement.

Knight's statement in the epigraph recognizes that in a strict sense such count measures always involve the adding up of unlikes but emphasizes their practical utility. (Recall George Box's dictum about models'

---

[1] Hoaglin and Tukey, 2006, provide a useful overview of BCO data.

[2] The paper uses the conventions that angle brackets $\langle \bullet \rangle$ indicate ordered tuples, curly brackets $\{\bullet\}$ indicate sets, and square brackets $[\bullet]$ indicate vectors. In some cases the distinction between tuples and sets is unimportant while in others it is useful to draw.

[3] The paper assumes that M is the same for everyone. While it would be straightforward to allow M to vary exogenously across the population the additional notational clutter is probably not warranted.

[4] In some applications score-type measures are computed as weighted sums of components, $\sum_{m \in E} w_m y_m$ (e.g. polygenic risk scores; Khera et al., 2016).

correctness versus usefulness.) Yet such practical utility is question-dependent. Researchers should ponder whether BCO measures meaningfully address their questions or whether—in Ravallion's, 2012, apt terminology—they are just "mashup indexes," convenient to analyze but of dubious scientific value.[5] (See also Maasoumi and Racine, 2016, and Sims et al., 2008.)

In some applications defining s to have nonnegative integer values in V is an arbitrary coding of an (M+1)-category fundamentally ordinal phenomenon. A prominent example in health applications is the mapping of $\{E, V, G, F, P\}$ health-status self reports into s-measures like $\{1, 2, 3, 4, 5\}$ and then treating s as if it possessed ratio-scale properties (e.g. computing moments). Among the perils of analyzing ordinal-measure outcomes is that individuals may differ in how they report experiences (health, happiness, etc.; Bond and Lang, 2019), a concern that does not plague true ratio-scale count measures.

No matter how the binary outcomes y_m are defined this paper proceeds for the most part by assuming that, to the researcher's satisfaction, their sum is a coherent ratio-scale count measure of a phenomenon of interest having "practical utility" for addressing the questions at hand.[6] In what follows it will be noted when ratio-scale measurement is not formally required for particular results to hold.

Equivalence Classes

Define

$$Y_m = \left\{ \mathbf{y} \middle| \sum_{j \in E} y_j = m \right\}, \quad m \in V . \tag{1}$$

[5] A motivation for using data on multiple y_m and summing them to obtain s is often to provide richer characterization and/or more precise estimates of some phenomenon of interest (welfare, health, etc.) than would be obtained by appealing to only one y_m. The more similar are the y_m the more reasonable it may be to sum them yet the less incremental information about the phenomenon may be gained. Conversely the more unlike are the y_m the less reasonable it may be to sum them even though the full ensemble of the y_m may offer rich information about the phenomenon/a of interest.

[6] While it is sometimes germane in applications to consider a general ordered structure with each $y_m \in \langle 0, 1, \ldots, R \rangle$ the paper focuses on binary y_m. One familiar case with R>1 is where each y_m is ordinal, defined on a Likert scale (e.g. Apgar scores used to assess neonates' health status (AAP/ACOG, 2006) use M=5 and R=2 while PHQ-2 scores (figure 1(d); Fleishman et al., 2014) use M=2 and R=3). While contrary arguments are sometimes advanced, such measurements are at best ordinal in nature and, as such, their mapping into non-negative integer values $\langle 0, 1, \ldots, R \rangle$ is arbitrary. While the measurement properties of quantities arising from summing M arbitrarily coded variables to obtain s are unclear, assertions that s measures thus derived possess ratio-scale properties are dubious.

The $\binom{M}{m}$ vector elements of each Y_m thus form an equivalence class (see Halmos, 1960) with respect to their elements' sums. That is, in each Y_m all $\mathbf{y}'$ and $\mathbf{y}''$ satisfy $\mathbf{y}' \sim_s \mathbf{y}''$ where the equivalence relationship $\sim_s$ is "has elements whose sum is the same as the elements of."

A first-order consideration is that even though Y_m defines an equivalence class with respect to $\sim_s$ equivalence this does not imply equivalence with respect to other criteria. For instance even though $\mathbf{y}' = [1, 0, 0] \sim_s [0, 1, 0] = \mathbf{y}''$ it may be that $\mathbf{y}'$ and $\mathbf{y}''$ correspond to different levels of welfare or health. Conversely $[1, 0, 0] = \mathbf{y}' \not\sim_s \mathbf{y}''' [0, 1, 1]$ even though $\mathbf{y}'$ and $\mathbf{y}'''$ may correspond to equivalent levels of welfare or health. As such using values of s as barometers of welfare or health means maintaining $\sim_s$ equivalences of its underlying $\mathbf{y}$, which equivalences may or may not be credible to maintain. This is among the compelling reasons to question the merits of mashup indexes in applications.

Probabilities, Moments, Quantiles

Let $\Pr(\mathbf{y})$ denote the joint distribution of the y_m. The probabilities $\Pr(s = m)$ are given by

$$\Pr(s = m) = \sum_{\mathbf{y} \in Y_m} \Pr(\mathbf{y}), \quad m \in V, \tag{2}$$

indicating that properties of $\Pr(s = m)$ are inherited from $\Pr(\mathbf{y})$. Define the probability distribution P of s as the ordered (M+1)-tuple

$$P = \langle \Pr(s = 0), \ \ldots, \ \Pr(s = M) \rangle, \tag{3}$$

whose terms are non-negative and sum to one. Until section 7 no further assumptions are made about the structure of P.[7] P has corresponding cumulatives and cumulative distribution

[7] This characterization of a BCO holds for any finite M. For some purposes the paper's discussion may be more useful if M is "not too large," e.g. M=5 or M=10, but perhaps not M=50 or M=100, i.e. few valuable insights may be sacrificed by treating P as continuous. For others, however, decision- or welfare-relevant parameters (e.g. $\Pr(s = 0)$) may require P to be treated as discrete even when M is "large".

$$C(v) = \sum_{m=0}^{v} \Pr(s = m), \quad v \in V \tag{4}$$

and

$$C = \langle C(0),\ C(1),\ \ldots,\ C(M) \rangle = \langle \Pr(s = 0),\ \Pr(s \le 1),\ \ldots,\ 1 \rangle . \tag{5}$$

The mean of distribution P, $E_P(s)$, can be written in four equivalent ways:

$$\begin{aligned} E_P(s) &= \sum_{m \in V} m \cdot \Pr(s = m) \\ &= \sum_{m \in V} \sum_{\mathbf{y} \in Y_m} m \cdot \Pr(\mathbf{y}) \\ &= \sum_{m \in V} (1 - C(m)) \\ &= \sum_{m \in E} \Pr(y_m = 1) \end{aligned} \tag{6}$$

The mean is necessarily finite, $E_P(s) \in (0,\ M)$, with open-interval notation used to highlight that the mean would only be 0 or M with degenerate P. Analogous derivations apply for higher-order raw moments $E_P(s^r)$, e.g. $E_P(s^r) = \sum_{m \in V} m^r \cdot \Pr(s = m)$ with $E_P(s^r) \in (0,\ M^r)$ which will be finite for finite M and r. Unless helpful for clarity the subscript P on the expectation operator will be suppressed. The paper sometimes uses $\mu^r = E(s^r)$ and $\mu = E(s)$ as shorthand.

The structure of P can be further appreciated by considering its quantiles. The q-th quantile of P is

$$Q_P(q) = \arg\min_{v \in V} \{ C(v) \,|\, C(v) \ge q \}, \quad q \in (0,\ 1) . \tag{7}$$

The quantiles $Q_P(q)$ are thus integer-valued. For any BCO distribution P with support V the quantiles and the mean are related by:

$$(1 - q) \cdot Q(q) \ \le\ \mu\ \le\ q \cdot Q(q)\ +\ (1 - q) \cdot M , \tag{8}$$

where P subscripts on Q are suppressed. The mean is thus strictly bounded by functions of the quantiles since the bounds interval width is less than M. If $Q(q) \in \{0,\ M\}$ only one bound in (8) is informative about μ but if $Q(q) \in \{1,\ \ldots,\ M - 1\}$ both are. For the median (8) becomes

$$.5Q(.5) \;\le\; \mu \;\le\; .5\big(Q(.5)+M\big). \tag{9}$$

The bounds interval width in (9) is .5M.

Covariates and Partial Effects

In many instances conditional-on-covariates parameters will be of interest. Let $\mathbf{x}$ denote the generic (K+1)-vector $\big[1,\, \mathbf{x}_v\big]=\big[1,\, x_1,\, \ldots,\, x_K\big]$. Particular values of $\mathbf{x}$ are denoted $\mathbf{x}(j)$ with corresponding conditional and cumulative distributions

$$P^j = \Big\langle \Pr\big(s=0\big|\mathbf{x}(j)\big),\; \ldots,\; \Pr\big(s=M\big|\mathbf{x}(j)\big)\Big\rangle \tag{10}$$

and

$$C^j = \Big\langle \Pr\big(s=0\big|\mathbf{x}(j)\big),\; \Pr\big(s\le 1\big|\mathbf{x}(j)\big),\; \ldots,\; 1\Big\rangle. \tag{11}$$

Of interest then are parameters that include $\Pr\big(s=m\big|\mathbf{x}\big)$ and $\mu^r(\mathbf{x})$. As appropriate to the questions at hand it may sometimes be useful to assume that $\mathbf{x}$ is exogenous, as-if-randomly-assigned (AIRA), etc.[8]

Given $\mathbf{x}$ the partial effects (PEs) on various parameters are often of primary interest. Let D_k be the partial effect operator $\Delta/\Delta x_k$ or $\partial/\partial x_k$ depending on how x_k is measured. The PEs of interest will often be on the conditional moments, $D_k\mu^r(\mathbf{x}) = D_k E\big(s^r\big|\mathbf{x}\big)$, or conditional probabilities, $D_k \Pr\big(s=m\big|\mathbf{x}\big)$, with the latter satisfying $\sum_{m\in V} D_k \Pr\big(s=m\big|\mathbf{x}\big)=0$. In practice average partial effects (APEs) like $E_{\mathbf{x}} D_k\mu^r(\mathbf{x})$ or $E_{\mathbf{x}} D_k \Pr\big(s=m\big|\mathbf{x}\big)$ are often the target of estimation, a leading case being where $x_k \in \{0,\, 1\}$ is an AIRA treatment that identifies the marginal potential-outcome distributions $P^j = \Big\langle \Pr\big(s=0\big|x_k=j,\, \mathbf{x}_{-k}\big),\; \ldots,\; \Pr\big(s=M\big|x_k=j,\, \mathbf{x}_{-k}\big)\Big\rangle$, $j\in\{0,\, 1\}$. The APEs are typically estimated by appealing to analogy principles as $\sum_{n=1}^N w_n D_k \widehat{\mu^r}(\mathbf{x})$ or $\sum_{n=1}^N w_n D_k \widehat{\Pr}\big(s=m\big|\mathbf{x}_n\big)$, with $\sum_{n=1}^N w_n = 1$.

The cautions raised above about the nature of mashup indexes should also elicit concerns about

[8] If M varies exogenously in the population (see footnote 3) then the paper's results go through unchanged with conditioning on $(\mathbf{x}, M)$ instead of just $\mathbf{x}$.

the nature of partial effects like

$$D_k E\left(s\middle|\mathbf{x}\right) = \sum_{m\in E} D_k \Pr\left(y_m = 1\middle|\mathbf{x}\right) \tag{12}$$

from (6), and

$$D_k \Pr\left(s = m\middle|\mathbf{x}\right) = \sum_{\mathbf{y}\in Y_m} D_k \Pr\left(\mathbf{y}\middle|\mathbf{x}\right) \tag{13}$$

from (2). Whether the sum of the summands on the RHS of (12) or (13) answers a scientifically meaningful question about consequences of changing **x** should merit serious consideration.

What Data Are Available?

There are two typical situations regarding data availability. The first is where the sample provides information on s but not on its y_m components. The second is where a sample provides information on all M y_m outcomes and their empirical joint distribution, in which case s is obviously knowable.[9]

Observing the y_m outcomes may be useful or essential for some purposes (Mullahy, 2023b). For instance suppose only s is observed. Apart from the extreme cases s=0 and s=M knowing that s=v obviously means that there are v realizations $y_m = 1$ and M–v realizations $y_m = 0$ but which y_m assume these values is unknown. The absence of such information may be unimportant if the y_m are exchangeable in P, but in other cases particular patterns of the $y_m = 0$ and $y_m = 1$ outcomes in P may be relevant. (For instance with M=7 and s=2 having positive outcomes on Monday and Thursday may mean something different than having positive outcomes on Saturday and Sunday.) While such jointness considerations may be critical for addressing some questions this paper focuses henceforth on s and assumes analysts have access to samples of its realizations whether or not data on the y_m are available.

Context of and Plan for the Paper

For some purposes BCOs pose few novel analytical considerations when it comes to their use in

[9] An example of the former is the BRFSS HD30 measure whose sample distribution is depicted in panel (c) of figure 1. Two examples of the latter are the HSE VM7 measure and the BRFSS CC8 measure whose sample distributions are depicted in panels (a) and (b) of figure 1.

Box 1: A Running Example—The 2015 Health Survey of England

A sample of N=6,769 respondents ages 25+ in the 2015 Health Survey of England (henceforth HSE) provides data for several empirical examples running throughout the paper. The health outcome of interest—whose sample distribution is depicted in figure 1(a)—is VM7, the number of days in the recall week on which the respondent reported engaging in vigorous and/or moderate physical activity. Thus M=7 and $s_n \in V = \{0, \ldots, 7\}$ where $n \in \{1, \ldots, N\}$ indexes observations. s is computed as $s_n = \sum_{m \in E} y_{m,n}$ where the $y_{m,n}$ are binary indicators of whether the respondent engaged in vigorous and/or moderate exercise on day m of the recall period (m=1 denotes Monday, m=7 denotes Sunday, etc.). The sample joint distribution of $\mathbf{y} = [y_1, \ldots, y_7]$, $\Pr(\mathbf{y})$, is observable in the HSE.

In some examples covariates **x** will be of interest: age in years (Age, treated as continuously measured though actually constructed using interval midpoints) and binary indicators sex (Female=1), schooling (NVQ45=1 if NVQ4 or NVQ5 qualification), residential region (South=1 if residing in Strategic Health Authority regions East England, London, South East, South Central, or South West), and urban residence (Urban=1). As-if-randomly-assigned treatments based on interviews' calendar quarter are used as treatments in an example discussed in section 4, under the premise that with all else equal individuals may be more inclined to engage in vigorous or moderate exercise in the warm-weather months (quarters 2 and 3) than in the cool-weather months (quarters 1 and 4).

evaluations or to estimation of regression models to describe their stochastic features. For other purposes, however, BCOs' bounded-count nature raises distinct analytical issues. While it is the latter purposes to which the paper devotes most of its attention the paper devotes some attention to the former so that it might claim to offer a reasonably comprehensive treatment of the analysis of BCO data in empirical research.

The paper is organized into main two parts. Part I is devoted to analytical issues arising when BCOs are used in various evaluation contexts. Part II focuses on estimation of regression models (Manski, 1991) to describe BCOs' probability distributions and moment structures. While the two parts are logically distinct the paper highlights several areas where their themes intersect.

In part I, section 2 considers evaluation using point-identified marginal distributions, focusing in particular on the implications of BCOs' measurement structure for understanding various dominance relationships. Section 3 examines average and quantile treatment effects. Section 4 investigates (partial) identification of treatment-effect distributions, noting special considerations that arise with BCOs. In part II section 5 provides an overview of data and estimation issues. Section 6 considers estimation of conditional moment models $\mu^r(\mathbf{x})$. Section 7 discusses issues arising with estimation of conditional probability models $\Pr(s = m|\mathbf{x})$. That discussion is fairly extensive since the merits and shortcomings of

various parametric specifications used in empirical work to model $\Pr(s=m|\mathbf{x})$ may not be widely appreciated. Section 8 concludes. Most technical material is relegated to footnotes or appendixes.

Part I - Evaluation with Bounded Count Outcomes

Part I of the paper assesses the implications of BCO data in evaluation contexts that run the gamut from medical technology assessment to the quantification of health disparities and many others. In some cases methods familiar in the analysis of unbounded and/or continuously distributed outcomes can be applied without modification, in others they must be adapted to accommodate the bounded-count nature of the data, while in still others BCO data simplify analysis compared with unbounded and/or continuously distributed data. It is assumed in part I that as-if-randomly-assigned "treatments" or covariates $\mathbf{x}(0)$ and $\mathbf{x}(1)$ point identify the marginal distributions $P^j = P(\mathbf{x}(j))$ which in section 4 are considered marginals of potential-outcome joint distributions $\Pr(s(0), s(1))$. $p^j_m = \Pr^j(s=m)$ is useful notational shorthand with $P^j = \langle p^j_0, \ldots, p^k_M \rangle$.

Standard references provide a helpful baseline for the issues considered in part I. Levy, 1998, offers a useful overview of stochastic dominance. Imbens and Wooldridge, 2009, provide a comprehensive review of average treatment effect (ATE) and quantile treatment effect (QTE) methods. Fan and Park, 2010, is a key reference for partial identification of treatment-effect distributions.[10]

2. Evaluation Based on Dominance Criteria

Relevant throughout this discussion is the recognition that BCOs have discrete and finite largest (s=M) and smallest (s=0) values. Whether 0 or M is the "best" outcome will depend on context. With this in mind three dominance criteria[11] are considered here:

(a) First-Order Dominance (FD): P^1 FD P^0 if $C^0(v) \geq C^1(v)$ for all $v \in \{0, \ldots, M-1\}$ with at least one strict inequality. FD is strong if all M inequalities are strict.

(b) w-Tail Dominance (TD_w): For integer $w < .5M$ P^1 TD_w P^0 if (i) $p^0_m \geq p^1_m$ with at least one inequality strict for $m \in \{0, \ldots, w\}$ and (ii) $p^0_m \leq p^1_m$ with at least one inequality strict

[10] Partial identification of BCO probability and treatment effect distributions in the presence of misreporting and measurement error (Millimet, 2011; Molinari, 2008) is considered in Appendix 1.

[11] Foster and Shorrocks, 1988, is a classic reference on discrete-distribution dominance.

for $m \in \{M-w, \ldots, M\}$. TD_w is strong if all 2w+2 inequalities are strict. P^1 FD P^0 implies P^1 TD_0 P^0 but does not imply P^1 TD_w P^0 for w>0.

(c) Spread Dominance (SD): P^1 SD P^0 means that the signs of the sequence of differences $\{p_m^1 - p_m^0\}$ change (ignoring ties) exactly twice as m increases on V; that is:

$$\left\{1\left(p_0^1 > p_0^0\right), \ldots, 1\left(p_M^1 > p_M^0\right)\right\} = \left\{\underbrace{1, \ldots, 1}_{M_1 \text{ terms}}, \underbrace{0, \ldots, 0}_{M_2 \text{ terms}}, \underbrace{1, \ldots, 1}_{M_3 \text{ terms}}\right\},$$

where $1 \le M_j \le M-2$ and $M_1 + M_2 + M_3 = M$. SD is a form of dispersive ordering or variability ordering (see Navard et al., 1993, Shaked, 1982, and Whitt, 1985) and is sometimes called a two-crossings property.

