arXiv:2403.07236v2 [econ.EM] 27 Mar 2024

Partial Identification of Individual-Level Parameters Using Aggregate Data in a Nonparametric Binary Outcome Model*

Sarah Moon†

March 28, 2024

Abstract

It is well known that the relationship between variables at the individual level can be different from the relationship between those same variables aggregated over individuals. This problem of aggregation becomes relevant when the researcher wants to learn individual-level relationships, but only has access to data that has been aggregated. In this paper, I develop a methodology to partially identify linear combinations of conditional average outcomes from aggregate data when the outcome of interest is binary, while imposing as few restrictions on the underlying data generating process as possible. I construct identified sets using an optimization program that allows for researchers to impose additional shape and data restrictions. I also provide consistency results and construct an inference procedure that is valid with aggregate data, which only provides marginal information about each variable. I apply the methodology to simulated and real-world data sets and find that the estimated identified sets are too wide to be useful, but become narrower as more assumptions are imposed and data aggregated at a finer level is available. This suggests that to obtain useful information from aggregate data sets about individual-level relationships, researchers must impose further assumptions that are carefully justified or seek out data aggregated at the finest level possible.

Keywords: Aggregate data, partial identification, shape restrictions, nonparametric, binary outcome

*I am especially grateful to Edward Vytlacil and Isaiah Andrews for their guidance and advice. I thank Alberto Abadie, Anna Mikusheva, Hyungsik Roger Moon, Elie Tamer, and participants in the MIT econometrics lunch seminar for their helpful comments. All errors are my own.

†Department of Economics, Massachusetts Institute of Technology, sarahmn@mit.edu

1 Introduction

Researchers frequently use publicly available data in policy analysis, which is usually provided at the aggregate level due to individual privacy concerns. For example, statewide standardized exam results are reported in the form of school-wide or school district-wide pass rates, as opposed to pass result and demographic information (e.g. gender, race, family income) for each individual student. Election data is available as vote shares and voter demographics over voting districts, as opposed to the vote and demographics information for each individual voter. However, it is well known that aggregate variables can be related in ways that are different from the same variables at the individual level,[1] a problem known as an ecological fallacy.[2]

In this paper I consider the problem of identifying linear combinations of conditional mean outcomes, $\mathbb{E}[Y_i|X_{1i}, \ldots, X_{Li}]$, when individual-level outcome Y_i is binary and the only data that is observed is the marginal distribution of each individual-level variable over many groups, which I call aggregate data. I develop a partial identification methodology that constructs sharp bounds by solving an optimization problem that considers all underlying joint distributions of individual-level covariates that are consistent with the observed marginal distributions. I also derive closed-form bounds on each conditional mean outcome $\mathbb{E}[Y_i|X_{1i}, \ldots, X_{Li}]$. Finally, I show how further restrictions on the underlying data generating process, like shape restrictions, can be incorporated into the optimization problem to obtain sharp bounds under those restrictions. Since I do not observe individual-level joint distributions, I develop valid inference procedures on the identified set using marginal information only.

To demonstrate how informative these bounds can be, I apply this methodology to several different simulated aggregate data sets, calibrated to a Rhode Island standardized exam score data set used later in an empirical application. I find that bounds are relatively wide on marginal effect parameters of interest. Imposing monotonicity shape restrictions in the empirical application helps narrow bounds on test score gaps, and using additional data at a finer level of aggregation makes bounds on test score gaps even more narrow. This suggests that obtaining useful individual-level results from aggregate data is easier when the data is aggregated at a finer level and when the researcher can impose carefully justified assumptions about the individual-level data generating process.

The ecological fallacy has been studied across many fields over the past several decades, beginning with the seminal work of Robinson (1950), Duncan and Davis (1953), Theil (1954), and Shiveley (1974). These early methods demonstrated issues with interpreting parameters from regressions with aggregate data as individual-level effects, and provided simple bounds on individual-level effects with a binary outcome and single binary covariate. Since then, research has been conducted on consideration of the identification of parameters (Stoker, 1984, 1986), testing for aggregation bias

[1] Firebaugh (1978) notes that "aggregate" and "individual" are relative terms, as individual variables correspond to the individual unit of analysis and need not correspond to an individual person. In the econometrics literature, the terms "macro" and "micro" are also often used.

[2] This problem has also been referred to as the problem of aggregation (Stoker, 1984).

(Lee et al., 1990), empirical illustrations of inconsistent results when using aggregate as opposed to individual-level data (Hsiao et al., 2005), and many other topics in the sciences and social sciences beyond economics.

While the problem of aggregation has been widely acknowledged as a major issue by the literature, to the best of my knowledge much of the literature is concerned with point identification of individual-level parameters (King et al., 1999; Rosen et al., 2001).[3] When only aggregate information is available, knowledge about parameters at the individual level is limited, and point identification is often not achieved without further assumptions. Imposing more assumptions may allow for precise results, but such assumptions may be less plausible. For example, one popular method for analyzing aggregate data is the ecological inference method (King, 1997), used often in political science studies of elections (Burden and Kimball, 1998). This method relies on many assumptions, like imposing individual-level joint distributions and a lack of bias introduced by the aggregation, that often fail to hold in applied settings (Tam Cho, 1998; Cho and Gaines, 2004).

Even without these strong assumptions, partial identification is still possible and may reveal useful insights about the size or range of magnitude of parameters of interest. Partial identification in various contexts has been widely studied in the econometrics literature, especially over the last thirty years.[4] Relevant to this paper is partial identification with data combination, since when combining two different data sets data the joint distribution between the data are unobserved (Cross and Manski, 2002; Molinari and Peski, 2006; Ridder and Moffitt, 2007; Fan et al., 2014, 2016). However in the data combination literature the joint distribution within each data set is observed; in the aggregate data setting I instead assume that the joint distribution between every combination of variables is unobserved.

The rest of the paper proceeds as follows. Section 2 presents identified sets on the parameters of interest under a few different assumptions. Section 3 develops consistent estimation and inference procedures. Section 4 presents and discusses results from an empirical application using standardized exam data and simulation exercises calibrated to the dataset. Section 5 concludes.

2 Identification

2.1 Identified set

Let $(Y_i, X_{1i}, \dots, X_{Li}, G_i), i = 1, \dots, n$ be an i.i.d. sequence of random variables.[5] Suppose outcome Y_i is binary, with $Y_i \in \{0, 1\}$ and suppose $G_i \in \{1, \dots, G\}$ denotes the group of individual i. Further assume covariates $X_i \equiv (X_{1i}, \dots, X_{Li})'$ is discrete with known finite support $\{x_k\}_{k=1}^K \subseteq \mathbb{R}^L$.[6]

The goal is to construct bounds on linear combinations of $\mathbb{E}[Y_i|X_i = x_k]$, $\sum_{k=1}^K \lambda_k \mathbb{E}[Y_i|X_i = x_k]$,

[3] One exception is Jiang et al. (2020), who develop partial identification techniques in a setting with two binary covariates and linear individual-level relationships.

[4] See Tamer (2010), Ho and Rosen (2015), Molinari (2020), and Kline and Tamer (2023) for detailed surveys of partial identification in economics.

[5] I assume individual-level variables are i.i.d. for the sake of simplicity; this assumption can be relaxed and the consistency and inference results of Section 3 will still hold.

[6] The case with multinomial or continuous outcome and continuous covariates is beyond the scope of this paper.

for given weights $\{\lambda_k\}_{k=1}^K$. For example, if we are interested in the average marginal effect of changing X_i from x_{k_1} to x_{k_2} on Y_i, we would choose $\lambda_{k_2} = 1, \lambda_{k_1} = -1$, and $\lambda_k = 0$ for all other k. I will construct identified sets using only expressions for $\mathbb{E}[Y_i|G_i = g], \mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g]$, and $\mathbb{P}[G_i = g]$ for every $\ell = 1, \ldots, L, k = 1, \ldots, K, g = 1, \ldots, G$; the sample equivalents of these parameters are observed in aggregate data.

For example, in a data set of standardized exam results and demographics, G_i denotes student i's school district, Y_i is an indicator for whether student i passed the exam or not, and X_i are student i's demographics. We observe (sample estimates of) the pass rate for every school district $\mathbb{E}[Y_i|G_i]$, demographics of every school district $\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g]$, and the number of students enrolled in each school district, with which we can obtain (sample estimates of) $\mathbb{P}[G_i = g]$.

I summarize the previous statements in the following assumption:

Assumption 1. *For random variables Y_i, G_i, and L-dimensional random vector X_i, suppose*

1. *(Y_i, G_i, X_i) are i.i.d.*
2. *Y_i is binary.*
3. *G_i is discrete with finite support $\{1, \ldots, G\}$.*
4. *X_i is discrete with finite support $\{x_k\}_{k=1}^K \subseteq \mathbb{R}^L$.*
5. *(Y_i, G_i, X_i) are latent; instead, we observe (sample analogs of) $\mathbb{E}[Y_i|G_i = g], \mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g]$, and $\mathbb{P}[G_i = g]$ for every $\ell = 1, \ldots, L, k = 1, \ldots, K, g = 1, \ldots, G$.*

Because Y_i is binary, the law of total probability gives us

$$\mathbb{E}[Y_i|X_i = x_k] = \sum_{g=1}^G \mathbb{P}[G_i = g]\mathbb{E}[Y_i|X_i = x_k|G_i = g],$$

$$\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K \mathbb{E}[Y_i|X_i = x_k, G_i = g]\mathbb{P}[X_i = x_k|G_i = g],$$

$$\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}\mathbb{P}[X_i = x_j|G_i = g].$$

As in Cross and Manski (2002), these are the only relationships we can use to relate joint information of interest to the observed marginal information in the data without any further assumptions.

Let $\delta_k \equiv \mathbb{E}[Y_i|X_i = x_k]$, $\gamma_{kg} \equiv \mathbb{E}[Y_i|X_i = x_k, G_i = g]$, and $\pi_{kg} \equiv \mathbb{P}[X_i = x_k|G_i = g]$ denote the unobserved parameters in the above equations. Then we can rewrite the equations as

$$\delta_k = \sum_{g=1}^G \mathbb{P}[G_i = g]\gamma_{kg}, \tag{1}$$

$$\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K \gamma_{kg}\pi_{kg} \tag{2}$$

$$\mathbb{P}[X_{\ell i} = x_{k,\ell} | G_i = g] = \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} \pi_{jg}. \tag{3}$$

Note that $\delta_k, \gamma_{kg}, \pi_{kg} \in [0,1]$ for all k, g. Combining this with equations (1), (2), and (3) above, we can define the identified set for $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i | X_i = x_k]$:

Lemma 1. *Suppose Assumption 1 holds. Given $\{\lambda_k\}_{k=1}^{K} \in \mathbb{R}^K$, the sharp identified set for $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i | X_i = x_k]$ is given by*

$$D = \Bigg\{ \sum_{k=1}^{K} \lambda_k d_k \;\Bigg|\; 0 \le d_k \le 1 \;\forall k, \text{ and } \exists\, (p_{1g}, \ldots, p_{Kg}) \in [0,1]^K, (c_{1g}, \ldots, c_{Kg}) \in [0,1]^K \;\forall g$$

$$\text{s.t. } d_k = \sum_{g=1}^{G} \mathbb{P}[G_i = g] c_{kg} \;\forall k, \;\; \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg} \;\forall g,$$

$$\mathbb{P}[X_{\ell i} = x_{k,\ell} | G_i = g] = \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} \;\forall \ell, k, g, \text{ and } \sum_{k=1}^{K} p_{kg} = 1 \;\forall g \Bigg\}. \tag{4}$$

Proposition 1. $D = \left[\sum_{g=1}^{G} \mathbb{P}[G_i = g] L_g, \sum_{g=1}^{G} \mathbb{P}[G_i = g] U_g\right]$*, where*

$$L_g \equiv \min_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg},$$

$$U_g \equiv \max_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}, \text{ and}$$

$$P_g \equiv \operatorname*{arg\,min}_{\{p_{kg}\} \in [0,1]^K} \sum_{r=1}^{LK} v_r^+ + v_r^- \;\text{ s.t. } \sum_{k=1}^{K} p_{kg} = 1, \; p_{kg}, v_r^+, v_r^- \ge 0 \;\forall k, r, \text{ and}$$

$$\mathbb{P}[X_{\ell i} = x_{k,\ell} | G_i = g] - \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} = v_{K(\ell-1)+k}^+ - v_{K(\ell-1)+k}^- \;\forall \ell, k.$$

I defer all proofs to Appendix B. Proposition 1 states that we can equivalently express the identified set D given by (4) as the weighted sum of solutions to bilevel optimization problems. This formulation is helpful because it suggests how computation of the lower and upper bound might be performed. In particular, solving for

$$\min_{\{c_{kg}\} \in [0,1]^K} \; (\max)_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg} \tag{5}$$

given a particular $\{p_{kg}\} \in P_g$ is a linear program. Thus we can solve for L_g and U_g by searching for the minimum and maximum respectively of the linear program (5) over all $\{p_{kg}\} \in P_g$. This optimization problem has a nonconvex objective; I suggest using existing derivative-free nonconvex solvers with a coarse grid of starting points to solve the optimization problem.

