Moment Inequalities in Regressions with Incompletely Observed Covariates

Darwin Yang*

This version: November 1, 2026.

Click here for the latest version.

Abstract

In economic applications with missing or set-valued data, standard assumptions like missing-at-random are untestable and often difficult to justify. In this paper, I study identification in mean regression models with incomplete covariate data under "worst-case" conditions: the process determining observability is unknown and unrestricted. To make progress, I develop a new methodology inspired by the panel literature. I represent model features with minimal assumptions as unobserved heterogeneity, and I then target a minimal collection of functions which have non-negative expectations with respect to the heterogeneity. This approach turns out to be tractable, and I obtain conditional moment inequalities which completely characterize identification in the incomplete data models. The method is also versatile; I illustrate how it can be used in other settings. Finally, numerical exercises demonstrate the identification potential of the derived moment inequalities, as well as their feasibility for use in environments which are close to empirical practice.

*I am deeply grateful to my advisors, Bo Honoré and Kevin Dano, for their support and guidance. I also thank Michal Kolesár, Mark Watson, and Elena Manresa for their feedback and advice on this project. All errors are my own.

1 Introduction

In this paper, I study identification of structural parameters in mean regression models where some covariates are incompletely observed due to an unknown and arbitrary process:

$$Y = h(X, Z^*, \theta) + \varepsilon, \quad \mathbb{E}[\varepsilon | X, Z^*] = 0, \tag{1}$$

where Z^* denotes the affected covariates and θ the parameters of interest. For instance, if Z^* is sometimes missing, the missing-ness probability could be fixed (missing-completely-at-random, MCAR); depend on Y and X (missing-at-random, MAR); or also depend on Z^* itself (missing-not-at-random, MNAR). I allow for all these cases, and Z^* can also be known to lie in a set, rather than being completely missing. Alongside the general setting, I also highlight the special case of binary choice where Z^* is either missing or observed:

$$P(Y = 1 | X, Z^*; \theta) = G(h(X, Z^*, \theta)), \quad Y \in \{0, 1\}. \tag{2}$$

My core results for these models are analytical collections of conditional moment inequalities which exhaust all moment information for θ, yielding outer sets which are non-sharp only due to limit point technicalities. The moment functions are intuitive and well-behaved. As a result, researchers can use them with standard procedures to conduct robust inference on θ, or alternatively to test point-estimates obtained under stronger assumptions. Numerical illustrations show that the inequalities can be useful, even in realistic specifications.

To obtain my findings for (1) and (2), I formalize the idea of finding "non-redundant" moment inequalities in a new way, compared to existing identification approaches which use support functions (ex. Beresteanu, Molchanov, and Molinari (2011), Li (2026)) or core-determining sets (ex. Galichon and Henry (2011), Chesher and Rosen (2017)). In particular, I propose a general heterogeneity-based modeling framework inspired by the panel literature; then, I search for key moment functions which have non-negative expectations for all possible

heterogeneity. This turns out to be analytically tractable in some environments, such as (1) and (2), and I show later that my approach can also be used to study other settings, like games and quantile regression. In general, my framework allows for a wide array of possible heterogeneity, and it can provide a unified and flexible base for studying identification. With that said, final derivations are model-by-model; it is not automatic.

I briefly outline my ideas in the context of (2). The objective is to learn about θ under minimal assumptions on the missing-ness process. This framing leads me to draw a parallel with the fixed effects approach in the panel literature, which places no restrictions on individual-specific heterogeneity. Accordingly, I write the model in exactly the panel fashion: I include unobserved model features as latent heterogeneity, specify outcome probabilities given heterogeneity and parameters, and then view the observed outcome distribution as a mixture of individual-specific ones. In (2), for instance, I specify unobserved heterogeneity as (Z^*, S_0, S_1), where $S_d \in [0, 1]$ is the probability of Z^* being observed when $Y = d$.

To the extent that the panel and microeconometrics literatures are viewed as distinct, my bridge between the two seems to be new. Nevertheless, it is a useful connection, as it allows me to invoke techniques developed for panels. One important method, for example, is Bonhomme (2012)'s functional differencing, which generates moment equality conditions by orthogonalizing out the panel fixed effects. This technique naturally translates within my framework, and it can apply even if heterogeneity is non-standard and infinite-dimensional.

However, there are no moment equalities in (2), and θ is only partially identified. As a result, I instead target a moment inequality characterization based on a *dual cone*: the moment functions which have non-negative expectations for all possible heterogeneity. I show that this approach treats a proposed θ as valid iff the observed outcome distribution belongs to a closed convex hull of the ones which are possible under the model and θ.

The problem thus reduces to characterizing the dual cone. Two of its characteristics are convenient for analysis. First, due to the cone structure, it is only necessary to extract a subset of the moment inequalities; this is analogous to core-determining classes in other

approaches. Second, the essential condition on the functions is a non-negativity one, which is often easy to work with. I exploit these features to obtain the relevant inequalities for (2).

The derived moment inequalities have clean interpretations. Each targets a different subset of the data, and they check whether the partially missing data can rationalize deviations from the binary choice model in the fully-observed parts. As such, one may have originally conjectured them as restrictions. Using my approach, though, I establish the converse as well: this class is exhaustive for identification, and I pinpoint the ones to use.

Lastly, one can also study functionals which jointly depend on both θ and unobservables, despite my focus on the former. Consider a population average marginal effect in (2). This depends on all realizations of Z^*, including unobserved ones. It is possible to characterize the effect's identified set via a two-step approach, using my moment inequalities as the first step. This is standard, so I do not discuss it further. Alternatively, if the two-step method is impractical, one can still perform valid, albeit conservative, inference, as I show in numerical illustrations.

1.1 Related Literature

Many papers have studied robust identification with incompletely observed data since the pioneering work by Horowitz and Manski (1998); a small selection is Horowitz and Manski (2000), Horowitz and Manski (2006), and Franguridi and Moon (2026). See Manski (2003) and Molinari (2020) for reviews. The core motivation for this literature is the fact that standard assumptions like MAR are often inherently untestable. As a result, more robust methods for extracting information can be desirable.

Within this literature, several papers have studied the related problem of best-linear-prediction with incompletely observed data, ex. Horowitz, Manski, Ponomareva, and Stoye (2003), Horowitz and Manski (2006), and Beresteanu, Molchanov, and Molinari (2011). However, this deals with unconditional moment restrictions, ex. $\mathbb{E}[X\varepsilon] = 0$ and $\mathbb{E}[Z^*\varepsilon] = 0$ in (1), rather than conditional ones. For conditional moment contexts, I mention two previous

papers: Manski and Tamer (2002) and Aucejo, Bugni, and Hotz (2017). Manski and Tamer (2002) consider regression with interval-valued covariates, but they add a substantive mean independence assumption. Aucejo, Bugni, and Hotz (2017) drop this assumption and study conditional moment models with missing covariates, which encompasses (2) and partially (1). They constructively characterize identification through being able to find a compatible distribution for the missing data. However, as this characterization is too complex to use, they ultimately work with outer sets instead. In comparison, I explicitly derive key moment inequalities.

In recent contemporaneous work, Fankhauser and Peled (2026) also study conditional moment models with incompletely observed data. They obtain sharp identification by extending Beresteanu, Molchanov, and Molinari (2011)'s method to handle an infinite number of moments. It turns out that my characterization for (1) resembles their main identification result.[1] Nevertheless, there are differences between our papers, which vary in substantiveness. A fundamental one is methodology; I use a completely different approach.

My paper adds to a vast body of work on identification in econometrics. Some references are Ekeland, Galichon, and Henry (2010), Beresteanu, Molchanov, and Molinari (2011), Schennach (2014), Chesher and Rosen (2017), Torgovitsky (2019), Botosaru, Loh, and Muris (2024), and Li (2026). Many of the existing papers already propose duality-based methods, like support functions, and a few have also mentioned dual cones.[2] Nevertheless, my particular framework and approach seem to be new. I comment on differences in more detail later in the paper. For now, I mention that I do not use random set theory, and I do not focus on moments. Instead, I work primarily with distributions for the observables, which is closest to the approach in Botosaru, Loh, and Muris (2024).

Finally, my framework and techniques draw on ideas from the fixed effects tradition in the panel literature. As functional differencing is the main antecedent for what I do here,

[1] Compare equation (8) and Theorem 1 in Fankhauser and Peled (2026) with Theorem 5.1 here.

[2] Mbakop (2023) invokes polyhedral dual cones, as certain semiparametric panel models exhibit model probabilities which form a polyhedron. Dobronyi, Gu, Russell, and Kim (2025) study fixed effects logit-type models, and they mention the dual cone as part of the moment problem.

I mention some core papers in its development: Bonhomme (2012), Honoré and Weidner (2025), and Bonhomme, Dano, and Graham (2026). The last of these, in particular, extends analysis to more general forms of heterogeneity, namely feedback processes, and also considers duality-based identification. As such, it can be viewed as another concurrent counterpart of mine. However, I focus on non-panel settings with different forms of heterogeneity, and the main focus in Bonhomme, Dano, and Graham (2026) is still moment equalities.

1.2 Organization of the Paper

The rest of the paper proceeds as follows. In Section 2, I present a motivating example for the case of binary choice with a missing covariate; then, I use a stylized setting to highlight the intuition and issues behind identification. The discussion suggests a need for a more systematic approach. Therefore, in Section 3, I outline my general framework, after which I present my specific results for (2) and (1) in Sections 4 and 5, respectively. In Section 6, I illustrate the flexibility of my framework by considering its use in some additional models. Finally, Section 7 contains some numerical illustrations for the binary choice setting, and Section 8 concludes. Appendix A contains proofs for the main text, Appendix B contains details on the numerical section, and Appendix C discusses some related topics.

2 Motivation and Discussion

To motivate the contents of this paper, I use the example of statistical binary choice models for outcomes like labor force participation, which are frequently estimated in the labor economics literature. Then, I consider identification in a stylized example with a single binary covariate. The discusion ultimately leads to my methodology and results in later sections.

2.1 Motivating Example

A classic question in economics is the effect of human capital on labor market outcomes. The following specification, for instance, is a simple statistical model which might be used to study the relationship between education and labor force participation:

$$lfp_i = \mathbf{1}\{\gamma \times educ_i + X_i'\beta + \varepsilon_i \geq 0\}, \tag{3}$$

where X_i is a vector of controls. This model may be of independent interest; more commonly, though, it is used as a first-stage in a larger specification, ex. to generate the correction in a Heckman (1979) selection model. A recent example is Borjas and Edo (2026).

Estimation of γ and β in (3) is textbook, given access to complete data on a random sample of individuals. Unfortunately, missing data is widespread in empirical applications (Abrevaya and Donald (2017)). Common responses are to drop observations or perform imputation; other techniques are also available. The validity of these approaches generally rests on strong assumptions, however, which has led to various critiques, ex. Manski (2025).

In (3), it can be unclear if missing-ness in variables should satisfy assumptions like MAR. For instance, highly educated individuals may be more willing to disclose their status. Alternatively, if opportunity costs of time vary with education, this could induce differences in item nonresponse. To be robust against all scenarios, it is necessary to study what can be learned about parameters from the data, absent any further assumptions on the missing-ness.

2.2 Illustration with a Binary Covariate: Intuition and Issues

I now provide an initial discussion of parameter identification when no further assumptions are made. It is possible to derive various restrictions by direct deduction. Nevertheless, I argue that it is difficult to achieve sharpness, or generalize, in this way.

To abstract from other complications and focus purely on the missing data, I examine a

stylized setup with only a single binary explanatory variable, which is potentially missing:

$$Y_i = \mathbf{1}\{\beta_0 + \gamma_0 Z_i^* + \varepsilon \geq 0\}, \quad Y_i, Z_i^* \in \{0,1\}, \quad -\varepsilon \sim G,$$

with G a known continuous and strictly increasing CDF, ex. the standard Normal. In the context of (3), for instance, Z_i^* could be an indicator for being highly educated.

Ignoring estimation issues and dropping i-subscripts, as is standard in identification, I assume that the available data consists of the true joint distribution of the observables, (Y, Z), where $Z = \varnothing$ if Z^* is missing, Z^* otherwise. There are six possible outcomes, given by $\{0,1\} \times \{0,1,\varnothing\}$. For each, the associated probability can be expressed like as follows:

$$P(Y=1, Z=0) = P(Z^*=0) \times \underbrace{P(Y=1|Z^*=0;\beta_0,\gamma_0)}_{G(\beta_0)} \times P(Z \neq \varnothing | Y=1, Z^*=0).$$

The middle term on the RHS is specified by the model and parameters. However, because Z^* is sometimes missing, the first term, i.e. the marginal distribution of Z^*, and the last one, i.e. the missing-ness process, are both unknown. Assumptions like M(C)AR can constrain these latent objects. Otherwise, though, they are completely unrestricted, as I consider here.

As the data is generated by the model and (β_0, γ_0), it is possible to find values for the unknown objects which rationalize the observed probabilities. This is a constructive identification approach. However, it is not practical with more covariates or larger supports.[3] As such, I explore testable restrictions which must hold when (β_0, γ_0) are the true parameters.

To derive some initial restrictions, suppose that $P(Y=1|Z=0)/P(Y=0|Z=0)$ exceeds the ratio which is predicted by the model, assuming exogenous missing-ness. That is, the ratio exceeds $G(\beta_0)/[1-G(\beta_0)]$, and there is a deficit of 0-outcomes in the $Z=0$ data. In that case, the displaced 0-outcomes must be "somewhere else." In particular, they

[3]This type of issue was previously noted by Aucejo, Bugni, and Hotz (2017).

must be in the missing segment. I codify this into the following restriction:

$$\frac{P(Y=1,Z=0)}{P(Y=0,Z=0)+P(Y=0,Z=\varnothing)} \leq \frac{G(\beta_0)}{1-G(\beta_0)},$$

where there is an inequality because $P(Y=0,Z=\varnothing)$ may contain 0-outcomes generated by $Z^*=1$ as well. Looking at the opposite case, $Z=1$, and $Z=0$ and $Z=1$ simultaneously yields additional inequalities in a similar way. Broadly speaking, these restrictions ask whether there is sufficient probability mass in the $Z=\varnothing$ segment to correct any deviations from the model in the $Z\neq\varnothing$ parts.

One might now conjecture that the class above is sharp, i.e. it embodies all model restrictions. In the usual identification program, one would then try to verify that, if a generic observed distribution satisfies the restrictions, then it is always possible to find values for the latent objects which can rationalize the data. However, the conjecture turns out to be false. To see this, consider an observed distribution with $P(Y=1|Z=z^*)/P(Y=0|Z=z^*)$ equal to the model-predicted ratio for all $z^*\in\{0,1\}$, along with $P(Y=1,Z=\varnothing)>0$ and $P(Y=0,Z=\varnothing)=0$. Then, the data trivially satisfies the prior restrictions. However, it is inconsistent with the model and (β_0,γ_0), as the missing segment cannot be explained on its own: the model probability of a 0-outcome is non-zero, and the segment also cannot contain excess 1-outcomes from the $Z\neq\varnothing$ parts, as those already align exactly with the model.

Therefore, there are additional constraints on the missing data. Finding these is more complex. For instance, the following candidate, suggested by the previous example,

$$\underbrace{\min\left\{\frac{G(\beta_0)}{1-G(\beta_0)},\frac{G(\beta_0+\gamma_0)}{1-G(\beta_0+\gamma_0)}\right\}}_{G_{rmin}} \leq \frac{P(Y=1,Z=\varnothing)}{P(Y=0,Z=\varnothing)} \leq \underbrace{\max\left\{\frac{G(\beta_0)}{1-G(\beta_0)},\frac{G(\beta_0+\gamma_0)}{1-G(\beta_0+\gamma_0)}\right\}}_{G_{rmax}},$$

is not valid, as the middle ratio can be generated arbitrarily through appropriate choices of $P(Z=\varnothing|Y,Z^*)$. Accordingly, the additional restrictions must involve probabilities from

cases where Z^* is observed. One example is

$$P(Y=1,Z=0)+P(Y=1,Z=1)+P(Y=1,Z=\varnothing)$$

$$\geq \frac{G(\beta_0)}{1-G(\beta_0)}P(Y=0,Z=0)+\frac{G(\beta_0+\gamma_0)}{1-G(\beta_0+\gamma_0)}P(Y=0,Z=1)+G_{rmin}P(Y=0,Z=\varnothing).$$

This stipulates that the total mass of 1-outcomes must match a lower bound associated with observed 0-outcomes. There are still more restrictions, though, as my later results will show.

Taking a step back, the present example should be stylized enough for a researcher to progressively enumerate valid restrictions while checking for sharpness. There are also core ideas which might generalize beyond the stylized setup. However, rigorously performing this generalization is tedious and potentially difficult, and moreover, it is not clear how to further extend to regression with set-valued data. Therefore, there is a limit on the potential of the ad-hoc method. This leads to the alternative approach summarized in the next subsection.

2.3 An Alternative Approach: Overview

The previous discussion indicates that it may be useful to study identification more systematically in this setting. However, many of the existing identification approaches do not supply satisfactory solutions, especially when attempting to allow for eventual generalizations. Some computational methods are intractable, for instance, when Z^* is continuous.

Instead, I extend the previous subsection in a more direct way. The prior restrictions are linear in observed probabilities. Thus, they are moment inequalities: $\mathbb{E}[m(Y,Z;\beta_0,\gamma_0)]\geq 0$. The ad-hoc approach involves finding a collection of such inequalities and then checking its sharpness. A natural alternative, instead, is to just search for and use all possible functions $m:\{0,1\}\times\{0,1,\varnothing\}\to\mathbb{R}$ which satisfy $\mathbb{E}[m(\cdot)]\geq 0$ if the data is consistent with the model and (β_0,γ_0). This ensures no restrictions are ignored, at least of this type.

At first glance, this idea seems to merely add further work to the ad-hoc approach. However, two features make it potentially tractable. First, it is not necessary to check all

possible m's because some are redundant. For instance, if m generates a valid inequality, then $\widetilde{m}+1$ does as well; however, it is clearly extraneous to verify $\mathbb{E}[\widetilde{m}(\cdot)] \geq 0$ in addition. This means it is possible to focus on a *core-determining* class of functions. Second, the functions are defined by a non-negativity property. This turns out to be highly convenient.

I now summarize how the procedure operates within the stylized binary choice model. If an m is valid with respect to the model and (β_0, γ_0), then it must satisfy

$$\sum_{z^* \in \{0,1\}} P(Z^* = z^*) \Bigg[\sum_{y \in \{0,1\}} \underbrace{P(Y = y | Z^* = z^*)}_{G(\beta_0+\gamma_0 z^*)^y [1-G(\beta_0+\gamma_0 z^*)]^{1-y}} \Bigg[m(y, z^*) P(Z \neq \varnothing | Y = y, Z^* = z^*) + m(y, \varnothing) P(Z = \varnothing | Y = y, Z = z^*) \Bigg] \Bigg] \geq 0,$$

where the latent objects are unrestricted in $[0, 1]$, apart from being probabilities. As I am interested in non-negativity, because the relevant values enter the formula linearly, it is enough to verify non-negativity only for the degenerate configurations, ex. $P(Z^* = 0) \in \{0, 1\}$. Writing out the corresponding conditions leads to the following constraints on m:

$$m(0, \varnothing) \times [1 - G(\beta_0 + \gamma_0 z^*)] + m(1, \varnothing) \times G(\beta_0 + \gamma_0 z^*) \geq 0,$$

$$m(0, \varnothing) \times [1 - G(\beta_0 + \gamma_0 z^*)] + m(1, \varnothing) \times G(\beta_0 + \gamma_0 z^*) \geq 0,$$

$$m(0, z^*) \times [1 - G(\beta_0 + \gamma_0 z^*)] + m(1, \varnothing) \times G(\beta_0 + \gamma_0 z^*) \geq 0,$$

$$m(0, z^*) \times [1 - G(\beta_0 + \gamma_0 z^*)] + m(1, z^*) \times G(\beta_0 + \gamma_0 z^*) \geq 0,$$

for all $z^* \in \{0, 1\}$. To obtain a core collection of m's which satisfy these constraints, I then apply a chain of arguments, summarized below.

1. Suppose $m(y, z) < 0$ for some (y, z); otherwise, it always has a non-negative expectation and is redundant. Then, either $m(0, \varnothing) > 0$ or $m(1, \varnothing) > 0$ by the constraints.

2. Because $m(0, \cdot)$ is only constrained by $m(1, \cdot)$ and vice-versa, I can split the valid

m's into two cases: either $m(0, \varnothing) > 0$ and $m(1, \varnothing) \leq 0$, or the reverse.[4] Letting $m(0, \varnothing) > 0 \geq m(1, \varnothing)$ for concreteness, I normalize $m(0, \varnothing) = 1$, as scale is irrelevant.

3. Using lower bounds from the non-negativity constraints and upper bounds from redundancy considerations, I pin down key values (or sets of values) for m. For example, the third constraint in the previous system, coupled with $m(0, \varnothing) = 1$, implies that

$$m(1, z^*) \geq -\min\{m(0, z^*), 1\} \times \frac{1 - G(\beta_0 + \gamma_0 z^*)}{G(\beta_0 + \gamma_0 z^*)}$$

for all $z^* \in \{0, 1\}$. This means that for any fixed value of $m(0, z^*)$, it is redundant to set $m(1, z^*)$ above the RHS above. I am also able to exclude the case $m(0, z^*) > 1$, as setting a value above one does not allow for a larger choice of $m(1, z^*)$.

4. From the previous step, I find that certain values of m lie in intervals, while ultimately entering any expectation of m in a linear manner. As such, by extreme point arguments, when attempting to verify that expectations are non-negative for all relevant moment functions, I can ignore the m's which do not use the interval endpoints for the values. For example, since I have excluded $m(0, z^*) > 1$, the RHS of the previous display resolves to $-m(0, z^*) \times \frac{1-G}{G}$. The term $m(0, z^*)$ thus enters linearly in $\mathbb{E}[m(\cdot)]$.