The definition of FD is standard. For P^1 FD P^0 to hold with BCOs two necessary conditions are that $p_0^0 \ge p_0^1$ and $p_M^0 \le p_M^1$. TD_w provides a weaker alternative to FD for formalizing the notion that P^1 is more likely to have large values and less likely to have small values than P^0. SD is an intuitive way to formalize the idea that one BCO distribution has greater spread or heavier tails than a comparison distribution. Note that SD rules out both FD and TD_w, and vice versa. See Appendix 2 for further discussion of dominance relationships with BCOs.

In one important respect BCOs' bounded character facilitates consideration of FD compared to outcomes having unbounded support, whether the latter are continuously or count-measured. With $s \in V$ it is obvious that there are finite discrete smallest and largest outcomes in the population. This does not hold for an outcome z having unbounded population support $V_U = [0, \infty)$ or $V_U = \{0, 1, \ldots\}$. If P^1 FD P^0 is defined as $Pr^0(z \le v) - Pr^1(z \le v) \ge 0$ for all values v in in the distribution support V_U then regardless of sample size it will never be possible to ascertain whether the difference $Pr^0(z \le v) - Pr^1(z \le v)$ changes sign on V_U even when its sign may not vary on the common sample support of z. This impossibility has motivated consideration of various forms of restricted stochastic dominance (Davidson and Duclos, 2013; Osuna, 2013). With BCOs such problems do not arise since the $Pr^j(s \le m)$ are point identified for all m in the bounded population support V.

Note that ratio-scale measurement of s is not required for coherent evaluations based on dominance criteria.

3. Average and Quantile Treatment Effects with BCOs

A familiar measure of the difference between two marginal distributions P^0 and P^1 is the average treatment effect $ATE = E_{P^1}(s) - E_{P^0}(s)$ where $E_{P^j}(s)$ is the mean of P^j, defined in (6). Beyond standard considerations there is nothing about BCOs requiring special treatment in the definition of ATEs although recognizing that the ATE is bounded in $(-M, M)$ may be useful for some purposes.

Another common measure of the difference between P^0 and P^1 is a quantile treatment effect $QTE(q) = Q_{P^1}(q) - Q_{P^0}(q)$ where $Q_{P^j}(q)$ is the q-th quantile of P^j, defined in (7). $QTE(q)$ is a ΔD treatment effect in the terminology of Manski, 1997, i.e. it is a contrast between features of two marginal distributions. Commonly encountered in practice are median treatment effects, $QTE(.5)$. Again there is nothing about BCOs requiring special treatment in defining $QTE(q)$ apart perhaps from acknowledging its integer-valued nature, $QTE(q) \in \{-M, \ldots, M\}$.

However for some purposes it may be useful to appreciate that owing to its integer-valued nature, and particularly with "small" M, there may be a non-trivial probability that $QTE(q) = 0$, i.e. the q-th quantile of both P^0 and P^1 is the same value in V. Studies whose evaluation criteria are specified in pre-analysis protocols may find it useful to consider ex ante the prospects of such "ties" and perhaps appeal to alternative or broader criteria for assessing the extent to which P^0 and P^1 differ. Dominance criteria are one obvious alternative. Another alternative is to report a set of QTEs instead of a single QTE, e.g. $\{QTE(.25), QTE(.5), QTE(.75)\}$ (see Imbens and Wooldridge, 2009).

Finally note that ratio-scale measurement of s is necessary for coherent interpretation of an ATE. Whether it is necessary for coherent interpretation of a QTE is debatable: Since quantiles are equivariant under increasing monotone transformations $t(\bullet)$ then $t\left(Q_{P^1}(q)\right) - t\left(Q_{P^0}(q)\right)$ will always have the same sign as $Q_{P^1}(q) - Q_{P^0}(q)$ even though the magnitudes of the two QTEs will depend on $t(\bullet)$.

4. Partial Identification of BCO Treatment Effect Distributions

This section considers s as the observed counterpart of count-valued potential outcomes $s(0)$ and $s(1)$ with each $s(\bullet) \in V$, i.e. for AIRA treatment $\mathbf{x}$, $s = 1(\mathbf{x} = \mathbf{x}(0)) \cdot s(0) + 1(\mathbf{x} = \mathbf{x}(1)) \cdot s(1)$. The potential outcomes' joint distribution $\Pr(s(0) = i,\ s(1) = j) = p_{ij}$ has support $V \times V$. The unobservable treatment effect (TE) is $\Delta = s(1) - s(0)$. The difference $s(1) - s(0)$ has coherent meaning if the $s(\bullet)$ have

ratio scale measures but it risks being meaningless if the $s(\bullet)$ are arbitrarily-labeled ordinal-scale outcomes.

Fan and Park, 2010, (FP) provide partial identification (PI) results for continuous outcomes' TE distributions. For BCOs a subtlety that can be ignored for continuous outcomes must be addressed. Specifically, defining best bounds on BCOs' TE distributions requires distinguishing strict from weak inequalities since for integers v $\Pr(z \le v) = 1 - \Pr(z \ge v)$ holds for continuous but not count-valued z.[12]

The support of the TE cumulative distribution $F_\Delta(\delta) = \Pr(\Delta \le \delta)$ is $T = \{-M, \ldots, M\}$. As is standard $F_\Delta(\delta)$ is defined based on a weak inequality with respect to its support points in T. To illustrate, for M=3 figure 2 depicts the sample space $V \times V$, TEs, and joint distribution. The sum of the blue dots' p_{ij} $(0 \le j \le i \le M)$ equals $\Pr(\Delta \le \delta)$ for $\delta = 0$.

The goal is to identify $\Pr(\Delta \le \delta)$ for all $\delta \in T$. While $\Pr(\Delta \le \delta)$ cannot generally be point identified—though $\Pr(\Delta \le M) = 1$ is trivially point identified—informative partial identification (PI) may be possible. Like FP, assume the marginal distributions $\Pr(s(\bullet))$ are point identified. PI of $\Pr(\Delta \le \delta)$ follows from derivation of its Boole-Fréchet lower and upper bounds (LB, UB) that are based on the identified $\Pr(s(\bullet))$. Thus, for each $\delta \in T$ the greatest identifiable LB on $\Pr(\Delta \le \delta)$ can be shown to be

$$LB^*(\Pr(\Delta \le \delta)) = \max_{t \in V} \max\{0,\ \Pr(s(1) \le t) - \Pr(s(0) + \delta < t)\}, \tag{14}$$

while the smallest identifiable UB on $\Pr(\Delta \le \delta)$ is

$$UB^*(\Pr(\Delta \le \delta)) = 1 - \max_{t \in V} \max\{0,\ \Pr(s(0) \le t) - \Pr(s(1) \le t + \delta)\} \tag{15}$$

$$= 1 - \max_{t \in V} \max\{0,\ \Pr(s(0) < t) - \Pr(s(1) < t + \delta)\}. \tag{16}$$

Proofs are presented in appendix 3, where the equivalence of (15) and (16) is shown (see Mullahy, 2023a, for further details).

As an illustrative example, subjects' interview dates in the HSE survey may plausibly be treated

[12] Machado and Santos Silva, 2005, and Manski, 2007 (page 40), raise related considerations in the context of quantile TEs.

AIRA. To gauge the "effect" of season on weekly exercise frequency assume $s(1)$ is observed if the interview took place in the second or third quarter of 2015 (spring and summer, N=3,167) and $s(0)$ is observed otherwise (winter and fall, N=3,602). The PI exercise is to compute the best bounds on $\Pr(\Delta \leq \delta)$ over $T = \{-7, \ldots, 7\}$. These bounds are shown in columns 2 and 3 of table 1.[13]

Part II - Estimation of BCO Regression Models

5. Overview

There is longstanding familiarity in applied microeconometrics with regression models[14] of *unbounded* count-valued outcomes' probability and moment structures (Cameron and Trivedi, 2013) and, perhaps slightly less so, with bounded *continuous* outcomes (Papke and Wooldridge, 1996). Less attention has been devoted to their intersection, BCOs, and any attention has been largely ad hoc rather than systematic. Part II of this paper is an attempt to fill some of this gap.

Because of their widespread use in applied work the focus here is on parametric specifications of conditional probabilities and moments. This by no means suggests that nonparametric specifications of $P(\mathbf{x})$ and $\mu^r(\mathbf{x})$ are uninteresting, especially since it will be noted at several junctures in what follows that these may have various attractive properties.

In health economics applications involving BCOs various parametric estimation strategies have been used with comparably varying focus on particular parameters. In addition to linear regression models of BCOs (e.g. Schwartz et al., 2018) researchers have considered unbounded count models (Poisson, negative binomial, and zero-inflated versions thereof, e.g. Anand, 2014, Holliday et al., 2010, Khan et al., 2008, McGinley et al., 2015), ordered regression models (ordered logit, ordered probit, e.g. Atlas and Skinner, 2010, Busch et al., 2014), multivariate probit (e.g. Conigliani et al., 2015), and other strategies. Dichotomization of s based on some threshold, resulting in binary composite outcome measures, is also a popular strategy in applied research (e.g. Fleishman et al., 2014, Geronimus et al., 2006). Applications of beta-binomial models in health economics are scant. Perhaps best known is Liu et al., 2013, who consider beta-binomial models of healthcare utilization distributions. (Heckman and Willis, 1977, is a well-known application of beta-binomial models in labor economics.) Beta-binomial probability models are discussed extensively in section 7.

---

[13] It may sometimes be possible to obtain tighter bounds by making (unverifiable) assumptions about features of the joint distribution $\Pr(s(0), s(1))$. See Mullahy, 2023a.

[14] This paper uses "regression" in the same sense as Manski, 1991: "A regression of y on x is any feature of the probability distribution of y conditional on x, considered as a function of x."

Heaps, Holes, Clumps, and Spills

Four features that may characterize some BCO data merit consideration before turning attention to specification of moment and probability models in sections 6 and 7.

Heaps

Heaping is a familiar phenomenon with count-valued outcome data (Cummings et al., 2015). Informally heaps might be conceived as local modes in an empirical distribution or as forms of empirical excess in some subset of V relative to some maintained marginal or conditional probability model. For instance, in figure 1(a) one might be inclined to report a heap at s=2 whereas in figure 1(c) heaps at $s \in \{5, 10, 15, 20, 25\}$ might be claimed. Without further assumptions or structure, however, the definition of a heap is to some degree subjective. For instance, whether a (local) mode of $\Pr(s)$ occurring at 0 or M constitutes a heap is unclear (and recalls discussion in the literature of so-called "excess zeros" in count data models) (see Roberts and Brewer, 2001).

One prominent reason ostensible heaps may arise in the empirical distribution of s is due to reporting errors or rounding (Bar and Lillard, 2012; Giustinelli et al., 2022; Pudney, 2008). For instance it might be reasonable to suggest that the pattern of heaps seen in figure 1(c) arises from a nontrivial fraction of respondents rounding their survey responses, due perhaps to imperfect recall.

A separate but perhaps less-widely appreciated reason heaps may arise owes to the joint dependence structure of $\Pr(\mathbf{y})$. That is, some y_m may tend to cluster together in the sense that some subvectors of $\mathbf{y}$ may be positively quadrant- or orthant-dependent (Denuit and Scaillet, 2004). Respondents who report $y_m = 1$ may be disproportionately inclined to also report $y_{m'} = 1$ and perhaps $y_{m''} = 1$. For instance, in the HSE sample the day having the largest (tetrachoric) correlation with Saturday exercise is Sunday, the days showing the largest correlations with Monday exercise are Wednesday and Friday, and the day showing the largest correlation with Tuesday exercise is Thursday. In each of these three cases the two or three binary outcomes are strongly positive quadrant/orthant dependent. These correlations suggest patterns of weekend exercising, three-weekdays-a-week exercising, and two-weekdays-a-week exercising that may give rise to the observed heap at s=2.[15,16]

[15] Figure 3 displays histograms of s arising from underlying $\mathbf{y}$ that in turn are binary indicators of $\text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ variates crossing zero thresholds (i.e. $\Pr(\mathbf{y})$ is multivariate probit). Panel (a) of figure 3 has M=8 and specifies the 1×8 mean vector $\boldsymbol{\mu}$ to have all elements equal to –.5 and the 8×8 block-diagonal covariance=correlation matrix $\boldsymbol{\Sigma}$ to contain four identical 2×2 blocks $\begin{bmatrix} 1 & .9 \\ .9 & 1 \end{bmatrix}$. It is evident that these (cont.)

Holes

Let the support of the sample distribution of s be denoted $V_N \subseteq V$; that is, V_N contains those elements of V having positive sample frequency. Holes in the sample distribution are present when $\#V_N < M+1$, that is, one or more values in the theoretical support are not observed in the sample. Empirical holes may occur at any value of s in V, including 0 and M.

Such holes will often be small-sample phenomena. For example in a subsample of females aged 18-19 (N=4,550) from the BRFSS sample (see figure 1(c)) holes are seen in the distribution of healthy days at $s \in \{1, 4, 6, 11\}$. With theoretical support V and with a sufficiently large sample one would in most instances expect to observe all M+1 values in V with some positive frequency.

In section 7 it will be shown that sample holes pose no analytical issues for some parametric models of $\Pr(s = m|\mathbf{x})$ as these models essentially "smooth" estimated distributions over the holes. For other parametric probability models, however, sample holes are problematic as such models are not designed to do such smoothing. (Note that a nonparametric or analogy estimate of the marginal or conditional probability that s takes on a hole value will simply and intuitively be zero.)

Clumps

In some cases data that naturally have BCO structures will be available to researchers only in coarsened or grouped form, owing either to the nature of survey instruments or to post-survey recoding. That is, instead of observing the counts themselves researchers have access only to data on clumps of counts. For example, the U.S. Youth Risk Behavior Survey asks of respondents: *During the past 30 days, on how many days did you (use an electronic vapor product/smoke cigarettes/have at least one drink of alcohol/etc.)?* Possible responses are: 0; 1 or 2; 3 to 5; 6 to 9; 10 to 19; 20 to 29; and All 30.

Suppose $s \in V$ is measured coarsely into G<M+1 groups or categories,

---

(cont.)
large correlations give rise to heaps at 0, 2, and 4. Panel (b) of figure 3 sets M=9 and specifies the 1×9 mean vector $\boldsymbol{\mu}$ to have all elements equal to –.5 and the 9×9 block-diagonal covariance=correlation matrix $\boldsymbol{\Sigma}$ to contain three identical 3×3 blocks $\begin{bmatrix} 1 & .9 & .9 \\ .9 & 1 & .9 \\ .9 & .9 & 1 \end{bmatrix}$. These large correlations give rise to heaps observed at 0, 3, and 6.

[16] Note that heaps in $\Pr(s)$ and $\Pr(s|\mathbf{x})$ are different phenomena.

$$
\begin{aligned}
g_1 &= \langle 0,\ \ldots,\ m_1 \rangle, \\
g_2 &= \langle m_1 + 1,\ \ldots,\ m_2 \rangle, \\
&\ldots \\
g_G &= \langle m_{G-1} + 1,\ \ldots,\ M \rangle,
\end{aligned}
\tag{17}
$$

where each g_j is a tuple of consecutive integers (though some g_j may be singletons), $\bigcup_{j=1}^{G} g_j = V$, and $g_j \cap g_{j'} = \varnothing$ for all $j \neq j'$. For instance the seven YRBS groups are $g_1 = \langle 0 \rangle$, $g_2 = \langle 1,\ 2 \rangle$, … $g_7 = \langle 30 \rangle$.

The implications of such coarsened BCO data for identification and estimation of conditional moments and probabilities will be discussed in sections 6 and 7. To preview: (a) the conditional group probabilities $\Pr(s \in g_j | \mathbf{x}) = \sum_{m \in g_j} \Pr(s = m | \mathbf{x})$ will generally be point identified, parametrically and nonparametrically; (b) the conditional moments $E(s^r | \mathbf{x})$ will not generally be point identified but can in general be informatively bounded either nonparametrically or parametrically; and (c) conditional probabilities $\Pr(s = m | \mathbf{x})$ will not generally be nonparametrically point identified but will for some, although not all, parametric probability models and some groupings be point identified.

Spills

Unlike heaps, holes, and clumps there is no excuse for spills. Spills can be conceived in two distinct ways. The first is that a sample contains values of s outside its theoretical support V (e.g. a respondent reports eight as the number of exercise days in the previous week). This form of misreporting or miscoding may be of interest in some settings but is beyond this paper's scope. The second type of spill is seen sometimes in applied research and can be problematic for different reasons. This type of spill arises when a maintained parametric model of P admits positive probability for values of m outside of V, most notably for m>M. For example P may be specified as a Poisson-type distribution that assigns positive probability to all nonnegative integers even when values of s must by definition reside in V.

One obvious reason this is problematic is that estimates will satisfy $\sum_{m \in V} \widehat{\Pr}(s = m | \mathbf{x}) < 1$ and thus implied estimated conditional means $\sum_{m \in V} m \cdot \widehat{\Pr}(s = m | \mathbf{x})$ will diverge from the parametric model's estimated conditional mean (e.g. $\exp(\mathbf{x}\hat{\boldsymbol{\beta}})$ for a Poisson distribution). While Poisson QML

estimation, for example, guarantees that the sample average estimate $\frac{1}{N}\sum_{n=1}^{N}\exp\left(\mathbf{x}_n\hat{\boldsymbol{\beta}}\right)$ will be in $\left(0,\ M\right)$ if $\mathbf{x}$ contains a constant and all s are in V, for some values of $\mathbf{x}$ in the sample $\exp\left(\mathbf{x}\hat{\boldsymbol{\beta}}\right)$ may exceed M (a phenomenon akin to the possibility that estimated linear probability models' predicted values may be outside $\left(0,1\right)$).[17] Whether such issues are worrisome in any particular setting will depend on the questions at hand. Yet spillage concerns are fully dismissible by judicious specification of $\Pr\left(s=m\middle|\mathbf{x}\right)$.

6. BCO Conditional Moment Models

Conditional means and conditional variances or other higher-order moments are sometimes a consideration in empirical work. The discussion in this section will be relatively brief since for the most part BCOs present few novel specification and estimation challenges that have not already been well resolved for bounded continuously distributed outcomes. While it might be more logical to discuss BCO conditional probability models (as in section 7) before a discussion of their conditional moment structures, that discussion involves many more novel considerations and is thus best deferred until then.

To keep notation from becoming unwieldy the various parametric probability and moment models considered here will use generic notation for key parameters. α, β, γ, and ξ will suffice to distinguish special cases. Notation like $\boldsymbol{\beta}=\left[\beta_0,\ \boldsymbol{\beta}_v^{\top}\right]^{\top}$ will denote the parameters multiplying $\mathbf{x}=\left[1,\ \mathbf{x}_v\right]$. Parameters that vary over outcomes $m\in V$ are subscripted by m (e.g. $\boldsymbol{\beta}_m$).