Remark 1. I show in Appendix A.1 that the problem of solving the linear program (5) given any particular $\{p_{kg}\} \in P_g$ has an analytical solution. While computing the analytical solution for each given $\{p_{kg}\}$ is fast, computing the solution from the linear program formulation is also fast and either method can be used.

Remark 2. Without further restrictions on the underlying data generating process, it will always be the case that $\sum_{g=1}^{G} \mathbb{P}[G_i = g] \sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|G_i = g] \in D$. To see why, note that, inspecting the optimization problems of Proposition 1, for any g and any $\{p_{kg}\} \in P_g$, letting $c_{kg} = \mathbb{E}[Y_i|G_i = g]$ for all $k = 1, \ldots, K$ satisfies the constraint that $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}$ because $\sum_{k=1}^{K} p_{kg} = 1$. This is relevant for average marginal effect parameters because the weights $\{\lambda_k\}$ are such that $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|G_i = g] = 0$ for each g, and we thus cannot rule out a zero average marginal effect.

2.2 Additional shape restrictions

There may be situations in which the researcher is willing to assume polyhedral shape restrictions on the conditional expectation function over groups $\mathbb{E}[Y_i|X_i, G_i]$. Examples of such shape restrictions include convexity, concavity, and monotonicity.[7]

I impose the shape constraints as an additional assumption:

Assumption 2. *For each $g = 1, \ldots, G$,*

$$S_g Y_{X,g} \leq a_g,$$

where $Y_{X,g} \equiv (\mathbb{E}[Y_i|X_i = x_1, G_i = g], \ldots, \mathbb{E}[Y_i|X_i = x_K, G_i = g])'$, $S_g \in \mathbb{R}^{s_g \times K}$ are known fixed matrices, and a_g are known fixed vectors.

We can simply add this shape restriction to the constraints of the optimization problems solving L_g and U_g in Proposition 1, as in Freyberger and Horowitz (2015), to obtain sharp bounds on $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|X_i = x_k]$:

Proposition 2. *Suppose Assumptions 1 and 2 hold. Given $\{\lambda_k\}_{k=1}^{K} \in \mathbb{R}^K$ and $S_g \in \mathbb{R}^{s_g \times K}$ for each g, the sharp identified set for $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|X_i = x_k]$ is given by*

$$D = \left[\sum_{g=1}^{G} \mathbb{P}[G_i = g] L_g, \sum_{g=1}^{G} \mathbb{P}[G_i = g] U_g \right],$$

where, defining $c_g \equiv (c_{1g}, \ldots, c_{Kg})'$,

$$L_g \equiv \min_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } S_g c_g \leq a_g \text{ and } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg},$$

[7] See the reviews written by Matzkin (1994) and Chetverikov et al. (2018) for other examples of shape restrictions that have been used in econometric models.

$$U_g \equiv \max_{\{c_{kg}\}\in[0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } S_g c_g \leq a_g \text{ and } \exists\{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg},$$

$$\begin{aligned} P_g \equiv \operatorname*{arg\,min}_{\{p_{kg}\}\in[0,1]^K} \sum_{r=1}^{LK} v_r^+ + v_r^- \ \text{ s.t. } \sum_{k=1}^{K} p_{kg} = 1,\ p_{kg}, v_r^+, v_r^- \geq 0\ \forall k, r, \text{ and} \\ \mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] - \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} = v^+_{K(\ell-1)+k} - v^-_{K(\ell-1)+k}\ \forall \ell, k. \end{aligned}$$

Again, given a particular $\{p_{kg}\} \in P_g$ the solution to

$$\min/\max_{\{c_{kg}\}\in[0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } S_g c_g \leq a_g \text{ and } \exists\{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg} \tag{6}$$

is a linear program. We can again solve for L_g and U_g by searching for the minimum and maximum respectively of (6), which is fast to compute, over all $\{p_{kg}\} \in P_g$. The suggested method of using a nonconvex solver with a coarse grid of starting points to solve the problem is still valid.

2.3 Additional aggregate data at a finer level

In some situations the research has access to additional data that is aggregated at a finer level than by groups G_i. In this section I will consider the case when $\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$ is observed by the researcher. For example, in a data set of standardized exam results and demographics, the researcher may have access to average pass results by race in some school districts, as long as the population of students of that racial identity in the district is large enough to avoid the risk of loss of privacy.

Assumption 3. *We observe (sample analogs of)* $\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$ *for all* $(\ell, k) \in F_g$*, where* F_g *is a (possibly empty) set of indices for each group* $G_i = g$*.*

By the definition of conditional probability and the law of total probability, for each $(\ell, k) \in F_g$,

$$\begin{aligned} &\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] \\ &= \mathbb{P}[Y_i = 1, X_{\ell i} = x_{k,\ell}|G_i = g] \\ &= \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g] \\ &= \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}\mathbb{E}[Y_i|X_i = x_k, G_i = g]\mathbb{P}[X_i = x_k|G_i = g] \\ &= \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}\gamma_{kg}\pi_{kg}. \end{aligned}$$

Again, we can easily add these restrictions to the constraints of the optimization problem solving L_g and U_g in Proposition 1 to obtain sharp bounds on $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|X_i = x_k]$:

Proposition 3. *Suppose Assumptions 1 and 3 hold. Given $\{\lambda_k\}_{k=1}^K \in \mathbb{R}^K$, the sharp identified set for $\sum_{k=1}^K \lambda_k \mathbb{E}[Y_i|X_i = x_k]$ is given by*

$$D = \left[\sum_{g=1}^G \mathbb{P}[G_i = g]L_g, \sum_{g=1}^G \mathbb{P}[G_i = g]U_g\right],$$

where

$$\begin{aligned}
L_g \equiv \min_{\{c_{kg}\}\in[0,1]^K} \sum_{k=1}^K \lambda_k c_{kg} \text{ s.t. } \exists\{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg}p_{kg} \text{ and}\\
\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c_{kg}p_{kg} \ \forall(\ell, k) \in F_g,\\
U_g \equiv \max_{\{c_{kg}\}\in[0,1]^K} \sum_{k=1}^K \lambda_k c_{kg} \text{ s.t. } \exists\{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg}p_{kg} \text{ and}\\
\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c_{kg}p_{kg} \ \forall(\ell, k) \in F_g,
\end{aligned}$$

$$\begin{aligned}
P_g \equiv \operatorname*{arg\,min}_{\{p_{kg}\}\in[0,1]^K} \sum_{r=1}^{LK} v_r^+ + v_r^- \text{ s.t. } \sum_{k=1}^K p_{kg} = 1,\ p_{kg}, v_r^+, v_r^- \geq 0 \ \forall k, r, \text{ and}\\
\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] - \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_{jg} = v_{K(\ell-1)+k}^+ - v_{K(\ell-1)+k}^- \ \forall \ell, k.
\end{aligned}$$

Since the additional constraints are all linear constraints, the solution to L_g and U_g for a particular $\{p_{kg}\} \in P_g$ is still a linear program. The previous discussion about how to compute the identified set applies here.

Remark 3. The sharp identified set D under Assumptions 1, 2, and 3 is given by adding both of the restrictions $S_g c_g \leq a_g$ and

$$\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c_{kg}p_{kg} \ \forall(\ell, k) \in F_g$$

to the optimization problems of L_g and U_g in Proposition 1.

Remark 4. This methodology is flexible and can easily be adjusted to accommodate further assumptions beyond Assumptions 2 and 3 on the underlying individual-level model through addition restrictions to the optimization problem. For example, if some of the covariances between covariates can be estimated or bounded using a separate data set, this could be incorporated as polyhedral restrictions on the joint support of the covariates in the P_g optimization problem. Restrictions on the underlying distribution of X_i can also be incorporated through specification of the support.

Remark 5. If we impose the assumption that the $Y_i, X_{1i}, \ldots, X_{Li}$ are mutually independent conditional on G_i we would obtain point identification of $\mathbb{E}[Y_i|X_i = x_k]$ and thus point identification of $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|X_i = x_k]$. If we only impose that the $X_{1i}, \ldots, X_{Li}$ were mutually independent conditional on G_i we would obtain point identification of the joint distribution of $X_i|G_i$. Sharp bounds on $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i|X_i = x_k]$ would then follow from Cross and Manski (2002): we can consider the L_g and U_g optimization problems with p_{kg} the joint distribution of X_i.

3 Estimation and Inference

3.1 Estimation

In practice, we observe sample analogs of the population values $\mathbb{E}[Y_i|G_i = g], \mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g], \mathbb{P}[G_i = g]$ in the aggregate data set, along with sample analogs of $\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$ if available. For all $\ell = 1, \ldots, L, j = 1, \ldots, K, g = 1, \ldots, G$, denote

$$\bar{Y}_g = \frac{1}{n} \sum_{i=1}^{n} Y_i \mathbb{1}\{G_i = g\} \tag{7}$$

$$\widehat{Pr}[X_{\ell i} = x_{j,\ell}|G_i = g] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_{\ell i} = x_{j,\ell}\} \mathbb{1}\{G_i = g\} \tag{8}$$

$$\widehat{Pr}[G_i = g] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{G_i = g\} \tag{9}$$

$$\bar{Y}_g|X_{\ell i} = x_{k,\ell} = \frac{1}{n} \sum_{i=1}^{n} Y_i \mathbb{1}\{X_{\ell i} = x_{k,\ell}\} \mathbb{1}\{G_i = g\}. \tag{10}$$

$\bar{Y}_g, \widehat{Pr}[X_{\ell i} = x_{j,\ell}], \widehat{Pr}[G_i = g]$, and $\bar{Y}_g|X_{\ell i} = x_{k,\ell}$ all converge to the respective population values by the law of large numbers. Thus we can construct a plug-in estimator, denote $\hat{D}$, for the sharp identified set by replacing all population values in the optimization problems of Propositions 1, 2, 3, or Remark 3 with their sample estimates. For example, the estimated sharp identified set discussed in Remark 3 looks like:

$$\hat{D} \equiv \left[\hat{L}, \hat{U}\right] \equiv \left[\sum_{g=1}^{G} \widehat{Pr}[G_i = g]\hat{L}_g, \sum_{g=1}^{G} \widehat{Pr}[G_i = g]\hat{U}_g\right],$$

$$\hat{L}_g \equiv \min_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } S_g c_g \leq a_g \ \forall g \text{ and } \exists \{p_{kg}\} \in \hat{P}_g \text{ with } \bar{Y}_g = \sum_{k=1}^{K} c_{kg} p_{kg},$$

$$\bar{Y}_g|X_{\ell i} = x_{k,\ell} \times \widehat{Pr}[X_{\ell i} = x_{j,\ell}|G_i = g] = \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_{kg} p_{kg} \ \forall (\ell, k) \in F_g,$$

$$\hat{U}_g \equiv \max_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } S_g c_g \leq a_g \ \forall g \text{ and } \exists \{p_{kg}\} \in \hat{P}_g \text{ with } \bar{Y}_g = \sum_{k=1}^{K} c_{kg} p_{kg},$$

$$\bar{Y}_g|X_{\ell i} = x_{k,\ell} \times \widehat{Pr}[X_{\ell i} = x_{j,\ell}|G_i = g] = \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_{kg} p_{kg} \;\forall(\ell, k) \in F_g,$$

$$\hat{P}_g \equiv \arg\min_{\{p_{kg}\}} \sum_{r=1}^{LK} v_r^+ + v_r^- \;\text{ s.t.} \sum_{k=1}^{K} p_{kg} = 1; p_{kg}, v_r^+, v_r^- \geq 0 \;\forall k, r;\; \text{and}$$
$$\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g] - \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} = v^+_{K(\ell-1)+k} - v^-_{K(\ell-1)+k} \;\forall \ell, k.$$

The following proposition shows that the lower and upper bounds of the plug-in estimated set $\hat{D}$ are consistent.

Proposition 4. *Suppose Assumption 1 holds and that $\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]$ are valid marginal distributions for X_i with respect to the assumed support. If Assumption 2 holds, define $\hat{D}$ with respect to Proposition 2; if Assumption 3 holds, define $\hat{D}$ with respect to Proposition 3; and if Assumptions 2 and 3 hold, define $\hat{D}$ with respect to Remark 3.*

Then conditional on $\hat{D}$ being nonempty, $\hat{L}_g \xrightarrow{p} L_g$ and $\hat{U}_g \xrightarrow{p} U_g$ as $n \to \infty$ for all g. Furthermore, the lower and upper bounds of $\hat{D}$ converge to the lower and upper bounds of D.