5. Collecting the "extreme" functions together, I obtain one class of moment inequalities. There is also a symmetric class for the case $m(0, \varnothing) \leq 0$, $m(1, \varnothing) > 0$.

In the end, I arrive at a core collection of moment inequalities: if $\mathbb{E}[m(\cdot)] \geq 0$ for all m in this collection, then the same holds for all valid m's. For some initial intuition, note that the decomposition in Step 2 is consistent with my separate treatment of 0- and 1-outcomes in Section 2.2. Meanwhile, the collection also contains the Section 2.2 restrictions, alongside others with similar structures. For further details, see the general result in Section 4.

[4] Formally, if a valid m satisfies $m(0, \varnothing), m(1, \varnothing) > 0$, then it is possible to decompose it as $m = \widetilde{m} + \widehat{m}$, where $\widetilde{m}$ and $\widehat{m}$ are also valid while falling under different cases.

By design, the obtained collection utilizes all moment inequality information in the model. Nevertheless, some issues remain. First, it is still unknown exactly what the restrictions identify and if they are sharp. Second, there may be complications when attempting to extend to continuous Z^*'s which are not apparent in the stylized setup. Finally, even if I prove results on these for the binary choice model with missing-ness, I am also interested in general regression with set-valued data, and my approach could also be useful in other settings. Therefore, it would ideally not be necessary to establish results, case-by-case.

Accordingly, in the next section I develop a general framework which addresses these points. Within the framework, I establish that my exhaustive moment inequality approach essentially achieves sharp identification. Meanwhile, another result handles the extension from discreteness to continuous cases. The general theory thus provides the supporting structure for my results in Sections 4 and 5. There is also discussion of some technical issues in identification, and in Section 6 I consider potential applications to other models.

3 General Framework and Methodology

In this section, I outline the formalities of my approach.

As a brief prelude, I note that I develop my framework in a general way, rather than tying it to specific settings. This lets it accommodate a variety of possible model features. At the same time, however, the generality can make the ideas seem disconnected from applications. I therefore use a running example, linear regression, to help keep the discussion concrete:

$$Y = W'\theta + \varepsilon, \quad \mathbb{E}[\varepsilon|W] = 0. \tag{4}$$

I use this, rather than the stylized model from Section 2, for two reasons. First, unlike the binary choice setting, (4) has a nonparametric component, i.e. the conditional distribution of ε given W. Second, it has an expected answer: the model is defined by a moment condition, so a search for non-redundant inequalities should return the moment. Accordingly, (4)

provides a basic check for my approach, and it also previews some aspects of how I later handle general regression. I return to the incomplete data contexts in Sections 4 and 5.

As a last preliminary, I state some notational conventions. I generally use stylized letters for sets, capital letters for random variables, and lowercase letters for realized values; one common exception is F, which I reserve for CDFs and distributions, often used interchangeably. Given two sets $\mathcal{R}$ and $\mathcal{S}$, $\mathcal{S}^{\mathcal{R}}$ is the set of functions from $\mathcal{R}$ to $\mathcal{S}$. If $\mathcal{R}$ is a measurable space, then $\mathcal{M}(\mathcal{R})$ is the space of finite signed measures on $\mathcal{R}$, with $\mathcal{M}_1(\mathcal{R})$ the subset of probability measures. The σ-algebra of $\mathcal{R}$ is usually arbitrary and left implicit, unless $\mathcal{R}$ is a metric or topological space. In that case, it is always endowed with the Borel σ-algebra, and $\mathcal{M}(\mathcal{R})$ consists of Borel measures. A product space always has the product topology. I state any other notation (or deviations from my conventions) as needed.

3.1 Model

I begin my exposition by describing the overarching setup.

An econometrician observes the joint distribution of outcome variables Y and conditioning variables W, which respectively take values in a metric space $\mathcal{Y}$ and measurable space $\mathcal{W}$.[5] This joint distribution is equivalently expressed as the collection of conditional outcome distributions $\{F_{Y|W=w}\}_{w\in\mathcal{W}}$, which I work with henceforth.[6]

The key econometric assumption is that the conditional outcome distributions depend on a structural parameter of interest, $\theta \in \Theta$, and unobserved individual-specific heterogeneity, $A \in \mathcal{A}$, where $\mathcal{A}$ is a metric space. In particular, a model consists of two objects: a "model map," $\mu^0 : \mathcal{W} \times \mathcal{A} \times \Theta \to \mathcal{M}_1(\mathcal{Y})$, and a collection of "admissible heterogeneity,"

[5] The main implication I use from $\mathcal{Y}$ being a metric space is that its singleton sets are measurable.

[6] One can impose standard assumptions on $\mathcal{Y}$ and $\mathcal{W}$ to ensure that a regular conditional distribution exists. I will only work with the individual conditional distributions.

$\mathcal{P}_A^0 \subseteq \mathcal{M}_1(\mathcal{A})$,[7] which each have the natural meanings given their names. The model map,

$$(w, a, \theta) \mapsto \mu^0(w, a, \theta) \equiv \mu^0_{w,a,\theta},$$

captures the imposed parametric structure: it specifies a probability measure on $\mathcal{Y}$ for every possible (w, a, θ). Meanwhile, $\mathcal{P}_A^0$ defines the possible distributions of interest for the individual-specific unobservables. Each such heterogeneity distribution, $\nu \in \mathcal{P}_A^0$, generates an outcome distribution under μ^0 as follows, which I denote by $\mu^0_{w,\nu,\theta} \in \mathcal{M}_1(\mathcal{Y})$:

$$\mu^0_{w,\nu,\theta}(\mathcal{S}) := \int_{\mathcal{A}} \mu^0_{w,a,\theta}(\mathcal{S}) d\nu(a), \quad \forall \text{ measurable } \mathcal{S} \subseteq \mathcal{Y},$$

assuming that $a \mapsto \mu^0_{w,a,\theta}(\mathcal{S})$ is always measurable. Thus, a model $(\mu^0, \mathcal{P}_A^0)$ induces the following set of possible outcome distributions for every w and θ:

$$\mathcal{P}_Y^0(w, \theta) \equiv \mathcal{P}_Y^0(w, \theta; \mu^0, \mathcal{P}_A^0) := \left\{ \mu^0_{w,\nu,\theta} \;\middle|\; \nu \in \mathcal{P}_A^0 \right\}.$$

I will often leave out μ^0 and $\mathcal{P}_A^0$ in my notation, unless I am comparing different ones.

I condense the relevant setup details into the following assumption.

Assumption 3.1. *(i) $\mathcal{Y}$ is a metric space, $\mathcal{W}$ is a measurable space, $\mathcal{A}$ is a metric space, and Θ is a set. (ii) There exists a known model map, $\mu^0 : \mathcal{W} \times \mathcal{A} \times \Theta \to \mathcal{M}_1(\mathcal{Y})$. (iii) There exists a known collection of admissible heterogeneity, $\mathcal{P}_A^0 \subseteq \mathcal{M}_1(\mathcal{A})$. (iv) For all $(w, \theta) \in \mathcal{W} \times \Theta$, the map $a \mapsto \mu^0_{w,a,\theta}(\mathcal{S})$ is measurable for all measurable $\mathcal{S} \subseteq \mathcal{Y}$.*

Condition (iv) is a basic regularity condition which generally holds. In this paper, I verify it directly for a few models by examining $a \mapsto \mu^0_{w,a,\theta}(\mathcal{S})$ for a sufficient class of $\mathcal{S}$'s.[8][9]

[7] In principle, $\mathcal{A}$ and $\mathcal{P}_A^0$ can depend on w and θ, with minimal adjustments to other statements. This is not necessary for the models in this paper, though, so I avoid introducing the additional notation.

[8] It suffices to check measurability for a class of generators of the σ-algebra which is closed under finite intersections and which contains $\mathcal{S}_1, \mathcal{S}_2, ...$ with $\mathcal{S}_k \uparrow \mathcal{Y}$. See, for instance, Remark 8.26 in Klenke (2020).

[9] If (iv) is difficult to establish, or if it is not necessary to consider uncountable heterogeneity mixtures, then my framework can still be applied by restricting attention to finite or countable mixtures. In such cases, $\mu^0_{w,\nu,\theta}$ remains well-defined as a sum. Some results become weaker, though.

For some intuition, consider the running example: $Y = W'\theta + \varepsilon$ and $\mathbb{E}[\varepsilon|W] = 0$. Fix realizations w for the covariates. Then, θ and a mean-zero conditional CDF for ε imply a conditional distribution for Y in the obvious way. Mapping into the general setting, the regression coefficients correspond to θ, the error CDF to A, and the $w'\theta$-translation of the error distribution to μ^0. Moreover, letting F_ε denote a generic mean-zero CDF equipped with the supremum norm, the function $F_\varepsilon \mapsto \mu^0_{w,F_\varepsilon,\theta}((-\infty, c]) = F_\varepsilon(c - w'\theta)$ is continuous for all $c \in \mathbb{R}$. As a result, (iv) holds, and (4) can be represented under Assumption 3.1.

I now discuss the role of A and $\mathcal{P}^0_A$. I view A as representing features about which the econometrician wishes to make minimal assumptions. Accordingly, I consider $\mathcal{P}^0_A$'s which at least contain all Dirac measures, and I also allow for A to vary in an unrestricted way across w's.[10] There are two interpretations of this, which affects the choice of $\mathcal{P}^0_A$. One is that heterogeneity indeed exists in the DGP. In (4), for instance, one individual might draw a Gaussian ε, while another might draw a logistic one. The observed distribution then integrates over heterogeneity. This view is the panel one, and it suggests a rich convex $\mathcal{P}^0_A$.

A second interpretation is that A consist of nuisance objects against which the econometrician wishes to be robust. In the regression model, every individual might draw from the same error distribution. However, the distribution may be difficult to estimate or uninteresting, so that determining it is irrelevant. One can then set $\mathcal{P}^0_A$ to be the Dirac measures.

The two views above are sometimes econometrically equivalent; convexity of $\mathcal{A}$ is one ingredient. However, it is not always the case. In this paper, I mostly follow the first perspective, which I consider to be closer in spirit to the minimal assumptions paradigm. As such, I freely treat $\mathcal{P}^0_Y(w, \theta)$ and its convex hull as essentially identical. If one prefers the second perspective instead, my methodology discards some information. Regardless, though, I consider the following assumption to be reasonable.

Assumption 3.2. *Admissible heterogeneity $\mathcal{P}^0_A$ contains all Dirac measures on $\mathcal{A}$.*

[10] That is, the heterogeneity distribution at $W = w$ is completely unrelated to the one at $W = w'$. Meanwhile, as a reminder, a Dirac measure is just a point mass distribution.

As a last comment, it is useful to specify $\mathcal{P}_A^0$, rather than picking a default one like $\mathcal{M}_1(\mathcal{A})$, for two reasons. First, it is possible to allow for too many or too few mixtures, loosely speaking, on a model-by-model basis. For example, a pathology that arises in unrestricted regression is that an infinite mixture of mean-zero distributions can have no moments,[11] whereas no such problem occurs in binary choice. As a result, flexibility in $\mathcal{P}_A^0$ can be helpful. The second reason is that I will later connect identification results across different $\mathcal{P}_A^0$'s. Flexibility in it thus enables me to state results in greater generality.

3.2 Identified Set

I now define the (sharp) identified set for the parameters θ, ex. the regression coefficients in (4). Note that it is without loss to perform the identification analysis conditional on W, due to the lack of restrictions on A across w's. Accordingly, I condition on $W = w$ from now on for some generic w, and I work with conditional identified sets.

Formally, I define the identified set, Θ_I, at w as follows:

$$\Theta_I(F_{Y|W=w}|w) \equiv \Theta_I(F_{Y|W=w}|w; \mu^0, \mathcal{P}_A^0) := \left\{ \theta \in \Theta \;\middle|\; F_{Y|W=w} \in \mathcal{P}_Y^0(w, \theta; \mu^0, \mathcal{P}_A^0) \right\}.$$

The identified set contains the θ's for which the observed data is consistent with the model. In the running example with $\mathcal{P}_A^0$ set to the Dirac measures, for instance, this requires $F_{Y|W=w}$ to be rationalizable by some mean-zero conditional error distribution, given $w'\theta$.

My chosen definition is natural within my framework. Nevertheless, there are alternatives. A prominent one is to take a closure of the model distributions under some topology,

$$\left\{ \theta \in \Theta \;\middle|\; F_{Y|W=w} \in \overline{\mathcal{P}_Y^0(w, \theta)} \right\}.$$

One example of this is Botosaru, Loh, and Muris (2024), who work with the topology on $\mathcal{M}(\mathcal{Y})$ induced by the total variation (TV) norm (albeit not necessarily in a heterogeneity-

[11] For example, consider an infinite sequence of two-point distributions which expands outwards.

based setting). An important justification for this is given by Bertanha and Moreira (2020), who showed that TV-separation is essential for the existence of tests with non-trivial power.

As my methodology also ends up checking membership in a set which can be viewed as a closure, it is not drastically different from the other approach. Nevertheless, it is unappealing to ignore the heterogeneity structure. Also, by not taking an initial stand in my definition, I am able to highlight a technical point: the choice of topology is substantive. I show in Appendix C that the TV-norm-closure identified set for model (4) is trivial, i.e. Θ, although θ is generally point-identified by the moment condition. In contrast, the topology induced under my approach can be finer, and in (4) I recover the moment restriction. My approach thus presents an option which tries to further separate identification from inference.

3.3 Dual Cone Moment Inequalities

The definition of Θ_I suggests a direct approach to identification: given θ, try to find a heterogeneity distribution in $\mathcal{P}_A^0$ which generates $F_{Y|W=w}$. This is a constructive approach, first explored in fixed effects panel settings by Honoré and Tamer (2006). However, it is impractical in the cases I consider, where heterogeneity is often complex; in (4), $\mathcal{A}$ is a set of CDFs. The intractability of the direct approach is a major reason for alternative methods.

My dual cone approach provides such an alternative. I formalize it and characterize its output relative to Θ_I. In the next subsection, I discuss why it might be attractive to use.

As the goal is to be robust to A, I propose to use all moment inequalities with that property. Fix some θ in addition to w, and let $\mathcal{L}^1(\mathcal{P}_Y^0(w,\theta))$ be the set of measurable real-valued functions of $\mathcal{Y}$ which are (absolutely) integrable with respect to all model probabilities:[12]

$$\mathcal{L}^1(\mu) := \left\{ m : \mathcal{Y} \to \mathbb{R} \;\middle|\; m \text{ measurable and } \int |m| d|\mu| < \infty \right\}, \quad \text{for } \mu \in \mathcal{M}(\mathcal{Y}),$$

$$\mathcal{L}^1(\mathcal{P}_Y^0(w,\theta)) := \bigcap_{\mu \in \mathcal{P}_Y^0(w,\theta)} \mathcal{L}^1(\mu).$$

[12] In this paper, integrability always refers to absolute integrability.

From this collection of "model-integrable" test funtions, I further isolate the subset

$$\mathcal{D}(w,\theta) := \left\{ m \in \mathcal{L}^1(\mathcal{P}_Y^0(w,\theta)) \;\middle|\; \int m d\mu \geq 0, \quad \forall \mu \in \mathcal{P}_Y^0(w,\theta) \right\}$$

for which the model-related integrals are always non-negative. I then finally suggest to use this subset of test functions as moment inequalities:

$$\Theta_I^+(F_{Y|W=w}|w) := \left\{ \theta \in \Theta \;\middle|\; F_{Y|W=w} \in \mathcal{M}_{w,\theta} \text{ and } \int m dF_{Y|W=w} \geq 0, \quad \forall m \in \mathcal{D}(w,\theta) \right\},$$

where $\mathcal{M}_{w,\theta} \subseteq \mathcal{M}(\mathcal{Y})$ is the subspace of measures with respect to which all functions in $\mathcal{L}^1(\mathcal{P}_Y^0(w,\theta))$ are integrable:

$$\mathcal{M}_{w,\theta} := \left\{ \mu \in \mathcal{M}(\mathcal{Y}) \;\middle|\; m \in \mathcal{L}^1(\mu), \quad \forall m \in \mathcal{L}^1(\mathcal{P}_Y^0(w,\theta)) \right\}.$$

The set Θ_I^+ is the product of my approach. In words, I include θ in Θ_I^+ iff it is consistent with moment information which is available under the model.

By construction, if $F_{Y|W=w}$ belongs to the set of model distributions $\mathcal{P}_Y^0(w,\theta)$, then $F_{Y|W=w} \in \mathcal{M}_{w,\theta}$, and all functions in $\mathcal{D}(w,\theta)$ have a non-negative expectation. This means that Θ_I^+ is at least an outer set: $\Theta_I(F_{Y|W=w}|w) \subseteq \Theta_I^+(F_{Y|W=w}|w)$. For a more complete comparison, though, I provide the proposition below. It arises because $\mathcal{D}(w,\theta)$ is a type of dual cone. This also explains my chosen terminology.

Proposition 3.1. *Consider the setting defined by Assumptions 3.1 and 3.2, and fix $(w,\theta) \in \mathcal{W} \times \Theta$. Then, for any $w \in \mathcal{W}$ and $\mu \in \mathcal{M}_1(\mathcal{Y})$,*

$$\begin{aligned}
\Theta_I^+(\mu|w) &= \left\{ \theta \in \Theta \;\middle|\; \mu \in \mathcal{M}_{w,\theta} \text{ and } \sup_{a \in \mathcal{A}} \int m d\mu_{w,a,\theta}^0 \geq \int m d\mu, \quad \forall m \in \mathcal{L}^1(\mathcal{P}_Y^0(w,\theta)) \right\} \\
&= \left\{ \theta \in \Theta \;\middle|\; \mu \in \overline{co}(\{\mu_{w,a,\theta}^0 \mid a \in \mathcal{A}\}) \right\},
\end{aligned}$$

where, in the second line of the display, the closure of the convex hull is taken in $\mathcal{M}_{w,\theta}$ under

the weak topology induced by $\mathcal{L}^1(\mathcal{P}_Y^0(w,\theta))$.[13]

The first equality gives a support-function-esque representation. The second says that, alongside $\mathcal{P}_Y^0(w,\theta)$, limit points are used which cannot be distinguished via the test functions.

There are two points to address about Θ_I^+ and Proposition 3.1. The first concerns the requirement $F_{Y|W=w} \in \mathcal{M}_{w,\theta}$. It says that the observed distribution should have the same "integration profile" as the model. For example, suppose that all model distributions have finite support; then, all measurable $m : \mathcal{Y} \to \mathbb{R}$ are model-integrable. However, many of these are not integrable with respect to continuous distributions. A basic example is the Uniform distribution on $\mathcal{Y} = [0,1]$, alongside the function which equals $1/y$ if $y > 0$, zero otherwise. Continuous distributions are thus ruled out.

An explicit verification of the integrability condition is infeasible. Nevertheless, I view the condition as a mild one which can generally be assumed, as long as one chooses a suitable μ^0 and $\mathcal{P}_A^0$. If a researcher observes a rich range of outcomes, then they should not specify a $\mathcal{P}_Y^0(w,\theta)$ which only contains discrete measures. Beyond such heuristics, I do not consider the integrability issue further; I instead focus on the moment inequality conditions.

The second point to address about Θ_I^+ relates to the limit points which it includes; as mentioned, I view convexity as natural. An immediate question is whether the limit points admit an interpretation beyond their definitions. Unfortunately, the generic answer to this question is no. Therefore, I give the following defense instead: I do not include limit points which are excluded by other approaches. For example, Botosaru, Loh, and Muris (2024)'s methodology ultimately uses a smaller collection of test functions, the measurable and bounded ones. As a result, Θ_I^+ is weakly smaller than the set which their method would produce. Per the earlier discussion, there is thus a sense in which Θ_I^+ cannot be statistically distinguished from Θ_I.[14] Accordingly, I view Θ_I^+ as functionally replicating the identified set, with an approximation at least as justifiable as other ones in the literature.

[13] This topology is the coarsest one on $\mathcal{M}_{w,\theta}$ such that the linear integral functionals $\mu \mapsto \int m d\mu$, $m \in \mathcal{L}^1(\mathcal{P}_Y^0(w,\theta))$, are continuous.

[14] Botosaru, Loh, and Muris (2024) additionally assume an absolute continuity condition for the outcome distributions.

3.4 Discussion and Implementation

It remains to describe how the dual cone methodology can be used in practice, along with why it can sometimes have advantages over competing alternatives. This reproduces some of the discussion in Section 2.3, but I nevertheless repeat some points for emphasis.

The most important point is that it is not necessary to actually check all functions in $\mathcal{D}(w,\theta)$ as moment inequalities. This parallels the core-determining classes which have appeared in other approaches, ex. Galichon and Henry (2011). A succinct explanation invokes the fact that $\mathcal{D}(w,\theta)$ is a cone; therefore, it suffices to check only the extreme rays or generators. More explicitly, observe that it is superfluous to check any $m \geq 0$, as the expectation is automatically non-negative. Also, suppose that m_1 and m_2 are both in $\mathcal{D}(w,\theta)$, but $m_1 \geq m_2$, i.e. m_1 (weakly) point-wise dominates m_2. Then, it is evident that it is only necessary to check m_2.

I formalize the previous considerations through the following definition. The choice of terminology is to highlight the analogy with the other approaches in the literature.

Definition 3.1. *Consider the setting defined by Assumptions 3.1 and 3.2. Then, for $(w,\theta) \in \mathcal{W} \times \Theta$, a subcollection $\mathcal{D}^*(w,\theta)$ of $\mathcal{D}(w,\theta)$ is* ***core-determining*** *at (w,θ) if, for any $\mu \in \mathcal{M}_{w,\theta} \cap \mathcal{M}_1(\mathcal{Y})$,*

$$\int m^* d\mu \geq 0 \quad \forall m^* \in \mathcal{D}^*(w,\theta) \quad \Rightarrow \quad \int m d\mu \geq 0 \quad \forall m \in \mathcal{D}(w,\theta).$$

As $\mathcal{D}^*(w,\theta) \subseteq \mathcal{D}(w,\theta)$, the other direction of the display is immediate. Therefore, assuming $F_{Y|W=w} \in \mathcal{M}_{w,\theta}$, it is necessary and sufficient for identification, in terms of Θ_I^+, to check the moment inequalities in $\mathcal{D}^*(w,\theta)$. This simplification is essential for practical use.