Conditional Moments

With outcomes $s\in V$ the r-th raw conditional moment of P is $\mu^r(\mathbf{x})=\sum_{m\in V}m^r\cdot\Pr\left(s=m\middle|\mathbf{x}\right)$

[17] A simple Stata example demonstrates such spillage (M is the largest value of the simulated outcome):

```
set seed 234
set obs 1000
qui gen s=runiform()>.9
qui gen x=s*20*runiform()+(1-s)*runiform()
qui gen y=rpoisson(.5+.5*x)
qui sum y
loc maxy=r(max)
poisson y x, vce(robust)
qui predict py
qui sum py
loc maxpy=r(max)
di _n "Max observed y = " `maxy' _n "Max predicted y = " `maxpy'
```

which will be finite in $\left[0, M^r\right]$ for any finite r>0. In particular the conditional mean satisfies $\mu(\mathbf{x}) \in \left[0, M\right]$ and in all but trivial cases will satisfy $\mu(\mathbf{x}) \in \left(0, M\right)$.

Doubly bounded continuously distributed outcomes residing in $\left[0, 1\right]$ and their corresponding conditional means residing in $\left(0, 1\right)$ motivated the parametric fractional regression models developed by Papke and Wooldridge (PW), 1996. With straightforward modifications the PW framework is directly applicable to BCO models where M is any finite positive integer. If parametric estimation of $\mu(\mathbf{x})$ is of interest it may thus be reasonable to specify $\mu(\mathbf{x}) = M \cdot F(\mathbf{x}\boldsymbol{\alpha})$ where $F(\bullet)$ is some continuous cumulative distribution function like a normal or logistic as in PW. Estimation of parameters and partial effects is straightforward and estimated conditional means satisfy outcomes' natural bounds, i.e. $\hat{\mu}(\mathbf{x}) = M \cdot F(\mathbf{x}\hat{\boldsymbol{\alpha}}) \in \left(0, M\right)$. No additional considerations arise due to BCOs' count-valued structure.[18,19]

With coarsened BCO data (17) the conditional moments $E\left(s^r \middle| \mathbf{x}\right)$ are generally informatively bounded (extending an argument in Manski, 1989):

---

[18] Indeed, with a logistic specification for $F(\bullet)$ the implied fractional regression conditional mean $M \cdot \exp(\mathbf{x}\boldsymbol{\alpha}) / \left(1 + \exp(\mathbf{x}\boldsymbol{\alpha})\right)$ is precisely that implied from specifying what in many contexts would be the canonical conditional probability model with data on M trials, binomial$\left(M, p(\mathbf{x})\right)$, as well as variants thereon like beta-binomial models (see section 7). See Wooldridge, 2010 (p. 739), for discussion of conditional mean estimation via QML under binomial assumptions.

[19] When data on $\mathbf{y}$ are available one may specify and estimate models of $E\left[s^r \middle| \mathbf{x}\right]$ that respect the underlying dependence structure of $\mathbf{y}$. Suppose latent $\mathbf{y}^* = \left[y_1^*, \ldots, y_M^*\right] \sim \text{MVN}\left(\mathbf{x}\boldsymbol{\Gamma}, \boldsymbol{\Sigma}\right)$ and that observed binary outcomes $\mathbf{y} = \left[y_1, \ldots, y_M\right]$ arise from threshold-crossings of $\left[y_1^*, \ldots, y_M^*\right]$, i.e. $\Pr\left(\mathbf{y} \middle| \mathbf{x}\right)$ has a multivariate probit structure with marginals $\Pr\left(y_m = 1 \middle| \mathbf{x}\right) = \Phi\left(\mathbf{x}\boldsymbol{\gamma}_m\right)$ for each m (Mullahy, 2016). With $E\left(y_m \middle| \mathbf{x}\right) = \Phi\left(\mathbf{x}\boldsymbol{\gamma}_m\right)$ then $E\left(s \middle| \mathbf{x}\right) = \sum_{m=1}^{M} \Phi\left(\mathbf{x}\boldsymbol{\gamma}_m\right)$ (note that for the first moment only the marginal distributions of $\Pr\left(\mathbf{y} \middle| \mathbf{x}\right)$ feature). In this case fractional probit is an approximation to the sum-of-probits specification, i.e. taking a first-order expansion of $E\left(s \middle| \mathbf{x}\right)$ around $\boldsymbol{\gamma}_m = \boldsymbol{\gamma}$ gives

$$\sum_{m=1}^{M} \Phi\left(\mathbf{x}\boldsymbol{\gamma}_m\right) \approx \sum_{m=1}^{M} \left(\Phi\left(\mathbf{x}\boldsymbol{\gamma}\right) + \phi\left(\mathbf{x}\boldsymbol{\gamma}\right)^{\top} \mathbf{x}\left(\boldsymbol{\gamma}_m - \boldsymbol{\gamma}\right)\right) = M \cdot \Phi\left(\mathbf{x}\boldsymbol{\gamma}\right) + \text{remainder}$$

The term $M \cdot \Phi\left(\mathbf{x}\boldsymbol{\gamma}\right)$ is the generalized PW fractional regression conditional mean with probit link.

$$\mathrm{LB}\left(\mathrm{E}\left(s^{r}\middle|\mathbf{x}\right)\right)=\sum_{j=1}^{G}\min\left(g_{j}\right)^{r}\cdot\Pr\left(s\in g_{j}\middle|\mathbf{x}\right) \tag{18}$$

and

$$\mathrm{UB}\left(\mathrm{E}\left(s\middle|\mathbf{x}\right)\right)=\sum_{j=1}^{G}\max\left(g_{j}\right)\cdot\Pr\left(s\in g_{j}\middle|\mathbf{x}\right). \tag{19}$$

Conditional Variances

Estimation of higher-order conditional moments may sometimes be of interest. Given BCO data all positive raw moments are bounded, $\mu^{r}(\mathbf{x})\in\left(0,\,M^{r}\right)$ so PW fractional regression specifications $\mu^{r}(\mathbf{x})=M^{r}\cdot F\left(\mathbf{x}\boldsymbol{\alpha}^{r}\right)$ may provide reasonable models of raw conditional moments of any order.

Consider specification and estimation of models of conditional variances, which may be of interest in welfare assessments or for other purposes. For any P the conditional variance is bounded, $\sigma^{2}\left(\mathbf{x}\right)=\mathrm{Var}\left(s\middle|\mathbf{x}\right)\in\left(0,\,.25M^{2}\right)$: the upper bound arises since the maximum-variance distribution assigns .5 probability mass to 0 and M and zero mass to all other s in V; the lower bound arises when all probability mass is concentrated on a single value of s in V. Invoking the same bounded-moment idea used to motivate fractional conditional mean models one might consider parametric estimation of a fractional regression model of $\sigma^{2}(\mathbf{x})$ using a GMM approach with moment equations:

$$\left\{\begin{array}{l}\sum_{n=1}^{N}\mathbf{x}_{n}^{\top}\left(s_{n}-M\cdot F\left(\mathbf{x}_{n}\boldsymbol{\alpha}\right)\right)\\ \sum_{n=1}^{N}\mathbf{x}_{n}^{\top}\left(\left(s_{n}-M\cdot F\left(\mathbf{x}_{n}\boldsymbol{\alpha}\right)\right)^{2}-.25M^{2}\cdot G\left(\mathbf{x}_{n}\boldsymbol{\beta}\right)\right)\end{array}\right\}\;=\;\mathbf{0}_{2(K+1)\times1} \tag{20}$$

where $G(\bullet)$ is some continuous distribution function. This specification enforces the dual requirements that $\mu(\mathbf{x})\in\left(0,\,M\right)$ and $\sigma^{2}(\mathbf{x})=.25M^{2}\cdot G\left(\mathbf{x}\boldsymbol{\beta}\right)\in\left(0,\,.25M^{2}\right)$. Without compelling reason to specify that F and G are different families of distributions using the same family of distribution functions to describe both the conditional mean and variance is reasonable:

$$\left\{\begin{array}{l}\sum_{n=1}^{N}\mathbf{x}_{n}^{\top}\left(s_{n}-M\cdot F\left(\mathbf{x}_{n}\boldsymbol{\alpha}\right)\right)\\ \sum_{n=1}^{N}\mathbf{x}_{n}^{\top}\left(\left(s_{n}-M\cdot F\left(\mathbf{x}_{n}\boldsymbol{\alpha}\right)\right)^{2}-.25M^{2}\cdot F\left(\mathbf{x}_{n}\boldsymbol{\beta}\right)\right)\end{array}\right\}\;=\;\mathbf{0}_{2(K+1)\times1} \tag{21}$$

with $\mu(\mathbf{x}) = M \cdot F(\mathbf{x}\boldsymbol{\alpha})$ and $\sigma^2(\mathbf{x}) = .25M^2 \cdot F(\mathbf{x}\boldsymbol{\beta})$. Alternatively one might specify the second moment equation as

$$\left\{ \begin{array}{l} \sum_{n=1}^{N} \mathbf{x}_n^\top \left(s_n - M \cdot F(\mathbf{x}_n\boldsymbol{\alpha}) \right) \\ \sum_{n=1}^{N} \mathbf{x}_n^\top \left(s_n^2 - M^2 \cdot F(\mathbf{x}_n\boldsymbol{\alpha})^2 - .25M^2 \cdot F(\mathbf{x}_n\boldsymbol{\beta}) \right) \end{array} \right\} = \mathbf{0}_{2(K+1)\times 1} \tag{22}$$

For either specification GMM estimation of the mean and variance parameters $[\boldsymbol{\alpha}, \boldsymbol{\beta}]$ is straightforward using, e.g., Stata's `gmm` command. See appendix 5 for details.[20]

Empirical Examples

Table 2 summarizes the point estimates of the conditional moment models' APEs for the HSE sample. The first four columns of results compare the APE estimates from nonparametric regression (using Stata's `npregress kernel` command), OLS regression, PW fractional logit regression, and the conditional mean (also a fractional logit) from ML estimation of the BBIN-2 beta-binomial specification described in section 7. The APE estimates are quite similar in magnitude and sign (especially linear and fractional logit) although there are a few discrepancies. The fifth and six columns of table 2 present the

[20] While 0 and $.25M^2$ globally bound $\sigma^2(\mathbf{x})$ for a given M, $\sigma^2(\mathbf{x})$ is bounded further by functions of $\mu(\mathbf{x})$. Specifically, letting $\lfloor \bullet \rfloor$ and $\lceil \bullet \rceil$ denote floor (largest integer not larger than •) and ceiling (smallest integer not smaller than •) the smallest possible $\sigma^2(\mathbf{x})$ for a given $\mu(\mathbf{x})$ arises from assigning probability $1-\lambda$ to $s = \lfloor \mu(\mathbf{x}) \rfloor$, probability λ to the adjacent or same value of $s = \lceil \mu(\mathbf{x}) \rceil$, and probability 0 to all other values in V, where $\lambda = \text{mod}(\mu(\mathbf{x}), 1) \in [0, 1)$ with $\text{mod}(u, v) = u - v \cdot \lfloor u/v \rfloor$. Thus $\lambda \cdot (1-\lambda) \cdot \left(\lfloor \mu(\mathbf{x}) \rfloor^2 + \lceil \mu(\mathbf{x}) \rceil^2 - 2\lfloor \mu(\mathbf{x}) \rfloor \cdot \lceil \mu(\mathbf{x}) \rceil \right)$ is a lower bound on $\sigma^2(\mathbf{x})$, equal to zero if $\mu(\mathbf{x})$ happens to be an integer. Correspondingly the largest possible variance for a given $\mu(\mathbf{x})$ is $\mu(\mathbf{x}) \cdot (M - \mu(\mathbf{x}))$, which arises from assigning probability $\mu(\mathbf{x})/M$ to s=M, probability $1 - (\mu(\mathbf{x})/M)$ to s=0, and probability 0 to all other values of s in V. In any event $\sigma^2(\mathbf{x})$ can never exceed $.25M^2$, which arises if $\mu(\mathbf{x}) = .5M$ with half the probability assigned to 0 and half to M.

For the HSE sample's conditional mean and variance estimates (whose APEs are reported in table 2) each of the sample's N estimated variances $\widehat{\sigma^2}(\mathbf{x}_n) = .25M^2 \cdot F(\mathbf{x}_n\hat{\boldsymbol{\beta}})$ respects these lower and upper bounds with $\hat{\mu}(\mathbf{x}_n) = F(\mathbf{x}_n\hat{\boldsymbol{\alpha}})$. It is unknown whether such respect is numerically guaranteed with this variance function specification or is merely a feature of this particular sample and model specification.

point estimates (obtained using Stata's `gmm` command) of the conditional variance APEs for the two alternative second-moment specifications (21) and (22). There are no appreciable differences between these APE estimates.

7. BCO Conditional Probability Models

If interest is confined to P's moments then specifying and estimating P simply to learn about its moments is generally unnecessary and often results in non-robust estimates. But in some cases a researcher's questions concern P itself rather than just its moments. This section explores such questions.

The parametric conditional probability specifications considered here are, in an important sense, ad hoc rather than implied by economic theory (although see Heckman and Willis, 1977). Moreover, one could specify a "structural" model of $\Pr(\mathbf{y}|\mathbf{x})$ as the foundation of a model of $\Pr(s|\mathbf{x})$ since $\Pr(s=m|\mathbf{x}) = \sum_{\mathbf{y}\in Y_m} \Pr(\mathbf{y}|\mathbf{x})$, $m \in V$, the multivariate probit model described in footnote 15 being an obvious choice. However unless $\Pr(\mathbf{y}|\mathbf{x})$ is specifically of interest beyond the presumed interest in $\Pr(s|\mathbf{x})$ then modeling $\Pr(\mathbf{y}|\mathbf{x})$ solely to model $\Pr(s|\mathbf{x})$ will generally be unnecessary and, if estimated parametrically, will typically involve estimation of many parameters (e.g. a multivariate probit model has $M\cdot(K+1)+.5M\cdot(M-1)$ free parameters whereas the *least* parsimonious parametric models for $\Pr(s|\mathbf{x})$ considered below involve either M+K or $2(K+1)$ parameters).

Several features of any candidate parametric model of P may be relevant in applications. The first feature (F1) is a candidate model's flexibility in describing the full empirical probability sequence P ("goodness of fit"). One prominent consideration is how many times the sign of $\Pr(s=m+1|\mathbf{x}) - \Pr(s=m|\mathbf{x})$ can change as m increases on $\langle 0, \ldots, M-1\rangle$; such sign patterns are relevant in describing the mode structure of P as well as the extent to which a candidate model can, for instance, describe heaps in P.

The second feature (F2) is a candidate model's flexibility in describing the sequence of PEs $\left\langle D_k \Pr(s=0|\mathbf{x}), \ldots, D_k \Pr(s=M|\mathbf{x})\right\rangle$, with a prominent consideration being how many times a model allows the sign of each $D_k \Pr(s=m|\mathbf{x})$ to change as m increases on V. For instance, a candidate model must allow at least two PE sign changes for a change in **x** to result in a change in spread as defined in section 2. Note that such sign change properties apply to the conditional PEs, $D_k \Pr(s=m|\mathbf{x})$, but not necessarily to their corresponding estimated sample APEs, $\frac{1}{N}\sum_{n=1}^{N} D_k \widehat{\Pr}(s=m|\mathbf{x}_n)$, the latter being the

quantities reported by Stata's `margins, dydx(*)` command (see appendix 4 for discussion).

Nonparametric models perform well on both F1 and F2 but at the cost of greater computational complexity and challenges in applying estimated models beyond the estimation sample. How the various parametric models perform will be considered below.

This section discusses three classes of parametric probability models that may be of interest in applied work: binomial models; beta-binomial models (several versions); and ordered regression models like ordered probit and ordered logit. The discussion here does not include models for $\Pr(s = m|\mathbf{x})$ derived from underlying models for $\Pr(\mathbf{y}|\mathbf{x})$ like multivariate probit specifications although in some instances when $\mathbf{y}$ is observed such specifications may merit consideration. Details on the models' PE structures, $D_k \Pr(s = m|\mathbf{x})$, are presented in Appendix 4.

Binomial Probability Models

The binomial $(M, p(\mathbf{x}))$ probability model, perhaps the canonical model for BCO data, is

$$\Pr(s = m|\mathbf{x}) = \binom{M}{m} \cdot p(\mathbf{x})^m \cdot (1 - p(\mathbf{x}))^{M-m}, \quad m \in V \tag{23}$$

where $p(\mathbf{x}) \in (0, 1)$ is typically but not necessarily specified as a logit function $\exp(\mathbf{x}\boldsymbol{\xi}) / (\exp(\mathbf{x}\boldsymbol{\xi}) + 1)$. The conditional mean and variance are

$$E(s|\mathbf{x}) = M \cdot \exp(\mathbf{x}\boldsymbol{\xi}) / (\exp(\mathbf{x}\boldsymbol{\xi}) + 1) \tag{24}$$

and

$$\mathrm{Var}(s|\mathbf{x}) = M \cdot \exp(\mathbf{x}\boldsymbol{\xi}) / (\exp(\mathbf{x}\boldsymbol{\xi}) + 1)^2. \tag{25}$$

The PEs $D_k \Pr(s = m|\mathbf{x})$ change sign only once as m increases on V (see Appendix 4).[21]

[21] The nature of s as a sum of the events y_m contraindicates a multinomial probability model for the corresponding outcome vector $[1(s = 0), \ldots, 1(s = M)]$. That said, such a model might be expected to have good "fit" properties (e.g. a multinomial logit specification would involve $K \cdot M$ parameters).

Beta-Binomial Probability Models

A general form of the beta-binomial probability model[22] conditioned on $\mathbf{x}$ is

$$\Pr\left(s=m\middle|\mathbf{x}\right)=\binom{M}{m}\cdot\frac{B\left(a+m,b+M-m\right)}{B\left(a,b\right)},\quad m\in V \tag{26}$$

$$=\binom{M}{m}\cdot\frac{\Gamma\left(a+b\right)\cdot\Gamma\left(a+m\right)\cdot\Gamma\left(b+M-m\right)}{\Gamma\left(a\right)\cdot\Gamma\left(b\right)\cdot\Gamma\left(a+b+M\right)}$$

where $a=a\left(\mathbf{x}\right)$, $b=b\left(\mathbf{x}\right)$, and $B\left(a,b\right)$ is the beta function $\Gamma\left(a\right)\cdot\Gamma\left(b\right)/\Gamma\left(a+b\right)$. One parameterization, referred to as BBIN-2, is $a\left(\mathbf{x}\right)=\exp\left(\mathbf{x}\boldsymbol{\alpha}\right)$ and $b\left(\mathbf{x}\right)=\exp\left(\mathbf{x}\boldsymbol{\beta}\right)$ (e.g. Heckman and Willis, 1977). In general $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ (or a and b) are separately identified. An alternative parameterization, BBIN-1, that restricts $\boldsymbol{\beta}_v=\mathbf{0}$ in $b\left(\mathbf{x}\right)=\exp\left(\beta_0+\mathbf{x}_v\boldsymbol{\beta}_v\right)$ is discussed below as is a distinct parameterization used in Stata's `betabin` procedure.