3.2 Inference

Existing inference methods for partially identified estimated sets usually require knowledge of the joint distribution of the individual-level data, to estimate a covariance matrix used in constructing critical values or test statistics for valid coverage. However, in my setting I only observe marginal distributions of each variable. I point out why existing methods cannot be applied for a few, by no means representative, examples below:

Example. Horowitz and Manski (2000) derive analytic bounds on conditional mean outcomes and point out that the delta method delivers asymptotic normality of the lower and upper bound estimators. The paper bootstraps the asymptotic covariance matrix to obtain a confidence interval that contains the identified set with correct asymptotic coverage. Putting aside that I do not have analytic bounds in my setting, being able to derive the asymptotic covariance of bounds $\hat{L}_g$ and $\hat{U}_g$ requires that I know the covariances between, for example, Y_i and any $X_{\ell i}$ or any $X_{\ell_1 i}$ and $X_{\ell_2 i}$ given G_i. However I only observe sample marginal distributions of each variable and thus cannot hope to estimate the covariance matrix. Bootstrapping will also not be possible because I do not observe the individual-level data and so cannot generate a bootstrap sample that reflects the dependence between all of the variables.

Example. Imbens and Manski (2004) provide a method for constructing confidence intervals on the parameter value of interest instead of on the entire identified set. This method chooses critical values for correct coverage by again relying on joint asymptotic normality of the lower and upper bound estimators, $\hat{L}_g$ and $\hat{U}_g$ in my setting. As discussed in the above example, I cannot hope to estimate the variances of the lower and upper bounds with the marginal information observed in aggregate data alone.

Example. Hsieh et al. (2022) construct confidence intervals for identified sets of solutions to convex optimization programs, specifically linear and quadratic optimization programs with estimated coefficients, exploiting the necessary and sufficient optimality conditions. Putting aside that the optimization program is nonconvex in my setting, implementation of this inference method requires the asymptotic covariance matrix of the estimated covariates of the optimization problem, which in my setting rely on $\bar{Y}_g, \widehat{Pr}[X_{\ell i} = x_{j,\ell}|G_i = g]$, and $\widehat{Pr}[G_i = g]$. Thus I again need to know the covariances between, for example, Y_i and any $X_{\ell i}$ or any $X_{\ell_1 i}$ and $X_{\ell_2 i}$ given G_i.

Therefore in this setting I am forced to rely on inference methods that require only marginal information of each variable. I choose to use the Bonferroni correction to make marginal confidence intervals on each sample observation jointly valid across the whole sample. On the other hand, note that each aggregate observation is the sample average of a binary random variable, as can be seen from equations (7), (8), (9), and (10). Thus I can use finite-sample valid binomial proportion confidence intervals for each sample observation, instead of relying on normal approximations. I choose to use Clopper-Pearson intervals to construct marginal confidence intervals on each observation.

Suppose we also observe n in the data set, where n is the number of individuals across all groups over which we aggregate to obtain the aggregate data. Let M be the total number of observations in the aggregate data set, that is, the total number of $\bar{Y}_g, \widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g], \widehat{Pr}[G_i = g]$, and $\bar{Y}_g|X_{\ell i} = x_{k,\ell}$ (if observed) observations in the data sets across all groups g, support points k and covariates ℓ.

The inference procedure is as follows:

1. For every sample observation $\hat{p}$, construct two-sided level $1 - \frac{\alpha}{M}$ Clopper-Pearson CIs, denoted $[p_L, p_U]$. For $X = N\hat{p}$, the Clopper-Pearson CI is determined by quantiles of the beta distribution:

$$[p_L, p_U] = \left[B\left(\frac{\alpha}{2M}, X, N - X + 1\right), B\left(1 - \frac{\alpha}{2M}, X + 1, N - X\right)\right]$$

The resulting confidence intervals are $\left[\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]_L, \widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]_U\right]$ for each $\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]$, $\left[\bar{Y}_{g,L}, \bar{Y}_{g,U}\right]$ for each $\bar{Y}_g$, $\left[\widehat{Pr}[G_i = g]_L, \widehat{Pr}[G_i = g]_U\right]$ for each $\widehat{Pr}[G_i = g]$, and (if observed) $\left[\bar{Y}_g|X_{\ell i} = {x_{k,\ell}}_L, \bar{Y}_g|X_{\ell i} = {x_{k,\ell}}_U\right]$ for each $\bar{Y}_g|X_{\ell i} = x_{k,\ell}$.

2. Solve the optimization programs of $\hat{D}$ for all values of sample observations within the marginal confidence intervals constructed in step 1. For example, under both Assumptions 1, 2, and 3, we solve:

$$a) \quad \hat{P}_{g,CI} \equiv \arg\min_{\{p_{kg}\}} \sum_{r=1}^{2LK} v_r^+ + v_r^- \text{ s.t.} \sum_{k=1}^{K} p_{kg} = 1; p_{kg}, v_r^+, v_r^- \geq 0 \ \forall k, r;$$

$$\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]_L - \sum_{k=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} \leq v^+_{K(\ell-1)+k} - v^-_{K(\ell-1)+k} \ \forall \ell, k, \text{ and}$$

$$\widehat{Pr}[X_{\ell i} = x_{k,\ell} | G_i = g]_U - \sum_{k=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} \geq v^+_{K(L+\ell-1)+k} - v^-_{K(L+\ell-1)+k} \; \forall \ell, k.$$

b) $$\hat{L}_{g,CI} \equiv \min_{\{c_{gk}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{gk} \text{ s.t. } S_g c_g \leq a_g \; \forall g \text{ and } \exists \{p_{gk}\} \in \hat{P}_{g,CI}$$
$$\text{with } \bar{Y}_{g,L} \leq \sum_{k=1}^{K} c_{gk} p_{gk}, \bar{Y}_{g,U} \geq \sum_{k=1}^{K} c_{gk} p_{gk},$$
$$\bar{Y}_g | X_{\ell i} = x_{k,\ell_L} \times \widehat{Pr}[X_{\ell i} = x_{j,\ell} | G_i = g]_L \leq \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_{kg} p_{kg} \; \forall (\ell, k) \in F_g,$$
$$\text{and } \bar{Y}_g | X_{\ell i} = x_{k,\ell_U} \times \widehat{Pr}[X_{\ell i} = x_{j,\ell} | G_i = g]_U \geq \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_{kg} p_{kg} \; \forall (\ell, k) \in F_g$$

c) $$\hat{U}_{g,CI} \equiv \max_{\{c_{gk}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{gk} \text{ s.t. } S_g c_g \leq 0 \; \forall g \text{ and } \exists \{p_{gk}\} \in \hat{P}_{g,CI}$$
$$\text{with } \bar{Y}_{g,L} \leq \sum_{k=1}^{K} c_{gk} p_{gk}, \bar{Y}_{g,U} \geq \sum_{k=1}^{K} c_{gk} p_{gk},$$
$$\bar{Y}_g | X_{\ell i} = x_{k,\ell_L} \times \widehat{Pr}[X_{\ell i} = x_{j,\ell} | G_i = g]_L \leq \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_{kg} p_{kg} \; \forall (\ell, k) \in F_g,$$
$$\text{and } \bar{Y}_g | X_{\ell i} = x_{k,\ell_U} \times \widehat{Pr}[X_{\ell i} = x_{j,\ell} | G_i = g]_U \geq \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_{kg} p_{kg} \; \forall (\ell, k) \in F_g.$$

3. The confidence interval is given by

$$\hat{D}_{CI} \equiv \left[\sum_{g=1}^{G} \widehat{Pr}[G_i = g]_L \hat{L}_{g,CI}, \sum_{g=1}^{G} \widehat{Pr}[G_i = g]_U \hat{U}_{g,CI} \right].$$

Proposition 5. *Suppose Assumption 1 holds. If Assumption 2 holds, define $\hat{D}_{CI}$ with respect to Proposition 2; if Assumption 3 holds, define $\hat{D}_{CI}$ with respect to Proposition 3; and if Assumptions 2 and 3 hold, define $\hat{D}_{CI}$ with respect to Remark 3.*

Suppose also that the $\widehat{Pr}[X_{\ell i} = x_{k,\ell} | G_i = g]$ are valid marginal distributions for X_i with respect to the assumed support. Then $\mathbb{P}[D \subseteq \hat{D}_{CI}] \geq 1 - \alpha$.

Since $\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i | X_i = x_k] \in D$, this means that $\mathbb{P}[\sum_{k=1}^{K} \lambda_k \mathbb{E}[Y_i | X_i = x_k] \in \hat{D}_{CI}] \geq 1 - \alpha$, as discussed in Imbens and Manski (2004). Proposition 5 says that the confidence interval $\hat{D}_{CI}$ has correct coverage for the identified set and thus for the identified parameter.

4 Simulations and Empirical Application

One setting in which publicly available data is in aggregate form is standardized exam data. In this section I will apply the methodology developed in the previous sections to construct bounds on conditional exam pass rates. In this application I focus on exam pass rates for English and math RICAS exams and student demographic information for the state of Rhode Island in spring of 2018 and spring of 2019 over all students in grades 3-8.[8]

In my application I produce sharp bounds on the average pass rate conditional on three covariates: race (white versus non-white), eligibility for free and reduced price lunch (FRPL), and English-language learner (ELL) status. In Section 4.1 I first explore what causes the width of the bounds to vary in simulations calibrated to the Rhode Island aggregate data. In Section 4.2 I then present the empirical application, where I estimate bounds with and without monotonicity shape restrictions and additional pass rate data for subgroups.

4.1 Simulations

I present three different simulation exercises. In all exercises I have three binary covariates $white_i, FRPL_i$, and ELL_i, and a binary outcome $pass_i$. There are 58 aggregate groups in each example, where in the Rhode Island data a group is a school district-year pair (each school district in a given year), and in all simulation exercises I assume there are 2000 individuals in each group.

In the first exercises, I choose a joint distribution such that the marginal distribution over groups of each covariate approximately matches the marginal distribution over groups of each aggregate-level covariate in the Rhode Island data, as plotted in Figure 1.

Figure 1: Sample distributions of aggregate variables in simulation exercise 1

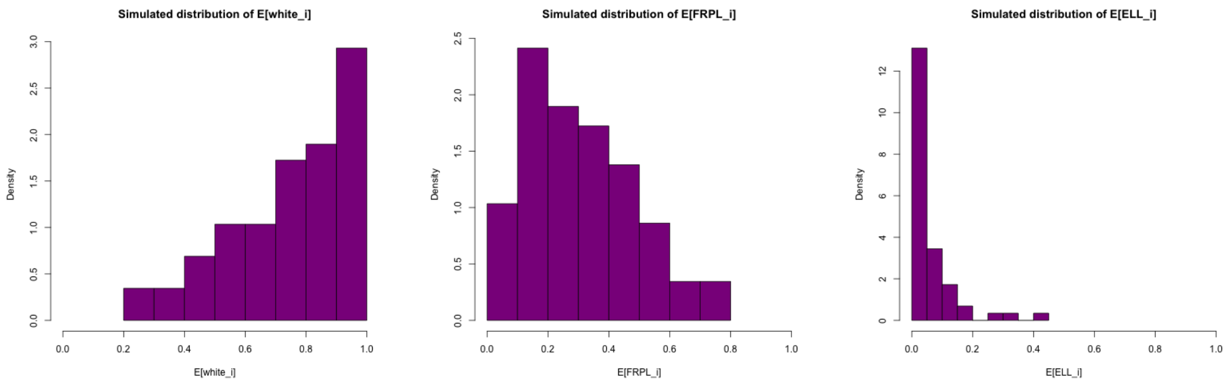


Table 1 presents results from the first simulation exercise. Estimated bounds in Column 2 are incredibly wide, with most of them being uninformative (equal to $[0, 1]$). The 95% confidence intervals presented in Column 3 are also wide, but not too much wider than the estimated bounds themselves. The parameters for which I obtain informative bounds seem to be those where the conditioning population is well-represented in the data: the simulated data has lots of groups with a high fraction of white students and low fractions of FRPL and ELL students, and the parameter for which I obtain the tightest bounds is the average pass rate among white, non-FRPL, non-ELL students. Sharp bounds on the difference between parameters are very similar to the Minkowski set difference between bounds on each of the parameters. In particular, as noted in Remark 2 the

[8] Much of this data is a subset of data from Jack et al. (2023) in their paper studying the effect of online learning during the COVID-19 pandemic on exam results. I thank the authors for their assistance with obtaining the data.

bounds contain 0. In fact, Remark 2 states that all bounds on the difference between parameters will contain 0 as I do not impose additional assumptions.