I can now explain a certain extension procedure which will let me handle continuous variables in the incomplete data models. To prove that a class of functions is core-determining, it is often easier to establish the following sufficient property:

Definition 3.2. *Consider the setting defined by Assumptions 3.1 and 3.2. Then, for* $(w, \theta) \in \mathcal{W} \times \Theta$*, a subcollection* $\mathcal{D}^*(w, \theta)$ *of* $\mathcal{D}(w, \theta)$ *is* ***strongly core-determining*** *at* (w, θ) *if, for any* $m \in \mathcal{D}(w, \theta)$ *and* $\mu \in \mathcal{M}_1(\mathcal{Y})$*,*

$$\mathcal{D}^*(w, \theta) \cup \{m\} \subseteq \mathcal{L}^1(\mu) \quad \text{and} \quad \int m^* d\mu \geq 0 \quad \forall m^* \in \mathcal{D}^*(w, \theta) \quad \Rightarrow \quad \int m d\mu \geq 0.$$

In this definition, for any fixed $m \in \mathcal{D}(w, \theta)$, μ-integrability of m and $\mathcal{D}^*(w, \theta)$ is enough for non-negativity to "pass through" from $\mathcal{D}^*(w, \theta)$ to m. In contrast, the definition of core-determining restricts attention to μ's in $\mathcal{M}_{w,\theta}$. A core-determining class thus relies on the model in a stricter manner than a strongly core-determining one.

By partially decoupling from $\mathcal{P}_A^0$, a strongly core-determining class allows for linkages across different choices for admissible heterogeneity. This is useful due to the proposition below. I re-introduce dependence on the model here as I am comparing two different ones.

Proposition 3.2. *Consider two identical settings defined by Assumptions 3.1 and 3.2, distinguished only by their admissible heterogeneity,* $\mathcal{P}_A^0$ *and* $\widetilde{\mathcal{P}}_A^0$*. Fix some* $(w, \theta) \in \mathcal{W} \times \Theta$*, and suppose that* $\mathcal{D}^*(w, \theta)$ *is strongly core-determining at* (w, θ) *in the* $\mathcal{P}_A^0$ *model. If* $\mathcal{P}_A^0 \subseteq \widetilde{\mathcal{P}}_A^0$ *and* $\mathcal{D}^*(w, \theta) \subseteq \mathcal{L}^1(\mathcal{P}_Y^0(w, \theta; \mu^0, \widetilde{\mathcal{P}}_A^0))$*, then* $\mathcal{D}^*(w, \theta)$ *is also strongly core-determining at* (w, θ) *in the* $\widetilde{\mathcal{P}}_A^0$ *model.*

The main condition is that $\mathcal{D}^*(w, \theta)$ must remain integrable, which is easy to check.[15]

Proposition 3.2 suggests a standard program for studying identification. One first investigates the model using a simple choice for $\mathcal{P}_A^0$. Then, the forms of the derived strongly core-determining functions define a range of admissible heterogeneity for which they are still sharp. If the functions are all bounded, for instance, then the identification results apply to all $\widetilde{\mathcal{P}}_A^0$ with $\mathcal{P}_A^0 \subseteq \widetilde{\mathcal{P}}_A^0 \subseteq \mathcal{M}_1(\mathcal{A})$. In the following sections, I implement this program to

[15] To give some intuition for Proposition 3.2, observe that expanding admissible heterogeneity leads to a larger set of model outcome distributions. This shrinks the set of test functions and thus the dual cone. The shrinkage direction aligns with the needed one for inheritance of the strong core-determining property.

If $\widetilde{\mathcal{P}}_A^0$ is instead a subset, the claim inverts. In that case, integrability and validity of $\mathcal{D}^*(w, \theta)$ in the $\widetilde{\mathcal{P}}_A^0$ model is automatic. However, the dual cone may grow, which can break the core-determining property.

obtain my results. My starting choice for $\mathcal{P}_A^0$ is generally the set of finite mixtures. This ensures that I only need to check conditions with respect to individual $\mu_{w,a,\theta}^0$'s, i.e. individual possible values for heterogeneity. I then push results forward through Proposition 3.2.

The remaining issue is how one can derive a dual cone, $\mathcal{D}(w,\theta)$, and core-determining class, $\mathcal{D}^*(w,\theta)$, in the first place. This process is necessarily model-specific. Nevertheless, the dual cone approach has an attractive feature, which is that the defining property is a non-negativity one. Non-negativity is well-suited for analytical results; for example, scaling factors can be normalized. It also helps when reducing to a core-determining class. As was mentioned, one can ignore the non-negative orthant and look for point-wise domination.

As an alternative to direct analysis, in some cases it is also possible to draw on existing results for non-negative functions in other parts of mathematics. In a concurrent paper, for instance, I study a class of semiparametric panel models using my dual cone approach. I show that model probabilities in those settings often have polynomial representations. I can then appeal to the rich algebraic geometry results which exist for non-negative polynomials.

As one more remark, it may be possible to mechanically obtain valid moment inequalities by projecting functions onto the dual cone. This extends Bonhomme (2012)'s original idea of projecting onto an orthogonal complement in functional differencing, and it may represent the automatic computational version of my approach. I do not pursue this thought in the paper, but it may be an interesting direction for further research.

To make the above points concrete, I return once more to model (4), with finite mixture heterogeneity imposed with respect to the set of mean-zero error CDFs. Fix some realization w for the covariates, as well as some θ for the regression coefficients. Then, the non-negativity dual cone condition on $m : \mathcal{Y} \to \mathbb{R}$ becomes $\int m(w'\theta + e)dF_\varepsilon(e) \geq 0$ for all mean-zero CDFs F_ε. The scaled residual functions, $m(y; w, \theta) = r(y - w'\theta)$, $r \in \mathbb{R}$, satisfy this condition, but many other functions do as well, such as when a positive constant is added. Nevertheless, one can show that a valid m cannot "dip" below a residual function anywhere, via appealing to two-point distributions for F_ε. Similarly, one can eliminate as redundant all functions

which "peak" anywhere above a residual function. After normalizing scale, I thus obtain the pair $m(y; w, \theta) = \pm(y - w'\theta)$ as core-determining, which confirms that Θ_I^+ reduces to checking the moment condition $\mathbb{E}[Y - W'\theta|W] = 0$. See Section 5 for additional details.

With the prior discussion complete, I can now finally make some closer comparisons with other approaches in the literature. In particular, consider support function techniques, which provide an obvious counterpart due to the characterization of Θ_I^+ in Proposition 3.1. Although the methods are all mathematically connected by being duality-based, the preceding discussion indicates that there can sometimes be reasons to prefer my approach. By targeting analytical tractability, my methodology can be useful in situations where implementations of the support function method, which are often computational, are infeasible, ex. in some cases with continuous variables. Moreover, I eliminate extraneous restrictions through checking for point-wise domination. The support function counterpart of this is to determine superfluous directions. However, this seems to be a more difficult task.

To conclude the section, I address final implementation by applied researchers once identification has been studied. This is straightforward: like many other methods in the literature, mine produces conditional moment inequalities. As a result, subsequent estimation and inference can proceed via standard methods. A few references on conditional moment inequality methods are Andrews and Shi (2013), Andrews and Shi (2017), and Andrews, Roth, and Pakes (2023). Unconditional moment inequality methods are also usable.

With my framework in hand, I now return to studying the specific regression settings with incomplete data. There is one more idea related to my approach, which involves formalizing the common wisdom that point-identification fails when no moment conditions exist. This is outside the primary focus of the paper, though, so I leave it in Appendix C.

4 Binary Choice with a Missing Covariate

In this section, I use my framework to investigate the general version of the stylized setup from Section 2. I do this by following the identification program described in Section 3.4.

To be precise, I study the following binary choice model:

$$P(Y = 1|X, Z^*; \theta) = G(X'\beta + \gamma Z^*), \tag{5}$$

where the parameters of interest are the regression coefficients $\theta = (\beta, \gamma) \in \mathbb{R}^{d_X+1}$. The variable $Y \in \{0, 1\}$ is the binary outcome, $X \in \mathcal{X} \subseteq \mathbb{R}^{d_X}$ and $Z^* \in \mathcal{Z}^* \subseteq \mathbb{R}$ are real-valued covariates, and $G : \mathbb{R} \to (0, 1)$ is a known link function.[16] I consider a linear index for simplicity, but my results still apply if one uses a more general form, $h(X, Z^*, \theta)$; θ can even be infinite-dimensional. To handle this, replace $x'\beta + \gamma z^*$ with $h(x, z^*, \theta)$ everywhere.

The central model feature here is that, although Y and X are always observed, Z^* sometimes is not. Moreover, this missing-ness is allowed to depend on all other objects in an arbitrary way. Thus, for instance, the data can be missing-not-at-random.

Due to the previous discussion in Section 2, I proceed quickly through some details in my exposition. The main result is a (strongly) core-determining collection of conditional moment inequalities. I conduct some related numerical exercises in Section 7.

4.1 Model Setup

I start my analysis by formulating the model as in Assumption 3.1. Let $\mathbb{R}_\varnothing$ denote the augmentation of $\mathbb{R}$ with the isolated point $\varnothing$. I endow $\mathbb{R}_\varnothing$ with the usual metric, extended so that the distance between $\varnothing$ and any $c \in \mathbb{R}$ is always one. The Borel sets of $\mathbb{R}_\varnothing$ are then just unions of the usual ones in $\mathbb{R}$ with either $\{\varnothing\}$ or $\emptyset$.

Assumption 4.1. *Fix $d_X \in \mathbb{N}$, let $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ and $\mathcal{Z}^* \subseteq \mathbb{R}$ be non-empty closed sets, and let $G : \mathbb{R} \to (0, 1)$ be a measurable function. Define $\mathcal{Z} = \mathcal{Z}^* \cup \{\varnothing\} \subseteq \mathbb{R}_\varnothing$. Then, in the notation*

[16] This accommodates linear probabilities, as long as all predicted means are indeed probabilities.

of Assumption 3.1:

- $\mathcal{Y}$ *corresponds to* $\{0,1\} \times \mathcal{Z}$, $\mathcal{W}$ *corresponds to* $\mathcal{X}$, $\mathcal{A}$ *corresponds to* $\mathcal{Z}^* \times [0,1] \times [0,1]$, *and* $\Theta \subseteq \mathbb{R}^{d_X} \times \mathbb{R}$.

- *The model map* $\mu^0 : \mathcal{X} \times \mathcal{Z}^* \times [0,1] \times [0,1] \times \Theta \to \mathcal{M}_1(\{0,1\} \times \mathcal{Z})$ *is defined as*

$$\begin{aligned}\mu^0_{x,z^*,s_0,s_1,\theta}(\{y,z\}) &\equiv \mu^0(\{y,z\}|x,z^*,s_0,s_1;\theta) \\ &= \mathbf{1}\{z = z^*\} \times s_y \times G(x'\beta + \gamma z^*)^y [1 - G(x'\beta + \gamma z^*)]^{1-y} \\ &\quad + \mathbf{1}\{z = \varnothing\} \times (1 - s_y) \times G(x'\beta + \gamma z^*)^y [1 - G(x'\beta + \gamma z^*)]^{1-y}.\end{aligned}$$

- *Admissible heterogeneity is* $\mathcal{P}^0_A = co(\{\delta_{z^*,s_0,s_1} \mid (z^*,s_0,s_1) \in \mathcal{Z}^* \times [0,1] \times [0,1]\})$, *where* δ_a *denotes the Dirac measure (i.e. the point mass distribution) at* a.

For verification of measurability condition (iv) from Assumption 3.1, see Lemma A.1 in Appendix A. I assume that $\mathcal{X}$ and $\mathcal{Z}^*$ are closed because they are supports, but this only plays a technical role. Also, $\mathcal{A}$ is naturally convex in this setting, so there is minimal difference between the two interpretations of heterogeneity from Section 3.2.

As stated in the introduction, s_0 and s_1 can be understood as the realizations of random variables S_0 and S_1, where S_d represents the conditional probability of observing Z^* when $Y = d$. I restrict G to non-degenerate probabilities, as is standard, but this is not crucial; the core ideas extend to the degenerate cases. Nevertheless, these cases require effort to state properly, so it is convenient to exclude them. Finally, $\mathcal{Z}^*$ is a set which represents the assumed support of Z^*. Ideally, the researcher uses the true support. However, resulting moment inequalities are still valid if a superset is used instead.

In Assumption 4.1, I set $\mathcal{P}^0_A$ to only contain finite heterogeneity mixtures. As μ^0 always outputs discrete distributions itself, it follows that the model distributions under Assumption 4.1 are all discrete as well. On one hand, this enables me to treat the model like the one in Section 2. On the other hand, this choice rules out continuous outcome distributions and

thus continuous Z^*'s, per the integrability discussion in Section 3.3. However, this is not a problem. Proposition 3.2 will let me extend my results to continuous mixtures.

4.2 Dual Cone

Condition on $X = x$ for some $x \in \mathcal{X}$, and fix some proposed $\theta \in \Theta$. To study identification, I now turn my attention to the dual cone, $\mathcal{D}(x, \theta)$, under Assumption 4.1. Recall that this is the set of functions which have non-negative expectations with respect to all possible model distributions. The idea is then to use these functions as moment inequalities. This formalizes the outline which I originally gave in Section 2.3 for the stylized setup.

As all model distributions have finite support, I can ignore integrability issues. Therefore, by the assumption of finite heterogeneity mixtures, the dual cone in this setting is

$$\mathcal{D}(x,\theta) = \Bigg\{ m : \{0,1\} \times \mathcal{Z} \to \mathbb{R} \;\Bigg|\; m \text{ meas. and } \int m(y,z,x,\theta) d\mu^0(y,z|x,z^*,s_0,s_1;\theta) \geq 0, \\ \forall (z^*, s_0, s_1) \in \mathcal{Z}^* \times [0,1] \times [0,1] \Bigg\},$$

as it suffices to impose the non-negativity condition only for Dirac heterogeneity measures. Meanwhile, s_0 and s_1 enter linearly into probabilities; therefore, I can further reduce the constraints to only the ones involving the extreme points, $(s_0, s_1) \in \{0,1\}^2$, which correspond to degenerate missing-ness processes. With this, the non-negativity constraints on m transform into the following system of linear inequalities:

$$m(0,\varnothing,x,\theta)[1 - G(x'\beta + \gamma z^*)] + m(1,\varnothing,x,\theta)G(x'\beta + \gamma z^*) \geq 0,$$

$$m(0,\varnothing,x,\theta)[1 - G(x'\beta + \gamma z^*)] + m(1,z^*,x,\theta)G(x'\beta + \gamma z^*) \geq 0,$$

$$m(0,z^*,x,\theta)[1 - G(x'\beta + \gamma z^*)] + m(1,\varnothing,x,\theta)G(x'\beta + \gamma z^*) \geq 0,$$

$$m(0,z^*,x,\theta)[1 - G(x'\beta + \gamma z^*)] + m(1,z^*,x,\theta)G(x'\beta + \gamma z^*) \geq 0,$$

for all $z^* \in \mathcal{Z}^*$. Note that only the first and last lines would appear if missing-ness were exogenous. However, the lack of restrictions produces the additional middle constraints. These represent the entanglement between the different parts of the data.

Subsequent analysis of the cone then proceeds along the outline which I gave in Section 2.3. As such, I move straight into discussing the results.

4.3 Results

I now define the relevant functions for a strongly core-determining class in this model, along with some intuition for their structure. Then, I state the formal identification theorem. I conclude with a short discussion.

I express the key moment functions through two collections which are mirror images of one another. Each collection is indexed by two objects: a binary index, $q_\varnothing \in \{0,1\}$, and a measurable index function, $q : \mathcal{Z}^* \to \{0,1\}$. The index $q_\varnothing$ controls the behavior of the moment function when $Z = \varnothing$, while $q(z^*)$ adjusts the behavior when $Z = z^*$. As $q_\varnothing$ and q range over their possible values, the moment function shifts to target different subsets of the data. This will become clearer after the definitions.

I start with the first collection. For a fixed $q_\varnothing \in \{0,1\}$ and measurable $q \in \{0,1\}^{\mathcal{Z}^*}$, I define $m_{q_\varnothing,q} : \{0,1\} \times \mathcal{Z} \to \mathbb{R}$ in a sequential fashion, with dependence on x and θ suppressed:

1. $m_{q_\varnothing,q}(0,\varnothing) = 1$.

2. If $q_\varnothing = 0$, $m_{q_\varnothing,q}(1,\varnothing) = 0$. Otherwise,

$$m_{q_\varnothing,q}(1,\varnothing) = \sup_{z^* \in \mathcal{Z}^*} -\frac{1 - G(x'\beta + \gamma z^*)}{G(x'\beta + \gamma z^*)} = -\frac{1 - \overline{G}(x,\theta)}{\overline{G}(x,\theta)},$$

where $\overline{G}(x,\theta) := \sup_{z^* \in \mathcal{Z}^*} G(x'\beta + \gamma z^*)$ is the maximum model-predicted probability.

3. For any $z^* \in \mathcal{Z}^*$: If $q(z^*) = 1$, $m_{q_\varnothing,q} = 1$. Otherwise,

$$m_{q_\varnothing,q}(0, z^*) = -m_{q_\varnothing,q}(1, \varnothing) \times \frac{G(x'\beta + \gamma z^*)}{1 - G(x'\beta + \gamma z^*)},$$

where $m_{q_\varnothing,q}(1, \varnothing)$ is the value selected in the previous step.

4. Finally, for any $z^* \in \mathcal{Z}^*$: $m_{q_\varnothing,q}(1, z^*) = -m_{q_\varnothing,q}(0, z^*) \times \frac{1-G(x'\beta+\gamma z^*)}{G(x'\beta+\gamma z^*)}$.

As an example of a resulting moment inequality, consider $q_\varnothing = 0$ and $q(z^*) = 1$ for all $z^* \in \mathcal{Z}^*$. Then, if $\mathcal{Z}^*$ is finite, $\mathbb{E}[m_{q_\varnothing,q}(Y, Z, X, \theta)|X = x] \geq 0$ translates into

$$P(Y = 0|X = x) \geq \sum_{z^* \in \mathcal{Z}^*} \left[\frac{1 - G(x'\beta + \gamma z^*)}{G(x'\beta + \gamma z^*)} \times P(Y = 1, Z = z^*|X = x) \right].$$

In Section 2, I derived an inequality of this type for the stylized model. It checks whether the mass of 0-outcomes in the incomplete data can rationalize deficits in the complete data.

In general, I provide three pieces of intuition for the form of the $m_{q_\varnothing,q}$ functions. First, observe that $m_{q_\varnothing,q}$ assigns non-negative values to outcome 0 and non-positive values to outcome 1. Since the underlying requirement is non-negativity of the expectation, these functions check whether there is sufficient mass of 0-outcomes in the partially missing data, as mentioned in my previous comment. Second, note the appearance of the term $\overline{G}(x, \theta)$. This term arises because of the global role of the missing data; it is entangled with all z^*'s. Finally, $m_{q_\varnothing,q}(0, z^*)$ and $m_{q_\varnothing,q}(1, z^*)$ always have the usual ratio associated with the standard moment condition under link function G (see step 4). In the special case where data is never missing, for instance, identification collapses to the usual conditional moments at each z^*.

As an illustration of the role of the indices, fix $q_\varnothing = 1$. Then, for any $z^* \in \mathcal{Z}^*$,

$$m_{q_\varnothing,q}(0, z^*) = 1 \quad \text{or} \quad m_{q_\varnothing,q}(0, z^*) = \frac{G(1 - \overline{G})}{\overline{G}(1 - G)} \leq 1,$$

depending on $q(z^*)$. These two values are the endpoints of an interval of possible weights on the $Z = z^*$ data. The lower bound is imposed by the non-negativity constraints, while

the upper bound arises from redundancy considerations. Since the weight enters in a linear fashion, it suffices to check only the interval endpoints, which is exactly done by setting $q(z^*)$ to either zero or one. This can be interpreted as needing to rationalize model deviations in either "direction," with the critical direction depending on whether $P(Y = 1, Z = z^*|X = x)/P(Y = 0, Z = z^*|X = x)$ exceeds, or is less than, than the ratio predicted by the model.

Finally, given the model's structure, thre is also a symmetric class of functions which places positive weights on the 1-outcomes instead, with outcome 1 swapped with 0 and G swapped with $1 - G$. The construction of a new class of $\widetilde{m}_{q_\varnothing,q}$ functions proceeds in this way based on $m_{q_\varnothing,q}$. In the scenario where the link function satisfies $G(c) = 1 - G(-c)$ for all $c \in \mathbb{R}$, there is an especially simple formula: $\widetilde{m}_{q_\varnothing,q}(y, z, x, \theta) = m_{q_\varnothing,q}(1 - y, z, x, -\theta)$. I re-introduce x and θ because the sign of the entire index should be flipped.

I can now state the main result. Let $\{0,1\}^{\mathcal{Z}^*}_m$ denote the subset of $\{0,1\}^{\mathcal{Z}^*}$ which is measurable, and let $\mathcal{Q}$ equal $\{0,1\} \times \{0,1\}^{\mathcal{Z}^*}_m$.

Theorem 4.1. *Consider the setting defined by Assumption 4.1. Fix $(x, \theta) \in \mathcal{X} \times \Theta$. Define the class of functions*

$$\mathcal{D}^*(x,\theta) := \left\{ m_{q_\varnothing,q}(\cdot, x, \theta) \;\middle|\; (q_\varnothing, q) \in \mathcal{Q} \right\} \bigcup \left\{ \widetilde{m}_{q_\varnothing,q}(\cdot, x, \theta) \;\middle|\; (q_\varnothing, q) \in \mathcal{Q} \right\}.$$

$\mathcal{D}^*(x, \theta)$ is strongly core-determining at (x, θ).