The conditional mean and variance are

$$E\left(s\middle|\mathbf{x}\right)=M\cdot a\big/\left(a+b\right) \tag{27}$$

and

$$\mathrm{Var}\left(s\middle|\mathbf{x}\right)=\frac{M\cdot a\cdot b\cdot\left(a+b+M\right)}{\left(a+b+1\right)\cdot\left(a+b\right)^2} \tag{28}$$

The beta-binomial probability distribution P can assume many different shapes depending on the values of a and b: U-shape, inverse-U-shape, J-shape, reverse-J-shape, or everywhere increasing or decreasing as m increases on V. Panels (a) and (b) of figure 4 depict P for M=4, a=2, b=3 and for M=10, a=.5, b=.8, respectively. Panels (c) and (d) of figure 4 illustrate the regions in (a,b)-space that result in beta-binomial distributions having various shapes for M=4 and M=10, respectively.

The BBIN-2 PEs $D_k\Pr\left(s=m\middle|\mathbf{x}\right)$ can change sign once or twice as m increases on V while the BBIN-1 PEs change sign only once. For BBIN-2 a necessary but not sufficient condition for $D_k\Pr\left(s=0\middle|\mathbf{x}\right)$ and $D_k\Pr\left(s=M\middle|\mathbf{x}\right)$ to have the same sign with a continuous x_k (i.e. two sign changes)

[22] This is known also as a negative hypergeometric distribution (see Johnson et al., 2005, p. 252).

is that $D_k a(\mathbf{x})$ and $D_k b(\mathbf{x})$ have the same sign.[23] Roughly speaking the closer are $a(\mathbf{x})$ and $b(\mathbf{x})$ and the closer are $D_k a(\mathbf{x})$ and $D_k b(\mathbf{x})$ the more likely that $D_k \Pr(s=0|\mathbf{x})$ and $D_k \Pr(s=M|\mathbf{x})$ have the same signs. See figure 5 for an illustration. Further results on the beta-binomial PEs are in Appendix 4.

The parameterization of the beta-binomial model used in Stata's `betabin` procedure (Hardin and Hilbe, 2014, henceforth HH) differs from the BBIN-2 and BBIN-1 specifications. Using this paper's notation HH begin with the specification (26) but reparameterize as

$$p(\mathbf{x}) = a/(a+b) = \exp(\mathbf{x}\boldsymbol{\xi})/(\exp(\mathbf{x}\boldsymbol{\xi})+1) \tag{29}$$

and

$$\upsilon = 1/\sigma = a+b, \tag{30}$$

yielding the probability model

$$\Pr(s=m|\mathbf{x}) = \binom{M}{m} \cdot \frac{B(\upsilon\cdot p(\mathbf{x})+m,\ \upsilon\cdot(1-p(\mathbf{x}))+M-m)}{B(\upsilon\cdot p(\mathbf{x}),\ \upsilon\cdot(1-p(\mathbf{x})))}, \quad m\in V \tag{31}$$

(There is a typo in the HH paper's definition of σ.) While (31) is a proper probability model HH's parameterization in terms of $(p(\mathbf{x}),\ \upsilon)$ does not allow both parameters to be specified as functions of $\mathbf{x}$ since υ (i.e. σ) is treated as a constant. Moreover in (31) $\mathbf{x}$ is introduced in a manner that does not follow from specifying either a or b in (26) as functions of $\mathbf{x}$.[24] The PEs $D_k \Pr(s=m|\mathbf{x})$ change sign only once as m increases on V (see Appendix 4); these differ from the BBIN-2 and BBIN-1 PEs.

Ordered Probability Models

The probability structure of ordered probability regression models typically assumes a latent outcome s* whose conditional distribution $F(s^*|\mathbf{x})$ has support $(-\infty, \infty)$. $F(\bullet)$ is assumed to be

[23] For example with M=7, $D_k a(\mathbf{x}) = 1$, and $D_k b(\mathbf{x}) = 1.5$, $D_k \Pr(s=0|\mathbf{x})$ and $D_k \Pr(s=M|\mathbf{x})$ have the same signs with $a(\mathbf{x}) = b(\mathbf{x}) = 1$ but have opposite signs with $a(\mathbf{x}) = 2$ and $b(\mathbf{x}) = 1$.

[24] It might also be noted that the HH specification is not the same as the BBIN-1 specification that restricts the BBIN-2 specification's $\beta_v = 0$.

continuous with corresponding unimodal density $f(\bullet) = F'(\bullet)$. The mapping from the unobserved s* to the observed s is given by $s = m \cdot 1\left(s^* \in \left(\kappa_m,\ \kappa_{m+1}\right)\right)$, where

$$\Pr\left(s = m \middle| \mathbf{x}\right) = F\left(\kappa_{m+1} - \mathbf{x}_v\boldsymbol{\beta}_v\right) - F\left(\kappa_m - \mathbf{x}_v\boldsymbol{\beta}_v\right), \quad m \in V, \tag{32}$$

where setting $\kappa_0 = -\infty$, and $\kappa_{M+1} = +\infty$ is standard. $F(\bullet)$ is typically assumed to be normal or extreme value, yielding ordered probit or ordered logit structures for $\Pr\left(s = m \middle| \mathbf{x}\right)$. The conditional mean is

$$E\left(s \middle| \mathbf{x}\right) = M - \sum_{m=1}^{M} F\left(\kappa_m - \mathbf{x}_v\boldsymbol{\beta}_v\right). \tag{33}$$

From (33) it is evident that the PEs $D_k E\left(s \middle| \mathbf{x}\right)$ will have the same sign as β_k. The PEs $D_k \Pr\left(s = m \middle| \mathbf{x}\right)$, discussed in Appendix 4, change sign only once as m increases on V.

The constant terms κ_m typically provide flexibility for the estimated ordered regression models to have good, albeit imperfect, predictive fit. That is, the sample average predictions

$$\tfrac{1}{N}\sum_{n=1}^{N} \widehat{\Pr}\left(s = m \middle| \mathbf{x}\right) = \tfrac{1}{N}\sum_{n=1}^{N} F\left(\hat{\kappa}_{m+1} - \mathbf{x}_{v,n}\hat{\boldsymbol{\beta}}_v\right) - F\left(\hat{\kappa}_m - \mathbf{x}_{v,n}\hat{\boldsymbol{\beta}}_v\right) \tag{34}$$

will not match exactly the sample frequencies $\tfrac{1}{N}\sum_{n=1}^{N} 1\left(s = m\right)$, an intuitive explanation being that the ML estimate of each κ_m is involved in fitting both $\Pr\left(s = m \middle| \mathbf{x}\right)$ and $\Pr\left(s = m - 1 \middle| \mathbf{x}\right)$ and will thus not result in a perfect sample average fit of either.[25]

One drawback of ordered probability models for BCOs is that they are unable to estimate the conditional probability of values in V that do not appear in the sample, i.e. holes, as discussed in section 5.[26] Other parametric probability models that treat s as having ratio-scale measure on the nonnegative

[25] The lack of precise "adding-up" is because ML estimation of an ordered regression model does not involve score equations like $\sum_{n=1}^{N} 1\left(s_n = m\right) - \left(F\left(\hat{\kappa}_{m+1} - \mathbf{x}_{v,n}\hat{\boldsymbol{\beta}}_v\right) - F\left(\hat{\kappa}_m - \mathbf{x}_{v,n}\hat{\boldsymbol{\beta}}_v\right)\right) = 0$.

[26] To appreciate this note that for ordered probability models sample outcomes that take on the five values {0,1,2,4,5} (a "hole" at S=3) are indistinguishable from those taking on the five values {0,1,2,3,4} when (cont.)

integers (e.g. binomial and beta-binomial) respect the theoretical support and will estimate positive conditional probability of the hole's or holes' value(s). Conversely, using analogy principle arguments the ordered regression models will effectively, although not explicitly, assign the absent value zero probability (as would a nonparametric estimator $\frac{1}{\#N(\mathbf{x})}\sum_{n\in N(\mathbf{x})} 1\left(s_n = m\right)$ with $N\left(\mathbf{x}\right)$ being the index set for observations with covariate value **x**). Whether the true $\Pr\left(s = m\middle|\mathbf{x}\right)$ is zero or positive when there is a sample hole at m for a given **x** is, of course, an open question.

Conditional Probability Models with Clumped BCO Data

To appreciate the distinctions between fully observed and coarsened BCO data, note that the sample likelihood when s is fully observed is

$$\prod_{n=1}^{N}\prod_{m\in V}\Pr\left(s = m\middle|\mathbf{x}_n\right)^{1\left(s_n = m\right)} \tag{35}$$

while when s is measured coarsely the likelihood is

$$\prod_{n=1}^{N}\prod_{j=1}^{G}\Pr\left(s \in g_j\middle|\mathbf{x}_n\right)^{1\left(s_n \in g_j\right)} = \prod_{n=1}^{N}\prod_{j=1}^{G}\left(\sum_{m\in g_j}\Pr\left(s = m\middle|\mathbf{x}_n\right)\right)^{1\left(s_n \in g_j\right)}. \tag{36}$$

Unless $g_j = \{m\}$ for some j then $\Pr\left(s = m\middle|\mathbf{x}\right)$ at that value of m will not generally be nonparametrically point identified with coarse measurement. It is also the case that $\Pr\left(s = m\middle|\mathbf{x}\right)$ will not generally be parametrically point identified for parametric probability models whose specifications involve m-specific parameters, most notably for present purposes the κ_m parameters in the ordered regression models (32). For other parametric probability models, however, $\Pr\left(s = m\middle|\mathbf{x}\right)$ may be point identified for all $m \in V$.

(cont.)
the theoretical support of $\Pr\left(s\middle|\mathbf{x}\right)$ is $V = \{0,1,2,3,4,5\}$. That is, the theoretical support is not recognized by ordered probability models since only outcomes' orderings matter. In a formal sense ordered probability models would treat a hole at s=m as if $\kappa_{m+1} = \kappa_m$ (or $\kappa_1 = \kappa_0 = -\infty$ at m=0 and $\kappa_{M+1} = \kappa_M = \infty$ at M=1), so that $\Pr\left(s = m\middle|\mathbf{x}\right) = F\left(\kappa_m - \mathbf{x}_v\boldsymbol{\beta}_v\right) - F\left(\kappa_m - \mathbf{x}_v\boldsymbol{\beta}_v\right) = 0$. But because ordered regression estimation algorithms ignore outcomes' ratio-scale measurement these issues go unrecognized.

Consider the binomial and beta-binomial specifications. Ignoring covariates for the moment, the binomial specification (23) involves only one parameter, p, and with as few as two groups maximization of (36) will point identify $\Pr(s=m)$ for all $m \in V$. Only one value of p can solve the moment equation

$$\tfrac{1}{N}\sum_{n=1}^{N}\left(1(s_n \in g_1) - \sum_{m \in g_1}\binom{M}{m}\cdot p^m \cdot (1-p)^{M-m}\right) = 0, \tag{37}$$

since the subtrahend in (37) is montonic in p. However, the beta-binomial specification involves two parameters, a and b, and in general there are multiple solutions to the moment equation

$$\tfrac{1}{n}\sum_{n=1}^{N}\left(1(s_n \in g_1) - \sum_{m \in g_1}\binom{M}{m}\cdot \frac{\Gamma(a+b)\cdot\Gamma(a+m)\cdot\Gamma(b+M-m)}{\Gamma(a)\cdot\Gamma(b)\cdot\Gamma(a+b+M)}\right) = 0. \tag{38}$$

Thus at least three groups (G≥3) are required for maximization of (36) to point identify $\Pr(s=m)$ for all $m \in V$ through unique solutions for a and b.[27]

Parametric Probability Models and Dominance Considerations

Should one contemplate estimation of a parametric conditional probability model as an ingredient in a conditional stochastic dominance evaluation involving $\Pr(s|\mathbf{x}(0))$ and $\Pr(s|\mathbf{x}(1))$ (recall section 2) then it merits emphasis that distributions like the binomial, BBIN-1, and ordered probability models that allow only one PE sign change as m increases on V will automatically show FD and can never demonstrate SD. Among the parametric probability models considered here only the BBIN-2 specification is sufficiently flexible to be able to rule out FD and accommodate SD. A larger question is whether nonparametric probability models might be a better alternative in such evaluations.

[27] To illustrate, suppose M=3 and consider two sets of beta-binomial parameters $(a_1=1,\ b_1=2)$ and $(a_2=2,\ b_2=3.8)$ corresponding to distributions $P^1=\langle .4,\ .3,\ .2,\ .1\rangle$ and $P^2=\langle .344,\ .356,\ .222,\ .078\rangle$. Consider groupings $g_1=\langle 0,\ 1\rangle$ and $g_2=\langle 2,\ 3\rangle$. The probability distributions for the grouped outcomes are identical, $P^1_g = P^2_g = \langle .7,\ .3\rangle$. Two groupings thus do not suffice to distinguish P^1 from P^2. Intuitively solution of the binomial specification moment equations involves solving one independent equation (i.e. at least two groups) in one unknown while solution of the beta-binomial specification moment equations involves solving two independent equations (i.e. at least three groups) in two unknowns.

Empirical Results

Table 3 presents the sample average probability predictions for the probability models considered above. With reference to column 1 heaps at m=2, m=3, and m=7 might be claimed. Unsurprisingly the M+1 nonparametric and linear probability models provide close fit to the empirical probabilities (the linear probability models' fits are exact). The estimates of the ordered probit specification also provide close fit to the empirical probabilities. Given this data structure the binomial model fit is poor while the three beta-binomial specifications all yield U-shaped patterns of prediction, predicting reasonably well the ostensible "heap" at m=7 but by their nature unable to also predict the "heaps" at m=2 and m=3.

Table 4 presents the point estimates of the sample averages of the PEs $D_k \Pr(s = m|\mathbf{x})$. While for some point estimates there is moderate similarity across the various probability models, the main takeaway message is that the choice of probability model matters a great deal. In particular the estimated APE sign changes as m increases on V are notable. The nonparametric and linear specifications can describe any pattern of PE sign changes. But among the other probability models only the BBIN-2 specification can accommodate more than one sign change as m increases on V (note the point estimates for Female and NVQ45, which suggest spread-reducing effects of changes in x_k).

Summary Properties of Parametric Probability Models

This table summarizes key attributes of the conditional probability models discussed above.

Property	Non-parametric	Binomial	BBIN-2	BBIN-1	Stata's `betabin`	Ordered Regression
Necessary FD?	N	Y	N	Y	Y	Y
Necessary w-TD?	N	Y	N	Y	Y	Y
Possible SD?	Y	N	Y	N	N	N
>1 $D_k \Pr(s = m\|\mathbf{x})$ Sign Change Possible?	Y	N	Y	N	N	N
Point-ID $\Pr(s = m\|\mathbf{x})$ with Heaps?	Y	N	N	N	N	Y
Point-ID $\Pr(s = m\|\mathbf{x})$ with Holes?	N	Y	Y	Y	Y	N
Point-ID $\Pr(s = m\|\mathbf{x})$ with Clumps?	N	Y	Y	Y	Y	N

8. Conclusions

BCO data are encountered often in empirical health economics and other microeconometric research. This paper has attempted to provide a comprehensive assessment of considerations arising in the analysis of BCO data with emphasis on applications. Even so many topics have been treated only cursorily, if at all, with nonparametric probability models and probability models for multivariate BCOs being two of the many topics shortchanged here that deserve detailed assessment elsewhere.

Acknowledgments

Thanks are owed to João Santos Silva, Jon Skinner, and Jeff Wooldridge for helpful comments on an earlier version of this paper. Beyond this paper many colleagues and seminar participants have previously provided helpful comments on various aspects of the ideas presented here. Some of the work was supported by an Evidence for Action grant (#73336) from RWJF to UW-Madison.

References

Alzer, H. and G. Jameson. 2017. "A Harmonic Mean Inequality for the Digamma Function and Related Results." *Rendiconti del Seminario Matematico della Università di Padova* 137: 203-209.

American Academy of Pediatrics and American College of Obstetricians and Gynecologists (AAP/ACOG). 2006. "The Apgar Score." *Pediatrics* 117: 1444-1447.

Atlas, S.J. and J. Skinner. 2010. "Education and the Prevalence of Pain." in D.A. Wise, ed., *Research Findings in the Economics of Aging*. Chicago: University of Chicago Press for NBER.

Bates, D.W. et al. 2023. "The Safety of Inpatient Care." *NEJM* 388: 142-153.

Bar, H.Y. and D.R. Lillard. 2012. "Accounting for Heaping in Retrospectively Reported Event Data—A Mixture-Model Approach." *Statistics in Medicine* 31: 3347-3365.

Bond, T.N., and K. Lang. 2019. "The Sad Truth About Happiness Scales." *Journal of Political Economy* 127: 1629–1640.

Busch, S. et al. 2014. "FDA and ABCs: The Unintended Consequences of Antidepressant Warnings on Human Capital." *Journal of Human Resources* 49: 540-571.

Cameron, A.C. and P.K. Trivedi. 2013. *Regression Analysis of Count Data*. Cambridge: Cambridge University Press.

Cebul, R.D. et al. 2011. "Electronic Health Records and Quality of Diabetes Care." *NEJM* 365: 825-833.

Conigliani, C. et al. 2015. "Prediction of Patient-Reported Outcome Measures via Multivariate Ordered Probit Models." *JRSS-A* 178: 567-591.

Cummings, T.H. et al. 2015. "Modeling Heaped Count Data." *The Stata Journal* 15: 457-479.

Davidson, R. and J.-Y. Duclos. 2013. "Testing for Restricted Stochastic Dominance." *Econometric Reviews* 32: 84-125.

Denuit, M. and O. Scaillet. 2004. "Nonparametric Tests for Positive Quadrant Dependence." *Journal of Financial Econometrics* 2: 422-450.

Fan, Y. and S.S. Park. 2010. "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference." *Econometric Theory* 26: 931-951.

Filmer, D. and K. Scott. 2012. "Assessing Asset Indices." *Demography* 49: 359-392.

Fleishman, J.A. et al. 2014. "Screening for Depression using the PHQ–2: Changes over Time in Conjunction with Mental Health Treatment." AHRQ Working Paper 14002.

Foster, J.E. and A.F. Shorrocks. 1988. "Poverty Orderings and Welfare Dominance." *Social Choice and Welfare* 5: 179-198.

Geronimus, A.T. et al. 2006. "'Weathering' and Age Patterns of Allostatic Load Scores among Blacks and Whites in the United States." *AJPH* 96: 826-833.

Giustinelli, P. et al. 2022. "Tail and Center Rounding of Probabilistic Expectations in the Health and

Retirement Study." *Journal of Econometrics* 231: 265-281.

Gruenwald, T.L. et al. 2006. "Combinations of Biomarkers Predictive of Later Life Mortality." *PNAS* 103: 14158–14163.

Halmos, P.R. 1960. *Naive Set Theory*. Princeton, NJ: D. Van Nostrand Company, Inc.

Hardin, J.W. and J.M. Hilbe. 2014. "Estimation and Testing of Binomial and Beta-Binomial Regression Models with and without Zero Inflation." *Stata Journal* 14: 292-303.

Heckman, J.J. and R.J. Willis. 1977. "A Beta-Logistic Model for the Analysis of Sequential Labor Force Participation by Married Women." *Journal of Political Economy* 85: 27-58.

Hoaglin, D.C. and J.W. Tukey. 2006. "Checking the Shape of Discrete Distributions." Chapter 9 in D.C. Hoaglin et al., Eds. *Exploring Data Tables, Trends, and Shapes*. New York: John Wiley & Sons.