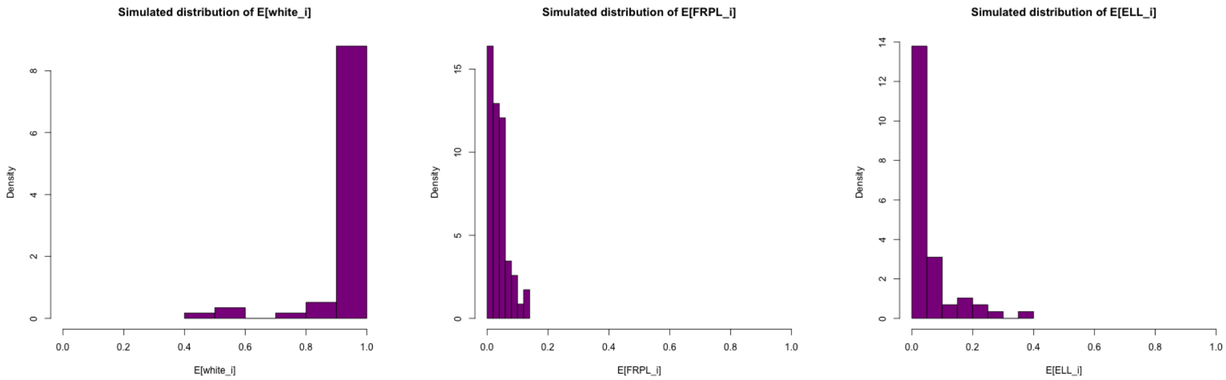
Table 1: Estimated bounds on conditional pass rate

	True Value	Estimated Bounds	95% CI	Bounds on Difference
Parameter	(1)	(2)	(3)	(4)
$\mathbb{E}\left[pass_i \middle| white_i = 1, FRPL_i = 0, ELL_i = 0\right]$	0.618	$[0.180, 0.860]$	$[0.173, 0.874]$	$[-0.820, 0.860]$
$\mathbb{E}\left[pass_i \middle| white_i = 0, FRPL_i = 0, ELL_i = 0\right]$	0.5	$[0, 1]$	$[0, 1]$	
$\mathbb{E}\left[pass_i \middle| white_i = 1, FRPL_i = 1, ELL_i = 0\right]$	0.274	$[0, 0.981]$	$[0, 0.981]$	$[-0.997, 0.981]$
$\mathbb{E}\left[pass_i \middle| white_i = 0, FRPL_i = 1, ELL_i = 0\right]$	0.184	$[0, 0.997]$	$[0, 1]$	
$\mathbb{E}\left[pass_i \middle| white_i = 1, FRPL_i = 0, ELL_i = 1\right]$	0.460	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}\left[pass_i \middle| white_i = 0, FRPL_i = 0, ELL_i = 1\right]$	0.345	$[0, 1]$	$[0, 1]$	
$\mathbb{E}\left[pass_i \middle| white_i = 1, FRPL_i = 1, ELL_i = 1\right]$	0.159	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}\left[pass_i \middle| white_i = 0, FRPL_i = 1, ELL_i = 1\right]$	0.097	$[0, 1]$	$[0, 1]$	

Notes: 95% CIs are confidence intervals on the estimated bounds in column 2. CIs are constructed using the method from Section 3.2, taking $\mathbb{P}[G_i = g]$ as observed instead of $\widehat{Pr}[G_i = g]$. Bounds on the difference are sharp bounds on the top minus the bottom parameter for which each set of bounds are reported.

To see if the estimated bounds could be narrower, in the second simulation exercise I choose a joint distribution that produces the same true conditional average pass rates, but such that the marginal distribution over groups of each covariate is even closer to either 0 or 1, as can be seen in Figure 2. Results are presented in Table 2. In Column 2, bounds on the parameter for which the conditioning population is well-represented in the data are narrower and more informative than in the first exercise, but bounds on all other parameters are uninformative, while in the first exercise some of the bounds on other parameters were informative. However, bounds on the difference in the first row of Column 4 are still wide, even though they are narrower than in the first exercise. This is likely because while there is more information about the first parameter, there is less information on the second parameter so that bounds on the difference are still wide.

Figure 2: Sample distributions of aggregate variables in simulation exercise 2



This suggests that if the conditioning populations for the first two parameters were well-represented, bounds on the difference might be narrower. In the third simulation exercise I choose a joint distribution that produces the same true conditional average pass rates and marginal distributions over groups for $\mathbb{E}[FRPL_i]$ and $\mathbb{E}[ELL_i]$, but I let some groups have $\mathbb{E}[white_i]$ close to 0

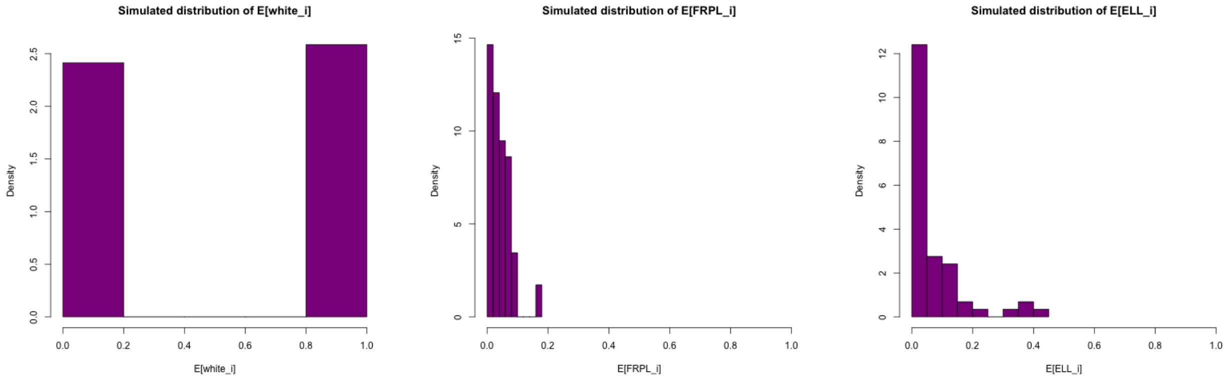
Table 2: Estimated bounds on conditional pass rate

	True Value	Estimated Bounds	95% CI	Bounds on Difference
Parameter	(1)	(2)	(3)	(4)
$\mathbb{E}\left[pass_i \middle\vert white_i = 1, FRPL_i = 0, ELL_i = 0\right]$	0.618	$[0.572, 0.749]$	$[0.564, 0.759]$	$[-0.428, 0.749]$
$\mathbb{E}\left[pass_i \middle\vert white_i = 0, FRPL_i = 0, ELL_i = 0\right]$	0.5	$[0, 1]$	$[0, 1]$	
$\mathbb{E}\left[pass_i \middle\vert white_i = 1, FRPL_i = 1, ELL_i = 0\right]$	0.274	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}\left[pass_i \middle\vert white_i = 0, FRPL_i = 1, ELL_i = 0\right]$	0.184	$[0, 1]$	$[0, 1]$	
$\mathbb{E}\left[pass_i \middle\vert white_i = 1, FRPL_i = 0, ELL_i = 1\right]$	0.460	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}\left[pass_i \middle\vert white_i = 0, FRPL_i = 0, ELL_i = 1\right]$	0.345	$[0, 1]$	$[0, 1]$	
$\mathbb{E}\left[pass_i \middle\vert white_i = 1, FRPL_i = 1, ELL_i = 1\right]$	0.159	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}\left[pass_i \middle\vert white_i = 0, FRPL_i = 1, ELL_i = 1\right]$	0.097	$[0, 1]$	$[0, 1]$	

Notes: 95% CIs are confidence intervals on the estimated bounds in column 2. CIs are constructed using the method from Section 3.2, taking $\mathbb{P}[G_i = g]$ as observed instead of $\widehat{Pr}[G_i = g]$. Bounds on the difference are sharp bounds on the top minus the bottom parameter for which each set of bounds are reported.

and others have $\mathbb{E}[white_i]$ close to 1, as can be seen in Figure 3. Results are presented in Table 3. As expected, bounds on each parameter in Column 3 are relatively narrow on the parameters for which the conditioning population is well-represented in the data. Notably relative to the second exercise, bounds on the first parameter are wider and bounds on the second parameter are narrower. This suggests that there is a trade-off between obtaining narrow bounds on a single parameter and obtaining narrow bounds on multiple parameters. Loosely, the groups with $\mathbb{E}[white_i]$ close to 0 help to make bounds on the second parameter narrow but make bounds on the first parameter wider.

Figure 3: Sample distributions of aggregate variables in simulation exercise 3



The bounds on the difference in the first row of Column 4 are narrower than in the first or the second exercise, but are still wide and, important to signing the parameter, contain 0. This means that, from this simulated data set, it would not be possible to say whether the white/non-white average pass rate gap among any group of FRPL/non-FRPL and ELL/non-ELL students is positive or negative. This suggests that without further restrictions, obtaining usefully informative bounds on average marginal effects is challenging.

Table 3: Estimated bounds on conditional pass rate

	True Value	Estimated Bounds	95% CI	Bounds on Difference
Parameter	(1)	(2)	(3)	(4)
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 0, ELL_i = 0]$	0.618	$[0.316, 0.860]$	$[0.312, 0.865]$	$[-0.222, 0.860]$
$\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 0, ELL_i = 0]$	0.5	$[0, 0.539]$	$[0, 0.540]$	
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 1, ELL_i = 0]$	0.274	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 1, ELL_i = 0]$	0.184	$[0, 1]$	$[0, 1]$	
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 0, ELL_i = 1]$	0.460	$[0, 1]$	$[0, 1]$	$[-0.897, 1]$
$\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 0, ELL_i = 1]$	0.345	$[0, 0.897]$	$[0, 0.911]$	
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 1, ELL_i = 1]$	0.159	$[0, 1]$	$[0, 1]$	$[-1, 1]$
$\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 1, ELL_i = 1]$	0.097	$[0, 1]$	$[0, 1]$	

Notes: 95% CIs are confidence intervals on the estimated bounds in column 2. CIs are constructed using the method from Section 3.2, taking $\mathbb{P}[G_i = g]$ as observed instead of $\widehat{Pr}[G_i = g]$. Bounds on the difference are sharp bounds on the top minus the bottom parameter for which each set of bounds are reported.

4.2 Empirical application

The results of the simulation exercise suggest that bounds on white/non-white average pass rate gaps will not be informative in the Rhode Island data. I thus consider imposing several additional assumptions to see whether more restrictions and information can help make bounds narrower.

4.2.1 Monotonicity restrictions

I first consider imposing several monotonicity shape restrictions. Motivated by test score gaps that have been documented between rich and poor students (Tavernise, 2012; Porter, 2015), I impose that for each value of $(white_i, ELL_i)$ and each the average pass rate is lower for FRPL students than for non-FRPL students:

$$\mathbb{E}[pass_i | FRPL_i = 1, white_i, ELL_i, G_i = g] - \mathbb{E}[pass_i | FRPL_i = 0, white_i, ELL_i, G_i = g] \leq 0. \quad (11)$$

I first present estimated bounds on math exam white/non-white average pass rate gaps in Table 4. As expected, sharp bounds on the white/non-white average pass rate gaps reported in Column 1 are wide and either uninformative or close to uninformative. Imposing the additional monotonicity restriction (11) helps narrow the bounds for some parameters, reported in Column 2, but bounds are still wide and contain 0.

For English exams, I impose an additional monotonicity restriction that for each value of $(white_i, FRPL_i)$ the average pass rate is lower for English-language learner students than for non-English-language learner students:

$$\mathbb{E}[pass_i | ELL_i = 1, white_i, FRPL_i, G_i = g] - \mathbb{E}[pass_i | ELL_i = 0, white_i, FRPL_i, G_i = g] \leq 0. \quad (12)$$

I present estimated bounds on English exam white/non-white average pass rate gaps in Table 5. Sharp bounds on the white/non-white average pass rate gaps reported in Column 1 are again wide and either uninformative or close to uninformative. Imposing both additional monotonicity

Table 4: Rhode Island white/non-white math exam pass rate differences

	Bounds without Monotonicity	Monotonicity Bounds
Parameter	(1)	(2)
$\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 0, ELL_i = 0]$ $-\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 0, ELL_i = 0]$	$[-0.940, 0.850]$	$[-0.872, 0.850]$
$\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 1, ELL_i = 0]$ $-\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 1, ELL_i = 0]$	$[-0.821, 0.994]$	$[-0.691, 0.596]$
$\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 0, ELL_i = 1]$ $-\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 0, ELL_i = 1]$	$[-1, 1]$	$[-1, 1]$
$\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 1, ELL_i = 1]$ $-\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 1, ELL_i = 1]$	$[-1, 1]$	$[-1, 1]$

Notes: Bounds on difference impose no shape restrictions; monotonicity bounds impose monotonicity restriction (11), as discussed in Section 4.2.

restrictions of (11) and (12) helps to narrow the bounds on all parameters, reported in Column 2, but all bounds are still wide and contain 0.

Table 5: Rhode Island white/non-white English exam pass rate differences

	Bounds without Monotonicity	Monotonicity Bounds
Parameter	(1)	(2)
$\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 0, ELL_i = 0]$ $-\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 0, ELL_i = 0]$	$[-0.904, 0.923]$	$[-0.823, 0.923]$
$\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 1, ELL_i = 0]$ $-\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 1, ELL_i = 0]$	$[-0.852, 1]$	$[-0.759, 0.706]$
$\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 0, ELL_i = 1]$ $-\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 0, ELL_i = 1]$	$[-1, 1]$	$[-1, 0.914]$
$\mathbb{E}[pass_i \| white_i = 1, FRPL_i = 1, ELL_i = 1]$ $-\mathbb{E}[pass_i \| white_i = 0, FRPL_i = 1, ELL_i = 1]$	$[-1, 1]$	$[-0.759, 0.706]$

Notes: Bounds on difference impose no shape restrictions; monotonicity bounds impose monotonicity restrictions (11) and (12), as discussed in Section 4.2.