Per the contents of Section 3, $\mathcal{D}^*(x, \theta)$ contains all moment information in the model, and I have completely characterized identification when Z^* can only have a discrete distribution.[17]

To move beyond discreteness, I inspect $\mathcal{D}^*(x, \theta)$ and apply Proposition 3.2. To start, suppose that G is monotone and $\mathcal{Z}^*$ is bounded; then, model probabilities are bounded away from zero and one. Therefore, all functions in $\mathcal{D}^*(x, \theta)$ are bounded for any specific (x, θ), and they are integrable with respect to all probability measures on $\{0, 1\} \times \mathcal{Z}$. Proposition 3.2 then immediately yields the corollary below, which I state without proof.

[17] A few elements of $\mathcal{D}^*(x, \theta)$ are also redundant, ex. when $q_\varnothing = 0$ and $q = 1$, but I avoid adding this in order to keep statements simpler.

Corollary 4.1. *Consider the setting defined by Assumption 4.1. Fix* $(x, \theta) \in \mathcal{X} \times \Theta$*, and suppose that* G *is monotone and* $\mathcal{Z}^*$ *is bounded. Then, for any* $\widetilde{\mathcal{P}}_A^0$ *such that* $\mathcal{P}_A^0 \subseteq \widetilde{\mathcal{P}}_A^0 \subseteq \mathcal{M}_1(\mathcal{Z}^* \times [0,1] \times [0,1])$*,* $\theta \in \Theta_I^+(F_{(Y,Z)|X=x}|x; \mu^0, \widetilde{\mathcal{P}}_A^0)$ *iff*

$$F_{(Y,Z)|X=x} \in \mathcal{M}_{x,\theta} \text{ and } \int m^* dF_{(Y,Z)|X=x} \geq 0 \quad \forall m^* \in \mathcal{D}^*(x,\theta),$$

where $\mathcal{M}_{x,\theta}$ *is the integrability-related subspace of* $\mathcal{M}(\{0,1\} \times \mathcal{Z})$ *defined in Section 3.2.*

This corollary extends my results to continuous and mixed Z^*'s.

In the scenario where $\mathcal{Z}^*$ is unbounded, the moment functions can become unbounded below.[18] In that case, one option is to restrict attention to heterogeneity distributions under which the functions remain integrable, i.e. where the tails of $\mathcal{Z}^*$ are thin enough. Alternatively, it is possible to replace individual functions in $\mathcal{D}^*(x,\theta)$ with certain sequences of truncations, ex. $\max\{m_{q_\varnothing,q}, -k\}$, $k \in \mathbb{N}$. With careful selection of these sequences, identification power can be preserved while making each element of the sequence integrable. As this is unlikely to be used in applications, however, I do not explore this.

On the practical implications of Theorem 4.1, an important point is that the number of moment inequalities depends on the cardinality of $\mathcal{Z}^*$. If this set is finite, then there are finitely many conditional moment inequalities. However, the number grows exponentially in $|\mathcal{Z}^*|$, and using the entire collection becomes intractable if Z^* is continuous. In that case, it may be useful to select some subcollection of q's to use. This can be potentially guided by the setting, ex. if there are subpopulations for which it is important to check the model.[19]

Finally, the result above is for the case where only a single covariate can be missing. It is clear that the claim extends if Z^* is a random vector which is either completely missing or completely observed. If the covariates can be missing in various configurations, though, then sharpness becomes more complex. Accordingly, I defer discussion of this to the next

[18] Some of the moment inequalities can also collapse together. If G is a CDF and $\sup_{z^* \in \mathcal{Z}^*} \gamma z^* = \infty$, for example, then $\overline{G}(x,\theta) = 1$, and $m_{q_\varnothing,q}(1,\varnothing) = 0$ regardless of $q_\varnothing$.

[19] Another option is to use methods developed for infinitely many moment inequalities, ex. Andrews and Shi (2017).

section, where relevant adjustments are more immediately apparent.

5 Mean Regression with a Set-Valued Covariate

I now turn to the most general setting: unrestricted (linear) mean regression,[20]

$$Y = X'\beta + \gamma Z^* + \varepsilon, \quad \mathbb{E}[\varepsilon | X, Z^*] = 0,$$

where Z^* can be known to lie in a particular set, rather than only being completely observed or missing. I allow the process determining this set to be completely unrestricted.

Results here are weaker than the ones in Section 4. I discuss the relationship at the end of the section. Furthermore, the presence of continuous outcomes leads to a larger core-determining class, which can be difficult to use in practice. Nevertheless, the functions still have an intuitive structure. In applications, I would suggest to use a tractable subset of them, although this loses some identification power.

5.1 Model Setup

The setup here has the same qualitative structure as in Section 4. However, objects are now more complicated. In particular, I again denote the support of Z^* as $\mathcal{Z}^*$. On top of this, however, I then let $\mathcal{K} \equiv \mathcal{K}(\mathcal{Z}^*)$ denote some countable collection of its subsets which satisfies

$$\mathcal{Z}^* \in \mathcal{K}, \quad |K| \geq 2 \quad \forall K \in \mathcal{K}.$$

These represent the possible observed sets for Z^*, and like with $\varnothing$ in Section 4, I add them to $\mathbb{R}$ as isolated "points": $\mathbb{R}_{\mathcal{K}} = \mathbb{R} \cup \mathcal{K}$.[21] I endow this augmented set with the standard

[20] I use the linearity at times for technical purposes, such as measurability. However, it otherwise does not play a substantive role. As such, it is possible to replace $X'\beta + \gamma Z^*$ with a generic $h(X, Z^*, \theta)$ like in Section 4, especially if h is well-behaved.

[21] The countability of $\mathcal{K}$ ensures that $\mathbb{R}_{\mathcal{K}}$ is a Polish space (as well as $\mathcal{Z}$, later), which I use for technical reasons. However, this assumption may not be necessary.

Euclidean metric, again extended so that the distance between any $K \in \mathcal{K}$ and anything else is one. Finally, let $\mathcal{F}_0 \subseteq [0,1]^{\mathbb{R}}$ be the set of CDFs on $\mathbb{R}$ which correspond to mean-zero distributions, equipped with the supremum norm.

The main model assumption is then as follows. I use $\mathcal{Z}_m^{\mathbb{R}}$ for the set of measurable functions from $\mathbb{R}$ to $\mathcal{Z}$, coupled with the supremum norm. Also, to streamline statements, I sometimes treat $\mathbb{R}_{\mathcal{K}}$ as a set of sets; for $z^* \in \mathcal{Z}^*$, I will say $z^* \in z$ to encompass both the case $z^* = z$ (if $z \in \mathcal{Z}^*$) and $z^* \in z$ (if $z \in \mathcal{K}$).

Assumption 5.1. *Fix $d_X \in \mathbb{N}$, let $\mathcal{X} \subseteq \mathbb{R}^{d_X}$ and $\mathcal{Z}^* \subseteq \mathbb{R}$ be non-empty closed sets, and let $\mathcal{K}$ be a countable collection of subsets of $\mathcal{Z}^*$ which satisfies $\mathcal{Z}^* \in \mathcal{K}$ and $|K| \geq 2$ for all $K \in \mathcal{K}$. I define $\mathcal{Z} := \mathcal{Z}^* \cup \mathcal{K} \subseteq \mathbb{R}_{\mathcal{K}}$. Then, in the notation of Assumption 3.1:*

- *$\mathcal{Y}$ corresponds to $\mathbb{R} \times \mathcal{Z}$, $\mathcal{W}$ corresponds to $\mathcal{X}$,*

$$\mathcal{A} := \left\{ (z^*, F, s) \in \mathcal{Z}^* \times \mathcal{F}_0 \times \mathcal{Z}_m^{\mathbb{R}} \;\middle|\; z^* \in s(c), \quad \forall c \in \mathbb{R} \right\},$$

  and $\Theta \subseteq \mathbb{R}^{d_X} \times \mathbb{R}$.

- *The model map $\mu^0 : \mathcal{X} \times \mathcal{A} \times \Theta \to \mathcal{M}_1(\mathbb{R} \times \mathcal{Z})$ is defined as*

$$\mu^0_{x,z^*,F,s,\theta}((-\infty, c] \times \{z\}) = \mathbf{1}\{z^* \in z\} \int_{-\infty}^{c - x'\beta - \gamma z^*} \mathbf{1}\{s(e) = z\} dF(e)$$

  for all $c \in \mathbb{R}$ and $z \in \mathcal{Z}$.

- *Admissible heterogeneity is $\mathcal{P}_A^0 = co(\{\delta_{z^*,F,s} \mid (z^*, F, s) \in \mathcal{A}\})$.*

In words, Assumption 5.1 says that a heterogeneity realization consists of z^*, the realization of Z^*; F, the conditional distribution of ε; and s, the observability process which defines the observed set of values given (y, x, z^*). Missing data corresponds to outcome $(y, \mathcal{Z}^*)$ and complete observability to (y, z^*). I then define the model map in the obvious way.

It is more natural to define s, in Assumption 5.1, as mapping into distributions over possible observed sets. However, my choice anticipates the fact that it suffices to focus on degenerate processes; heterogeneity convexification fills in the intermediate probabilities.

5.2 Dual Cone

I now outline how I analyze the dual cone $\mathcal{D}(x, \theta)$. I describe the central insight carefully, but I reserve most other details for the proof in the appendix.

Fix some x and θ. I ignore integrability for now, as well as an additional assumption that I later impose. Then, for the dual cone, the non-negativity constraints on a measurable $m : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ are

$$\int m(x'\beta + \gamma z^* + e, s(e))dF(e) \geq 0,$$

for all $(z^*, F, s) \in \mathcal{A}$. By Winkler (1988), it suffices to consider the extreme points of $\mathcal{F}_0$, which correspond to the extreme points of the set of mean-zero probability measures on $\mathbb{R}$: the one- and two-point distributions

$$\{\delta_0\} \cup \left\{ \frac{c_+}{-c_- + c_+}\delta_{c_-} + \frac{-c_-}{-c_- + c_+}\delta_{c_+} \;\middle|\; c_- < 0, c_+ > 0 \right\}.$$

Through this, I can reduce s to $z \in \mathcal{Z}$ or $(z_-, z_+) \in \mathcal{Z} \times \mathcal{Z}$, as long as they contain z^*; these are the evaluations of s at the support points of the error distribution. Consequently, the primary dual cone conditions become as follows, at least for the two-point error distributions:

$$\frac{c_+}{-c_- + c_+}m(x'\beta + \gamma z^* + c_-, z_-) + \frac{-c_-}{-c_- + c_+}m(x'\beta + \gamma z^* + c_+, z_+) \geq 0,$$

for $z^* \in \mathcal{Z}^*$, $c_- < 0$, $c_+ > 0$, and $z_-, z_+ \in \mathcal{Z}$ with $z^* \in z_-$, $z^* \in z_+$. Re-writing yields

$$\frac{m(x'\beta + \gamma z^* + c_-, z_-)}{c_-} \leq \frac{m(x'\beta + \gamma z^* + c_+, z_+)}{c_+}. \tag{6}$$

Meanwhile, the one-point error distributions impose non-negativity directly at certain points.

The key observation now is that, since the above constraints must hold for all heterogeneity, they must also hold for $z_- = z_+ = z^*$, for generic $z^* \in \mathcal{Z}^*$. Then, for constraints in (6) which involve $m(\cdot, z^*)$, the values of $m(y, z^*)$ effectively enter only through

$$r_{m,-}(z^*) := \sup_{c_-<0} \frac{m(x'\beta + \gamma z^* + c_-, z^*)}{c_-} \quad \text{and} \quad r_{m,+}(z^*) := \inf_{c_+>0} \frac{m(x'\beta + \gamma z^* + c_+, z^*)}{c_+}.$$

These are minimal and maximal slopes for y-lines which lie below $m(\cdot, z^*)$ everywhere to the right or left of $x'\beta + \gamma z^*$, while passing through zero at $y = x'\beta + \gamma z^*$. Therefore, when looking for a core-determining class, it is redundant to consider any m where $m(y, z^*)$ is strictly above its associated lower bound line for some $y \in \mathbb{R}$, once $r_{m,-}(z^*)$ and $r_{m,+}(z^*)$ are fixed. I can thus restrict attention to m's where $m(\cdot, z^*)$ is piece-wise linear,

$$m(y, z^*) = \max\{r_{m,-}(z^*)[y - x'\beta - \gamma z^*], r_{m,+}(z^*)[y - x'\beta - \gamma z^*]\}.$$

In other words, I can condense the behavior of core m's, for the completely observed data, into a collection of slopes: $r_{m,-} : \mathcal{Z}^* \to \mathbb{R}$ and $r_{m,+} : \mathcal{Z}^* \to \mathbb{R}$, with $r_{m,-} \leq r_{m,+}$.

The reduction of $m(\cdot, z)$ to slope terms is the main tool of the proof, and such terms are the main objects which I use in the arguments. To outline the remaining steps: I next establish that it is redundant to consider m's with $r_{m,-}(z^*) \neq r_{m,+}(z^*)$ for any $z^* \in \mathcal{Z}^*$. As such, I can instead refer to $r : \mathcal{Z}^* \to \mathbb{R}$, and the previous display transforms into $m(y, z^*) = r(z^*)[y - x'\beta - \gamma z^*]$. Under an additional assumption, the slopes must also be uniformly bounded; so, I can re-scale r to take values in $[-1, 1]$. Finally, I exhibit a lower bound on $m(y, z)$ for $z \in \mathcal{K}$, where setting m equal to the bound does not violate the constraint. This bound turns out to be simple: $m(y, z) = \sup_{z^* \in z} m(y, z^*)$. It follows that a core m uses a weighted residual for all $z^* \in \mathcal{Z}^*$, if the true covariate value is observed. Otherwise, it uses the supremum of the possible weighted residuals over the observed set z.

The upcoming result should now be apparent. To provide some further intuition before-

hand, I consider two special cases. First, suppose that there is no incompleteness in the data: $P(Z \in \mathcal{Z}^* | X = x) = 1$. Then, checking the core moment inequalities reduces to verifying

$$\mathbb{E}[r(z^*)(Y - X'\beta - \gamma Z^*) | X = x, Z = z^*] \geq 0$$

for all $z^* \in \mathcal{Z}^*$ and $r(z^*) \in [-1, 1]$. This is equivalent to $\mathbb{E}[Y - X'\beta - \gamma z^* | X = x, Z = z^*] = 0$.

Now, consider the opposite extreme where there are no X's and Z^* is always missing:

$$Y = \gamma Z^* + \varepsilon, \quad \mathbb{E}[\varepsilon | Z^*] = 0, \quad P(Z = \mathcal{Z}^* | X = x) = 1.$$

I assume that $\mathcal{Z}^*$ is finite to make a point. Then, although Z^* is never observed, I can still specify $r : \mathcal{Z}^* \to [-1, 1]$ and define a corresponding m as described above.

Some thought shows that it is extraneous to consider an induced m which is not the point-wise supremum of at most two lines. To specify this more explicitly, suppose $\gamma > 0$, and label the elements of $\mathcal{Z}^*$ as $z_1^*, ..., z_L^*$, where $L = |\mathcal{Z}^*|$. Then, the case with one line is given by making m negative on either $(-\infty, \gamma z_1^*)$ or $(\gamma z_L^*, \infty)$, alongside $r = 1$ and $r = -1$ respectively. The resulting moment functions are $m_{-\infty}(y) = y - \gamma z_1^*$ and $m_{+\infty}(y) = \gamma z_L^* - y$. Along with this, though, there are also intermediate cases where m is negative on an interval, $(\gamma z_l^*, \gamma z_{l+1}^*)$ for some $l \in \{1, ..., L-1\}$. Then, with re-scaling, one can choose $r(z_l^*) = -1$ and $r(z_{l+1}^*) \in (0, \infty)$, with other values of r aligned to never enter the supremum, so that

$$m(y) = \begin{cases} \gamma z_l^* - y & \text{if } y \leq \frac{\gamma z_l^* + r(z_{l+1}^*)\gamma z_{l+1}^*}{1 + r(z_{l+1}^*)}, \\ r(z_{l+1}^*)[y - \gamma z_{l+1}^*] & \text{otherwise.} \end{cases}$$

Relative to the general case, additional slopes are irrelevant since they can only increase $m(y)$; there are no fully-observed parts of the data for them to weigh up or down.

If $\mathcal{Z}^*$ is replaced with a compact interval here, then $m_{-\infty}$ and $m_{+\infty}$ remain. If $\mathcal{Z}^* = \mathbb{R}$, then there are no informative restrictions. However, if there are gaps in the support, then

restrictions can arise, even when $\mathcal{Z}^*$ is unbounded. To the best of my knowledge, this is a new insight; most other papers assume that observed sets are intervals.

5.3 Results

I now present the formal result. Let $\mathcal{R}$ be the set of functions from $\mathcal{Z}^*$ to $[-1, 1]$ with a supremum norm of one. For any $r \in \mathcal{R}$, I define $m_r : \mathbb{R} \times \mathcal{Z} \times \mathcal{X} \times \Theta \to \mathbb{R}$ as follows:

$$m_r(y, z, x, \theta) = \begin{cases} \sup_{z^* \in z} r(z^*)[y - x'\beta - \gamma z^*] & \text{if } z \in \mathcal{K}, \\ r(z)[y - x'\beta - \gamma z] & \text{otherwise.} \end{cases}$$

On top of Assumption 5.1, I assume that $\mathcal{Z}^*$ is bounded. This ensures that the supremums in m_r are finite everywhere for all $r \in \mathcal{R}$, and it is also sufficient for the integral of $m_r(\cdot, \mathcal{Z}^*, x, \theta)$ to be finite for all mean-zero error distributions. If $\mathcal{Z}^*$ is unbounded, then it can still be possible to find restrictions through a subcollection of r's; recall the previous special case.

The theorem below is then the counterpart of Theorem 4.1. For technical reasons, as r may be non-measurable, I introduce minimal measurable majorants: For $r \in \mathcal{R}$ and a probability measure μ on $\mathbb{R} \times \mathcal{Z}$, let $\mathcal{C}(m_r, \mu)$ be the set of measurable functions which point-wise dominate m_r, while also being point-wise dominated μ-almost-surely by all other such functions. Intuitively, $\mathcal{C}(m_r, \mu)$ contains tight measurable approximations from above.[22]

Theorem 5.1. *Consider the setting defined by Assumption 5.1, and suppose that $\mathcal{Z}^*$ is bounded. Fix $(x, \theta) \in \mathcal{X} \times \Theta$. Define the class of functions*

$$\begin{aligned} \mathcal{D}^*(x, \theta) := &\left\{ m_r(\cdot, x, \theta) \;\middle|\; r \in \mathcal{R},\ r \text{ measurable} \right\} \bigcup \\ &\left\{ m^* : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}, \quad m^*(y, z) = \min\{m(y, z), \sup_{z^* \in \mathcal{Z}^*} |y - x'\beta - \gamma z^*|\} \right. \\ &\left. \;\middle|\; \exists \mu \in \mathcal{M}_1(\mathbb{R} \times \mathcal{Z}), r \in \mathcal{R} \text{ s.t. } m \in \mathcal{C}(m_r(\cdot, x, \theta), \mu) \right\}. \end{aligned}$$

[22] For example, see Section 1.2 in van der Vaart and Wellner (2023).

Then, $\mathcal{D}^*(x,\theta)$ is strongly core-determining at (x,θ).

The statement is complex, but the main message is that $\{m_r\}_{r\in\mathcal{R}}$ comprises the key functions.

Regarding the potential for extension via Proposition 3.2, observe that the functions in $\mathcal{D}^*(x,\theta)$ are generally model-integrable only if the induced distribution for Y has a finite moment. Therefore, my results do not accommodate mixtures which lead to heavy tails in the outcome variable. I do not view this as a negative, though, because a researcher will likely only use the mean regression model if they are willing to assume the existence of a moment. Meanwhile, Proposition 3.2 handles the extension to continuous Z^*'s.

5.4 Discussion

The m_r functions have an underlying linear structure, which is practically useful. However, at the same time, there are a huge number of them. One option is to use a subset of r's which lead to moment functions with a nicer form. For example, if one sets $r(z^*) = 1 - 2 \times \mathbf{1}\{z^* \geq z_0\}$ for some $z_0 \in \mathcal{Z}^*$, $m_r(\cdot, z)$ is either linear in y or piecewise linear with at most two parts.

At a theoretical level, it is possible to generalize the setup and Theorem 5.1 in various ways. One candidate is to allow for incompletely observed outcomes as well. I can handle this within my framework by including the true outcome realization, y^*, in heterogeneity, alongside z^*. Then, in the arguments of Section 5.2, I observe that outcome uncertainty acts similarly to uncertainty in a covariate. This suggests a natural extension of Theorem 5.1. Similar reasoning applies to the case where multiple covariates can be incompletely observed.

In the opposite direction, which relates to the next section, one can impose additional restrictions on the observability process. A nonparametric example is to assume that the observability of Z^* is "monotone" in the error: larger error realizations correspond to more information on Z^*. To encode this formally, I can add the following constraint to the allowed observability processes in Assumption 5.1: for $c, \widetilde{c} \in \mathbb{R}$, $c \leq \widetilde{c}$ implies $s(c) \supseteq s(\widetilde{c})$, where s is

the process. This ends up modifying the dual cone non-negativity constraint (6),

$$\frac{m(x'\beta + \gamma z^* + c_-, z_-)}{c_-} \leq \frac{m(x'\beta + \gamma z^* + c_+, z_+)}{c_+},$$

so that it need only hold for $z_-, z_+ \in \mathcal{Z}$ with $z_- \supseteq z_+$. I do not provide concrete results for this extension, but I do note that the monotonicity assumption seems to have identifying power. This is due to the following example: Fix x and θ, and let $\mathcal{Z}^*$ be a two-point support such that $x'\beta + \gamma z^* \in \{0, 1\}$. Consider $r : \mathcal{Z}^* \to [-1, 1]$ such that $r(z^*) = 0.5 + 0.5 \times \mathbf{1}\{x'\beta + \gamma z^* = 1\}$. Then, in the base model, m_r is not point-wise dominated in the dual cone, with $m_r(y, \mathcal{Z}^*) = \max\{0.5y, y - 1\}$. When the monotonicity assumption is added, however, the modified function $\widetilde{m}_r$, with $\widetilde{m}_r(y, \mathcal{Z}^*) = 0.5y$, now becomes valid,[23] with $\widetilde{m}_r(y, z) \leq m_r(y, z)$ everywhere and a strict inequality when $y > 2$ and $z = \mathcal{Z}^*$. The dual cone under monotonicity is thus strictly larger, i.e. there are more moment inequalities.