Holliday, K.L. et al. 2010. "Genetic Variation in the Hypothalamic-Pituitary-Adrenal Stress Axis Influences Susceptibility to Musculoskeletal Pain: Results from the EPIFUND Study." *Annals of Rheumatic Diseases* 69: 556-560.

Imbens, G.W. and J.M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47: 5-86.

Johnson, N.L., A.W. Kemp, and S. Kotz. 2005. *Univariate Discrete Distributions,* Third Edition. Hoboken, NJ: Wiley.

Jones, J. et al. 2023. "Selective Inhibition of $Na_v1.8$ with VX-548 for Acute Pain." *NEJM* 389: 393-405.

Khan, N. et al. 2008. "Effect of Prescription Drug Coverage on Health of the Elderly." *Health Services Research* 43: 1576-1597.

Khera, A.V. et al. 2016. "Genetic Risk, Adherence to a Healthy Lifestyle, and Coronary Disease." *NEJM* 375: 2349-2358.

Kroenke, K. et al. 2001. "The PHQ-9: Validity of a Brief Depression Severity Measure." *Journal of General Internal Medicine* 16: 606-613.

Levy, H. 1998. *Stochastic Dominance: Investment Decision Making under Uncertainty.* Norwell, MA: Kluwer.

Liu, C.-F. et al. 2013. "Beta-Binomial Regression and Bimodal Utilization." *Health Services Research* 48: 1769-1778.

Maasoumi, E. and J.S. Racine. 2016. "A Solution to Aggregation and and Application to Multidimensional 'Well-Being' Frontiers." *Journal of Econometrics* 191: 374-383.

Machado, J.A.F. and J.M.C. Santos Silva. 2005. "Quantiles for Counts." *JASA* 100: 1226-1237.

Manski, C.F. 1989. "Anatomy of the Selection Problem." *Journal of Human Resources* 24: 343-360.

Manski, C.F. 1991. "Regression." *Journal of Economic Literature* 29: 34-50.

Manski, C.F. 1997. "Monotone Treatment Response." *Econometrica* 65: 1311-1334.

Manski, C.F. 2007. *Identification for Prediction and Decision*. Cambridge, MA: Harvard University Press.

Marschak, J. 1950. "Rational Behavior, Uncertain Prospects, and Measurable Utility." *Econometrica* 18: 111-141.

McGinley, J.S. et al. 2015. "A Novel Modeling Framework for Ordinal Data Defined by Collapsed Counts." *Statistics in Medicine* 34: 2312-2324.

McKenzie, D.J. 2005. "Measuring Inequality with Asset Indicators." *Journal of Population Economics* 18: 229-260.

Millimet, D.L. 2011. "The Elephant in the Corner: A Cautionary Tale About Measurement Error in Treatment Effects Models." *Advances in Econometrics* 27A: 1-39.

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

Mullahy, J. 2016. "Estimation of Multivariate Probit Models via Bivariate Probit." *Stata Journal* 16: 37-51.

Mullahy, J. 2018b. "Treatment Effects with Multiple Outcomes." NBER w.p. 25307.

Mullahy, J. 2022. "Investigating Health-Related Time Use with Partially Observed Data." *Review of Economics of the Household* 20: 103-121.

Mullahy, J. 2023a. "Partial Identification of Treatment-Effect Distributions with Count-Valued Outcomes." NBER Working Paper 31005.

Mullahy, J. 2023b. "Investigating Health Outcomes Defined by Multiple Chronic Conditions." in *Essays in Honor of Andrew M. Jones.* B. Baltagi and F. Moscone, Eds. Bingley, UK: Emerald Publishing Limited (in press).

Navard, S.E. et al. 1993. "A Characterization of Discrete Unimodality with Applications to Variance Upper Bounds." *Annals of the Institute of Statistical Mathematics* 45: 603-614.

Nolan, B. and C.T. Whelan. 2010. "Using Non-Monetary Deprivation Indicators to Analyze Poverty and Social Exclusion: Lessons from Europe?" *Journal of Policy Analysis and Management* 29: 305-325.

Osuna, E.E. 2013. "Tail-Restricted Stochastic Dominance." *IMA Journal of Management Mathematics* 24: 21-44.

Papke, L.E. and J.M. Wooldridge. 1996. "Econometric Methods for Fractional Response with an Application to 401(K) Plan Participation Rates." *Journal of Applied Econometrics* 11: 619-632.

Papke, L.E. and J.M. Wooldridge. 2008. "Panel Data Methods for Fractional Response Variables with an Application to Test Pass Rates." *Journal of Econometrics* 145: 121-133.

Pogue J. et al. 2012. "Designing and Analyzing Clinical Trials with Composite Outcomes: Consideration

of Possible Treatment Differences between the Individual Outcomes." *PLoS ONE* 7: e34785.

Pudney, S. 2008. "Heaping and Leaping: Survey Response Behaviour and the Dynamics of Self-Reported Consumption Expenditure." Institute for Social & Economic Research Working Paper 2008-09.

Ravallion, M. 2012. "Mashup Indices of Development." *The World Bank Research Observer* 27: 1-32.

Schwartz, A.L. et al. 2018. "Low-Value Service Use in Provider Organizations." *Health Services Research* 53: 87-119.

Roberts, Jr., J.M. and D.D. Brewer. 2001. "Measures and Tests of Heaping in Discrete Quantitative Distributions." *Journal of Applied Statistics* 28: 887-896.

Shaked, M. 1982. "Dispersive Ordering of Distributions." *Journal of Applied Probability* 19: 310-320.

Sims, T. et al. 2008. "Simple Counts of ADL Dependencies Do Not Adequately Reflect Older Adults' Preferences toward States of Functional Impairment." *Journal of Clinical Epidemiology* 61: 1261-1270.

Stevens, S.S. 1946. "On the Theory of Scales of Measurement." *Science* (New Series) 103: 677-680.

U.S. Food and Drug Administration. 2017. "Multiple Endpoints in Clinical Trials—Guidance for Industry." USDHHS/FDA CBER/CDER.

Whitt, W. 1985. "Uniform Conditional Variability Ordering of Probability Distributions." *Journal of Applied Probability* 22: 619-633.

Wooldridge, J.M. 2010. *Econometric Analysis of Cross Section and Panel Data*, 2nd Edition. Cambridge, MA: MIT Press.

Appendix 1: Partial Identification of P with Outcome Reporting Errors

Reporting or other measurement errors involving the outcomes $s \in V$ result in the probability distribution of the observed outcomes, $P = \langle p_0, p_1, \ldots, p_M \rangle$, differing from the probability distribution of the true outcomes s*, $P^* = \langle p_0^*, p_1^*, \ldots, p_M^* \rangle$. Since V is bounded the magnitudes of reporting errors are bounded and this can be exploited to informatively partially identify true probabilities and corresponding partial effects. Since V contains a finite number of values it is possible to consider misreporting probabilities at each value in V. This appendix explores such partial identification and how it can be refined by assumptions on the nature of misreporting.[28]

The main ideas can be illustrated using Marschak probability triangles to depict the case with $V = \langle 0, 1, 2 \rangle$; results for general M≥2 follow. Consider panel (a) in figure A.1. The vector of probabilities governing realizations of the observed and possibly misreported outcomes corresponding to probability distribution P is denoted $\mathbf{p} = [\Pr(s=0), \Pr(s=2)]$. With no assumptions (denoted A0) on the nature of reporting errors the true probabilities $\mathbf{p}^*$ (the graphical representation of P*) can be anywhere in the triangle. Let J*=J*(A•) denote the identified set for $\mathbf{p}^*$ under assumption A•. Thus J*(A0) is the entire Marschak triangle.

Consider now an assumption (A1) that the probability of overreporting a true zero s*=0 exceeds the probability of underreporting. For instance, larger values of s* may correspond to outcomes considered increasingly socially undesirable or stigmatized. Under A1 J* shrinks to the light blue quadrilateral in panel (b). If it is assumed additionally (A2) that the probability of underreporting s*=2 exceeds the probability of overreporting then J* shrinks further to the dark blue quadrilateral in panel (c). If it is also assumed (A3') that the probability of underreporting s*=1 exceeds the probability of overreporting then J* shrinks yet further to the dark blue triangle denoted B in panel (d), whereas if it is assumed (A3") that the probability of overreporting s*=1 exceeds the probability of underreporting then J* shrinks to the dark blue parallelogram denoted C in panel (d).

Symmetric results follow when the misreporting directions are reversed. Underreporting of s*=0 and overreporting of s*=M might arise, for instance, when larger values of s are considered socially desirable. Assume again that M=2. Panel (e) shows J* with no assumptions on misreporting (A0), as in

[28] The assumptions are probabilistic in the sense that assuming (say) underreporting of some particular value s*=m does not imply that all individuals underreport that value but rather that the probability of underreporting exceeds that of overreporting, a more credible assumption than a blanket assumption of no overreporting. For instance, suppose .3 of the population overreports s*=0, .1 underreports s*=0, and .6 accurately reports s*=0. Then $p_0^* = .7 < .9 = p_0$ so on balance there is overreporting, i.e. $p_0^* < p_0$.

figure A.1's panel (a). If it is assumed (A4) that the probability of underreporting s*=0 exceeds the probability of overreporting then J* shrinks to the light blue quadrilateral shown in panel (f). Additionally assuming (A5) that the probability of overreporting s*=2 exceeds the probability of overreporting further shrinks J* to the dark blue quadrilateral in panel (g). Assuming still further (A6') that the probability of underreporting s*=1 exceeds the probability of overreporting shrinks J* to the dark blue triangle denoted B in panel (h), whereas assuming (A6") that the probability of overreporting s*=1 exceeds the probability of underreporting shrinks J* to the dark blue parallelogram denoted C.

For concreteness consider a coarsened version the vigorous-moderate exercise outcome in the HSE sample where the intermediate outcomes $1 \le S \le 6$ are collapsed into a single category. Thus the probability distribution of the observed outcomes is the triple $P = \langle p_0, p_{1...6}, p_7 \rangle = \langle .332, .509, .159 \rangle$. If vigorous-moderate exercise is considered socially desirable it might be credible to assume that s*=0 is on balance underreported while $s^* \in \{1, ..., 6\}$ and s*=7 are on balance overreported. In this instance J*=J*(A6") is shown by the dark blue parallelogram in panel (i) of figure A.1.

For general M≥2, define

$$K_0^+ = \left\{ P^* \middle| p_0^* \ge p_0 \right\}, \qquad K_0^- = \left\{ P^* \middle| p_0^* \le p_0 \right\}$$

and (A1.1)

$$K_m^+ = \left\{ P^* \middle| p_m^* \ge p_m \right\}, \quad K_m^- = \left\{ P^* \middle| p_m^* \le p_m \right\}, \quad m \in \{1, \ldots, M\}$$

Consider two cases. K_0^- and K_0^+ correspond to $J^*(A1)$ and $J^*(A4)$ for any M≥2. The intersections $K_0^- \bigcap \left(\bigcap_{m=1}^M K_m^+ \right)$ and $K_0^+ \bigcap \left(\bigcap_{m=1}^M K_m^- \right)$ generalize J*(A3') and J*(A6") to higher dimensions. For assumptions beyond those considered here J* will be generally be defined by various set intersections.

One might also consider bounds on $E(s^* | \mathbf{x})$. Since s* is bounded the basic results follow from Manski, 1989. To determine $UB\left(E(s^* | \mathbf{x})\right)$ allocate $\max\left\{ p_M^* \middle| p_M^* \in J^*(A\bullet) \right\} = g(M)$ to s*=M, then of the remaining probability allocate $\max\left\{ p_{M-1}^* \middle| p_{M-1}^* \in J^*(A\bullet),\ p_{M-1}^* \le 1 - g(M) \right\} = g(M-1)$ to s*=M–1, etc. until all probability mass is allocated. Using the pseudo probabilities $g(m)$ $UB\left(E(s^* | \mathbf{x})\right)$ is computed as $\sum_{m=0}^M m \cdot g(m)$. Inverting the process with pseudo probabilities $h(m)$ —i.e. allocate $\max\left\{ p_0^* \middle| p_0^* \in J^*(A\bullet) \right\} = h(0)$ to s*=0, allocate $\max\left\{ p_1^* \middle| p_1^* \in J^*(A\bullet),\ p_1^* \le 1 - h(0) \right\} = h(1)$ to s*=1, etc.—

gives $LB\left(E\left(s^*\middle|\mathbf{x}\right)\right)$ as $\sum_{m=0}^{M} m \cdot h(m)$.

Partial Effects

To learn the true PEs

$$D_k \Pr\left(s^* = m\middle|\mathbf{x}\right) = \Pr\left(s^* = m\middle|x_k = 1,\ \mathbf{x}_{-k}\right) - \Pr\left(s^* = m\middle|x_k = 0,\ \mathbf{x}_{-k}\right) \tag{A1.2}$$

in the presence of reporting errors let $P^0 = P\left(x_k = 0,\ \mathbf{x}_{-k}\right)$ and $P^1 = P\left(x_k = 1,\ \mathbf{x}_{-k}\right)$ be the observed probability distributions at the two values of the treatment. Under any of the assumptions A• described above let J_0^* and J_1^* denote the identified sets for the true probabilities $P^*\left(x_k = 0,\ \mathbf{x}_{-k}\right)$ and $P^*\left(x_k = 1,\ \mathbf{x}_{-k}\right)$. Bounds on the partial effects $D_k \Pr\left(s^* = m\middle|\mathbf{x}\right)$ are then defined by: (a) determining the largest value of $p_{m,1}^*$ in J_1^* and subtracting from it the smallest value of $p_{m,0}^*$ in J_0^* (UB); and (b) determining the smallest value of $p_{m,1}^*$ in J_1^* and subtracting from it the largest value of $p_{m,0}^*$ in J_0^* (LB). Bounds on conditional mean partial effects

$$D_k E\left(s^*\middle|\mathbf{x}\right) = E\left(s^*\middle|x_k = 1,\ \mathbf{x}_{-k}\right) - E\left(s^*\middle|x_k = 0,\ \mathbf{x}_{-k}\right) \tag{A1.3}$$

follow using the same logic.

Appendix 2: Further Remarks on Dominance Relationships with BCOs

Figure A.2(a) uses a Marschak probability triangle to illustrate the range of possible dominance relationships when M≥2. The case M=3 suffices to illustrate the key points.[29] Let the vector $\mathbf{p}^0$ in $\left(\Pr(s=0),\ \Pr(s=M)\right)$-space be specified as $\mathbf{p}^0=\left[p_0^0,\ p_3^0\right]=[.5,\ .2]$. The possible FD relationships between P^0 (which is partially but not fully defined by the values of $\mathbf{p}^0$) and possible values of P^1 depend on the location of the corresponding $\mathbf{p}^1=\left[p_0^1,\ p_3^1\right]$ relative to $\mathbf{p}^0$. P^1 FD P^0 may hold if $\mathbf{p}^1$ is located NW of $\mathbf{p}^0$ in either of the orange-shaded areas and will hold if $\mathbf{p}^1$ sits in the dark-orange triangle. Similarly P^0 FD P^1 may hold if $\mathbf{p}^1$ is located SE of $\mathbf{p}^0$ in either of the blue-shaded areas and will hold if $\mathbf{p}^1$ sits in the dark-blue triangle. The intuition relies on the horizontal or vertical distance from any $\mathbf{p}$ to the hypotenuse being $\sum_{m=1}^{M-1}\Pr(s=m)$. If $\mathbf{p}^1$ sits in the dark orange triangle the largest possible value of $\sum_{m=0}^{M-1}p_m^1$ is thus p_0^0. As such $C^0(s)\geq C^1(s)$ for all $s\in V$ so P^1 FD P^0.[30]

Figure A.2(c) shows SD relationships for general M≥2. From the baseline $\mathbf{p}^0$ it may but will not necessarily hold that P^0 SD P^1 for $\mathbf{p}^1$ in the blue region and it may hold that P^1 SD P^0 for $\mathbf{p}^1$ in the light-orange region. P^1 SD P^0 for $\mathbf{p}^1$ on the dark-orange segment of the hypotenuse. There can be no SD relationship in the gray regions.

---

[29] With M=2 $\mathbf{p}^1$ sitting anywhere NW of $\mathbf{p}^0$ (i.e. in either of the orange-shaded areas) implies P^1 FD P^0 and likewise that $\mathbf{p}^1$ sitting any where SE of $\mathbf{p}^0$ (in either blue-shaded area) implies P^0 FD P^1.

[30] Consider a numerical example. In figure A.2(b) $\mathbf{p}^0=[.5,\ .2]$ as above. Suppose $P^1=\langle .4,\ .2,\ .1,\ .3\rangle$ so that $\mathbf{p}^1=\left[p_0^1,\ p_3^1\right]=[.4,\ .3]$. Consider two possible P^0 consistent with $\mathbf{p}^0$: $P^{0A}=\langle .5,\ .1,\ .2,\ .2\rangle$ and $P^{0B}=\langle .5,\ 0,\ .3,\ .2\rangle$. P^1 FD P^{0A}, P^1 does not FD P^{0B}, but note that P^1 TD P^{0A} and P^1 TD P^{0B} both hold.

s	P^{0A}	C^{0A}	P^{0B}	C^{0B}	P^1	C^1
0	.5	.5	.5	.4	.4	.4
1	.1	.6	0	.5	.2	.6
2	.2	.8	.3	.8	.1	.7
3	.2	1	.2	1	.3	1

Appendix 3: Proofs of Expressions (14) and (15)-(16) (Section 4)

Lower Bounds (Expression (14)). For each $\delta \in T$:

$$\Pr(\Delta \le \delta) = \Pr(s(1) - s(0) \le \delta) \tag{A3.1}$$

$$= \Pr(s(1) \le s(0) + \delta) \tag{A3.2}$$

$$\ge \Pr(s(1) \le t \le s(0) + \delta) \quad \text{for any } t \in V \tag{A3.3}$$

$$= \Pr(s(1) \le t \ \wedge \ t \le s(0) + \delta) \tag{A3.4}$$

$$\ge \max\{0,\ \Pr(s(1) \le t) + \Pr(s(0) + \delta \ge t) - 1\} \tag{A3.5}$$

$$= \max\{0,\ \Pr(s(1) \le t) + (1 - \Pr(s(0) + \delta < t)) - 1\} \tag{A3.6}$$

$$= \max\{0,\ \Pr(s(1) \le t) - \Pr(s(0) + \delta < t)\}, \tag{A3.7}$$

where (A3.5) is a Boole-Fréchet LB on (A3.4) for any $t \in V$, which is a valid LB since $\Pr(\Delta \le \delta) \ge \max\{0,\ \Pr(s(1) \le t) - \Pr(s(0) + \delta < t)\}$ and which is identifiable since it relies only on the identified marginal distributions $\Pr(s(\bullet))$. This proves part (a) of expression (14).