4.2.2 Additional pass rates among subgroups

Rhode Island makes available RICAS assessment results aggregated by student subgroup publicly available on their data portal. In particular, exam pass rates at the district level for white/nonwhite students are available for all school districts.

I present bounds using the additional subgroup data on the math exam white/non-white average pass rate gaps in Table 6. Bounds using the additional subgroup data on the English exam white/non-white average pass rate gaps are presented in Table 7. I present bounds using the additional data both without and with the monotonicity assumptions of Section 4.2.1.

Without the monotonicity assumptions, using the additional subgroup data makes the estimated bounds slightly narrower than without the additional subgroup data, as we can see from comparing Column 1 of Tables 4 and 6 and comparing Column 1 of Tables 5 and 7. Bounds under the monotonicity assumptions are also narrower with the additional subgroup data than without, as

Table 6: Rhode Island white/non-white math exam pass rate differences, with subgroup data

	Bounds without Monotonicity	Monotonicity Bounds
Parameter	(1)	(2)
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 0, ELL_i = 0]$ $-\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 0, ELL_i = 0]$	$[-0.852, 0.746]$	$[-0.772, 0.654]$
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 1, ELL_i = 0]$ $-\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 1, ELL_i = 0]$	$[-0.760, 0.977]$	$[-0.269, 0.499]$
$\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 0, ELL_i = 1]$ $-\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 0, ELL_i = 1]$	$[-1, 1]$	$[-1, 1]$
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 1, ELL_i = 1]$ $-\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 1, ELL_i = 1]$	$[-1, 0.887]$	$[-1, 0.887]$

Notes: Bounds on difference impose no shape restrictions but use additional pass rate by subgroup data; monotonicity bounds impose monotonicity restriction (11) in additional to additional pass rate by subgroup data. See Section 4.2 for more details.

we can see from comparing Column 2 of Tables 4 and 6 and comparing Column 2 of Tables 5 and 7. The additional monotonicity assumption for English exam data appears to make bounds significantly narrower, especially in combination with the additional subgroup data. Although all bounds still contain 0, these results seem to suggest that the value of having additional information in the form of finer levels of aggregation goes some way towards obtaining more information about marginal effect parameters of interest, especially in combination with other shape restrictions.

Table 7: Rhode Island white/non-white English exam pass rate differences, with subgroup data

	Bounds without Monotonicity	Monotonicity Bounds
Parameter	(1)	(2)
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 0, ELL_i = 0]$ $-\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 0, ELL_i = 0]$	$[-0.803, 0.807]$	$[-0.656, 0.585]$
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 1, ELL_i = 0]$ $-\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 1, ELL_i = 0]$	$[-0.793, 0.990]$	$[-0.349, 0.575]$
$\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 0, ELL_i = 1]$ $-\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 0, ELL_i = 1]$	$[-1, 1]$	$[-0.996, 0.856]$
$\mathbb{E}[pass_i \vert white_i = 1, FRPL_i = 1, ELL_i = 1]$ $-\mathbb{E}[pass_i \vert white_i = 0, FRPL_i = 1, ELL_i = 1]$	$[-1, 0.910]$	$[-0.349, 0.485]$

Notes: Bounds on difference impose no shape restrictions but use additional pass rate by subgroup data; monotonicity bounds impose monotonicity restrictions (11) and (12) in additional to additional pass rate by subgroup data. See Section 4.2 for more details.

5 Conclusion

In this paper I present sharp bounds on individual-level parameters of interest when the only available data is aggregate data, and I develop a valid inference method relying only on the available marginal information of each covariate. In simulations and an empirical application I show that the sharp bounds are too wide to be useful with realistic aggregate data. Both additional shape restrictions and using additional data at a finer level of aggregation help to make sharp bounds narrower and together significantly narrow the sharp bounds. However, sharp bounds are unable to pin down, for example, the signs of marginal effects. This suggests that individual-level analyses using aggregate data are more precise when the aggregate data is available at a finer level of aggregation and when there is more underlying structure known about the individual-level data generating process. Researchers hoping to use aggregate data to make individual-level claims should try to obtain additional data at as fine of a level as possible and impose carefully justified assumptions in order to obtain useful results.

References

Burden, B. C. and D. C. Kimball (1998). A new approach to the study of ticket splitting. *American Political Science Review 92*(3), 533–544.

Chetverikov, D., A. Santos, and A. M. Shaikh (2018). The econometrics of shape restrictions. *Annual Review of Economics 10*, 31–63.

Cho, W. K. T. and B. J. Gaines (2004). The limits of ecological inference: The case of split-ticket voting. *American Journal of Political Science 48*(1), 152–171.

Cross, P. J. and C. F. Manski (2002). Regressions, short and long. *Econometrica 70*(1), 357–368.

Duncan, O. D. and B. Davis (1953). An alternative to ecological correlation. *American Sociological Review 18*(6), 665–666.

Fan, Y., R. Sherman, and M. Shum (2014). Identifying treatment effects under data combination. *Econometrica 82*(2), 811–822.

Fan, Y., R. Sherman, and M. Shum (2016). Estimation and inference in an ecological inference model. *Journal of Econometric Methods 5*(1), 17–48.

Firebaugh, G. (1978). A rule for inferring individual-level relationships from aggregate data. *American Sociological Review*, 557–572.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon Sc. 4*, 53–84.

Freyberger, J. and J. L. Horowitz (2015). Identification and shape restrictions in nonparametric instrumental variables estimation. *Journal of Econometrics 189*(1), 41–53.

Ho, K. and A. M. Rosen (2015). Partial identification in applied research: Benefits and challenges. Working Paper 21641, National Bureau of Economic Research.

Horowitz, J. L. and C. F. Manski (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association 95*(449), 77–84.

Hsiao, C., Y. Shen, and H. Fujiki (2005). Aggregate vs. disaggregate data analysis—a paradox in the estimation of a money demand function of Japan under the low interest rate policy. *Journal of Applied Econometrics 20*(5), 579–601.

Hsieh, Y.-W., X. Shi, and M. Shum (2022). Inference on estimators defined by mathematical programming. *Journal of Econometrics 226*(2), 248–268.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Jack, R., C. Halloran, J. Okun, and E. Oster (2023). Pandemic schooling mode and student test scores: evidence from US school districts. *American Economic Review: Insights 5*(2), 173–190.

Jiang, W., G. King, A. Schmaltz, and M. A. Tanner (2020). Ecological regression with partial identification. *Political Analysis 28*(1), 65–86.

King, G. (1997). *A solution to the ecological inference problem: Reconstructing individual behavior from aggregate data.* Princeton University Press.

King, G., O. Rosen, and M. A. Tanner (1999). Binomial-beta hierarchical models for ecological inference. *Sociological Methods & Research 28*(1), 61–90.

Kline, B. and E. Tamer (2023). Recent developments in partial identification. *Annual Review of Economics 15*, 125–150.

Lee, K. C., M. H. Pesaran, and R. G. Pierse (1990). Testing for aggregation bias in linear models. *The Economic Journal 100*(400), 137–150.

Matzkin, R. L. (1994). Restrictions of economic theory in nonparametric methods. *Handbook of econometrics 4*, 2523–2558.

Molinari, F. (2020). Microeconometrics with partial identification. *Handbook of Econometrics 7*, 355–486.

Molinari, F. and M. Peski (2006). Generalization of a result on "Regressions, short and long". *Econometric Theory 22*(1), 159–163.

Porter, E. (2015). Education gap between rich and poor is growing wider. *The New York Times*. `https://www.nytimes.com/2015/09/23/business/economy/education-gap-between-rich-and-poor-is-growing-wider.html`.

Ridder, G. and R. Moffitt (2007). The econometrics of data combination. *Handbook of econometrics 6*, 5469–5547.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review 15*(3), 351–357.

Rosen, O., W. Jiang, G. King, and M. A. Tanner (2001). Bayesian and frequentist inference for ecological inference: The R×C case. *Statistica Neerlandica 55*(2), 134–156.

Shiveley, W. P. (1974). Utilizing external evidence in cross-level inference. *Political Methodology*, 61–73.

Stoker, T. M. (1984). Completeness, distribution restrictions, and the form of aggregate functions. *Econometrica 52*(4), 887–907.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica 54*(6), 1461–1481.

Tam Cho, W. K. (1998). Iff the assumption fits...: A comment on the King ecological inference solution. *Political Analysis 7*, 143–163.

Tamer, E. (2010). Partial identification in econometrics. *Annual Review of Economics 2*(1), 167–195.

Tavernise, S. (2012). Education gap grows between rich and poor, studies say. *The New York Times*. `https://www.nytimes.com/2012/02/10/education/education-gap-grows-between-rich-and-poor-studies-show.html`.

Theil, H. (1954). Linear aggregation of economic relations.

A Appendix: Additional Results

A.1 Closed-form solution for L_g and U_g problems

Not only is (5) a linear program, the following corollary shows (5) has a closed-form solution given any $\{p_{kg}\} \in P_g$:

Corollary A.1. *For given weights $\lambda_1, \ldots, \lambda_K$ and any fixed $\{p_{kg}\} \in P_g$, relabel the indices $k = 1, \ldots, K$ so that $\frac{\lambda_1}{p_{1g}} \geq \cdots \geq \frac{\lambda_K}{p_{Kg}}$, where if $p_{kg} = 0$ we define $\frac{\lambda_k}{p_{gk}} \equiv +\infty$. Then*

$$\min_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}$$

is attained by letting

$$c_{1g} = \begin{cases} \max\left\{0, \frac{\mathbb{E}[Y_i|G_i=g]-1+p_{1g}}{p_{1g}}\right\} & p_{1g} > 0 \\ 0 & p_{1g} = 0 \end{cases},$$

$$c_{kg} = \begin{cases} 0 & \sum_{j=1}^{k} p_{jg} \leq 1 - \mathbb{E}[Y_i|G_i = g] \\ \frac{\mathbb{E}[Y_i|G_i=g]-1+\sum_{j=1}^{k} p_{jg}}{p_{kg}} & \sum_{j=1}^{k-1} p_{jg} \leq 1 - \mathbb{E}[Y_i|G_i = g] < \sum_{j=1}^{k} p_{jg} \\ 1 & \sum_{j=1}^{k-1} p_{jg} > 1 - \mathbb{E}[Y_i|G_i = g] \end{cases},$$

and

$$\max_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}$$

is attained by letting

$$c_{1g} = \begin{cases} \min\left\{1, \frac{\mathbb{E}[Y_i|G_i=g]}{p_{1g}}\right\} & p_{1g} > 0 \\ 1 & p_{1g} = 0 \end{cases}$$

$$c_{kg} = \begin{cases} 1 & \sum_{j=1}^{k} p_{jg} \leq \mathbb{E}[Y_i|G_i = g] \\ \frac{\mathbb{E}[Y_i|G_i=g]-\sum_{j=1}^{k-1} p_{jg}}{p_{kg}} & \sum_{j=1}^{k-1} p_{jg} \leq \mathbb{E}[Y_i|G_i = g] < \sum_{j=1}^{k} p_{jg} \\ 0 & \sum_{j=1}^{k-1} p_{jg} > \mathbb{E}[Y_i|G_i = g] \end{cases}.$$

Proofs are collected in Appendix B.

A.2 Fréchet inequalities

In Section 2.1 I derived sharp bounds for a linear combination of $\mathbb{E}[Y_i|X_i = x_k]$, defined using the solutions to optimization programs. In this section I show that if we are interested in obtaining sharp bounds on each $\mathbb{E}[Y_i|X_i = x_k]$ parameter under Assumption 1 alone, we can obtain closed-form bounds using Fréchet inequalities.

The Fréchet inequalities, explicitly derived by Fréchet (1951), state that if there are N events $A_1, \ldots, A_N$, it holds that

$$\max\left\{1 - N + \sum_{i=1}^{N} Pr[A_i], 0\right\} \leq Pr\left[\bigcap_{i=1}^{N} A_i\right] \leq \min\left\{Pr[A_1], \ldots, Pr[A_N]\right\}. \tag{A.1}$$

Therefore, for y in the support of Y_i and x_k in the support of X_i, it follows from the Fréchet inequalities that

$$\begin{aligned} L_g(y, x_k) &\equiv \max\left\{\mathbb{P}[Y_i = y|G_i = g] + \sum_{\ell=1}^{L} \mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] - L, 0\right\} \\ &\leq \mathbb{P}\left[Y_i = y, X_i = x_k|G_i = g\right] \\ &\leq \min\left\{\mathbb{P}[Y_i = y|G_i = g], \mathbb{P}[X_{1i} = x_{k,1}|G_i = g], \ldots, \mathbb{P}[X_{Li} = x_{k,L}|G_i = g]\right\} \\ &\equiv U_g(y, x_k). \end{aligned} \tag{A.2}$$

Note that we only require knowledge of the marginal probabilities of each random variable in order to be able to calculate the lower and upper bounds; joint probabilities are not needed. If we are only given $\mathbb{P}[Y_i = y|G_i = g], \mathbb{P}[X_{1i} = x_1|G_i = g], \ldots, \mathbb{P}[X_{Li} = x_L|G_i = g]$ and nothing else, then it is well known that the Fréchet inequalities are sharp; that is, they are the tightest possible bounds given the assumptions. Situations in which the Fréchet inequalities are not sharp include those in which we know variables are independent or if the (known) support of the variables is such that knowledge of a marginal probability provides knowledge of the joint probability.