As a last point, I elaborate on the relationship between Theorem 5.1 and Theorem 4.1. Observe that the arguments in Section 5.2 do not rely on the fact that ε has unrestricted support on $\mathbb{R}$. Accordingly, Theorem 5.1 extends to any mean regression model, including the binary choice setting. Indeed, inspecting the moment functions of Theorem 4.1 reveals that they have the form proposed in this section. For example, consider $q_\varnothing = 1$ and $q(z^*) = 1$ for all $z^* \in \mathcal{Z}^*$. Then, the function $m_{q_\varnothing, q}$ was previously defined as

$$m_{q_\varnothing,q}(y, z, x, \theta) = \begin{cases} 1 & \text{if } z = \varnothing, y = 1, \\ -\frac{1-\overline{G}(x,\theta)}{\overline{G}(x,\theta)} & \text{if } z = \varnothing, y = 0, \\ \frac{G(x'\beta+\gamma z)}{1-G(x'\beta+\gamma z)} \times \frac{1-\overline{G}(x,\theta)}{\overline{G}(x,\theta)} & \text{if } z \neq \varnothing, y = 1, \\ -\frac{1-\overline{G}(x,\theta)}{\overline{G}(x,\theta)} & \text{otherwise.} \end{cases}$$

For $y \in \{0, 1\}$, $m_{q_\varnothing,q}(y, \varnothing, x, \theta)$ is the supremum of $m_{q_\varnothing,q}(y, z^*, x, \theta)$ over $z^* \in \mathcal{Z}^*$. Also, for

[23] Under monotonicity, the values of $m(y, \mathcal{Z}^*)$, when $y > 2$, only face non-negativity constraints from $m(\widetilde{y}, \mathcal{Z}^*)$, $\widetilde{y} < 2$, rather than $m(\widetilde{y}, z^*)$ also, for $z^* \in \mathcal{Z}^*$. The steeper bounding line $y - 1$, from $x'\beta + \gamma z^* = 1$, thus no longer places a restriction on $m(y, \mathcal{Z}^*)$ when $y > 2$.

fixed $z^* \in \mathcal{Z}^*$, $m_{q\varnothing,q}(\cdot, z^*, x, \theta)$ is the standard moment condition for the model.

With that said, I present the binary choice model with missing-ness to start, and separately, for several reasons. First, it is an important special case. Second, I use more direct arguments in my proof of Theorem 4.1, which may provide a useful template elsewhere. Finally, there is additional content in Theorem 4.1. Binary choice has special features, such as being fully parametric with all means between the two outcomes, which enable the reduction to the smaller class in Section 4. The functions also have clean interpretations.

6 Other Models

Having presented my results for the target mean regression models, I now discuss the potential use of my framework and approach in other settings.

For many of the model features which appear here, there already exist extensive literatures which study them. I do not claim that my approach is superior to existing ones. Instead, my goal is to illustrate the flexibility of my approach. One can then use my framework to incorporate multiple, and varied, forms of heterogeneity at the same time. It is true that subsequent derivation of a core-determining class in a complex model may be difficult. Nevertheless, the unified structure can still be useful for studying identification.

6.1 Quantile Regression with Incompletely Observed Data

Alongside mean regression, researchers may also be interested in quantile regression with incomplete data:

$$Y = X'\beta + \gamma Z^* + \varepsilon, \quad \mathbb{E}[\mathbf{1}\{\varepsilon \leq 0\}|X, Z^*] = \alpha, \tag{7}$$

where $\alpha \in (0, 1)$ is some fixed number. For example, Beresteanu and Sasaki (2021) study quantile regression with set-valued outcomes using random set theory.

Within my proposed framework, analysis of (7) turns out to have a similar flavor to the general mean regression case studied in Section 5. The difference is that, rather than

working with bounding slopes, one instead focuses here on levels.

To illustrate, fix some x and θ, and suppose that missing-ness is all-or-nothing, as in Section 4. By looking at the two-point error distributions which satisfy the quantile constraint, I obtain the following non-negativity restrictions on a function m in the dual cone:

$$\alpha m(x'\beta + \gamma z^* + c_-, \varnothing) + (1-\alpha) m(x'\beta + \gamma z^* + c_+, \varnothing) \geq 0,$$

$$\vdots$$

$$\alpha m(x'\beta + \gamma z^* + c_-, z^*) + (1-\alpha) m(x'\beta + \gamma z^* + c_+, z^*) \geq 0,$$

for $z^* \in \mathcal{Z}^*$, $c_- \leq 0$, and $c_+ > 0$. Therefore, for particular z^*, the relevant values of $m(y, z^*)$ are its infimums over $y \leq x'\beta + \gamma z^*$ and over $y > x'\beta + \gamma z^*$. Based on this, arguments can then proceed as in previous sections, with appropriate adjustments made in statements.

6.2 Equilibrium Selection in Games, with Additional Features

In this subsection, I consider the Tamer (2003) entry game setting with two firms, which is a canonical example of an incomplete econometric model. I also discuss how one can then add various features on top of the base setup.

The setup is a standard one. There are two firms, $j = \{1, 2\}$, which decide whether to enter in various markets. Let $(Y, X) \in \{0, 1\}^2 \times \mathbb{R}^{2d_X}$ be the random vector of market-level data, where $Y = (Y_1, Y_2)$ consists of the entry decisions, i.e. $Y_j = 1$ iff firm j enters, and $X = (X_1, X_2)$ are market-level firm characteristics. Then, the assumed model is that, within each market, observed entry decisions correspond to a pure strategy Nash equilibrium with simultaneous moves and complete information:

$$Y_1 = \mathbf{1}\{X_1'\beta + \Delta Y_2 + \varepsilon_1 \geq 0\} \quad \text{and} \quad Y_2 = \mathbf{1}\{X_2'\beta + \Delta Y_1 + \varepsilon_2 \geq 0\}.$$

In the display, $\theta = (\beta, \Delta) \in \mathbb{R}^{d_X} \times \mathbb{R}_{--}$ are structural payoff parameters, and $\varepsilon = (\varepsilon_1, \varepsilon_2) \in \mathbb{R}^2$

are firm-market-specific shocks known by the firms but not the researcher. As Δ is negative, it has an interpretation as a competitive effect. Moreover, this effect induces equilibrium multiplicity: for certain ε realizations, both $Y = (1,0)$ and $Y = (0,1)$ are rationalizable. I assume for now that ε has a known distribution independent of X, denoted F_ε.

As in a large strand of literature, ex. Tamer (2003) and Ciliberto and Tamer (2009), I place no restrictions on the "selection mechanism," i.e. how equilibria are selected within the multiplicity region. As a result, choosing θ does not immediately yield a well-defined probability distribution for Y, given $X = x$. As is known, though, it is possible to complete the model via a random object which represents the mechanism, ex. see Section 3 in Beresteanu, Molchanov, and Molinari (2011). I formulate the setup within my framework via such a completion. I augment the model with scalar heterogeneity, $S \in [0,1]$, which represents the aggregate probability of observing outcome $(0,1)$ if ε realizes in the multiplicity region.[24] Then, for any $y \in \{0,1\}^2$, let $p_{y_1y_2}(x,\theta)$ denote the probability, under F_ε, of the ε-region where $Y = y$ is the unique outcome given $X = x$ and θ. The model map is then

$$\mu^0_{x,s,\theta}(\{y\}) := p_{y_1y_2}(x,\theta) + \mathbf{1}\{y_1 + y_2 = 1\} \times s^{\mathbf{1}\{y=(0,1)\}}[1-s]^{1-\mathbf{1}\{y=(0,1)\}} \times p_E(x,\theta),$$

where $p_E(x,\theta) := 1 - \sum_{y\in\{0,1\}^2} p_{y_1y_2}(x,\theta)$ is the probability of the multiplicity region.

For any x and θ, the dual cone restrictions on a function m translate into the following representation, which uses the extreme points of $[0,1]$, the support of S:

$$\begin{pmatrix} p_{00}(x,\theta) & p_{01}(x,\theta) + p_E(x,\theta) & p_{10}(x,\theta) & p_{11}(x,\theta) \\ p_{00}(x,\theta) & p_{01}(x,\theta) & p_{10}(x,\theta) + p_E(x,\theta) & p_{11}(x,\theta) \end{pmatrix} \begin{pmatrix} m(0,0) \\ m(0,1) \\ m(1,0) \\ m(1,1) \end{pmatrix} \geq 0.$$

[24] In principle, it is possible to write the selection mechanism as depending on all objects, including the specific realizations of ε. As the dual cone only cares about the observed outcomes, though, I can equivalently work with the integrated selection probability.

It is not too difficult to verify that a core-determining class of moment functions exists which corresponds to moment equality conditions for the probabilities of outcomes $(0,0)$ and $(1,1)$, alongside lower bound inequalities for the probabilities of $(0,1)$ and $(1,0)$. This reproduces the original results in Tamer (2003) and other papers.

Although the above specification is the most basic one in its class, it is possible to augment it with various features within my framework. To give some examples: One can incorporate restrictions on the selection mechanism, like $S \geq \alpha$ for some $\alpha \in [0,1]$. This affects the extreme selection mechanisms which are relevant for the dual cone. Another option is to allow for heterogeneity in the competitive environments across markets, ex. collusion or imperfect information. Researchers may also wish to be robust to distributional assumptions on ε; expanding the possible F_ε's is one way of doing this. Missing data is also feasible to include, via the methods used in earlier sections.

A last example is the use of random coefficients.[25] For instance, suppose that the competitive parameter Δ can now differ across markets, and that the random coefficient distribution is the object of interest, or functionals of it like moments. In that case, one can include these functionals in θ. For example, conditioning on $X = x$, one might study

$$Y_j = \mathbf{1}\{x_j'\beta + (\overline{\Delta}_x + \eta_x)Y_{3-j} + \varepsilon_j \geq 0\}, \quad j \in \{1,2\},$$

where $\overline{\Delta}_x$ is the mean, across markets of type x, of a competitive effect which is identical between the firms, and η_x has some mean-zero distribution within some class. If η_x is only nonparametrically restricted, for instance, then dual cone analysis for a particular $(\beta, \overline{\Delta}_x)$ might focus on certain finite-support distributions for η_x, like in Section 5.

[25] One paper studying identification in such settings is Dunker, Hoderlein, Kaido, and Sherman (2018).

6.3 Endogeneity and Instruments

Finally, I consider endogeneity and instruments. Unlike Chesher and Rosen (2017), my approach is not aimed at this class of models. Nevertheless, I can still make progress.

For concreteness, I examine the binary choice model $Y = \mathbf{1}\{X'\beta + \varepsilon \geq 0\}$ where the explanatory variables $X \in \mathbb{R}^{d_X}$ are arbitrarily endogenous with respect to ε. For identification purposes, a collection of instruments, $Z \in \mathcal{Z}$, is instead available. The assumption is that $\varepsilon | Z = z \sim F_\varepsilon$ for all $z \in \mathcal{Z}$, where F_ε is some known distribution (or part of θ). For prior work on this type of setup, see Chesher (2013) and Chesher, Rosen, and Smolinski (2013).

Fix β and condition on $Z = z$. In my framework, I specify heterogeneity as the conditional distribution of X given ε, $F_{X|\varepsilon}$. A natural model map outputs probability distributions for (Y, X). For a measurable $m : \{0,1\} \times \mathcal{X} \to \mathbb{R}$, the resulting dual cone condition is

$$\int \left[\int \left(m(0,x) \times \mathbf{1}\{x'\beta + e < 0\} + m(1,x) \times \mathbf{1}\{x'\beta + e \geq 0\} \right) dF_{X|\varepsilon=e}(x) \right] dF_\varepsilon(e) \geq 0,$$

for all possible $F_{X|\varepsilon}$. In the display, the inner integral is always minimized by $F_{X|\varepsilon=e}$ being a point mass that selects the minimal value for $m(y,x)$ which is possible across x's, given e.

For concreteness, suppose X is scalar and binary, along with $\beta > 0$. Then, for $e \in [0, \infty)$, only outcome 1 is possible; for $e \in [-1, 0)$, both outcomes are possible, depending on x; and for $e \in (-\infty, -1)$, only outcome 0 is possible. The previous display therefore simplifies to

$$\begin{aligned} &\min\{m(1,0), m(1,1)\} \times F_\varepsilon([0,\infty)) \\ &+ \min\{m(0,0), m(1,1)\} \times F_\varepsilon([-1,0)) \\ &+ \min\{m(0,0), m(0,1)\} \times F_\varepsilon((-\infty,-1)) \geq 0. \end{aligned}$$

Then, as the terms are all minimums multiplied by non-negative constants, it is possible to re-write the above as a system of inequalities, or equivalently a matrix inequality $Bm \geq 0$ like in the last subsection. One can study the relevant m's using convex analysis.

My approach is more cumbersome than Chesher and Rosen (2017)'s in this example. However, there may be some advantages if additional heterogeneity or structure is present.

7 Numerical Illustrations

As the last main content of the paper, I present some numerical illustrations for the binary choice moment inequalities of Section 4. There are two parts. First, I display the identification potental of the inequalities by considering stylized DGPs with two discrete covariates. This allows me to calculate identified sets and show them visually, alongside confidence regions. Then, to demonstrate that my results can actually be used in settings closer to empirical practice, I work with a selected sample from the Current Population Survey (CPS), from IPUMS (Flood et al. (2025)).[26] I show that robust inference is possible, even with a dozen regressors.

I exclusively use logit models in this section; Λ denotes the standard logistic CDF.

7.1 Simple DGPs

Consider the setup from Section 4, and let X, the always-observed covariate, and Z^*, the potentially missing one, be discrete scalar random variables. This lets me simulate population-level probabilities, and I can also use all moment inequalities from Theorem 4.1. As there are finitely many outcomes and moment inequalities, standard regularity conditions for inference procedures are generically satisfied; it is also helpful that the probability ratios in the $m_{q_\varnothing,q}$ functions are just exponentials. Finally, in the following DGPs, the true parameters are always $(\beta_0, \gamma_0) = (0, 1)$, so that the logit model for the outcomes is $P(Y = 1|X = x, Z^* = z^*) = \Lambda(z^*)$.

In Figure 1, I display some identified sets (or more precisely, Θ_I^+) for different supports of X and Z^*; the plots are grids of possible parameter values. In the first column, the

[26] The source data comes from the United States Census Bureau and Bureau of Labor Statistics.

support of X is $\mathcal{X} = \{1\}$; in the second column, $\mathcal{X} = \{-2, -1, 1, 2\}$. Meanwhile, the first row corresponds to $\mathcal{Z}^* = \{-1, 1\}$, while the second row has $\mathcal{Z}^* = \{-2, -1, 1, 2\}$. In all of the DGPs, Z^* is uniform on its support regardless of X, while the missing-ness process is $P(Z \neq \varnothing | Y = y, X = x, Z^* = z^*) = \Lambda(x + y \times z^*)$.

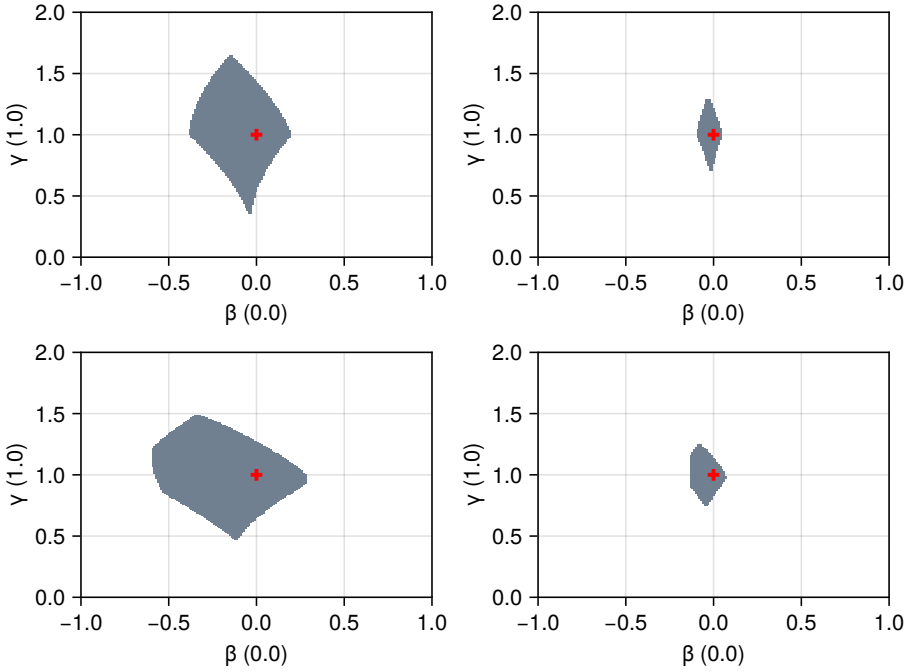


Figure 1: Identified sets for different $\mathcal{X}$'s and $\mathcal{Z}^*$'s.

In this specification, the projection of the identified sets along the β-dimension is smaller than the one for γ. Meanwhile, when $\mathcal{X}$ is larger and wider, the identified set can become quite small. The effect of changing $\mathcal{Z}^*$, however, is more ambiguous. In Figure 1, enlarging $\mathcal{Z}^*$ flattens the identified set out, which reflects less information on β and more on γ.

Moving on, I now show that my method can be robust while a feasible MLE based on MCAR, i.e. ignoring observations with missing data, is invalid.[27] I set $\mathcal{X} = \{-1, 0, 1\}$ and $\mathcal{Z}^* = \{-1, 1\}$, with Z^* again uniform conditional on X. The missing-ness process is

$$P(Z \neq \varnothing | Y = y, X = x, Z^* = z^*) = 1 - 0.5 \times [\mathbf{1}\{y = 1, z^* = -1\} + \mathbf{1}\{y = 0, z^* = 1\}].$$

[27] As the data is missing-not-at-random here, MAR methods should also fail, but I stick with MCAR for simplicity.

The first plot of Figure 2 displays resulting identified sets, where I generate the maximum likelihood one by checking the moment inequalities with the incomplete observations deleted; this is equivalent to verifying $\mathbb{E}[Y - \Lambda(\beta X + \gamma Z)|X, Z = z^*] = 0$ for $z^* \in \mathcal{Z}^*$, i.e. on the fully observed data only. Note that the ML identified set is small but far from the truth.

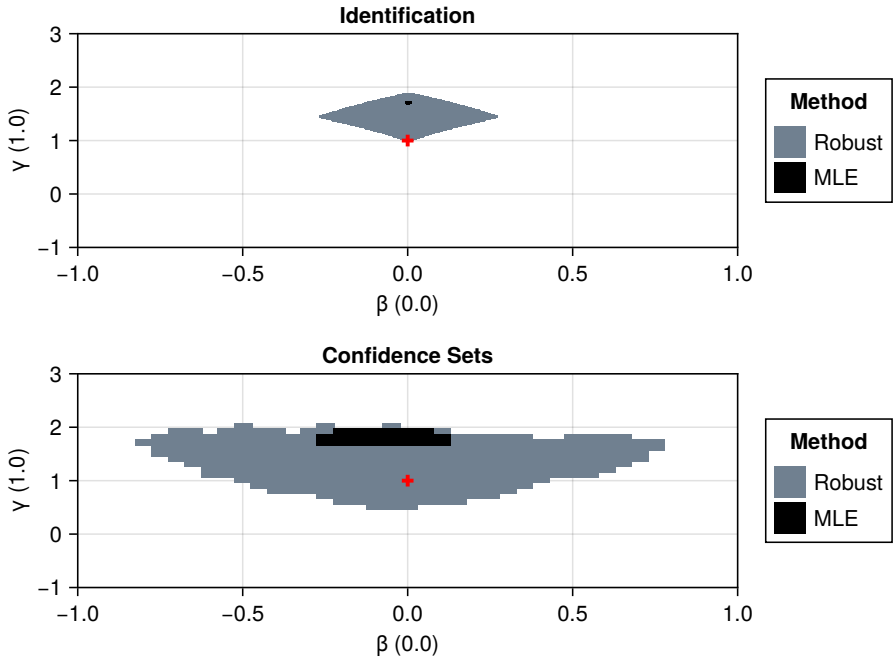


Figure 2: Comparison with maximum likelihood based on MCAR.

The invalidity of the MLE also extends to inference. In the second plot of Figure 2, I display 95% confidence sets for a single random sample of $N = 2000$ observations. For the moment inequalities, I compute the set by inverting the Andrews and Shi (2013) test. For the feasible MLE, the shown set is a confidence set based on the Wald test. As seen in the figure, the robust confidence set is large, but it contains the true value. Meanwhile, the ML confidence set is exactly the reverse. To check that this sample is not special, I also draw a hundred additional ones, checking in each whether the hypothesis tests reject the true value. I find that the Wald test always rejects the true parameters, while the moment inequality test never rejects.

Finally, I attempt to quantify the efficiency loss from allowing arbitrary missing-ness

when the data actually satisfies MCAR. To do this, I fix the missing-ness probability at 0.1, so that the feasible MLE is valid; I also reduce the sample size to $N = 500$. Then, for each sample, I invert the Andrews and Shi (2013) test on a grid of parameter values, after which I check the proportion of the robust set which also lies in the ML confidence set. Finally, I average the proportion over draws. The resulting average area-ratio is approximately 0.108.