Since (A3.7) is the largest valid bound at each $t \in V$ (see *Intuition* below) then the best LB is the maximum of (A3.7) over $t \in V$, i.e.

$$LB^*(\Pr(\Delta \le \delta)) = \max_{t \in V} \max\{0,\ \Pr(s(1) \le t) - \Pr(s(0) + \delta < t)\}. \tag{A3.8}$$

This proves part (b) of expression (14). ■

Intuition— Defining an identifiable best LB requires using only information contained in the marginal distributions $\Pr(s(\bullet))$ that are identified by assumption. The basic structure of any such bounds will be based on a difference in marginal probabilities of the sort shown in (A3.8). Any bound claiming to be a best bound will turn out to be the difference in the overall joint probability of a subset $R_{(+)}$ of the points $\{(s(0),\ s(1)) | \Delta \le \delta\}$ minus the joint probability of a subset $R_{(-)}$ of the points $\{(s(0),\ s(1)) | \Delta > \delta\}$, i.e.

$$\sum_{(s(0),\ s(1)) \in R_{(+)}} \Pr(s(0),\ s(1)) \ - \ \sum_{(s(0),\ s(1)) \in R_{(-)}} \Pr(s(0),\ s(1)). \tag{A3.9}$$

This difference cannot be larger than the overall joint probability of the points in $R_{(+)}$ which in turn is no larger than the overall joint probability of all points satisfying $\{(s(0),\ s(1)) | \Delta \le \delta\}$. It is thus a valid LB.

To appreciate that (A3.8) is the best such LB assume the existence of an even greater LB,

$$LB^{**}(\Pr(\Delta \le \delta)) = \max\{0,\ \Pr(A_1) - \Pr(A_0)\} \ge LB^{*}(\Pr(\Delta \le \delta)) \tag{A3.10}$$

for observable events A_j in the marginal distributions $\Pr(s(\bullet))$. For LB** to exceed LB* it must hold either that (a) $\Pr(A_1) > \Pr(s(1) \le t)$ (e.g. $\Pr(A_1) = \Pr(s(1) \le t + k)$, $k \in \{1, 2, \ldots\}$), or that (b) $\Pr(A_0) < \Pr(s(0) + \delta < t)$ (e.g. $\Pr(A_0) < \Pr(s(0) + \delta < t - k)$, $k \in \{1, 2, \ldots\}$) so that for index sets $Q_{(+)}$ and $Q_{(-)}$ defined implicitly by $\Pr(A_0)$ and $\Pr(A_1)$,

$$\begin{aligned} &\sum_{(s(0),\, s(1)) \in Q_{(+)}} \Pr(s(0),\, s(1)) \ - \sum_{(s(0),\, s(1)) \in Q_{(-)}} \Pr(s(0),\, s(1)) \ \ge \\ &\sum_{(s(0),\, s(1)) \in R_{(+)}} \Pr(s(0),\, s(1)) \ - \sum_{(s(0),\, s(1)) \in R_{(-)}} \Pr(s(0),\, s(1)) \end{aligned} \quad . \tag{A3.11}$$

It is straightforward to show, however, that for either (a) or (b) the difference $\Pr(A_1) - \Pr(A_0)$ results in at least one point in $\{(s(0),\, s(1)) | \Delta > \delta\}$ being included in the index set $Q_{(+)}$ in (A1.11) so that LB** cannot be a valid LB since it may exceed $\Pr(\Delta \le \delta)$.

Upper Bounds (Expressions (15)-(16)). For each $\delta \in T$:

$$\Pr(\Delta \le \delta) = \Pr(s(1) - s(0) \le \delta) \tag{A3.12}$$

$$= 1 - \Pr(s(1) - s(0) > \delta) \tag{A3.13}$$

$$= 1 - \Pr(s(1) - \delta > s(0)) \tag{A3.14}$$

$$\le 1 - \Pr(s(1) - \delta \ge^{a} t \ge^{b} s(0)) \quad \text{for any } t \in V. \tag{A3.15}$$

If one and only one of the inequalities $\ge^a$ or $\ge^b$ in (A3.15) is strict then (A3.15) necessarily follows from (A3.14) since then $\Pr(s(1) - \delta > t \ge s(0)) \le \Pr(s(1) - \delta > s(0))$ or $\Pr(s(1) - \delta \ge t > s(0)) \le \Pr(s(1) - \delta > s(0))$. If neither of the inequalities $\ge^a$ or $\ge^b$ in (A3.15) is strict then (A3.15) will not necessarily follow from (A3.14). If both inequalities $\ge^a$ and $\ge^b$ in (A3.15) are strict then (A3.15) will be too large to serve as the basis of a best UB relative to other identifiable alternatives. The relevant cases for determining a best UB thus involve one and only one of $\ge^a$ or $\ge^b$ being strict. It is shown below that $UB^{*}(\Pr(\Delta \le \delta))$ is the same regardless of which one is strict.

Case 1: $\ge^a$ is strict, $\ge^b$ is weak. Then continuing from (A3.15):

$$1-\Pr\left(s(1)-\delta>t\geq s(0)\right)=1-\Pr\left(s(1)-\delta>t\ \wedge\ t\geq s(0)\right) \tag{A3.16}$$

$$\leq 1-\max\left\{0,\ \Pr\left(s(1)>t+\delta\right)+\Pr\left(t\geq s(0)\right)-1\right\} \tag{A3.17}$$

$$=1-\max\left\{0,\ \left(1-\Pr\left(s(1)\leq t+\delta\right)\right)+\Pr\left(s(0)\leq t\right)-1\right\} \tag{A3.18}$$

$$=1-\max\left\{0,\ \Pr\left(s(0)\leq t\right)-\Pr\left(s(1)\leq t+\delta\right)\right\}, \tag{A3.19}$$

where (A3.17) is one minus a Boole-Fréchet LB on the subtrahend in (A3.16). This proves expression (3.a) in part (a) of Proposition 2. Then:

$$UB^{*}\left(\Pr(\Delta\leq\delta)\right)=\min_{t\in V}\left\{1-\max\left\{0,\ \Pr\left(s(0)\leq t\right)-\Pr\left(s(1)\leq t+\delta\right)\right\}\right\} \tag{A3.20}$$

$$=1-\max_{t\in V}\max\left\{0,\ \Pr\left(s(0)\leq t\right)-\Pr\left(s(1)\leq t+\delta\right)\right\}. \tag{A3.21}$$

This proves expression (15).

Case 2: $\geq^{b}$ is strict, $\geq^{a}$ is weak. Then continuing from (A3.15):

$$1-\Pr\left(s(1)-\delta\geq t>s(0)\right)=1-\Pr\left(s(1)-\delta\geq t\ \wedge\ t>s(0)\right) \tag{A3.22}$$

$$\leq 1-\max\left\{0,\ \Pr\left(s(1)\geq t+\delta\right)+\Pr\left(t>s(0)\right)-1\right\} \tag{A3.23}$$

$$=1-\max\left\{0,\ \left(1-\Pr\left(s(1)<t+\delta\right)\right)+\Pr\left(s(0)<t\right)-1\right\} \tag{A3.24}$$

$$=1-\max\left\{0,\ \Pr\left(s(0)<t\right)-\Pr\left(s(1)<t+\delta\right)\right\}, \tag{A3.25}$$

where (A3.23) is one minus a Boole-Fréchet LB on the subtrahend in (A3.22). This proves expression (3.b) in part (a) of Proposition 2. Then:

$$UB^{*}\left(\Pr(\Delta\leq\delta)\right)=\min_{t\in V}\left\{1-\max\left\{0,\ \Pr\left(s(0)<t\right)-\Pr\left(s(1)<t+\delta\right)\right\}\right\} \tag{A3.26}$$

$$=1-\max_{t\in V}\max\left\{0,\ \Pr\left(s(0)<t\right)-\Pr\left(s(1)<t+\delta\right)\right\}. \tag{A3.27}$$

This proves expression (16).

The equivalence of (15) and (16) can be shown by noting that $\Pr\left(s(\bullet)<0\right)=0$ and $\Pr\left(s(\bullet)\leq M\right)=1$ so that the $\max_{t\in V}$ operation is running over the same arguments but offset by one unit. To see this concretely consider M=3 and $\delta=0$. Then from (A3.21),

$$1-\max_{t\in V}\max\left\{0,\ \Pr\left(s(0)\le t\right)-\Pr\left(s(1)\le t\right)\right\}=1-\max\begin{Bmatrix}\max\left\{0,\ \Pr\left(s(0)\le 0\right)-\Pr\left(s(1)\le 0\right)\right\},\\ \max\left\{0,\ \Pr\left(s(0)\le 1\right)-\Pr\left(s(1)\le 1\right)\right\},\\ \max\left\{0,\ \Pr\left(s(0)\le 2\right)-\Pr\left(s(1)\le 2\right)\right\},\\ 0\end{Bmatrix} \quad \text{(A3.28)}$$

since $\max\left\{0,\ \Pr\left(s(0)\le 3\right)-\Pr\left(s(1)\le 3\right)\right\}=\max\left\{0,\ 1-1\right\}=0$, whereas from (A3.27)

$$1-\max_{t\in V}\max\left\{0,\ \Pr\left(s(0)<t\right)-\Pr\left(s(1)<t\right)\right\}=1-\max\begin{Bmatrix}0,\\ \max\left\{0,\ \Pr\left(s(0)<1\right)-\Pr\left(s(1)<1\right)\right\},\\ \max\left\{0,\ \Pr\left(s(0)<2\right)-\Pr\left(s(1)<2\right)\right\},\\ \max\left\{0,\ \Pr\left(s(0)<3\right)-\Pr\left(s(1)<3\right)\right\}\end{Bmatrix}$$

$$=1-\max\begin{Bmatrix}0,\\ \max\left\{0,\ \Pr\left(s(0)\le 0\right)-\Pr\left(s(1)\le 0\right)\right\},\\ \max\left\{0,\ \Pr\left(s(0)\le 1\right)-\Pr\left(s(1)\le 1\right)\right\},\\ \max\left\{0,\ \Pr\left(s(0)\le 2\right)-\Pr\left(s(1)\le 2\right)\right\}\end{Bmatrix} \quad \text{(A3.29)}$$

since $\max\left\{0,\ \Pr\left(s(0)<0\right)-\Pr\left(s(1)<0\right)\right\}=\max\left\{0,\ 0-0\right\}=0$. ■

Appendix 4: Probability Model Partial Effects

This appendix derives formally and discusses further the PEs for the probability models in section 7. Among other things the appendix considers for each model the number of possible sign changes for $D_k \Pr(s = m|\mathbf{x})$ as m increases on V. (Recall that for any probability model the PEs must change sign at least once as m increases on V due to adding up, $\sum_{m \in V} D_k \Pr(s = m|\mathbf{x}) = 0$.) While either one or two changes of sign as m increases on V can be accommodated in the parametric models considered in the paper, nonparametric specifications can accommodate as many as M sign changes. As noted in the main text the ability of a probability model to accommodate more than one such sign change means that it can describe phenomena like spread-increasing or -reducing changes in **x**.

Binomial Models

For continuous x_k the PEs of the binomial $(M, p(\mathbf{x}))$ model (23) are given by

$$D_k \Pr(s = m|\mathbf{x}) = \Pr(s = m|\mathbf{x}) \cdot D_k p(\mathbf{x}) \cdot \left(\frac{m - p(\mathbf{x}) \cdot M}{p(\mathbf{x}) \cdot (1 - p(\mathbf{x}))} \right), \tag{A4.1}$$

with $D_k p(\mathbf{x}) = \partial p(\mathbf{x}) / \partial x_k$. It is easy to see that the fraction in parentheses increases from negative to positive as m increases on V. As such the sign of $D_k \Pr(s = m|\mathbf{x})$ changes once as m increases on V. Panel (a) of figure A.3 depicts this result for a discrete change in **x**.

Beta-Binomial Models

Consider first the BBIN-2 specification (26) with continuous x_k. Let $D_k a(\mathbf{x}) = \partial a(\mathbf{x}) / \partial x_k$ and $D_k b(\mathbf{x}) = \partial b(\mathbf{x}) / \partial x_k$ (e.g. $D_k a(\mathbf{x}) = \partial \exp(\mathbf{x}\boldsymbol{\alpha}) / \partial x_k = \alpha_k \exp(\mathbf{x}\boldsymbol{\alpha})$). Recall that $D_k b(\mathbf{x}) = 0$ in the BBIN-1 specification. Then the PE $D_k \Pr(s = m|\mathbf{x})$ is obtained using log-derivatives as $D_k \Pr(s = m|\mathbf{x}) = \Pr(s = m|\mathbf{x}) \cdot D_k \log(\Pr(s = m|\mathbf{x}))$, viz

$$D_k \Pr(s = m|\mathbf{x}) = \Pr(s = m|\mathbf{x}) \cdot \{ T_a(m) \cdot D_k a(\mathbf{x}) + T_b(m) \cdot D_k b(\mathbf{x}) \} \tag{A4.2}$$

where

$$T_a(m) = \psi(a + b) + \psi(a + m) - \psi(a) - \psi(a + b + M) \tag{A4.3}$$

and

$$T_b(m) = \psi(a + b) + \psi(b + M - m) - \psi(b) - \psi(a + b + M), \tag{A4.4}$$

where $\psi(\bullet)$ is the digamma function $d\log\Gamma(\bullet)/d(\bullet)$, and where $a(\mathbf{x})$ and $b(\mathbf{x})$ are written as a and b to reduce clutter. Re-write $T_a(m)$ as $T_a(m) = T_{a1} + T_{a2}(m)$ by adding and subtracting $\psi(a+M)$, giving

$$T_{a1} = \psi(a+b) + \psi(a+M) - \psi(a) - \psi(a+b+M) \tag{A4.5}$$

and

$$T_{a2}(m) = \psi(a+m) - \psi(a+M). \tag{A4.6}$$

Similarly re-write $T_b(m)$ as $T_b(m) = T_{b1} + T_{b2}(m)$ by adding and subtracting $\psi(b+M)$, giving

$$T_{b1} = \psi(a+b) + \psi(b+M) - \psi(b) - \psi(a+b+M) \tag{A4.7}$$

and

$$T_{b2}(m) = \psi(b+M-m) - \psi(b+M). \tag{A4.8}$$

T_{a1} and T_{b1} are positive since $\psi(\bullet)$ is a strictly concave function (Alzer and Jameson, 2017) and $g(u+v)+g(u+w) > g(u)+g(u+v+w)$ for strictly concave functions $g(\bullet)$. $T_{a2}(m)$ increases monotonically from negative to zero as m increases on V while $T_{b2}(m)$ decreases monotonically from zero to negative as m increases on V. Thus $T_a(m)$ increases montonically from negative to positive as m increases on V while $T_b(m)$ decreases montonically from positive to negative as m increases on V.

It follows from (A4.2) that if $D_k a(\mathbf{x})$ and $D_k b(\mathbf{x})$ have opposite signs then the PEs $D_k \Pr(s=m|\mathbf{x})$ will be either monotonically increasing or monotonically decreasing as m increases on V and will thus change sign only once. If $D_k a(\mathbf{x})$ and $D_k b(\mathbf{x})$ have the same sign then $D_k \Pr(s=m|\mathbf{x})$ will change sign either once or twice as m increases on V, depending on the particular values of $a(\mathbf{x})$, $b(\mathbf{x})$, $D_k a(\mathbf{x})$, and $D_k b(\mathbf{x})$. (Recall the example from section 7 showing the contrasting PE sign patterns when $a(\mathbf{x})=1$ vs. $a(\mathbf{x})=2$ with $b(\mathbf{x}) = D_k a(\mathbf{x}) = D_k b(\mathbf{x}) = 1$ and M=7.) In the BBIN-1 specification where $D_k b(\mathbf{x}) = 0$ it is immediately evident that $D_k \Pr(s=m|\mathbf{x})$ changes sign only once as m increases on V.

Stata's `betabin` Procedure

As noted in section 7 the beta-binomial specification of $\Pr(s=m|\mathbf{x})$ used for Stata's `betabin` procedure (31) differs from the BBIN-2 and BBIN-1 specifications (Hardin and Hilbe, 2014, henceforth HH). The PEs for a continuous x_k in the HH specification—those whose sample averages would be

reported by Stata's `margins` command after executing `betabin`—are given by

$$\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)=\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)\cdot\mathrm{D_k}\mathrm{p}(\mathbf{x})\cdot\upsilon\cdot\left\{\psi\left(\mathrm{p}(\mathbf{x})\cdot\upsilon+\mathrm{m}\right)\ +\ \psi\left(\left(1-\mathrm{p}(\mathbf{x})\right)\cdot\upsilon\right)\ -\right.$$
$$\left.\psi\left(\left(1-\mathrm{p}(\mathbf{x})\right)\cdot\upsilon+\mathrm{M}-\mathrm{m}\right)\ -\ \psi\left(\mathrm{p}(\mathbf{x})\cdot\upsilon\right)\right\} \tag{A4.9}$$

Evaluating this expression at m=0 and m=M shows that the sign of $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)$ at those values will be opposite since $\psi(\bullet)$ is increasing in its argument. The sign of $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)$ changes only once as m increases on V.

Ordered Probability Models

Recall from (32)

$$\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)=\mathrm{F}\left(\kappa_{\mathrm{m}+1}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)-\mathrm{F}\left(\kappa_{\mathrm{m}}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right),\quad \mathrm{m}\in\mathrm{V}\ , \tag{A4.10}$$

For the ordered regression model PEs, consider first the derivatives

$$\begin{aligned}\mathrm{D_K}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)&=-\boldsymbol{\beta}_{\mathrm{v,k}}\cdot\left(\mathrm{F}'\left(\kappa_{\mathrm{m}+1}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)-\mathrm{F}'\left(\kappa_{\mathrm{m}}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)\right)\\&=-\boldsymbol{\beta}_{\mathrm{v,k}}\cdot\left(\mathrm{f}\left(\kappa_{\mathrm{m}+1}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)-\mathrm{f}\left(\kappa_{\mathrm{m}}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)\right)\end{aligned} \tag{A4.11}$$

As m increases on V the difference $\mathrm{f}\left(\kappa_{\mathrm{m}+1}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)-\mathrm{f}\left(\kappa_{\mathrm{m}}-\mathbf{x}_{\mathrm{v}}\boldsymbol{\beta}_{\mathrm{v}}\right)$ will initially be positive then negative since $\mathrm{f}(\bullet)$ is unimodal (see Wooldridge, 2010, page 656, for related discussion). $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)$ will thus change sign only once as m increases on V. As such the PEs at m=0 and at m=M are necessarily of opposite signs.