For a given k we can define bounds on $\mathbb{E}[Y_i|X_i = x_k]$:

Proposition A.1. *Suppose Assumption 1 holds. Let*

$$D^F \equiv \left[\sum_{g=1}^{G} \mathbb{P}[G_i = g] L_g^F, \sum_{g=1}^{G} \mathbb{P}[G_i = g] U_g^F\right],$$

where

$$\begin{aligned} L_g^F &\equiv \sum_{k=1}^{K} \lambda_k \frac{L_g(1, x_k)}{L_g(1, x_k) + U_g(0, x_k)}, \\ U_g^F &\equiv \sum_{k=1}^{K} \lambda_k \frac{U_g(1, x_k)}{L_g(0, x_k) + U_g(1, x_k)}. \end{aligned}$$

Then $D \subseteq D^F$, and if the Fréchet inequalities on $\mathbb{P}[Y_i = y, X_i = x_k|G_i = g]$ are sharp for all $y \in \{0, 1\}, k = 1, \ldots, K, g = 1, \ldots, G$ then $D = D^F$.

To construct an estimator of the identified set D^F, the sample analogs of $\mathbb{P}[Y_i = y|G_i = g], \mathbb{P}[X_{1i} = x_1|G_i = g], \ldots, \mathbb{P}[X_{Li} = x_L|G_i = g]$ observed in the aggregate data can be plugged into the formula given in Proposition A.1. These estimated lower and upper bounds will be consistent because they are numerically equivalent to the estimated bounds of Proposition 1, which are consistent by Proposition 4.

To construct a valid confidence region for the identified set D^F, a similar approach to that of Section 3.2 can be used. If jointly valid marginal confidence intervals are constructed on each sample observation, calculating the estimated bounds, plugging lower confidence interval bounds into any lower bounds and upper confidence interval bounds into any upper bounds, will produce a confidence interval with correct coverage.

B Appendix: Proofs

B.1 Proof of Lemma 1

Proof. As argued in the main text, the only information we have in addition to Assumption 1 are equations (1), (2), (3), and that $\delta_k, \gamma_{kg}, \pi_{kg} \in [0,1]$ for all k, g. The set given by (4) imposes all of these restrictions and nothing more, and finds the set of $\sum_{k=1}^K \lambda_k \delta_k$ such that the restrictions are satisfied. Thus the set is sharp. □

B.2 Proof of Proposition 1

Proof. For any g, given any $\{p_{kg}\}_{k=1}^K$ such that $\sum_{k=1}^K p_{kg} = 1$ and $p_{kg} \geq 0$ $\forall k$, note that there exists $\{c_{kg}\}_{k=1}^K \in [0,1]^K$ such that $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg} p_{kg}$; namely $c_{kg} = \mathbb{E}[Y_i|G_i = g]$ $\forall k$.

Thus imposing the restriction $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg} p_{kg}$ does not further restrict the set of possible $\{p_{kg}\}_{k=1}^K$ if the following restrictions already hold: $\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg}$, $\sum_{k=1}^K p_{kg} = 1$ and $p_{kg} \geq 0$ $\forall k$. Finally note that all restrictions involving $\{p_{kg}\}_{k=1}^K$ only involve p_{kg} with the same index g. This means it is equivalent to write (4) as

$$D = \left\{ \sum_{k=1}^K \lambda_k d_k \,\middle|\, 0 \leq d_k \leq 1 \ \forall k, \text{ and } \exists\, (p_{1g}, \dots, p_{Kg}) \in P_g, (c_{1g}, \dots, c_{Kg}) \in [0,1]^K \ \forall g \right.$$
$$\left. \text{s.t. } d_k = \sum_{g=1}^G \mathbb{P}[G_i = g] c_{kg} \ \forall k,\ \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg} p_{kg} \ \forall g \right\}, \quad \text{(B.1)}$$

where

$$P_g = \left\{ (p_{1g}, \dots, p_{Kg}) \in [0,1]^K \,\middle|\, \mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} \ \forall \ell, k, \text{ and } \sum_{k=1}^K p_{kg} = 1 \right\}. \quad \text{(B.2)}$$

If $\mathbb{P}[G_i = g], c_{kg} \in [0,1]$ for all k and g then because $d_k = \sum_{g=1}^G \mathbb{P}[G_i = g] c_{kg}$ we know $d_k \in [0,1]$ for all k. Then we can get rid of d_k in (B.1) by plugging in one of the constraints like so:

$$D = \left\{ \sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c_{kg} \,\middle|\, \exists\, (p_{1g}, \dots, p_{Kg}) \in P_g, (c_{1g}, \dots, c_{Kg}) \in [0,1]^K \ \forall g \right.$$

$$\left. \text{s.t. } \mathbb{E}[Y_i | G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg} \ \forall g \right\}. \quad \text{(B.3)}$$

I next show that D is an interval. For any given $\{p_{kg}\} \in P_g$, we argued above that $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\mathbb{E}[Y_i|G_i = g] \in D$. For any g consider arbitary $\{c_{kg}\}, \{c'_{kg}\} \in [0,1]^K$ with corresponding $\{p_{kg}\}, \{p'_{kg}\} \in P_g$ such that $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}$ and $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c'_{kg} p'_{kg}$. We know $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c_{kg} \in D$ and $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c'_{kg} \in D$. For arbitrary $t \in [0,1]$, let

$$\tilde{c}_{kg} = t c_{kg} + (1-t)\mathbb{E}[Y_i|G_i = g]$$

for all k, g and let

$$\tilde{c}'_{kg} = t c'_{kg} + (1-t)\mathbb{E}[Y_i|G_i = g]$$

for all k, g. Then $\mathbb{E}[Y_i|G_i = g] = \sum_k p_{kg}\tilde{c}_{kg} = \sum_k p'_{kg}\tilde{c}'_{kg}$ for all g and $\tilde{c}_{kg}, \tilde{c}'_{kg} \in [0,1]$ for all k, g. Thus $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\tilde{c}_{kg} \in D$ and $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\tilde{c}'_{kg} \in D$. Note

$$\begin{aligned}
&\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\tilde{c}_{kg} \\
&= t\left(\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c_{kg}\right) + (1-t)\left(\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\mathbb{E}[Y_i|G_i = g]\right), \\
&\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\tilde{c}'_{kg} \\
&= t\left(\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c'_{kg}\right) + (1-t)\left(\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\mathbb{E}[Y_i|G_i = g]\right).
\end{aligned}$$

Since t was arbitrary between 0 and 1, D contains any value between $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c_{kg}$ and $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\mathbb{E}[Y_i|G_i = g]$. Similarly, D contains any value between $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c'_{kg}$ and $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g]\mathbb{E}[Y_i|G_i = g]$. In particular, D contains any value between $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c_{kg}$ and $\sum_{k=1}^{K} \lambda_k \sum_{g=1}^{G} \mathbb{P}[G_i = g] c'_{kg}$. Thus D is an interval.

This means that we can equivalently express (B.3) and (B.2) as optimization problems as follows: $D = [L, U]$, where

$$\begin{aligned}
L &\equiv \min_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \sum_{g=1}^{G} \mathbb{P}[G_i = g]\lambda_k c_{kg} \text{ s.t. } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}, \\
U &\equiv \max_{\{c_{kg}\} \in [0,1]^K} \sum_{k=1}^{K} \sum_{g=1}^{G} \mathbb{P}[G_i = g]\lambda_k c_{kg} \text{ s.t. } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}, \text{ and}
\end{aligned}$$

$$P_g = \underset{\{p_{kg}\}\in[0,1]^K}{\arg\min} \sum_{r=1}^{LK} v_r^+ + v_r^- \text{ s.t. } \sum_{k=1}^{K} p_{kg} = 1,\ p_{kg}, v_r^+, v_r^- \geq 0\ \forall k, r, \text{ and}$$
$$\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] - \sum_{j=1}^{K} \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} p_{jg} = v^+_{K(\ell-1)+k} - v^-_{K(\ell-1)+k}\ \forall \ell, k.$$

The slack variables in the P_g problem will all be equal to zero at optimum.

Finally, since the constraints in L and U are for each g, we can equivalently solve a program for each group g and take a weighted sum of the solutions to get lower and upper bounds for D, like so: $D = \left[\sum_{g=1}^{G} \mathbb{P}[G_i = g] L_g, \sum_{g=1}^{G} \mathbb{P}[G_i = g] U_g\right]$, where

$$L_g \equiv \min_{\{c_{kg}\}\in[0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg},$$
$$U_g \equiv \max_{\{c_{kg}\}\in[0,1]^K} \sum_{k=1}^{K} \lambda_k c_{kg} \text{ s.t. } \exists \{p_{kg}\} \in P_g \text{ with } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}.$$

□

B.3 Proof of Proposition 2

Proof. To prove sharpness of the identified set, it is sufficient to show that the following set is an interval: for $c_g \equiv (c_{1g}, \ldots, c_{Kg})'$,

$$\left\{ \sum_{k=1}^{K} \sum_{g=1}^{G} \mathbb{P}[G_i = g] \lambda_k c_{kg} \middle| S_g c_g \leq a_g\ \forall g \text{ and } \exists\, (p_{1g}, \ldots, p_{Kg}) \in P_g, (c_{1g}, \ldots, c_{Kg}) \in [0,1]^K\ \forall g \right.$$
$$\left. \text{s.t. } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg}\ \forall g \right\}. \quad \text{(B.4)}$$

We know that the set of $c_g \in [0,1]^{KG}$ such that $S_g c_g \leq a_g$ for all g is a convex, closed, and bounded set because it is the intersection of a closed, bounded, and convex set with a half-space. This means the set of values of $\sum_k \lambda_k c_{kg}$ such that $S_g c_g \leq a_g$ for all g is also a closed and bounded and convex set.

In the proof of Proposition 1 I argued that for any $c_g, c_g' \in [0,1]^K$ with corresponding $\{p_{kg}\}, \{p_{kg}'\} \in P_g$ that satisfy

$$\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg} p_{kg} \qquad \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^{K} c_{kg}' p_{kg}' \quad \text{(B.5)}$$
$$S_g c_g \leq a_g \qquad S_g c_g' \leq a_g, \quad \text{(B.6)}$$

we know $t c_{kg} + (1-t)\mathbb{E}[Y_i|G_i = g]$ and $t c_{kg}' + (1-t)\mathbb{E}[Y_i|G_i = g]$ also satisfy (B.5) in place of c_{kg} and c_{kg}' respectively, for any $t \in [0,1]$. Thus for any $\sum_{k=1}^{K} \lambda_k \tilde{c}_{kg}$ between $\sum_{k=1}^{K} \lambda_k c_{kg}$ and $\sum_{k=1}^{K} \lambda_k c_{kg}'$,

$\{\tilde{c}_{kg}\}$ satisfies $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K \tilde{c}_{kg}p_{kg}$. And that (B.6) holds implies $S_g(tc_g + (1-t)c'_g) \leq a_g$ for any $t \in [0,1]$, meaning that for any $\sum_{k=1}^K \lambda_k\tilde{c}_{kg}$ between $\sum_{k=1}^K \lambda_k c_{kg}$ and $\sum_{k=1}^K \lambda_k c'_{kg}$, $\{\tilde{c}_{kg}\}$ satisfies $S_g\tilde{c}_g \leq a_g$.

Thus any $\sum_{k=1}^K \sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k\tilde{c}_{kg}$ between $\sum_{k=1}^K \sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k c_{kg}$ and $\sum_{k=1}^K \sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k c'_{kg}$ is in the set given by (B.4), meaning (B.4) is an interval. Thus the sharp identified set is an interval. The rest of Proposition 2 can be proved following arguments similar to those used in the proof of Proposition 1. □

B.4 Proof of Proposition 3

Proof. To prove sharpness of the identified set, it is sufficient to show that the following set is an interval:

$$\Bigg\{\sum_{k=1}^K\sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k c_{kg}\Bigg|\ \exists\{p_{kg}\} \in P_g, \{c_{kg}\} \in [0,1]^K\ \forall g \text{ s.t. } \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg}p_{kg}\ \forall g,$$
$$\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c_{kg}p_{kg}\ \forall(\ell,k) \in F_g, \forall g\Bigg\}. \tag{B.7}$$

Note that the set of $\{c_{kg}\} \in [0,1]^{KG}$ such that

$$\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c_{kg}p_{kg}\ \forall(\ell,k) \in F_g, \forall g \tag{B.8}$$

is a convex, closed, and bounded set because it is the intersection of a closed, bounded, and convex half-space. This means the set of values of $\sum_k \lambda_k c_{kg}$ such that (B.8) holds is also a closed and bounded and convex set.