7.2 Empirical Illustration

In the previous subsection, I generated confidence sets by inverting tests for the entire parameter vector. This is infeasible in most empirical applications, which generally include a non-trivial number of additional covariates. I show here that it is nevertheless possible to use my moment inequalities in a more typical specification. I accomplish this through applying subvector inference procedures to individual elements of θ. Robust inference on average marginal effects (AMEs) is also feasible.

I work here with a cross section of $N = 4678$ working-age women from the Current Population Survey. I use an indicator for employment as the dependent variable. Meanwhile, for the explanatory variables, X consists of an intercept and ten covariates (ex. years of education and age variables), and Z^* is a lagged employment indicator.

In my chosen sample, there is no missing-ness. Instead, I approach this as a simulation. To start, I obtain initial logit maximum likelihood estimates on the full sample, which I treat as the true parameter values. Then, I generate replacement outcomes using the logit model, followed by artifical missing-ness in Z^* through a particular process.

In Table 1 below, I report estimation results for one draw where around 84% of the observations are completely observed. The columns correspond to the following objects: the true model parameter from the original logit MLE; the infeasible ML confidence interval (CI) which knows the true Z^*; the feasible ML CI which drops observations with missing-ness; and a CI based on my moment inequalities. All confidence levels are 95%. To construct the last CIs, I apply Bei (2024)'s subvector inference procedure for unconditional moment

inequalities to individual parameters. See Appendix B for more details.

I consider three target parameters: the regression coefficient on years of education, the regression coefficient on the lagged employment indicator, and the full-population average marginal effect for years of education, $\tau_{educyears}$. Note that this last object depends also on the missing covariate values; as such, the feasible MLE has no means of directly estimating it. I instead report the corresponding version which only uses the fully observed data. Meanwhile, to compute the robust CI for the AME, I add it as a parameter in θ. Then, I augment my collection of moment inequalities for the coefficients with two additional unconditional ones: lower and upper bounds for τ which impute worst- or best-case values for Z^* when it is missing. The robust CI is thus conservative.

	Truth	I-MLE	F-MLE	Robust
$educyears$	0.171	[0.138, 0.213]	[0.243, 0.346]	[0.022, 0.347]
$lagworking$	4.557	[4.336, 4.797]	[4.418, 5.082]	[1.731, 6.308]
$\tau_{educyears}$	0.012	[0.010, 0.016]	[0.013, 0.018]	[0.007, 0.026]

Table 1: Confidence intervals for regression coefficients and an AME.

With the caveat that Table 1 only represents a single draw, there are nevertheless some observations which align with expectations. The variable $educyears$ enters into my specified missing-ness process, while the reported feasible MLE CI for its coefficient is above the true value. This reflects the bias from placing an incorrect assumption on the missing-ness process. At the same time, the variable $lagworking$ does not affect missing-ness, which seems to be reflected in the feasible ML CI. In comparison with these, the robust CIs are meaningfully longer, which could reflect both statistical uncertainty and a loss of some identification power, as I am using an unconditional moment inequality procedure. Nevertheless, they seem to be valid for all three parameters, and in this instance they are still informative on effects.

More importantly, although the moment inequality procedure is computationally demanding, the above table indicates that it is nevertheless possible to perform robust inference even with various additional covariates. This is not immediate, as many moment inequality procedures suffer from a curse of dimensionality in the number of parameters.

8 Conclusion

In this paper, I study mean regression models with incompletely observed covariates where the observability process is completely unrestricted. I characterize the identified set for the parameters via collections of conditional moment inequalities. The inequalities have intuitive structures, and I show that they can be informative, as well as feasibly used in empirical applications.

To obtain my concrete results, I also develop a general framework and identification approach based on heterogeneity and a dual cone. I argue that the approach can sometimes have advantages over existing competitors, and I demonstrate how it might be used in models beyond mean regression with incomplete data. For example, quantile regression with incomplete data can be studied using the same arguments as in mean regression. Applying my framework in other models represents one potential direction for future research. For instance, in concurrent work, I use my dual cone approach to study a class of panel models.

References

Abrevaya, J., and S. G. Donald (2017): "A GMM Approach for Dealing with Missing Data on Regressors," *Review of Economics and Statistics*, 99(4), 657–662.

Aliprantis, C. D., and K. C. Border (1999): *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer-Verlag, Berlin, second edn.

Andrews, D. W. K., and X. Shi (2013): "Inference Based on Conditional Moment Inequalities," *Econometrica*, 81(2), 609–666.

——— (2017): "Inference Based on Many Conditional Moment Inequalities," *Journal of Econometrics*, 196(2), 275–287.

Andrews, I., J. Roth, and A. Pakes (2023): "Inference for Linear Conditional Moment Inequalities," *Review of Economic Studies*, 90(6), 2763–2791.

Aucejo, E. M., F. A. Bugni, and V. J. Hotz (2017): "Identification and Inference on Regressions with Missing Covariate Data," *Econometric Theory*, 33, 196–241.

Bahadur, R. R., and L. J. Savage (1956): "The Nonexistence of Certain Statistical Procedures in Nonparametric Problems," *Annals of Mathematical Statistics*, 27(4), 1115–1122.

Bei, X. (2024): "Local linearization based subvector inference in moment inequality models," *Journal of Econometrics*, 238, 1–35.

Beresteanu, A., I. Molchanov, and F. Molinari (2011): "Sharp Identification Regions in Models with Convex Moment Predictions," *Econometrica*, 79(6), 1785–1821.

Beresteanu, A., and Y. Sasaki (2021): "Quantile regression with interval data," *Econometric Reviews*, 40(6), 562–583.

Bertanha, M., and M. J. Moreira (2020): "Impossible Inference in Econometrics: Theory and Applications," *Journal of Econometrics*, 218(2), 247–270.

Besancon, M., T. Papamarkou, D. Anthoff, A. Arslan, S. Byrne, D. Lin, J. Pearson, and M. Schauer (2021): "Distributions.jl: Definition and Modeling of Probability Distributions in the JuliaStats Ecosystem," *Journal of Statistical Software*, 98(16), 1–30.

Bezanson, J., A. Edelman, S. Karpinski, and V. B. Shah (2017): "Julia: A fresh approach to numerical computing," *SIAM Review*, 59(1), 65–98.

Bonhomme, S. (2012): "Functional differencing," *Econometrica*, 80(4), 1337–1385.

Bonhomme, S., K. Dano, and B. Graham (2023): "Identification in a binary choice panel data model with a predetermined covariate," *SERIEs*, pp. 315–351.

——— (2026): "Moment Restrictions for Nonlinear Panel Data Models with Feedback," Working Paper.

Borjas, G. G., and A. Edo (2026): "Gender, Selection into Employment, and the Wage Impact of Immigration," *Journal of Labor Economics*, 44(2), 515–552.

Botosaru, I., I. Loh, and C. Muris (2024): "An Adversarial Approach to Identification," arXiv preprint arXiv:2411.04239v2 [econ.EM].

Chesher, A. (2013): "Semiparametric structural models of binary response: shape restrictions and identification," *Econometric Theory*, 29, 231–266.

Chesher, A., and A. M. Rosen (2017): "Generalized Instrumental Variable Models," *Econometrica*, 85(3), 959–989.

Chesher, A., A. M. Rosen, and K. Smolinski (2013): "An instrumental variable model of multiple discrete choice," *Quantitative Economics*, 4(2), 157–196.

Ciliberto, F., and E. Tamer (2009): "Market Structure and Multiple Equilibria in Airline Markets," *Econometrica*, 77(6), 1791–1828.

Danisch, S., and J. Krumbiegel (2021): "Makie.jl: Flexible high-performance data visualization for Julia," *Journal of Open Source Software*, 6(65), 3349.

Dobronyi, C. R., J. Gu, T. M. Russell, and K. i. Kim (2025): "Identification of Dynamic Panel Logit Models with Fixed Effects," arXiv preprint arXiv:2104.04590v3 [econ.EM].

Dunker, F., S. Hoderlein, H. Kaido, and R. Sherman (2018): "Nonparametric identification of the distribution of random coefficients in binary response static games of complete information," *Journal of Econometrics*, 206(1), 83–102.

Ekeland, I., A. Galichon, and M. Henry (2010): "Optimal transportation and the falsifiability of incompletely specified economic models," *Economic Theory*, 42, 355–374.

Fankhauser, M., and G. Peled (2026): "Sharp Identification for Regressions with Interval-Observed and Missing Covariates," Job Market Paper.

Flood, S., M. King, R. Rodgers, S. Ruggles, J. R. Warren, D. Backman, E. Breton, G. Cooper, J. Rivera Drew, S. Richards, D. Van Riper, and K. C. W. Williams (2025): "IPUMS CPS: Version 13.0 [dataset]," Minneapolis, MN: IPUMS, `https://doi.org/10.18128/D030.V13.0`.

Franguridi, G., and H. R. Moon (2026): "Generalized Method of Moments with Partially Missing Data," arXiv preprint arXiv:2511.21988v2 [econ.EM].

Galichon, A., and M. Henry (2011): "Set Identification in Models with Multiple Equilibria," *Review of Economic Studies*, 78(4), 1264–1298.

Goulart, P. J., and Y. Chen (2024): "Clarabel: An interior-point solver for conic programs with quadratic objectives," arXiv preprint arXiv:2405.12762 [math.OC].

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Honoré, B. E., and E. Tamer (2006): "Bounds on Parameters in Panel Dynamic Discrete Choice Models," *Econometrica*, 74(3), 611–629.

Honoré, B. E., and M. Weidner (2025): "Moment Conditions for Dynamic Panel Logit Models with Fixed Effects," *Review of Economic Studies*, 92, 3112–3137.

Horowitz, J. L., and C. F. Manski (1998): "Censoring of outcomes and regressors due to survey nonresponse: Identification and estimation using weights and imputations," *Journal of Econometrics*, 84, 37–58.

——— (2000): "Nonparametric Analysis of Randomized Experiments with Missing Covariate and Outcome Data," *Journal of the American Statistical Association*, 95(449), 77–84.

——— (2006): "Identification and estimation of statistical functionals using incomplete data," *Journal of Econometrics*, 132, 445–459.

Horowitz, J. L., C. F. Manski, M. Ponomareva, and J. Stoye (2003): "Computation of bounds on population parameters when the data are incomplete," *Reliable Computing*, 9, 419–440.

Johnson, S. G. (2007): "The NLopt nonlinear-optimization package," `https://github.com/stevengj/nlopt`.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, 70(4), 1397–1432.

Klenke, A. (2020): *Probability Theory*. Springer, Cham, Switzerland.

Li, L. (2026): "Identification and Counterfactual Analysis in Incomplete Models with Support and Moment Restrictions," arXiv preprint arXiv:2603.07722v1 [econ.EM].

Liu, D. C., and J. Nocedal (1989): "On the limited memory BFGS method for large scale optimization," *Mathematical Programming*, 45, 503–528.

Lubin, M., O. Dowson, J. Dias Garcia, J. Huchette, B. Legat, and J. P. Vielma (2023): "JuMP 1.0: Recent improvements to a modeling language for mathematical optimization," *Mathematical Programming Computation*, 15, 581–589.

Madsen, K., S. Zertchaninov, and A. Zilinskas (1998): "StoGO: A program for global optimization," `http://www.imm.dtu.dk/projects/scicomp/GlobOpt/opt.html`.

Manski, C. F. (2003): *Partial Idenitification of Probability Distributions*. Springer-Verlag, New York.

——— (2025): "Inference with Imputed Data: The Allure of Making Stuff Up," *Journal of Labor Economics*, 43(S1), S333–S350.

Manski, C. F., and E. Tamer (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

Mbakop, E. (2023): "Identification in Some Discrete Choice Models: A Computational Approach," arXiv preprint arXiv:2305.15691 [econ.EM].

Mogensen, P. K., and A. N. Riseth (2018): "Optim: A mathematical optimization package for Julia," *Journal of Open Source Software*, 3(24), 615.

Molinari, F. (2020): "Microeconometrics with partial identification," in *Handbook of Econometrics*, ed. by S. N. Durlauf, L. P. Hansen, J. J. Heckman, and R. L. Matzkin, vol. 7A of *Handbooks of Economics*, pp. 355–486. Elsevier, North Holland.

Powell, M. J. D. (2009): "The BOBYQA algorithm for bound constrained optimization without derivatives," Discussion Paper NA2009/06, Department of Applied Mathematics and Theoretical Physics, Cambridge University, Cambridge, UK.

Revels, J., M. Lubin, and T. Papamarkou (2016): "Forward-Mode Automatic Differentiation in Julia," arXiv preprint arXiv:1607.07892 [cs.MS].

Runarsson, T. P., and X. Yao (2005): "Search biases in constrained evolutionary optimization," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 35, 233–243.

Schaefer, H. H., and M. P. Wolff (1999): *Topological Vector Spaces*. Springer, New York, NY.

Schennach, S. M. (2014): "Entropic Latent Variable Integration via Simulation," *Econometrica*, 82(1), 345–385.

Singer, I. (1997): *Abstract Convex Analysis*. Wiley-Interscience, New York, NY.

Svanberg, K. (2002): "A class of globally convergent optimization methods based on conservative convex separable approximations," *SIAM Journal on Optimization*, 12, 555–573.

Tamer, E. (2003): "Incomplete Simultaneous Discrete Response Model with Multiple Equilibria," *Review of Economic Studies*, 70(1), 147–165.

Torgovitsky, A. (2019): "Partial identification by extending subdistributions," *Quantitative Economics*, 10, 105–144.

van der Vaart, A. W., and J. A. Wellner (2023): *Weak Convergence and Empirical Processes*. Springer, Cham, Switzerland, second edn.

Winkler, G. (1988): "Extreme Points of Moment Sets," *Mathematics of Operations Research*, 13(4), 581–587.

A Proofs

A.1 Proposition 3.1

Proof. Fix $(w, \theta) \in \mathcal{W} \times \Theta$ and a probability measure $\widetilde{\mu}$ on $\mathcal{Y}$. To reduce notational clutter, I will use $\mathcal{M}$ in place of $\mathcal{M}_{w,\theta}$ and $\mathcal{L}^1_Y$ in place of $\mathcal{L}^1(\mathcal{P}^0_Y(w, \theta))$. $\mathcal{M}_1$ denotes the intersection of $\mathcal{M}$ with the probability measures on $\mathcal{Y}$. (I do not use $\mathcal{M}(\mathcal{Y})$ or $\mathcal{M}_1(\mathcal{Y})$ in the proof, to reduce risk of confusion.)

I prove the first equality by using results on abstract convexity from Singer (1997). To start, observe from the definitions that $\theta \in \Theta^+_I(\widetilde{\mu}|w)$ iff

$$\widetilde{\mu} \in \left\{ \mu \in \mathcal{M}_1 \;\middle|\; \int m d\mu \geq 0, \quad \forall m \in \mathcal{D}(w, \theta) \right\}, \tag{8}$$

where I have used the fact that μ is a probability measure to replace $\mathcal{M}$ with $\mathcal{M}_1$. Now, define the "coupling function"

$$\varphi : \mathcal{M}_1 \times \mathcal{L}^1_Y \to \overline{R}, \quad \varphi(\mu, m) = \int m d\mu,$$

which is well-defined (and even real-valued) by the restriction of the integrating measures to a subset of $\mathcal{M}$. Regarding φ, note that

$$\{\varphi(\cdot, m) \mid m \in \mathcal{L}^1_Y\} + \mathbb{R} = \{\varphi(\cdot, m) \mid m \in \mathcal{L}^1_Y\}, \tag{9}$$

where $\mathbb{R}$ is identified with the constant functions on $\mathcal{Y}$, because $\varphi(\cdot, m) + c = \varphi(\cdot, m + c)$ for all $c \in \mathbb{R}$ by the first argument of φ being a probability measure. Finally, define the following "abstract dualities" from Singer (1997) based on φ: For $\mathcal{R} \subseteq \mathcal{M}_1$ and $\mathcal{R}' \subseteq \mathcal{L}^1_Y$,

$$\Delta_\varphi \mathcal{R} \equiv \Delta_\varphi(\mathcal{R}) := \{m \in \mathcal{L}^1_Y \mid \varphi(\mu, m) \leq 0, \quad \forall \mu \in \mathcal{R}\},$$

$$\Delta'_\varphi \mathcal{R}' \equiv \Delta'_\varphi(\mathcal{R}') := \{\mu \in \mathcal{M}_1 \mid \mathcal{R}' \subseteq \Delta_\varphi(\{\mu\})\} = \{\mu \in \mathcal{M}_1 \mid \varphi(\mu, m) \leq 0, \quad \forall m \in \mathcal{R}'\}.$$

These definitions imply the following equation:

$$\begin{aligned}
\Delta'_\varphi \Delta_\varphi \mathcal{R} &= \{\mu \in \mathcal{M}_1 \mid \Delta_\varphi \mathcal{R} \subseteq \Delta_\varphi(\{\mu\})\} \\
&= \{\mu \in \mathcal{M}_1 \mid \varphi(\mu, m) \leq 0, \quad \forall m \in \mathcal{L}_Y^1 \text{ with } \sup_{\widehat{\mu} \in \mathcal{R}} \varphi(\widehat{\mu}, m) \leq 0\} \\
&= \{\mu \in \mathcal{M}_1 \mid \varphi(\mu, m) \geq 0, \quad \forall m \in \mathcal{L}_Y^1 \text{ with } \inf_{\widehat{\mu} \in \mathcal{R}} \varphi(\widehat{\mu}, m) \geq 0\},
\end{aligned}$$

where the last equality uses the linearity of φ and $\mathcal{L}_Y^1$ being a vector space. Inspection of the last expression in the display shows that it is equal to the set in (8) when $\mathcal{R}$ is replaced by the set of model probabilities, $\mathcal{P}_Y^0(w, \theta)$.

The first equality in the proposition then follows by Proposition 6.5 in Singer (1997): By display (9), as well as Remark 6.14(f) in Singer (1997), Proposition 6.5 yields that

$$\Delta'_\varphi \Delta_\varphi \mathcal{P}_Y^0(w, \theta) = co_{\mathcal{L}_Y^1, \varphi} \mathcal{P}_Y^0(w, \theta),$$

where the RHS is (see display (2.196) in Singer (1997))

$$co_{\mathcal{L}_Y^1, \varphi}(\mathcal{P}_Y^0(w, \theta)) = \{\mu \in \mathcal{M}_1 \mid \nexists m \in \mathcal{L}_Y^1 \text{ such that } \sup_{\widehat{\mu} \in \mathcal{P}_Y^0(w, \theta)} \varphi(\widehat{\mu}, m) < \varphi(\mu, m)\}.$$

The first RHS in Proposition 3.1 results from re-writing the above display. I replace the condition $\mu \in \mathcal{M}_1$ with $\mu \in \mathcal{M} \equiv \mathcal{M}_{w,\theta}$ because $\widetilde{\mu}$ is known to be a probability measure, and I use Assumption 3.2 and the linearity of φ to reduce the supremum to one over the Dirac heterogeneity measures. (For a more explicit formalization of this last step, see the argument in the proof of Proposition 3.2, and take a supremum over the inner integral.)

The second equality of Proposition 3.1 is a standard alternate characterization of a support function formulation. I provide a formal argument for completeness. Let $\langle \cdot, \cdot \rangle : \mathcal{M} \times \mathcal{L}_Y^1 \to \mathbb{R}$ be the natural extension of φ in its first argument, i.e. $\langle \mu, m \rangle = \int m d\mu$,

and observe that it is a real-valued bilinear form on two real vector spaces. It also satisfies standard separation axioms: If $\langle \mu, m \rangle = 0$ for all $m \in \mathcal{L}_Y^1$, then μ is the zero measure due to the measurable indicator functions (which are bounded and thus always integrable), and if $\langle \mu, m \rangle = 0$ for all $\mu \in \mathcal{M}$, then m is the zero function due to the Dirac measures on $\mathcal{Y}$ (with respect to which all measurable real-valued functions are integrable). Therefore, $(\mathcal{M}, \mathcal{L}_Y^1, \langle \cdot, \cdot \rangle)$ comprises a (separated) dual system.

Now, endow $\mathcal{M}$ with the weak topology mentioned in the proposition. With this, $\mathcal{M}$ is a locally convex space. Also, by the result in Section 1.2, Chapter IV of Schaefer and Wolff (1999), the dual space of $\mathcal{M}$ under this topology is exactly $\{\langle \cdot, m \rangle \mid m \in \mathcal{L}_Y^1\}$. It follows that the set

$$\left\{ \mu \in \mathcal{M} \;\middle|\; \sup_{a \in \mathcal{A}} \int m d\mu_{w,a,\theta}^0 \geq \int m d\mu, \quad \forall m \in \mathcal{L}_Y^1 \right\},$$

which appears in the first RHS of the proposition, is the intersection of all closed half-spaces containing $\{\mu_{w,a,\theta}^0 \mid a \in \mathcal{A}\}$. I thus conclude that the set above is exactly $\overline{co}(\{\mu_{w,a,\theta}^0 \mid a \in \mathcal{A}\})$, where the closed convex hull operator is taken in $\mathcal{M}$ under the weak topology; for example, see Section 2.2b (and Section 2.1 for definitions) in Singer (1997).

□

A.2 Proposition 3.2

Proof. Fix $(w, \theta) \in \mathcal{W} \times \Theta$, along with $\mathcal{P}_A^0, \widetilde{\mathcal{P}}_A^0 \in \mathcal{M}_1(\mathcal{A})$. Let $\mathcal{D}(w, \theta)$ and $\widetilde{\mathcal{D}}(w, \theta)$ be the corresponding dual cones with respect to the common model map μ^0. Finally, let $\mathcal{D}^*(w, \theta) \subseteq \mathcal{D}(w, \theta)$ be strongly core-determining at (w, θ) in the $\mathcal{P}_A^0$ model.