Consider now the discrete difference

$$\begin{aligned}\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)=&\left(\mathrm{F}\left(\kappa_{\mathrm{m}+1}-\mathbf{x}_{\mathrm{v}}(1)\boldsymbol{\beta}_{\mathrm{v}}\right)-\mathrm{F}\left(\kappa_{\mathrm{m}}-\mathbf{x}_{\mathrm{v}}(1)\boldsymbol{\beta}_{\mathrm{v}}\right)\right)-\\&\left(\mathrm{F}\left(\kappa_{\mathrm{m}+1}-\mathbf{x}_{\mathrm{v}}(0)\boldsymbol{\beta}_{\mathrm{v}}\right)-\mathrm{F}\left(\kappa_{\mathrm{m}}-\mathbf{x}_{\mathrm{v}}(0)\boldsymbol{\beta}_{\mathrm{v}}\right)\right)\end{aligned} \tag{A4.12}$$

where the difference between $\mathbf{x}(0)$ and $\mathbf{x}(1)$ is due to a discrete change in $\mathrm{x_k}$. To appreciate that these discrete partial effects have the same structure as the derivatives consult panel (b) of figure A.3 which shows the shift in the density $\mathrm{f}\left(\mathrm{s}^*\middle|\mathbf{x}\right)$ due to a discrete change from $\mathbf{x}(0)$ to $\mathbf{x}(1)$, where $\boldsymbol{\beta}$ is defined such that $\mathbf{x}(0)\boldsymbol{\beta}<\mathbf{x}(1)\boldsymbol{\beta}$. Comparing areas under the curves between any two adjacent values of κ_{m} shows that initially $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)<0$ and then $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)>0$. In the example depicted in figure A.3 with M=7 $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)<0$ for $\mathrm{m}\in\{0,\ ...,\ 4\}$ while $\mathrm{D_k}\Pr\left(\mathrm{s}=\mathrm{m}\middle|\mathbf{x}\right)>0$ for $\mathrm{m}\in\{5,\ 6,\ 7\}$.

Conditional vs. Average PEs

For the parametric probability models that allow only one sign change for the PEs as m increases on V (specifically, binomial and ordered regression) it should be emphasized that the single sign change result characterizes the *conditional* PEs $D_k \Pr(s=m|\mathbf{x})$. The *average* PEs, $E_{\mathbf{x}} D_k \Pr(s=m|\mathbf{x})$, estimated as $\frac{1}{N}\sum_{n=1}^{N} D_k \widehat{\Pr}(s=m|\mathbf{x}_n)$ need not obey these restrictions on sign changes as m increases on V.[31]

However the sample APEs at m=0 and m=M will necessarily have opposite signs even if there is more than one sign change of the sample APEs as m increases on V. This holds because for any conditioning $\mathbf{x}$ $D_k \Pr(s=0|\mathbf{x}) > 0 > D_k \Pr(s=M|\mathbf{x})$ if $\beta_k < 0$ and $D_k \Pr(s=0|\mathbf{x}) < 0 < D_k \Pr(s=M|\mathbf{x})$ if $\beta_k > 0$. Unlike m=0 and m=M the sign of $D_k \Pr(s=m|\mathbf{x})$ for $m \in \{1, ..., M-1\}$ may vary with $\mathbf{x}$. As such the sample average of these always-negative or always-positive terms must be negative or positive. So in this sense for binomial and ordered regression models a change in $\mathbf{x}$ cannot result in a change in "spread"—in this context meaning simultaneous increase or simultaneous decrease in both $\Pr(s=0|\mathbf{x})$ and $\Pr(s=M|\mathbf{x})$—either conditionally on $\mathbf{x}$ or averaged over $\mathbf{x}$.

---

[31] This Stata example provides a numerical illustration for the binomial and ordered probit specifications:

```
set seed 2345
set obs 1000
loc M=7
loc M1=`M'+1
gen x1=runiform()>.5
gen x2=runiform()>.5
gen ys=x1 + 5*x2 + rnormal()
gen y=autocode(ys,`M1',0,`M1')-1

* binomial specification
qui glm y x1 x2, fam(binomial `M') link(logit) robust
est store eglm
forval j=0/`M' {
 qui est restore eglm
 qui margins, dydx(*) expression(comb(`M',`j')*(logistic(xb()))^(`j')*  ///
  (1-logistic(xb()))^(`M'-`j')) post
  est store dydx_`j'
}
est table dydx_*

* ordered probit specification
qui oprobit y x1 x2
margins, dydx(*)
```

Appendix 5: Estimation Strategies

This appendix discusses briefly strategies that can be used in Stata to estimate the models and corresponding APEs described in sections 6 and 7 and Appendix 4. The Stata code as well as the HSE estimation sample are available in this downloadable zip file.

Estimation of Conditional Moments and APEs via Fractional Regression and GMM

Fractional regression specifications (Papke and Wooldridge, 1996) provide a straightforward approach to estimation of conditional means or other raw moments using Stata. Focusing first on conditional means, estimation of conditional mean APEs for fractional probit or fractional logit specifications can be performed using either the `glm` command or, with modification, the `fracreg logit` and `fracreg probit` commands, as follows:

```
local M = <numerical value of M>
local xvars="x1 x2 ..."
local yvar="yvar"
glm `yvar' `xvars', link(logit) family(binomial `M') vce(robust)
margins, dydx(*)
```

For fractional probit specifications replace `link(logit)` with `link(probit)`. This `glm` command yields APE estimates identical to

```
tempvar tyvar
gen `tyvar' = `yvar'/`M'
fracreg logit `tyvar' `xvars', vce(robust)
margins, dydx(*) expr(`M'*logistic(xb()))
```

replacing `fracreg logit` with `fracreg probit` as appropriate.

As noted in the main text, simultaneous estimation of conditional means and conditional variances and corresponding APEs when each moment is specified as a fractional regression can be accomplished using Stata's `gmm` command. Consider for example the moment specification (21). Stata's `gmm` procedure can be used to estimate $[\alpha,\beta]$ and the corresponding APEs as follows:

```
gmm ( 1: `yvar' - `M'*logistic({xa:`xvars' _cons}) )                ///
    ( 2: (`yvar' - `M'*logistic({xa:}))^2 -                         ///
         .25*(`M'^2)*logistic({xb:`xvars' _cons}) ),                ///
    inst(`xvars') igmm vce(robust) winit(i) quickderivatives
margins, dydx(*) expression(`M'*logistic(xb(xa)))
margins, dydx(*) expression(.25*`M'^2*logistic(xb(xb)))
```

For probit links replace `logistic(...)` with `normal(...)`. The GMM estimates of the first-moment parameters α are identical those obtained using the GLM or fractional regression approaches described above. The code used to obtain the results reported in tables 2-4 is included in the downloadable zip file.

Estimation of Binomial and Beta-Binomial Probability Models and APEs

The binomial model (23) and its corresponding predictions and APEs can be estimated using Stata's `glm` command:

```
glm `yvar' `xvars', family(binomial `M') link(logit) vce(robust)
```

replacing `link(logit)` with `link(probit)` as appropriate. Computation of predicted probabilities and APEs requires additional postestimation programming. The code to do this is included in the downloadable zip file.

The BBIN-2 and BBIN-1 specifications (26) and their corresponding APEs and predicted probabilities can be estimated using an author-written Mata function `bbin()` that is included in the downloadable zip file.

Stata's version of the beta-binomial model (31) can be estimated using

```
betabin `yvar' `xvars', n(`M') vce(robust)
```

Computation of predicted probabilities and APEs requires additional postestimation programming. The code to do this is included in the downloadable zip file.

Estimation of Ordered Probability Models and APEs

Estimation of ordered probit or ordered logit probability models is straightforward using Stata's `oprobit` or `ologit` commands. Computation of predicted probabilities and APEs requires additional postestimation programming. The code to do this is included in the downloadable zip file.

Figure 1: Four Samples' Distributions

Panel (a): the number of days on which adult respondents reported engaging in vigorous or moderate exercise in the prior week (VM7; 2015 Health Survey of England (HSE))

Panel (b): the number of chronic conditions from a list of eight reported by adult respondents (CC8; 2021 Behavioral Risk Factors Surveillance System (BRFSS))

Panel (c): the number of physically healthy days in the previous 30 days reported by adult respondents (HD30; 2021 BRFSS)

Panel (d): the PHQ-2 depression screener score for adult respondents (PHQ2; 2020 Medical Expenditure Panel Survey (MEPS))

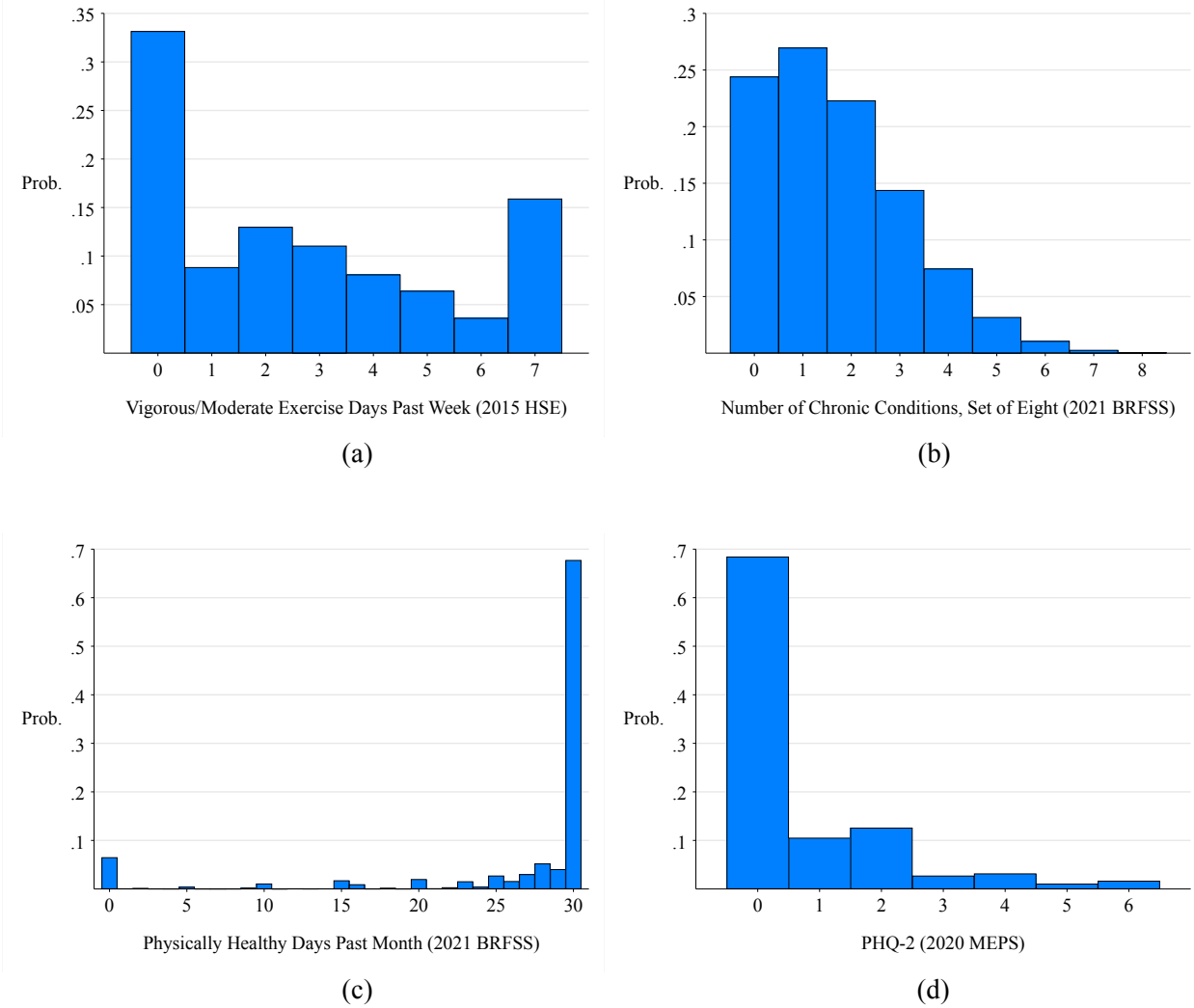


Figure 2: Sample Space, $V \times V$, TEs, and Joint Distribution for M=3

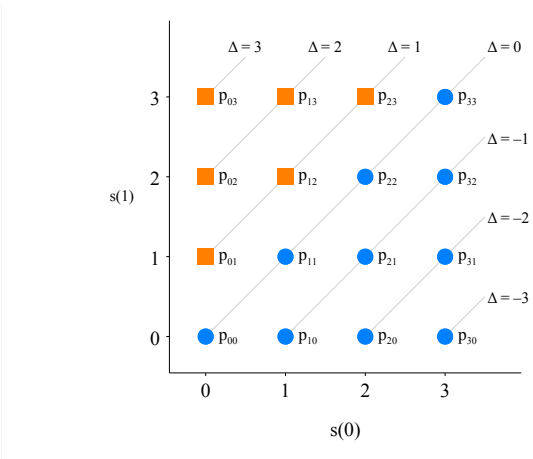
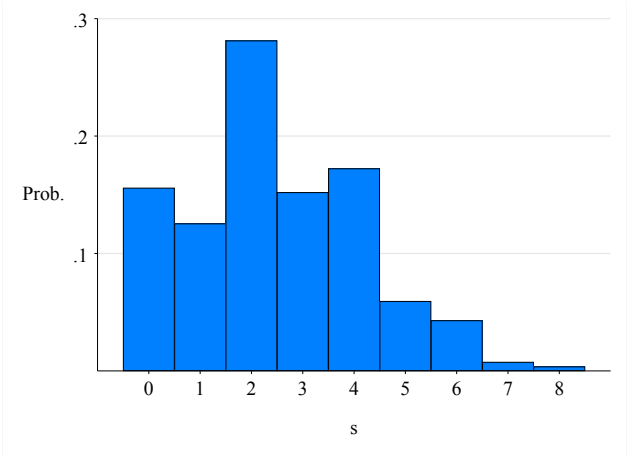
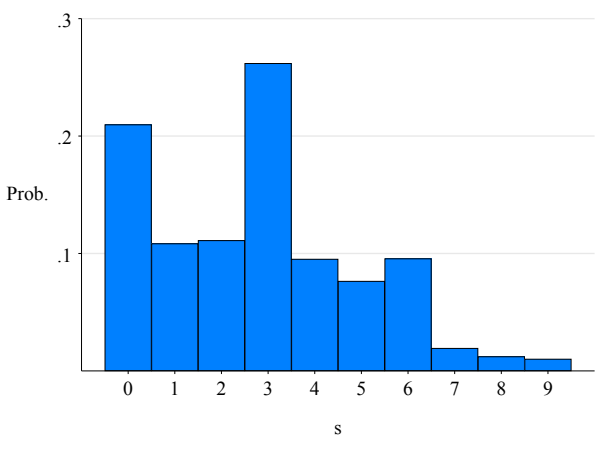


Figure 3: Multivariate Probit Heaped Histograms

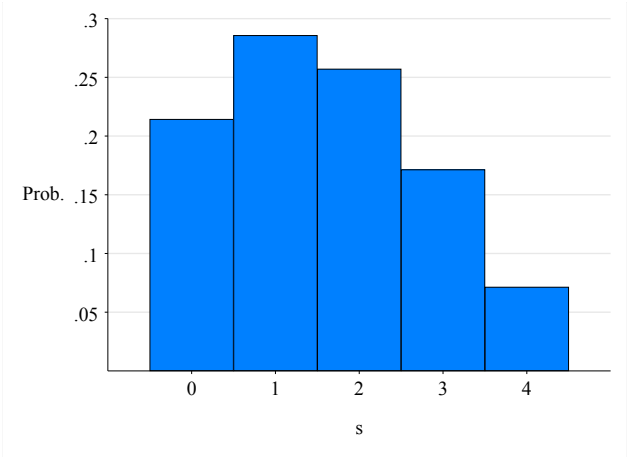


(a)

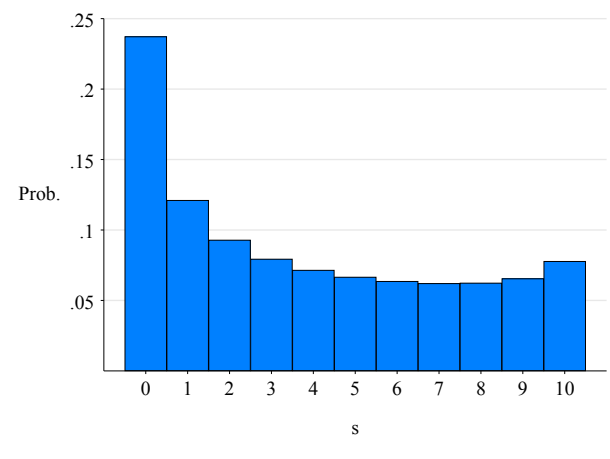


(b)

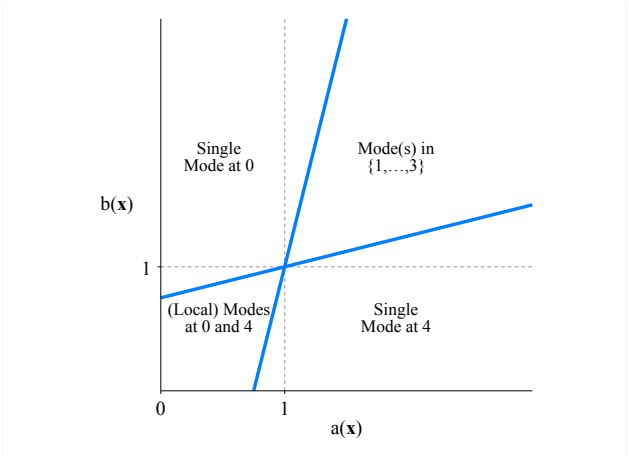
Figure 4: Examples of Beta-Binomial Distributions



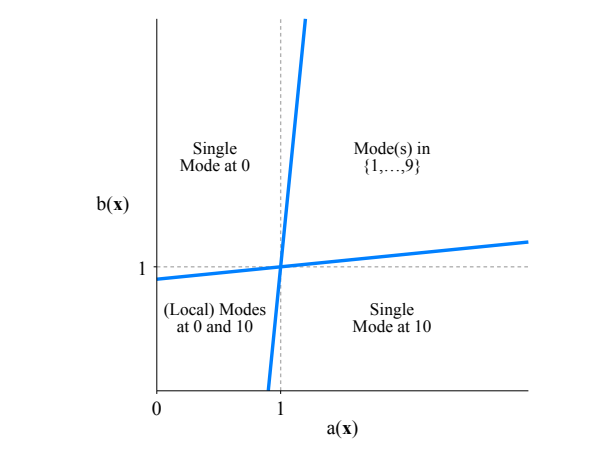
(a)



(b)



(c)



(d)