In the proof of Proposition 1 I argued that for any $\{c_{kg}\}, \{c'_{kg}\} \in [0,1]^K$ with corresponding $\{p_{kg}\}, \{p'_{kg}\} \in P_g$ that satisfy

$$\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c_{kg}p_{kg}, \qquad \mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K c'_{kg}p'_{kg}, \tag{B.9}$$

$$\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c_{kg}p_{kg}, \tag{B.10}$$

$$\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}c'_{kg}p'_{kg}, \tag{B.11}$$

we know $tc_{kg} + (1-t)\mathbb{E}[Y_i|G_i = g]$ and $tc'_{kg} + (1-t)\mathbb{E}[Y_i|G_i = g]$ also satisfy (B.9) in place of c_{kg} and c'_{kg} respectively, for any $t \in [0,1]$. Thus for any $\sum_{k=1}^K \lambda_k\tilde{c}_{kg}$ between $\sum_{k=1}^K \lambda_k c_{kg}$ and $\sum_{k=1}^K \lambda_k c'_{kg}$, $\{\tilde{c}_{kg}\}$ satisfies $\mathbb{E}[Y_i|G_i = g] = \sum_{k=1}^K \tilde{c}_{kg}p_{kg}$.

Note that, similar to the claim made in the proof of Proposition 1, since $\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i =$

$g] = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_{kg} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p'_{kg}$ it follows that (B.10) still holds when letting $c_{kg} = \mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$ and (B.11) still holds when letting $c'_{kg} = \mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$.

Thus $tc_{kg}+(1-t)\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$ in place of c_{kg} and $tc'_{kg}+(1-t)\mathbb{E}[Y_i|X_{\ell i} = x_{k,\ell}, G_i = g]$ in place of c'_{kg} also satisfy (B.10) and (B.11) respectively for any $t \in [0,1]$. So for any $\sum_{k=1}^K \lambda_k \tilde{c}_{kg}$ between $\sum_{k=1}^K \lambda_k c_{kg}$ and $\sum_{k=1}^K \lambda_k c'_{kg}$, $\{\tilde{c}_{kg}\}$ in place of $\{c_{kg}\}$ satisfies (B.10).

Thus any $\sum_{k=1}^K \sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k \tilde{c}_{kg}$ between $\sum_{k=1}^K \sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k c_{kg}$ and $\sum_{k=1}^K \sum_{g=1}^G \mathbb{P}[G_i = g]\lambda_k c'_{kg}$ is in the set given by (B.7), meaning (B.7) is an interval. Thus the sharp identified set is an interval. The rest of Proposition 3 can be proved following arguments similar to those used in the proof of Proposition 1. □

B.5 Proof of Proposition 4

Proof. I prove the result for $\hat{D}$ defined with respect to Proposition 1, then discuss the additional restrictions implied by Assumptions 2 and 3.

The result follows from the Theorem of the Maximum for the following correspondence:

$$\Gamma\left(y, \{g_{k,\ell}\}_{k,\ell}\right) = \underset{\{p_k, c_k\}_k}{\arg\max} \left(\underset{\{p_k, c_k\}_k}{\arg\min}\right) \Bigg\{ \sum_{k=1}^K \lambda_k c_k \text{ s.t. } y = \sum_{k=1}^K c_k p_k, \sum_{k=1}^K p_k = 1,$$
$$g_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_j, c_k \in [0,1]\ \forall k, p_k \geq 0\ \forall k \Bigg\}.$$

Since the $\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]$ are valid marginal distributions for X_i with respect to the assumed support, this correspondence is analogous to the set defined in Proposition 2; that is, the constraints in P_g hold with $v_r^+ = v_r^- = 0$.

To apply the Theorem of the Maximum, I must show Γ is a continuous and compact-valued correspondence. That Γ is compact-valued follows because the constraints are intersections of non-parallel planes with a curve ($y = \sum_{k=1}^K c_k p_k$) on a compact set ($c_k \in [0,1], p_k \geq 0, \sum_{k=1}^K p_k = 1$).

Let $(y^n, \{g^n_{k,\ell}\}_{k,\ell})$ be an arbitrary sequence such that $(y^n, \{g^n_{k,\ell}\}_{\ell,j}) \to (y, \{g_{k,\ell}\}_{k,\ell})$ as $n \to \infty$ and all $y^n, y \in [0,1]$ and $\{g^n_{k,\ell}\}_{k,\ell}$ are valid marginal distributions given the assumed support. Let $(\{p^n_k\}, \{c^n_k\})$ be an arbitrary sequence such that $(\{p^n_k\}, \{c^n_k\}) \to (\{p_k\}, \{c_k\})$ as $n \to \infty$ and for each n, $(\{p^n_k\}, \{c^n_k\}) \in \Gamma\left((y^n, \{g^n_{k,\ell}\}_{k,\ell})\right)$. If we can show $(\{p_k\}, \{c_k\}) \in \Gamma((y, \{g_{k,\ell}\}_{k,\ell}))$ then Γ is upper hemicontinuous.

Since $\sum_{k=1}^K p^n_k = 1$ for all n, $p^n_k \geq 0$ for all k, n, and $p^n_k \to p_k$ for all k, it follows that $\sum_{k=1}^K p_k = 1$ and $p_k \geq 0$ for all k. Since $c^n_k \in [0,1]$ for all k, n and $c^n_k \to c_k$ for all k, it follows that $c_k \in [0,1]$ for all k. Since $y^n = \sum_{k=1}^K p^n_k c^n_k$ for all k, n and $y^n \to y, d^n_k \to d_k, p^n_k \to p_k$ for all k, n, it follows that $y = \sum_k p_k d_k$. Since $g^n_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p^n_j$ for all k, ℓ, n and we have $p^n_k \to p_k, g^n_{k,\ell} \to g_{k,\ell}$ for all k, ℓ, it follows that $g_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_j$ for all k, ℓ. Thus $(\{p_k\}, \{c_k\}) \in \Gamma((y, \{g_{k,\ell}\}_{k,\ell}))$ and thus Γ is upper hemicontinuous.

Let $(y^n, \{g^n_{k,\ell}\}_{k,\ell})$ be an arbitrary sequence such that $(y^n, \{g^n_{k,\ell}\}_{\ell,j}) \to (y, \{g_{k,\ell}\}_{k,\ell})$ as $n \to \infty$ and all $y^n, y \in [0,1]$ and $\{g^n_{k,\ell}\}_{k,\ell}$ are valid marginal distributions given the assumed support. Let $(\{p_k\}, \{c_k\}) \in \Gamma(y, \{g_{k,\ell}\}_{k,\ell})$ be arbitrary. If we can show there exists a subsequence $(y^{n_t}, \{g^{n_t}_{k,\ell}\}_{k,\ell})$ and sequence $(\{p^t_k\}, \{c^t_k\})$ such that $(\{p^t_k\}, \{c^t_k\}) \to (\{p_k\}, \{c_k\})$ as $t \to \infty$ and $(\{p^t_k\}, \{c^t_k\}) \in \Gamma(y^{n_t}, \{g^{n_t}_{k,\ell}\}_{k,\ell})$ for all t, then Γ is lower hemicontinuous.

We know that $g_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_j$ and $\sum_{k=1}^K p_k = 1$. Because we assume each $\{g_{k,\ell}\}$ is a valid marginal distribution of covariates there are strictly less than K unique noncollinear equations in this system of equations. We can express these equations as a linear system $Ap = g$, where $p = \{p_k\}$ and $g = \{g_{k,\ell}\}_{k,\ell}$, and all solutions are given by $p = A^+g + (I - A^+A)w$, where A^+ is the Moore-Penrose inverse and w is any arbitrary vector of correct dimension.

Thus the solution $\{p_k\}$ to this system of linear equations is continuous in $\{g_{k,\ell}\}_{k,\ell}$. Therefore for each t there exists n_t large enough that we can find $\{p^t_k\}$ arbitrarily close to $\{p_k\}$ such that $g^{n_t}_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p^t_j$ and $\sum_{k=1}^K p^t_k = 1$. Such a solution exists because we assume $\hat{D}$ is always nonempty with a valid marginal distribution of covariates $\{g^{n_t}_{k,\ell}\}$.

Recall each $p_k \geq 0$. In fact, the proposition imposes as an assumption that we only need consider $\{p_k\}$ such that each $p_k > 0$. Thus we can make $p^t_k \geq 0$ for each k. Thus we can construct a sequence $\{p^t_k\}$ with corresponding subsequence $\{g^{n_t}_{k,\ell}\}_{k,\ell}$ such that $g^{n_t}_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p^t_j$ for all t, $\sum_{k=1}^K p^t_k = 1$ for all t, $p^t_k \geq 0$ for all k, and $p^t_k \to p_k$ for all k as $t \to \infty$.

Given this sequence $\{p^t_k\}$, we need to find a corresponding sequence $\{c^t_k\}$ and subsequence $\{y^{n_t}\}$ such that $y^{n_t} = \sum_{k=1}^K c^t_k p^t_k$ and $c^t_k \in [0,1]$ for all k, t. Note that we can write the equation $y = \sum_{k=1}^K c_k p_k$ as $p'c = y$ and consider the Moore-Penrose representation of the set of all solutions c. As a matrix algebra result, it is true that the Moore-Penrose inverse of p is continuous, that is $(p_t)^+ \to p^+$ as long as the rank of p_t is the same as the rank of p for all t. Since the rank of p_t and p is always 1, it follows that the solution c to the equation is continuous in p and y.

Therefore for each t there exists n_t large enough that we can find $\{\tilde{c}^t_k\}$ arbitrarily close to $\{c_k\}$ with $y^{n_t} = \sum_{k=1}^K \tilde{c}^t_k p^t_k$ (with some relabeling of the $\{p^t_k\}$ sequence indices as necessary). We know $c_k \in [0,1]$ for each k, but it may be the case that $\tilde{c}^t_k \notin [0,1]$ for some k. However for those k we will have $\tilde{c}^t_k$ arbitrarily close to $[0,1]$. As argued in the proof of Proposition 1 it also holds that $y^{n_t} = \sum_{k=1}^K y^{n_t} p^t_k$ (and we know $y^{n,t} \in [0,1]$), meaning there exists another feasible $\{c^t_k\}$ between $\{\tilde{c}^t_k\}$ and $(y^{n_t}, \ldots, y^{n_t})$ that is arbitrarily close to $\{\tilde{c}^t_k\}$ but with $c^t_k \in [0,1]$ for all k.

Thus we have sequence $(\{p^t_k\}, \{c^t_k\})$ with corresponding subsequence $(y^{n_t}, \{g^{n_t}_{k,\ell}\}_{k,\ell})$ such that $(\{p^t_k\}, \{c^t_k\}) \to (\{p_k\}, \{c_k\})$ as $t \to \infty$, and $(\{p^t_k\}, \{c^t_k\}) \in \Gamma(y^{n_t}, \{g^{n_t}_{k,\ell}\}_{k,\ell})$. Therefore Γ is lower hemicontinuous.

We have shown Γ is a continuous and compact-valued correspondence. This means that $\hat{L}_g$ and $\hat{U}_g$ are continuous functions of $\bar{Y}_g$ and all $\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]$. We also know that the population L_g and U_g is the same continuous function of $\mathbb{E}[Y_i|G_i = g]$ and all $\mathbb{P}[X_{\ell i} = x_{k,\ell}|G_i = g]$. Because $\bar{Y}_g$ and all $\widehat{Pr}[X_{\ell i} = x_{k,\ell}|G_i = g]$ are consistent by the weak law of large numbers, continuous mapping theorem gives us that $\hat{L}_g \xrightarrow{p} L_g$ and $\hat{U}_g \xrightarrow{p} U_g$ as $n \to \infty$ for all g.

The law of large numbers also gives us that $\widehat{Pr}[G_i = g]$ is consistent, so that by continuous mapping theorem again we have that the lower and upper bounds of $\hat{D}$ converge to the lower and upper bounds of D.

To accommodate the restriction that $S_g c \leq a_g$ under Assumption 2 in correspondence Γ, note that any sequence $\{c_k^n\} \to \{c_k\}$ with $(\{p_k^n\}, \{c_k^n\}) \in \Gamma\left((y^n, \{g_{k,\ell}^n\}_{k,\ell})\right)$ satisfies $S_g c^n \leq a_g$ and thus $S_g c \leq a_g$ by continuity as well. So upper hemicontinuity is maintained.

The assumption that $\hat{D}$ is nonempty and the observed marginals are valid with respect to the assumed support ensures that for $(y^n, \{g_{k,\ell}^n\}_{k,\ell}) \to (y, \{g_{k,\ell}\}_{k,\ell})$ there exists $(\{p_k\}, \{c_k\}) \in \Gamma(y, \{g_{k,\ell}\}_{k,\ell})$ and that $\Gamma(y^{n_t}, \{g_{k,\ell}^{n_t}\}_{k,\ell})$ is nonempty. Let $(\{p_k^t\}, \{c_k^t\}) \in \Gamma(y^{n_t}, \{g_{k,\ell}^{n_t}\}_{k,\ell})$ be a sequence such that $\max_k\{|p_k^t - p_k|, |c_k^t - c_k|\}$ is minimized for each t.