Suppose that $\mathcal{P}_A^0 \subseteq \widetilde{\mathcal{P}}_A^0$ and $\mathcal{D}^*(w, \theta) \subseteq \mathcal{L}^1(\mathcal{P}_Y^0(\cdot; \widetilde{\mathcal{P}}_A^0))$. There are two requirements which must be verified in order to conclude that $\mathcal{D}^*(w, \theta)$ is strongly core-determining in the $\widetilde{\mathcal{P}}_A^0$ model. First, I need to check that $\mathcal{D}^*(w, \theta) \subseteq \widetilde{\mathcal{D}}(w, \theta)$, which reduces to checking the non-negativity property because integrability is assumed. Then, I must prove that the strong core-determining property remains satisfied.

For the first requirement, let $\mu \in \mathcal{P}_Y^0(\cdot; \widetilde{\mathcal{P}}_A^0)$ with generating heterogeneity distribution $\nu \in \widetilde{\mathcal{P}}_A^0$. It suffices to show that $\int m^* d\mu \geq 0$ for any $m^* \in \mathcal{D}^*(w, \theta)$. Accordingly, let m^* be some arbitrary function in $\mathcal{D}^*(w, \theta)$.

I proceed using conventional measure theory arguments. By assumption, m^* is integrable with respect to μ. Let $m^* = m^{*+} - m^{*-}$ be the usual decomposition of m into its positive and negative parts, with integrability implying that

$$0 \leq \int m^{*+} d\mu < \infty, \quad 0 \leq \int m^{*-} d\mu < \infty.$$

Now, using standard arguments based on simple functions and monotone convergence, as well as the formula $\mu(\mathcal{S}) = \int \mu^0_{w,a,\theta}(\mathcal{S}) d\nu(a)$ for all measurable $\mathcal{S} \subseteq \mathcal{Y}$, it is straightforward to establish the equality

$$\int m^{*+} d\mu = \int \left[\int m^{*+} d\mu^0_{w,a,\theta} \right] d\nu(a),$$

and likewise for m^{*-}. Finiteness of the relevant integrals then yields the combined formula

$$\int m^* d\mu = \int \left[\int m^* d\mu^0_{w,a,\theta} \right] d\nu(a).$$

(Recall that m^* is integrable with respect to $\mu^0_{w,a,\theta}$ for all $a \in \mathcal{A}$, because the Dirac measures are always included in admissible heterogeneity by Assumption 3.2.) Finally, because $m^* \in \mathcal{D}^*(w, \theta) \subseteq \mathcal{D}(w, \theta)$, and $\mu^0_{w,a,\theta} \in \mathcal{P}_Y^0(\cdot; \mathcal{P}_A^0)$ by Assumption 3.2, the inner bracketed term on the RHS in the display is always non-negative. This implies that $\int m^* d\mu \geq 0$, as desired.

It remains to check the strong core-determining property. Let $\mu \in \mathcal{M}_1(\mathcal{Y})$ and $m \in \widetilde{\mathcal{D}}(w, \theta)$. Assume that all of $\mathcal{D}^*(w, \theta)$ and m are μ-integrable, and also that

$$\int m^* d\mu \geq 0 \quad \forall m^* \in \mathcal{D}^*(w, \theta).$$

Then, to prove $\int m d\mu \geq 0$, it is enough to show that $m \in \mathcal{D}(w,\theta)$, because $\mathcal{D}^*(w,\theta)$ is strongly core-determining in the $\mathcal{P}_A^0$ model; in other words, I show that $\widetilde{\mathcal{D}}(w,\theta) \subseteq \mathcal{D}(w,\theta)$.

To see why this claim is true, first observe that m satisfies the integrability condition for $\mathcal{D}(w,\theta)$ because

$$m \in \widetilde{\mathcal{D}}(w,\theta) \subseteq \mathcal{L}^1(\mathcal{P}_Y^0(\cdot;\widetilde{\mathcal{P}}_A^0)) \subseteq \mathcal{L}^1(\mathcal{P}_Y^0(\cdot;\mathcal{P}_A^0)),$$

where the last inclusion follows from $\mathcal{P}_A^0 \subseteq \widetilde{\mathcal{P}}_A^0$ and $\mathcal{P}_Y^0(\cdot;\mathcal{P}_A^0) \subseteq \mathcal{P}_Y^0(\cdot;\widetilde{\mathcal{P}}_A^0)$. (Therefore, the first $\mathcal{L}^1$ space is an intersection involving more individual $\mathcal{L}^1$ spaces, i.e. there are more measures which impose integrability constraints.) Finally, the non-negativity condition also holds because

$$\int m d\widetilde{\mu} \geq 0 \quad \forall \widetilde{\mu} \in \mathcal{P}_Y^0(\cdot;\widetilde{\mathcal{P}}_A^0) \quad \Rightarrow \quad \int m d\widehat{\mu} \geq 0 \quad \forall \widehat{\mu} \in \mathcal{P}_Y^0(\cdot;\mathcal{P}_A^0),$$

based on the same inclusions. □

A.3 Lemma A.1

Lemma A.1. *Consider the setting defined by Assumption 4.1. The model map μ^0 satisfies condition (iv) from Assumption 3.1.*

Proof. As $\mathbb{R}_\varnothing$ is a Polish space, $\mathcal{Z}$ is Polish as well, as a closed subset of $\mathbb{R}_\varnothing$. Therefore, to verify (iv), it suffices to check the condition for the measurable rectangles of $\{0,1\} \times \mathcal{Z}$. These rectangles can be generically written as

$$\mathcal{S} = \mathcal{S}_Y \times (\mathcal{S}_{Z^*} \cup \mathcal{S}_\varnothing),$$

where $\mathcal{S}_Y \subseteq \{0,1\}$, $\mathcal{S}_{Z^*} \subseteq \mathcal{Z}^*$ (measurable), and $\mathcal{S}_\varnothing \subseteq \{\varnothing\}$.

For such a measurable rectangle $\mathcal{S}$, the relevant function to examine for (iv) is

$$
\begin{aligned}
&(z^*, s_0, s_1) \mapsto \mu^0(\mathcal{S}|x, z^*, s_0, s_1; \theta) \\
&\qquad = \sum_{y \in \{0,1\}} \mathbf{1}\{y \in \mathcal{S}_Y\} \times \\
&\qquad\quad \Bigg\{ \Big[G(x'\beta + \gamma z^*) \Big]^y \times \Big[s_y \mathbf{1}\{z^* \in \mathcal{S}_{Z^*}\} + (1 - s_y)\mathbf{1}\{\varnothing \in \mathcal{S}_\varnothing\} \Big] \\
&\qquad\quad + \Big[1 - G(x'\beta + \gamma z^*) \Big]^{1-y} \times \Big[s_y \mathbf{1}\{z^* \in \mathcal{S}_{Z^*}\} + (1 - s_y)\mathbf{1}\{\varnothing \in \mathcal{S}_\varnothing\} \Big] \Bigg\}.
\end{aligned}
$$

This function is measurable in z^* for fixed (s_0, s_1) and continuous in (s_0, s_1) for fixed z^*. Accordingly, it is measurable as a Caratheodory function defined on a product of Polish spaces, ex. see Lemma 4.50 in Aliprantis and Border (1999). Alternatively, the last expression is an elementary composition (ex. addition and multiplication) of measurable functions, all of which depend on only one variable at most. □

A.4 Theorem 4.1

Proof. Fix $(x, \theta) \in \mathcal{X} \times \Theta$. I generally suppress dependence on (x, θ) to reduce notation, when this does not create risk of confusion. By the conditions of Assumption 4.1, all model outcome distributions have finite support. It follows that any measurable $m : \{0, 1\} \times \mathcal{Z} \to \mathbb{R}$ is model-integrable, and I need not consider integrability for the dual cone $\mathcal{D}(x, \theta)$.

The structure of the proof is as follows. I first use a series of arguments to arrive at $\mathcal{D}^*(x, \theta)$ as my candidate class. Then, I validate that $\mathcal{D}^*(x, \theta)$ is strongly core-determining. The latter step is the formal one which proves the claim.

Before proceeding, I reproduce the characterization of $\mathcal{D}(x, \theta)$ from the main text:

$$
\begin{aligned}
\mathcal{D}(x, \theta) = \Bigg\{ m : \{0, 1\} \times \mathcal{Z} \to \mathbb{R} \;\Bigg|\; & m \text{ meas. and } \int m d\mu^0_{x, z^*, s_0, s_1, \theta} \geq 0, \\
& \forall (z^*, s_0, s_1) \in \mathcal{Z} \times [0, 1] \times [0, 1] \Bigg\}.
\end{aligned}
$$

I can focus on Dirac heterogeneity measures because I have restricted admissible heterogeneity to finite mixtures. Now, the Dirac measures at $(z^*, s_0, s_1) \in \mathcal{Z}^* \times \{0,1\} \times \{0,1\}$ imply that the following are necessary conditions for a measurable m to belong to $\mathcal{D}(x,\theta)$:

$$m(0,\varnothing) \times [1 - G(x'\beta + \gamma z^*)] + m(1,\varnothing) \times G(x'\beta + \gamma z^*) \geq 0, \tag{10}$$

$$m(0,\varnothing) \times [1 - G(x'\beta + \gamma z^*)] + m(1,z^*) \times G(x'\beta + \gamma z^*) \geq 0, \tag{11}$$

$$m(0,z^*) \times [1 - G(x'\beta + \gamma z^*)] + m(1,\varnothing) \times G(x'\beta + \gamma z^*) \geq 0, \tag{12}$$

$$m(0,z^*) \times [1 - G(x'\beta + \gamma z^*)] + m(1,z^*) \times G(x'\beta + \gamma z^*) \geq 0, \tag{13}$$

for all $z^* \in \mathcal{Z}^*$. By extreme point arguments, it is straightforward to verify that these conditions are also sufficient. In my proof, I will work with (10) through (13) when handling the dual cone.

Derivation of $\mathcal{D}^*(w,\theta)$:

As the integral of any measurable non-negative function with respect to a probability measure is automatically non-negative, it suffices to restrict attention to m's which take at least one (strictly) negative value. Moreover, any valid function which point-wise dominates another valid function is redundant can be ignored; see the discussion in Section 3.4.

Step 1: Signs of $m(\cdot,\varnothing)$

I assume $m(y,z) < 0$ for some $(y,z) \in \{0,1\} \times \mathcal{Z}$. Then, at least one of $m(0,\varnothing)$ or $m(1,\varnothing)$ must be positive. If the mentioned z equals z^* for some $z^* \in \mathcal{Z}^*$, then $m(1-y,\varnothing) > 0$ by (11) or (12); otherwise, if z equals $\varnothing$, then (10) generates the positivity requirement.

Moreover, I argue that it is sufficient to consider only two distinct cases:

$$m(0,\varnothing) > 0, m(1,\varnothing) \leq 0 \quad \text{and} \quad m(0,\varnothing) \leq 0, m(1,\varnothing) > 0.$$

Suppose that $m \in \mathcal{D}(x,\theta)$ satisfies $m(1,\varnothing) > 0$ and $m(0,\varnothing) > 0$. Then, as I show, it is

possible to write m as $m = m_0 + m_1$, where m_0 falls under the first case and m_1 under the second. Moreover, m_0 and m_1 belong to $\mathcal{D}(x,\theta)$ as well. With $\int m_0 d\mu \geq 0$ and $\int m_1 d\mu \geq 0$ implying $\int m d\mu \geq 0$, it follows that such an m can be excluded from $\mathcal{D}^*(x,\theta)$.

To construct m_0, let $\mathcal{Z}_0^*$ be the subset of z^*'s for which $m(0,z^*) > 0$. Then, define m_0 as

$$m_0(0,\varnothing) = m(0,\varnothing), \quad m_0(1,\varnothing) = 0,$$

$$m_0(0,z^*) = m(0,z^*), \quad m_0(1,z^*) = m(1,z^*), \quad \forall z^* \in \mathcal{Z}_0^*,$$

$$m_0(\cdot) = 0 \quad \text{otherwise.}$$

As $\mathcal{Z}_0^*$ is measurable, m_0 is measurable. It also satisfies constraints (10) through (13) by inheritance from m. Meanwhile, m_1 is defined in an analogous manner, expect with respect to $\mathcal{Z}^* \setminus \mathcal{Z}_0^*$, and possesses the same properties.

Step 2: Initial values for $m(0,\varnothing)$ and $m(1,\varnothing)$

For concreteness, I focus on the case with $m(0,\varnothing) > 0$ and $m(1,\varnothing) \leq 0$; the other one will be symmetric. As scale is irrelevant, I normalize $m(0,\varnothing) = 1$.

With this choice, (10) imposes the constraint

$$m(1,\varnothing) \in \left[\sup_{z^* \in \mathcal{Z}^*} -\frac{1 - G(x'\beta + \gamma z^*)}{G(x'\beta + \gamma z^*)}, 0\right]$$

which is equivalent to

$$m(1,\varnothing) \in \left[-\frac{1 - \overline{G}(x,\theta)}{\overline{G}(x,\theta)}, 0\right].$$

Temporarily fix some value for $m(1,\varnothing)$ in this interval.

Step 3: Values for $m(1,z^*)$

I now consider the value to assign to $m(1,z^*)$, for generic $z^* \in \mathcal{Z}^*$. Constraints (11) and

(13) generate the following lower bound, which depends also on $m(0, z^*)$:

$$\begin{aligned} m(1, z^*) &\geq \max\left\{ -\frac{m(0, z^*)[1 - G(x'\beta + \gamma z^*)]}{G(x'\beta + \gamma z^*)}, -\frac{1 - G(x'\beta + \gamma z^*)}{G(x'\beta + \gamma z^*)} \right\} \\ &= -\min\{m(0, z^*), 1\} \times \frac{1 - G(x'\beta + \gamma z^*)}{G(x'\beta + \gamma z^*)}. \end{aligned}$$

I observe, for any $m(0, z^*)$, that it is redundant to set $m(1, z^*)$ to a value above the last expression above. In the original constraints, $m(1, z^*)$ interacts only with $m(0, \varnothing)$, which is fixed at one, and $m(0, z^*)$. Accordingly, choosing a larger value of $m(1, z^*)$ than the last expression in the display does not slacken any constraints.

To state this more formally, consider $\widetilde{m} \in \mathcal{D}(x, \theta)$ of the type that has been constructed so far, with $\widetilde{m}(1, z^*)$ greater than the displayed expression. Then, $\widetilde{m}$ point-wise dominates the function m which is equal everywhere except at $m(1, z^*)$, with $m(1, z^*)$ set equal to the bound. The adjusted function m remains measurable, and it still satisfies the non-negativity conditions of $\mathcal{D}(w, \theta)$ by construction. Therefore, I can exclude $\widetilde{m}$ from $\mathcal{D}^*(x, \theta)$.

Taking stock, I have shown that, in $\mathcal{D}^*(x, \theta)$, I am able to impose

$$m(1, z^*) = -\min\{m(0, z^*), 1\} \times \frac{1 - G(x'\beta + \gamma z^*)}{G(x'\beta + \gamma z^*)}.$$

Step 4: Values for $m(0, z^*)$

Given the previous choice for $m(1, z^*)$, constraint (13) is automatically satisfied for any value of $m(0, z^*)$. Therefore, the only remaining restriction on $m(0, z^*)$ is the following lower bound, which is set by (12) and which depends on $m(1, \varnothing)$:

$$m(0, z^*) \geq -\frac{m(1, \varnothing) G(x'\beta + \gamma z^*)}{1 - G(x'\beta + \gamma z^*)}.$$

Recall the interval constraint on $m(1, \varnothing)$ from Step 2. Inserting that constraint here, I obtain

that the lower bound lies in the interval

$$\left[0, \frac{G(x'\beta + \gamma z^*)}{\overline{G}(x, \theta)} \times \frac{1 - \overline{G}(x, \theta)}{1 - G(x'\beta + \gamma z^*)}\right],$$

whose upper endpoint is weakly less than one.

I now deduce that it is unnecessary to consider $m(0, z^*) > 1$. With fixed $m(1, \varnothing)$, increasing the value of $m(0, z^*)$ above its lower bound only potentially affects the value taken by $m(1, z^*)$ via (13); see the display at the end of the previous step. From inspection of that display, it is immediately apparent that setting $m(0, z^*) > 1$ generates no effect on $m(1, z^*)$. I thereby conclude that it suffices to restrict attention, in $\mathcal{D}^*(x, \theta)$, to

$$m(0, z^*) \in \left[-\frac{m(1, \varnothing)G(x'\beta + \gamma z^*)}{1 - G(x'\beta + \gamma z^*)}, 1\right].$$

(A similar formalization of this can be done as in Step 3, but it is essentially the same.)

With this, I have shown that for $m(0, \varnothing) = 1$ and a generic fixed $m(1, \varnothing)$, it is enough to restrict attention to m's with

$$m(0, z^*) \in \left[-\frac{m(1, \varnothing)G(x'\beta + \gamma z^*)}{1 - G(x'\beta + \gamma z^*)}, 1\right],$$

$$m(1, z^*) = -\min\{m(0, z^*), 1\} \times \frac{1 - G(x'\beta + \gamma z^*)]}{G(x'\beta + \gamma z^*)},$$

for all $z^* \in \mathcal{Z}^*$. The first line of the display implies, though, that $\min\{m(0, z^*), 1\}$ is always $m(0, z^*)$. Therefore,

$$m(1, z^*) = -\frac{m(0, z^*)[1 - G(x'\beta + \gamma z^*)]}{G(x'\beta + \gamma z^*)}.$$

From this formula, it further follows that, given this class of m's and an outcome probability measure μ (with generic measurability and integrability assumed), that the value of $m(0, z^*)$, when taken from the interval above with a fixed $m(1, \varnothing)$, enters the integral of m in an entirely linear manner: apart from itself, it only acts as a scaling factor on $m(1, z^*)$.

Accordingly, I can additionally focus only on the extreme points for $m(0, z^*)$, namely

$$m(0, z^*) = -\frac{m(1, \varnothing)G(x'\beta + \gamma z^*)}{1 - G(x'\beta + \gamma z^*)} \quad \text{or} \quad m(0, z^*) = 1.$$

If the integral is non-negative for both of these values of $m(0, z^*)$, then it should also be non-negative for any intermediate $m(0, z^*)$.

Step 5: Returning to $m(1, \varnothing)$

Finally, from Step 2, I have the following interval constraint on $m(1, \varnothing)$ from (10) and $m(1, \varnothing) \leq 0$:

$$m(1, \varnothing) \in \left[-\frac{1 - \overline{G}(x, \theta)}{\overline{G}(x, \theta)}, 0 \right].$$

I ensured in Step 4 that $m(0, z^*)$ was set to always satisfy (12).

Inspecting the last two displays from Step 4, it is evident that for any $m(1, \varnothing)$ chosen in the above interval, $m(0, z^*)$ and $m(1, z^*)$ are either unaffected by $m(1, \varnothing)$, or it acts as a common scaling factor on both of them. In other words, much like $m(0, z^*)$ above, $m(1, \varnothing)$ always enters linearly, and as such I can test non-negativity only at the endpoints

$$m(1, \varnothing) = -\frac{1 - \overline{G}(x, \theta)}{\overline{G}(x, \theta)} \quad \text{or} \quad m(1, \varnothing) = 0.$$

Step 6: Summary and measurability

Letting $q_\varnothing \in \{0, 1\}$ control the endpoint selected for $m(1, \varnothing)$ and $q : \mathcal{Z}^* \to \{0, 1\}$ the endpoints selected for $m(0, z^*)$, I arrive at the collection of $m_{q_\varnothing, q}$ functions in $\mathcal{D}^*(w, \theta)$.

As a technical condition, the resulting functions $m_{q_\varnothing, q}$ must be measurable. For this, it is necessary and sufficient that $q : \mathcal{Z}^* \to \{0, 1\}$ is measurable. If q is not measurable, then $q^{-1}(\{1\})$ is non-measurable with respect to the Borel σ-algebra on $\mathcal{Z}^*$. Then, the cross section of $m^{-1}_{q_\varnothing, q}(\{1\})$ at $y = 0$ is

$$\{\varnothing\} \cup \{z^* \in \mathcal{Z}^* \mid G(x'\beta + \gamma z^*) = \overline{G}(x, \theta)\} \cup q^{-1}(\{1\}).$$

As the first two sets on the RHS are measurable, the union is not, which means that $m_{q_\varnothing,q}$ is not a measurable function.

The $\widetilde{m}_{q_\varnothing,q}$ functions arise from considering the case $m(1,\varnothing) = 1$, $m(0,\varnothing) \leq 0$. In that case, the problem is identical to the one that I have considered so far, except with y swapped with $1-y$ and G with $1-G$. The resulting $\widetilde{m}_{q_\varnothing,q}$ functions then emerge as a mirror image of the ones that have already been constructed, using the same arguments.

Validation:

It remains to verify that the proposed collection $\mathcal{D}^*(x,\theta)$ is strongly core-determining. Let $m \in \mathcal{D}(x,\theta)$, and let μ be a probability measure on $\{0,1\} \times \mathcal{Z}$. Suppose that $\mathcal{D}^*(w,\theta) \cup \{m\} \subseteq \mathcal{L}^1(\mu)$ and that $\int m^* d\mu \geq 0$, for all $m^* \in \mathcal{D}^*(w,\theta)$. The goal is to show $\int m d\mu \geq 0$. The key property that I will implicitly invoke is that, because $m \in \mathcal{D}(x,\theta)$, it must satisfy (10) through (13). Observations from the prior analysis therefore apply.

If m is non-negative, then its expectation is immediately non-negative. Accordingly, suppose instead that m takes on a negative value. As mentioned in Step 1, it must therefore be the case that either $m(0,\varnothing) > 0$ or $m(1,\varnothing) > 0$, or both. Assume for concreteness that the first is true, while $m(1,\varnothing) \leq 0$. If $m(1,\varnothing)$ were also positive, then by Step 1 I can decompose m into two separate functions, one for each case. The second case would be essentially the same as the first, so I do not discuss it to avoid repetition.