Figure 5: Beta-Binomial Parameter Values and Dominance Relationships

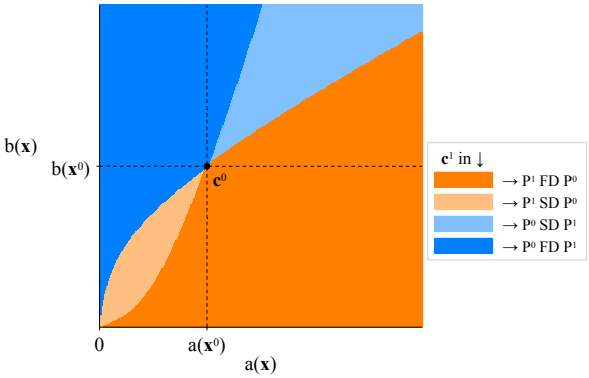
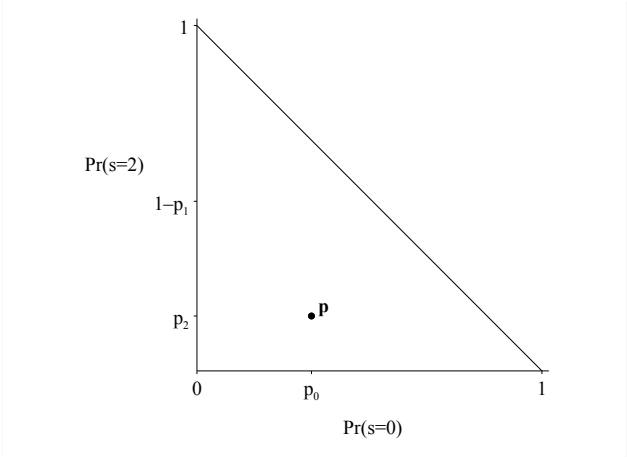
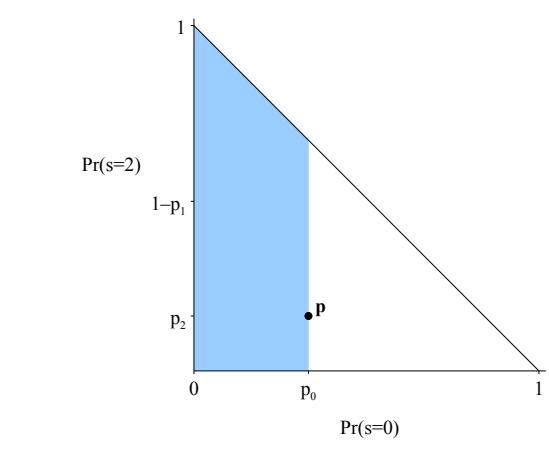


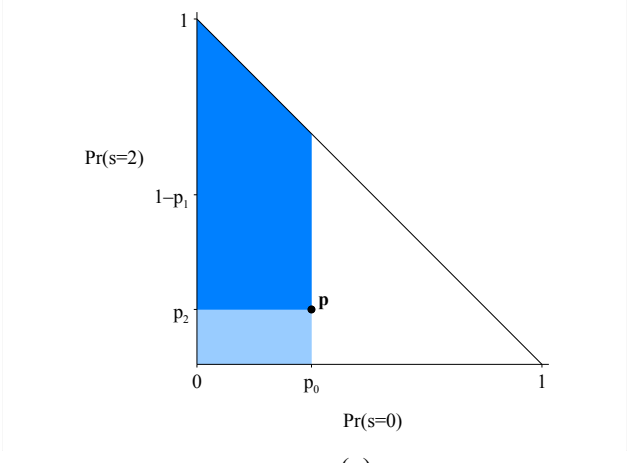
Figure A.1: Partial Identification of P with Misreporting



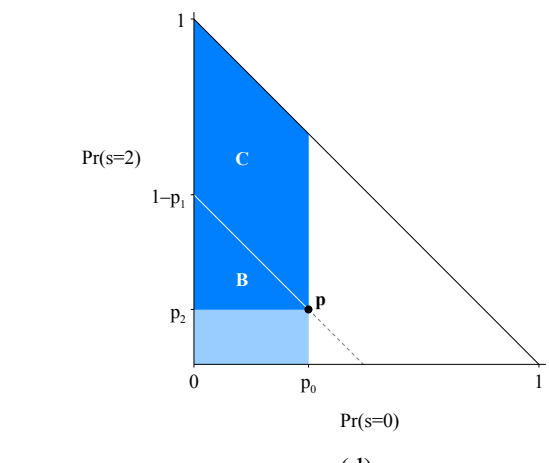
(a)



(b)



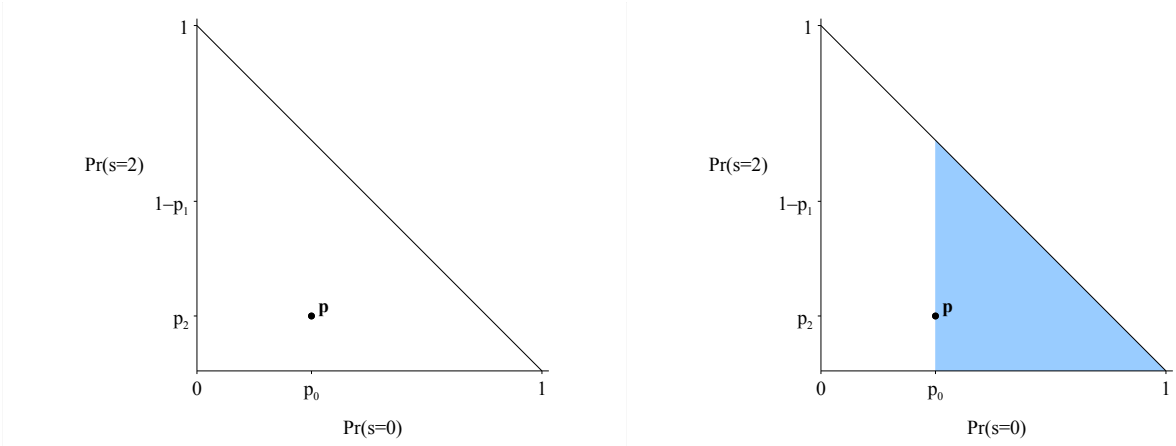
(c)



(d)

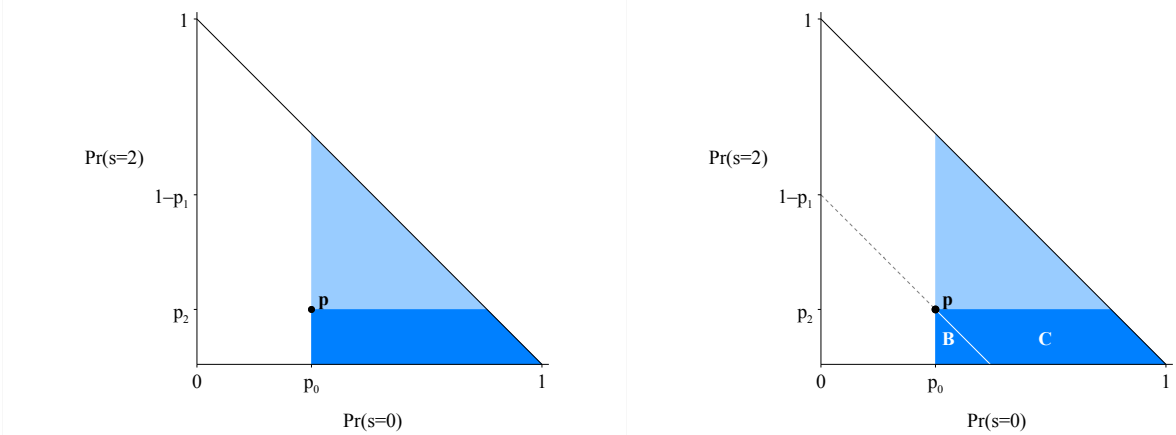
(cont.)

(cont.) Figure A.1: Partial Identification of P with Misreporting



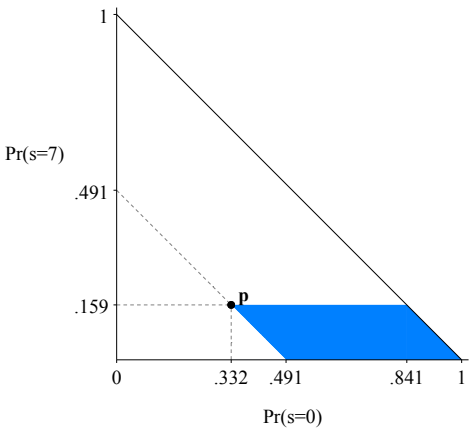
(e)

(f)



(g)

(h)



(i)

Figure A.2: Marschak Probability Triangles Illustrating Dominance Relationships

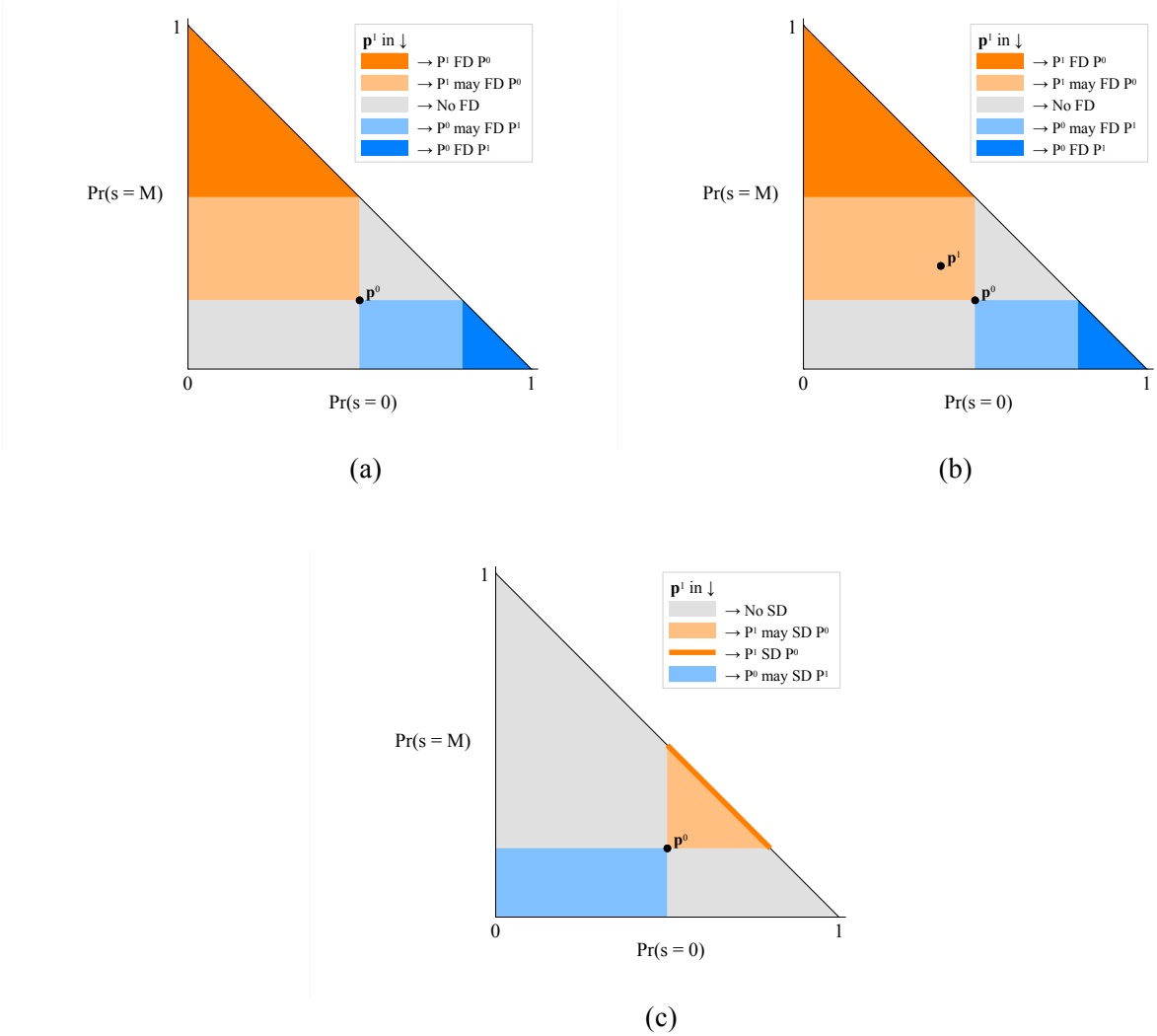
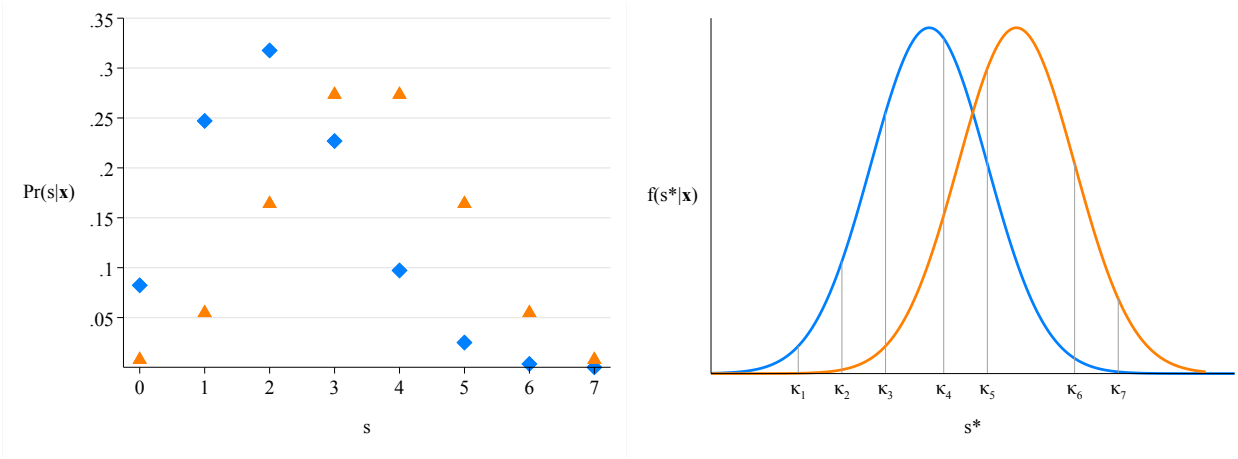


Figure A.3: Illustrating $D_k \Pr(s = m|\mathbf{x})$ for Binomial and Ordered Probit Models



(a) (b)

Table 1: $LB^*(\Pr(\Delta \le \delta))$ and $UB^*(\Pr(\Delta \le \delta))$ for Weekly Total Days of Vigorous or Moderate Exercise Given Season-of-Year Treatment (y_0 is Fall/Winter; y_1 is Spring/Summer)—2015 Health Survey of England

δ	$LB^*(\Pr(\Delta \le \delta))$	$UB^*(\Pr(\Delta \le \delta))$
-7	0	.147
-6	0	.180
-5	0	.240
-4	0	.318
-3	0	.425
-2	0	.556
-1	0	.643
0	.303	.946
1	.393	1
2	.521	1
3	.635	1
4	.720	1
5	.788	1
6	.827	1
7	1	1

Table 2: Point Estimates of Conditional Mean and Conditional Variance Average Partial Effects

$$\frac{1}{N}\sum_{n=1}^{N} D_k \hat{E}\left(s|\mathbf{x}\right),\ \frac{1}{N}\sum_{n=1}^{N} D_k \widehat{Var}\left(s|\mathbf{x}_n\right)$$

(Note: Fractional Logit-1 and Fractional Logit-2 refer to the fractional regression GMM moment specifications appearing in (21) and (22), respectively)

	$D_k E\left(s\|\mathbf{x}\right)$				$D_k Var\left(s\|\mathbf{x}\right)$	
Covariate	Non-parametric	Linear	Fractional Logit	BBIN-2	Fractional Logit-1	Fractional Logit-2
Age	−.023	−.022	−.023	−.021	−.013	−.001
Female	.015	−.026	−.028	.002	−.489	−.478
NVQ45	.315	.251	.247	.238	−.741	−.858
South	.113	.104	.104	.111	−.035	−.005
Urban	−.529	−.605	−.611	−.638	−.551	−.451

Table 3: Sample Average Probability Estimates $\frac{1}{N}\sum_{n=1}^{N}\widehat{\Pr}\left(s=m\middle|\mathbf{x}_n\right)$

m	True	Probability Model						
						Beta–binomial		
		Non-parametric	Linear	Ordered probit	Binomial	BBIN-2	BBIN-1	Stata version
0	.332	.333	.332	.333	.045	.332	.333	.333
1	.088	.086	.088	.089	.161	.121	.122	.125
2	.130	.129	.130	.130	.267	.091	.091	.092
3	.111	.109	.111	.110	.266	.080	.079	.079
4	.081	.081	.081	.080	.171	.076	.075	.074
5	.064	.064	.064	.063	.071	.077	.076	.075
6	.036	.036	.036	.035	.018	.087	.087	.086
7	.159	.156	.159	.159	.002	.136	.137	.137

Table 4: Estimates of Conditional Probability Average Partial Effects, $\frac{1}{N}\sum_{n=1}^{N} D_k \widehat{\Pr}\left(s = m \middle| \mathbf{x}_n\right)$
(Negative point estimates are displayed in red for clearer visualization.)

Covariate	m	Non-Parametric	Ordered Probit	Binomial	Beta-binomial		
					BBIN-2	BBIN-1	Stata Version
Age	0	.0054	.0037	.0015	.0045	.0044	.0038
	1	.00001	.0003	.0033	−.0002	−.00001	.0003
	2	−.0004	.00004	.0021	−.0005	−.0003	−.0001
	3	−.0010	−.0003	−.0011	−.0006	−.0005	−.0002
	4	−.0007	−.0004	−.0028	−.0007	−.0006	−.0003
	5	−.0010	−.0005	−.0021	−.0008	−.0007	−.0005
	6	−.0006	−.0003	−.0007	−.0008	−.0009	−.0008
	7	−.0010	−.0025	−.0001	−.0009	−.0015	−.0022
Female	0	−.0322	−.0035	.0018	−.0243	−.0169	−.0058
	1	.0118	−.0003	.0040	.0088	.0001	−.0004
	2	.0162	−.00004	.0026	.0091	.0013	.0001
	3	.0270	.0003	−.0014	.0084	.0018	.0003
	4	.0104	.0004	−.0035	.0074	.0022	.0005
	5	−.0145	.0005	−.0025	.0058	.0026	.0008
	6	−.0112	.0003	−.0009	.0023	.0033	.0012
	7	−.0008	.0023	−.0001	−.0175	.0056	.0033
NVQ45	0	−.1101	−.0470	−.0147	−.0970	−.0786	−.0485
	1	.0223	−.0040	−.0351	.0176	−.0010	−.0039
	2	.0166	−.0010	−.0251	.0236	.0057	.0006
	3	.0245	.0036	.0108	.0240	.0083	.0026
	4	.0361	.0054	.0312	.0228	.0102	.0043
	5	.0150	.0061	.0233	.0199	.0123	.0063
	6	.0022	.0043	.0083	.0131	.0158	.0103
	7	−.0103	.0326	.0012	−.0240	.0272	.0283
South	0	−.0383	−.0185	−.0067	−.0269	−.0258	−.0201
	1	.0176	−.0014	−.0150	.0021	.0001	−.0013
	2	.0037	−.0002	−.0098	.0038	.0020	.0004
	3	.0029	.0015	.0053	.0043	.0028	.0012
	4	−.0033	.0022	.0130	.0046	.0033	.0018
	5	.0072	.0024	.0094	.0047	.0040	.0026
	6	.0006	.0017	.0033	.0047	.0050	.0042
	7	.0082	.0124	.0005	.0026	.0086	.0112
Urban	0	.0879	.0921	.0313	.0938	.1020	.0955
	1	.0001	.0090	.0807	.0105	.0027	.0096
	2	.0045	.0038	.0674	.0010	−.0065	.0003
	3	−.0190	−.0058	−.0147	−.0039	−.0104	−.0041
	4	−.0092	−.0099	−.0735	−.0080	−.0132	−.0076
	5	−.0016	−.0118	−.0624	−.0129	−.0163	−.0120
	6	−.0005	−.0085	−.0247	−.0215	−.0212	−.0205
	7	−.0611	−.0688	−.0040	−.0591	−.0370	−.0613