I claim $(\{p_k^t\}, \{c_k^t\}) \to (\{p_k\}, \{c_k\})$. From the restrictions that $g_{k,\ell}^{n_t} = \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_j^t$ we know $\max_k \sum_{k=1}^K |p_k^t - p_k| \to 0$ as discussed above with the Moore-Penrose inverse solution to the linear system, so $\{p_k^t\} \to \{p_k\}$. And if $\max_k \sum_{k=1}^K |c_k^t - c_k| \not\to 0$ then $c_k^t p_k^t \not\to c_k p_k$, meaning the constraint $y = \sum_{k=1}^K c_k p_k$ cannot hold either as $y^{n_t} \to y$. Thus it must be that $\{c_k^t\} \to \{c_k\}$ and so lower hemicontinuity is also maintained because $S_g c^t \leq a_g$ by assumption.

To accommodate the restriction

$$y_{k,\ell} g_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_j p_j \tag{B.12}$$

under Assumption 3 in correspondence Γ, note that any sequence $(\{p_k^n\}, \{c_k^n\}) \to (\{p_k\}, \{c_k\})$ with $(\{p_k^n\}, \{c_k^n\}) \in \Gamma\left((y^n, \{g_{k,\ell}^n\}_{k,\ell}, \{y_{k,\ell}^n\}_{k,\ell})\right)$ satisfies

$$y_{k,\ell}^n g_{k,\ell}^n = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_j^n p_j^n,$$

and thus (B.12) holds by continuity as well. So upper hemicontinuity is maintained.

The assumption that $\hat{D}$ is nonempty and the observed marginals are valid with respect to the assumed support ensures that for $(y^n, \{g_{k,\ell}^n\}_{k,\ell}, \{y_{k,\ell}^n\}_{k,\ell}) \to (y, \{g_{k,\ell}\}_{k,\ell}, \{y_{k,\ell}\}_{k,\ell})$ there exists $(\{p_k\}, \{c_k\}) \in \Gamma(y, \{g_{k,\ell}\}_{k,\ell}, \{y_{k,\ell}\}_{k,\ell})$ and that $\Gamma(y^{n_t}, \{g_{k,\ell}^{n_t}\}_{k,\ell}, \{y_{k,\ell}^{n_t}\}_{k,\ell})$ is nonempty. Let $(\{p_k^t\}, \{c_k^t\}) \in \Gamma(y^{n_t}, \{g_{k,\ell}^{n_t}\}_{k,\ell}, \{y_{k,\ell}^{n_t}\}_{k,\ell})$ be a sequence such that $\max_k\{|p_k^t - p_k|, |c_k^t - c_k|\}$ is minimized for each t.

I claim $(\{p_k^t\}, \{c_k^t\}) \to (\{p_k\}, \{c_k\})$. From the restrictions that $g_{k,\ell}^{n_t} = \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\}p_j^t$ we know $\max_k \sum_{k=1}^K |p_k^t - p_k| \to 0$ as discussed above with the Moore-Penrose inverse solution to the linear system. And if $\max_k \sum_{k=1}^K |c_k^t - c_k| \not\to 0$ then $c_k^t p_k^t \not\to c_k p_k$, meaning the constraints $y = \sum_{k=1}^K c_k p_k$ and $y_{k,\ell} g_{k,\ell} = \sum_{j=1}^K \mathbb{1}\{x_{j,\ell} = x_{k,\ell}\} c_j p_j$ cannot hold either as $y^{n_t} \to y, g_{k,\ell}^{n_t} \to g_{k,\ell}, y_{k,\ell}^{n_t} \to y_{k,\ell}$. Thus it must be that $\{c_k^t\} \to \{c_k\}$ and so lower hemicontinuity is also maintained. □

B.6 Proof of Proposition 5

Proof. Since Clopper-Pearson intervals are finite-sample valid for each sample observation, using the Bonferroni correction means that the population analog of every single sample observation is contained in its Clopper-Pearson interval with joint probability greater than $1-\alpha$.

Note that if all population analogs of the sample observations are jointly in their respective Clopper-Pearson intervals then any $\{c_{kg}\}$ and $\{p_{kg}\}$ consistent with the population P_g and L_g, U_g constraints in the formulation of D are also consistent with the $\hat{P}_{g,CI}$ and $\hat{L}_{g,CI}, \hat{U}_{g,CI}$ constraints in the formulation of $\hat{D}_{CI}$. Thus the set of $\sum_{k=1}^{K} \lambda_k c_{kg}$ consistent with the population P_g and L_g, U_g constraints in the formulation of D is a subset of the set of $\sum_{k=1}^{K} \lambda_k c_{kg}$ consistent with the $\hat{P}_{g,CI}$ and $\hat{L}_{g,CI}, \hat{U}_{g,CI}$ constraints in the formulation of $\hat{D}_{CI}$. And if the true $\mathbb{P}[G_i = g]$ are all contained in their Clopper-Pearson intervals, it follows that $D \subseteq \hat{D}_{CI}$.

Since the event that all population analogs of the sample observations are jointly in their respective Clopper-Pearson intervals happens with probability at least $1-\alpha$, it follows that $\mathbb{P}[D \subseteq \hat{D}_{CI}] \geq 1-\alpha$. □

B.7 Proof of Corollary A.1

Proof. For given weights $\lambda_1, \ldots, \lambda_K$ and any fixed $\{p_{kg}\} \in P_g$, relabel the indices $k = 1, \ldots, K$ so that $\frac{\lambda_1}{p_{1g}} \geq \cdots \geq \frac{\lambda_K}{p_{Kg}}$, where if $p_{kg} = 0$ we define $\frac{\lambda_k}{p_{gk}} \equiv +\infty$.

Note that the proposed solution in Corollary A.1 is feasible. We will take advantage of strong duality (because Slater's condition holds) and use the joint feasiblity and satisfaction of complementary slackness for proposed solutions to the primal and dual problems to show optimality.

The dual of the minimization linear program is

$$\max_{u \in \mathbb{R}, v \in \mathbb{R}^{2K}} -\mathbb{E}[Y_i|G_i = g]u - \sum_{i=1}^{K} v_i \text{ s.t. } -p_{kg}u - v_k \leq \lambda_k \; \forall k = 1, \ldots, K, v_i \geq 0 \; \forall i = 1, \ldots, K.$$

Let k be such that $\sum_{j=1}^{k-1} p_{jg} \leq 1 - \mathbb{E}[Y_i|G_i = g] < \sum_{j=1}^{k} p_{jg}$ holds. Consider a solution to the dual where $v_i = 0$ for all $i \leq k$, and for $i \geq k$ we have that v_i satisfies $-p_{ig}u - v_i = \lambda_i$, meaning $u = -\frac{\lambda_k}{p_{kg}}$. Thus for $i > k$ we have $v_i > 0$ because $-p_{ig}u = \frac{\lambda_k p_{ig}}{p_{kg}} \geq \lambda_i$. Clearly this is a feasible solution.

We see that complementary slackness holds with the condition that $c_{ig} \leq 1$ for all i and $v_i \geq 0$ for all i because $c_{ig} \neq 1$ for all $i \leq k$ while $v_i = 0$ for all $i \leq k$. Complementary slackness holds with the condition that $c_{ig} \geq 0$ for all i and $-p_{ig}u - v_i \leq \lambda_i$ for all i because $c_{ig} \neq 0$ for all $i \geq k$ while $-p_{ig}u - v_i = \lambda_i$ for all $i \geq k$.

Thus the proposed solution for the minimization problem is optimal.

The dual of the maximization linear program is

$$\min_{u \in \mathbb{R}, v \in \mathbb{R}^{2K}} \mathbb{E}[Y_i|G_i = g]u + \sum_{i=1}^{K} v_i \text{ s.t. } p_k u + v_k \geq \lambda_k \; \forall k = 1, \ldots, K, v_i \geq 0 \; \forall i = 1, \ldots, K.$$

Let k be such that $\sum_{j=1}^{k-1} p_{jg} \leq \mathbb{E}[Y_i|G_i = g] < \sum_{j=1}^{k} p_{jg}$ holds. Consider a solution to the dual where $v_i = 0$ for all $i \geq k$, and for $i \leq k$ we have that v_i satisfies $p_{ig}u + v_i = \lambda_i$, meaning $u = \frac{\lambda_k}{p_{kg}}$. Thus for $i < k$ we have $v_i > 0$ because $p_{ig}u = \frac{\lambda_k p_{ig}}{p_{kg}} \leq \lambda_i$. Clearly this is a feasible solution.

We see that complementary slackness holds with the condition that $c_{ig} \leq 1$ for all i and $v_i \geq 0$ for all i because $c_{ig} \neq 1$ for all $i \geq k$ while $v_i = 0$ for all $i \geq k$. Complementary slackness holds with the condition that $c_{ig} \geq 0$ for all i and $p_{ig}u + v_i \geq \lambda_i$ for all i because $c_{ig} \neq 0$ for all $i \leq k$ while $p_{ig}u + v_i = \lambda_i$ for all $i \leq k$.

Thus the proposed solution for the minimization problem is optimal. □

B.8 Proof of Proposition A.1

Proof. First I show that $D \subseteq D^F$.

Since D is truly an interval, as proved when proving Proposition 1, there exists some $\{c_{kg}\}_{k,g}, \{p_{kg}\}_{k,g}$ that satisfy the constraints of L_g, U_g, and P_g where $\sum_{g=1}^{G} \mathbb{P}[G_i = g] \sum_{k=1}^{K} \lambda_k c_{kg} \in D$.

By an argument analogous to that given in the proof of Proposition 1, the set of each c_{kg} that satisfy the constraints is the sharp identified set for $\mathbb{E}[Y_i|X_i = x_k, G_i = g]$. Thus it is sufficient to show that for each k, g,

$$\frac{L_g(1, x_k)}{L_g(1, x_k) + U_g(0, x_k)} \leq \mathbb{E}[Y_i|X_i = x_k, G_i = g] \leq \frac{U_g(1, x_k)}{U_g(1, x_k) + L_g(0, x_k)}.$$

This means $\left[\frac{L_g(1,x_k)}{L_g(1,x_k)+U_g(0,x_k)}, \frac{U_g(1,x_k)}{U_g(1,x_k)+L_g(0,x_k)}\right]$ are also bounds, so it follows that

$$\frac{L_g(1, x_k)}{L_g(1, x_k) + U_g(0, x_k)} \leq c_{kg} \leq \frac{U_g(1, x_k)}{U_g(1, x_k) + L_g(0, x_k)}.$$

and thus $\sum_{g=1}^{G} \mathbb{P}[G_i = g] \sum_{k=1}^{K} \lambda_k c_{kg} \in D^F$.

Note

$$\begin{aligned}
\mathbb{E}[Y_i|X_i = x_k, G_i = g] &= \mathbb{P}[Y_i = 1|X_i = x_k, G_i = g] \\
&= \frac{\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g]}{\mathbb{P}[X_i = x_k|G_i = g]} \\
&= \frac{\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g]}{\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g] + \mathbb{P}[Y_i = 0, X_i = x_k|G_i = g]}.
\end{aligned}$$

One can check that the function $\frac{x}{x+y}$ is increasing in x and decreasing in y; thus $\mathbb{E}[Y_i|X_i = x_k, G_i = g]$ attains its minimum when $\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g]$ is as small as possible and $\mathbb{P}[Y_i = 0, X_i = x_k|G_i = g]$ is as large as possible. Similarly $\mathbb{E}[Y_i|X_i = x_k, G_i = g]$ attains its maximum when $\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g]$ is as large as possible and $\mathbb{P}[Y_i = 0, X_i = x_k|G_i = g]$ is as small as possible.

Since $\mathbb{P}[Y_i = 0, X_i = x_k|G_i = g] \in [L_g(0, x_k), U_g(0, x_k)]$ and $\mathbb{P}[Y_i = 1, X_i = x_k|G_i = g] \in [L_g(1, x_k), U_g(1, x_k)]$, it follows then that $\mathbb{E}[Y_i|X_i = x_k, G_i = g] \in \left[\frac{L_g(1,x_k)}{L_g(1,x_k)+U_g(0,x_k)}, \frac{U_g(1,x_k)}{U_g(1,x_k)+L_g(0,x_k)}\right]$.

Now note that if the Fréchet inequalities on $\mathbb{P}[Y_i = y, X_i = x_k | G_i = g]$ are sharp for all $y \in \{0,1\}, k = 1,\dots,K, g = 1,\dots,G$ then $\mathbb{P}[Y_i = 0, X_i = x_k | G_i = g]$ attains its minimum at $L_g(0, x_k)$ and its maximum at $U_g(0, x_k)$, and $\mathbb{P}[Y_i = 1, X_i = x_k | G_i = g]$ attains its minimum at $L_g(1, x_k)$ and its maximum at $U_g(1, x_k)$. Thus bounds $\left[\frac{L_g(1,x_k)}{L_g(1,x_k)+U_g(0,x_k)}, \frac{U_g(1,x_k)}{U_g(1,x_k)+L_g(0,x_k)}\right]$ on $\mathbb{E}[Y_i | X_i = x_k, G_i = g]$ are indeed sharp and so $D = D^F$. □