By re-scaling, I may assume $m(0,\varnothing) = 1$. Hold fixed for now the values $m(1,\varnothing)$ and $m(0,z^*)$ for all $z^* \in \mathcal{Z}^*$. Then, as $m \in \mathcal{D}(x,\theta)$, it must point-wise dominate the function which replaces the current value of $m(1,z^*)$, for all $z^* \in \mathcal{Z}^*$, by

$$-\min\{m(0,z^*),1\} \times \frac{1-G(x'\beta+\gamma z^*)}{G(x'\beta+\gamma z^*)},$$

due to the lower bound found in Step 3. As $m(0,z^*)$ is measurable in z^*, the lower bound function remains measurable. It also stays integrable: the change increases the negative part of the integral, but there exists a function in $\mathcal{D}^*(x,\theta)$ which sets $m(1,z^*)$ to $-\frac{1-G}{G}$ for

all $z^* \in \mathcal{Z}^*$, while being μ-integrable by assumption. Finally, the dual cone non-negativity conditions still hold for the lower bound function by construction. Therefore, I have a function in $\mathcal{D}(x, \theta)$ which is point-wise dominated by m; it suffices to verify non-negativity for the alternate function instead. As such, I proceed as if m is the adjusted function.

Now, I turn to the values for $m(0, z^*)$. Consider the set of z^*'s for which $m(0, z^*) > 1$. This is a measurable subset of $\mathcal{Z}^*$, as it is the $y = 0$ cross-section of $m^{-1}((1, \infty))$. Accordingly, by similar arguments as in the last paragraph, I can adjust these values downwards to $m(0, z^*) = 1$ without affecting measurability, integrability, or non-negativity. (Note that this does not change $m(1, z^*)$.) Consequently, by the dual cone non-negativity lower bound on $m(0, z^*)$,

$$m(0, z^*) \in \left[-\frac{m(1, \varnothing) G(x'\beta + \gamma z^*)}{1 - G(x'\beta + \gamma z^*)}, 1 \right], \quad \forall z^* \in \mathcal{Z}^*,$$

which means that

$$m(1, z^*) = -\frac{m(0, z^*)[1 - G(x'\beta + \gamma z^*)]}{G(x'\beta + \gamma z^*)}$$

for all $z^* \in \mathcal{Z}^*$. (Because $m(1, \varnothing) \leq 0$, $m(0, z^*) \geq 0$.)

To construct an additional adjustment, I now scale $m(0, z^*)$ and $m(1, z^*)$ based on the conditional expectation of $m(\cdot, z^*)$. Formally, consider the function $m_{0,q} \in \mathcal{D}^*(x, \theta)$ with $q = 1$, which was defined as

$$m_{0,q}(y, z) \begin{cases} 1 & \text{if } z = \varnothing, y = 0, \\ 0 & \text{if } z = \varnothing, y = 1, \\ 1 & \text{if } z \neq \varnothing, y = 0, \\ -\frac{1 - G(x'\beta + \gamma z)}{G(x'\beta + \gamma z)} & \text{otherwise.} \end{cases}$$

By assumption, this function is μ-integrable. The conditional expectation

$$\mathbb{E}_\mu[m_{0,q}(Y, Z)|Z]$$

is thus well-defined and finite μ-a.s. Now, let $\mathcal{Z}^*_-$ denote the subset of $\mathcal{Z}^*$ (ignoring the null set) for which the conditional expectation is negative, which is measurable. Then, I define the function $\widetilde{q} : \mathcal{Z}^* \to \{0, 1\}$ by $\widetilde{q}(z^*) = \mathbf{1}\{z^* \in \mathcal{Z}^*_-\}$. In my construction of $\mathcal{D}^*(x, \theta)$, the q index functions maximize or minimize the weights at different z^*'s. Understood in this way, $\widetilde{q}$ maximizes weights at which the conditional expectation of m is negative. The idea, then, is to replace $m(0, z^*)$ with its upper bound, namely 1, if the conditional expectation is negative, and to replace it with its lower bound, $-m(1, \varnothing)G/(1 - G)$, if it is non-negative. Values for $m(1, z^*)$ also change simultaneously, based on the previous formula. By monotonicity and iterated expectations, the adjusted function then has an unconditional μ-integral weakly less than the un-adjusted m. Measurability and integrability of the adjustment follow from $\mathcal{Z}^*_-$ being measurable.

I have thus arrived at a lower bound function (or, with lower bound integral), presently defined as

$$m(y, z) = \begin{cases} 1 & \text{if } z = \varnothing, y = 0 \\ m(1, \varnothing) & \text{if } z = \varnothing, y = 1 \\ 1 & \text{if } z \in \mathcal{Z}^*_-, y = 0 \\ -\frac{1-G(x'\beta+\gamma z)}{G(x'\beta+\gamma z)} & \text{if } z \in \mathcal{Z}^*_-, y = 1, \\ -\frac{m(1,\varnothing)G(x'\beta+\gamma z)}{1-G(x'\beta+\gamma z)} & \text{if } z \notin \mathcal{Z}^*_-, y = 0, \\ m(1, \varnothing) & \text{otherwise.} \end{cases}$$

By $m(1, \varnothing)$, I just mean the original value, which I have kept fixed up to this point. The last step, then, is to observe that $m(1, \varnothing)$ must belong to the interval

$$\left[-\frac{1 - \overline{G}(x, \theta)}{\overline{G}(x, \theta)}, 0 \right]$$

as discussed in Step 2; this is independent of all adjustments so far. As $m(1, \varnothing)$ enters $\int m d\mu$ in a linear way, by the familiar arguments a lower bound integral is provided by one of the

two adjusted functions which replace $m(1, \varnothing)$ with either $-\frac{1-\overline{G}}{\overline{G}}$ or 0, assuming that they are integrable. However, these two functions are just $m_{1,\widetilde{q}}$ and $m_{0,\widetilde{q}}$, which belong to $\mathcal{D}^*(x,\theta)$ as $\widetilde{q}$ is measurable. If both of their integrals are non-negative, then $\int m d\mu \geq 0$. This verifies the strong core-determining property and therefore concludes the proof. □

A.5 Theorem 5.1

Proof. Fix $(x,\theta) \in \mathcal{X} \times \Theta$, and assume that $\mathcal{Z}^*$ is bounded. If $\gamma = 0$, then the claim becomes trivial, as the incomplete observability has no effect. Accordingly, I henceforth suppose $\gamma \neq 0$. As in the proof of Theorem 4.1, I generally suppress dependence on (x,θ) to reduce notation.

As discussed in the main text, by the finite mixture heterogeneity in Assumption 5.1 and the results in Winkler (1988), the non-negativity dual cone constraints on a measurable and model-integrable m reduce to

$$m(x'\beta + \gamma z^*, z) \geq 0,$$

$$\frac{c_+}{-c_- + c_+} m(x'\beta + \gamma z^* + c_-, z_-) + \frac{-c_-}{-c_- + c_+} m(x'\beta + \gamma z^* + c_+, z_+) \geq 0,$$

for all $c_- < 0$, $c_+ > 0$, $z^* \in \mathcal{Z}^*$, and $z, z_-, z_+ \in \mathcal{Z}$ which contain z^* (abusing notation). Re-arranging the second display leads to

$$\sup_{c_-<0} \frac{m(x'\beta + \gamma z^* + c_-, z_-)}{c_-} \leq \inf_{c_+>0} \frac{m(x'\beta + \gamma z^* + c_+, z_+)}{c_+}, \tag{14}$$

which I make use of throughout.

Now, let $m : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$ be an element of the dual cone $\mathcal{D}(x,\theta)$. Per the definition of a strong-core determining class, let μ be a probability measure on $\mathbb{R} \times \mathcal{Z}$ with $\mathcal{D}^*(x,\theta) \cup \{m\} \subseteq \mathcal{L}^1(\mu)$. To prove the theorem, I will show that m must point-wise dominate a function within the class m_r, $r \in [-1,1]^{\mathcal{Z}^*} = \mathcal{R}$. Then, I handle two separate cases. First, for the case where

r is measurable, I show that $m_r \in \mathcal{D}(x, \theta)$. This explains one of the sets in the union which defines $\mathcal{D}^*(x, \theta)$. Second, for the case where r is non-measurable, I use the notion of minimal measurable majorants to surmount the difficulty. This explains the other set in the union.

Proceeding into the proof, I define the following two objects for all $z^* \in \mathcal{Z}^*$:

$$r_{m,-}(z^*) := \sup_{c_- < 0} \frac{m(x'\beta + \gamma z^* + c_-, z^*)}{c_-},$$

$$r_{m,+}(z^*) := \inf_{c_+ > 0} \frac{m(x'\beta + \gamma z^* + c_+, z^*)}{c_+}.$$

By condition (14) with $z_- = z_+ = z^*$, $r_{m,-}(z^*) \leq r_{m,+}(z^*)$ for all $z^* \in \mathcal{Z}^*$. Now, for generic z^*, the above definitions imply that

$$m(y, z^*) \geq r_{m,-}(z^*)[y - x'\beta - \gamma z^*]$$

for all $y < x'\beta + \gamma z^*$, as well as

$$m(y, z^*) \geq r_{m,+}(z^*)[y - x'\beta - \gamma z^*]$$

for all $y \geq x'\beta + \gamma z^*$, where I add the equality part via incorporating the one-point distribution condition. Furthermore, note that setting $m(y, z^*)$ exactly equal to these lower bounds does not affect any other non-negativity constraints, as $m(\cdot, z^*)$ effectively enters those conditions only through $r_{m,-}(z^*)$ and $r_{m,+}(z^*)$. As such, m point-wise dominates its modified version where the above two displays hold with equality for all $y \in \mathbb{R}$ and $z^* \in \mathcal{Z}^*$, and the latter also still satisfies (14). Accordingly, I treat m as if these equalities do hold for it, to avoid introducing a new function. In other words:

$$m(y, z^*) = \max\{r_{m,-}(z^*)[y - x'\beta - \gamma z^*], r_{m,+}(z^*)[y - x'\beta - \gamma z^*]\}.$$

Now, as an intermediate step, I establish a lower bound for $m(y, z)$, where $z \in \mathcal{K}$ and

$y \in \mathbb{R}$. Fix a generic such (y, z). Then, there are three possible cases for any $z^* \in z$. As the first case, suppose that z^* satisfies $y < x'\beta + \gamma z^*$. Then, by (14) with $z_- = z$ and $z_+ = z^*$, it must be that

$$\frac{m(y,z)}{y - x'\beta - \gamma z^*} \leq r_{m,+}(z^*) \quad \Rightarrow \quad m(y,z) \geq r_{m,+}(z^*)[y - x'\beta - \gamma z^*].$$

Alternatively, if $y > x'\beta + \gamma z^*$, plugging in $z_- = z^*$ and $z_+ = z$ yields

$$r_{m,-}(z^*) \leq \frac{m(y,z)}{y - x'\beta - \gamma z^*} \quad \Rightarrow \quad m(y,z) \geq r_{m,-}(z^*)[y - x'\beta - \gamma z^*].$$

Finally, if $y = x'\beta + \gamma z^*$ for some z^*, then $m(y,z)$ must be non-negative. Putting these scenarios together results in the following lower bound on $m(y,z)$:

$$\max\left\{\sup_{z^* \in z : y < x'\beta + \gamma z^*} r_{m,+}(z^*)[y - x'\beta - \gamma z^*], \sup_{z^* \in z : y \geq x'\beta + \gamma z^*} r_{m,-}(z^*)[y - x'\beta - \gamma z^*]\right\}.$$

One implication of this bound is that $r_{m,-}$ and $r_{m,+}$ must be uniformly bounded. Otherwise, the bound is infinite for some $y \in \mathbb{R}$ and $z = \mathcal{Z}^*$, using the boundedness of $\mathcal{Z}^*$, which contradicts m being real-valued. Therefore, by re-scaling I can assume that $|\max\{r_{m,-}, r_{m,+}\}|$ has a supremum norm of one. However, at this point I am not yet able to modify $m(y,z)$ to take the lower bound everywhere, as it only ensures $m(y,z)$ is greater than the minimum of $r_{m,-}(z^*)[y - x'\beta - \gamma z^*]$ and $r_{m,+}(z^*)[y - x'\beta - \gamma z^*]$, rather than the maximum.

Instead, I prove that for any $z^* \in \mathcal{Z}^*$ with $r_{m,-}(z^*) < r_{m,+}(z^*)$, I can construct a modification of m on the values of $m(\cdot, z^*)$, denoted $\widetilde{m}$, with $r_{m,-}(z^*) \leq r_{\widetilde{m},-}(z^*) = r_{\widetilde{m},+}(z^*) \leq r_{m,+}(z^*)$ while still satisfying the non-negativity constraints. (The values of $\widetilde{m}(y, z^*)$ are defined like before, relative to $r_{\widetilde{m},.}$.) By inspecting the previous formula

$$m(y, z^*) = \max\{r_{m,-}(z^*)[y - x'\beta - \gamma z^*], r_{m,+}(z^*)[y - x'\beta - \gamma z^*]\},$$

it is evident that such an $\widetilde{m}$ would satisfy $\widetilde{m}(y, z^*) \leq m(y, z^*)$ for all $y \in \mathbb{R}$.

To generate the $\widetilde{m}$ function, first define it as equal to m, then set $r_{\widetilde{m},-}(z^*)$ as the largest value which does not induce $\widetilde{m}$ to violate (14); this is well-defined in $[r_{m,-}(z^*), r_{m,+}(z^*)]$ because $r_{m,-}(z^*)$ is valid and $r_{m,+}(z^*)$ is an upper bound. Moreover, it is evident that the following formula holds, otherwise a larger value could be selected:

$$r_{\widetilde{m},-}(z^*) = \inf_{(y,z)\in\mathbb{R}\times\mathcal{Z}: z^*\in z, y > x'\beta+\gamma z^*} \frac{m(y,z)}{y - x'\beta - \gamma z^*}. \tag{15}$$

This means that for any $\epsilon > 0$, I can find a relevant (y_ϵ, z_ϵ) such that

$$\frac{m(y_\epsilon, z_\epsilon)}{y_\epsilon - x'\beta - \gamma z^*} \leq r_{\widetilde{m},-}(z^*) + \epsilon.$$

Now, suppose that additionally changing $r_{\widetilde{m},+}(z^*)$ from $r_{m,+}(z^*)$ to $r_{\widetilde{m},-}(z^*)$ produces a non-negativity violation. By the lower bound for $m(y, z)$ derived previously, there thus exists $(\widehat{y}, \widehat{z}) \in \mathbb{R} \times \mathcal{K}$ with $z^* \in \widehat{z}$ and $\widehat{y} < x'\beta + \gamma z^*$ such that

$$m(\widehat{y}, \widehat{z}) < r_{\widetilde{m},-}(z^*)[\widehat{y} - x'\beta - \gamma z^*] \quad \Rightarrow \quad \frac{m(\widehat{y}, \widehat{z})}{\widehat{y} - x'\beta - \gamma z^*} > r_{\widetilde{m},-}(z^*).$$

But this means that, for ϵ small enough,

$$\frac{m(\widehat{y}, \widehat{z})}{\widehat{y} - x'\beta - \gamma z^*} > \frac{m(y_\epsilon, z_\epsilon)}{y_\epsilon - x'\beta - \gamma z^*},$$

where $z^* \in \widehat{z}$, $z^* \in z_\epsilon$, $\widehat{y} < x'\beta + \gamma z^*$, and $y_\epsilon > x'\beta + \gamma z^*$. This implies that m violates constraint (14) with $z_- = \widehat{z}$ and $z_+ = z_\epsilon$, which is a contradiction. Therefore, the proposed function modification $\widetilde{m}$, with $r_{\widetilde{m},-}(z^*) = r_{\widetilde{m},+}(z^*)$, continues to satisfy the non-negativity conditions, while still being point-wise dominated by m.

By the previous paragraphs, I may assume that $r_{m,-}(z^*) = r_{m,+}(z^*)$ for all $z^* \in \mathcal{Z}^*$. Collect these values in $r : \mathcal{Z}^* \to [-1, 1]$. (The previous step may have made the supremum

norm of r strictly less than one. If so, I re-scale again.) I return now to considering $m(y,z)$ for $z \in \mathcal{K}$; recall that I have not modified these values so far. With the new $r(z^*)$, the earlier lower bound on $m(y,z)$, $z \in \mathcal{K}$, simplifies to

$$\sup_{z^* \in z} r(z^*)[y - x'\beta - \gamma z^*].$$

As m has continued to satisfy the non-negativity conditions through its progressive modifications, $m(y,z)$ must be weakly greater than the above bound everywhere. It also turns out that setting $m(y,z)$ exactly equal to the bound does not violate (14) either. This is immediately seen by observing that the above display is now just $\sup_{z^* \in z} m(y,z^*)$.

With this, I arrive at a function of the form m_r, $r \in \mathcal{R}$, which is everywhere point-wise dominated by the original version of m. It now remains to verify some technical issues which revolve around whether r is measurable; from (15), this is not guaranteed.

To start, I consider the case where r is measurable and verify that m_r is in $\mathcal{D}(x,\theta)$. This handles the first component of $\mathcal{D}^*(x,\theta)$. Progressing through the dual cone requirements: On measurability, I observe that m_r is continuous in y for fixed $z \in \mathcal{Z}$, as $m(\cdot,z)$ is either affine or a point-wise supremum of affine functions which is finite everywhere, due to r and $\mathcal{Z}^*$ being bounded. Meanwhile, for fixed $y \in \mathbb{R}$, $m(y,\cdot)$ is measurable when r is measurable. As $\mathbb{R}$ and $\mathcal{Z}$ are Polish spaces, joint measurability of m_r follows from being a Caratheodory function, along with Lemma 4.50 in Aliprantis and Border (1999).

Regarding model integrability, $r \in [-1,1]$ implies the following uniform bound:

$$|m_r(y,z)| \leq \sup_{z^* \in \mathcal{Z}^*} |y - x'\beta - \gamma z^*|$$

for all $y \in \mathbb{R}$ and $z \in \mathcal{Z}$. With $\mathcal{Z}^*$ being bounded, $m_r(\cdot,z)$ thus lies, for all $z \in \mathcal{Z}$, beneath a non-negative function which is the point-wise maximum of affine functions in y with unit slope, i.e. akin to an absolute value function, and also above the non-positive version of that function. As the bounding functions have finite integrals as long as Y has a finite

(conditional) mean, m_r is integrable under the model for any $(z^*, F, s) \in \mathcal{A}$.

Finally, the non-negativity conditions are easy to verify. It follows that $m_r \in \mathcal{D}(x, \theta)$.

Now, to complete the proof, I drop the assumption that r is measurable; then, m_r may not be measurable. Recall that the original dual cone function was denoted m, and I showed $m_r \leq m$ everywhere. Now, let $m_r^* : \mathbb{R} \times \mathcal{Z} \to \mathbb{R}$, $m_r^* \in \mathcal{C}(m_r, \mu)$ be a minimal measurable majorant of m_r under μ; see van der Vaart and Wellner (2023). This function is measurable and satisfies: (i) $m_r^* \geq m_r$ and (ii) $m_r^* \leq f$ μ-a.s. for all measurable f with $f \geq m_r$ μ-a.s. Note that the majorant can generally values in the extended reals. To constrain this (and large values in general), I define here one last function using the upper bound from the previous display,

$$m^*(y, z) := \min \left\{ m_r^*(y, z), \sup_{z^* \in \mathcal{Z}^*} |y - x'\beta - \gamma z^*| \right\},$$

which is real-valued everywhere. As the upper bound function is measurable, m^* is still measurable; it also continues to satisfy properties (i) and (ii). Since I included m^* in $\mathcal{D}^*(x, \theta)$ via the second set in the union, it is μ-integrable by assumption. Property (ii) then implies that $\int m^* d\mu \leq \int m d\mu$, so non-negativity of the former implies non-negativity of the latter.

It thus remains to show that $m^* \in \mathcal{D}(x, \theta)$. I have already mentioned measurability. Model-integrability follows from the same comments as before. Finally, non-negativity follows because $m^* \geq m_r$, and the latter satisfies the relevant conditions by construction.

□

B Details of Numerical Exercises

B.1 Empirical Illustration

I work with a sample of $N = 4678$ women from the IPUMS Current Population Survey. The dependent variable is an indicator for employment. In addition to an intercept, I include the following covariates: four race dummies, age (re-scaled) and its square, an indicator for being

married, number of children, number of children less than five years old, years of education, and lagged employment. I select the sample so that it has no missing-ness.

After estimating a logit model on the sample, I re-simulate the dependent variables based on the estimates. Then, I induce artificial missingness in $Z^* = lagworking$ based on the following specified process:

$$P(Z \neq \varnothing|\cdot) = \begin{cases} 0.4 & \text{if } y = 0 \text{ and } educyears \geq 12, \\ 0.4 & \text{if } y = 1 \text{ and } educyears < 12, \\ 1 & \text{otherwise.} \end{cases}$$

For the simulation draw that I consider, lagged employment is observed for 3929 individuals, or a proportion of around 0.840.

To compute confidence intervals for regression coefficients using my moment inequalities, I apply the subvector inference procedure (i.e. a hypothesis test) proposed by Bei (2024) for unconditional moment inequality models, nested within an outer search using Kaido, Molinari, and Stoye (2019)'s EAM algorithm. In this case, the relevant subvector is just the target coefficient of interest, and the relevant search space is an interval of candidate values for the coefficient. These two features help with surmounting the curse of dimensionality in the number of total model parameters.

To transform the conditional moment inequalities into unconditional ones, I generate hypercube instruments from the covariates, based on only a single dimension each, according to the procedure described in Andrews and Shi (2013). So as not to have too many moment inequalities relative to the sample size, I use a subset of the indicators.

B.2 References for Code

All code is written in Julia (Bezanson, Edelman, Karpinski, and Shah (2017)). I generate plots using CairoMakie (Danisch and Krumbiegel (2021)). I perform optimization in

environments from three packages: JuMP (Lubin, Dowson, Dias Garcia, Huchette, Legat, and Vielma (2023)), Optim (Mogensen and Riseth (2018)), and NLopt (Johnson (2007)). In the first, I use the Clarabel solver for quadratic programs (Goulart and Chen (2024)); in the second, I use the L-BFGS algorithm (Liu and Nocedal (1989)); and in the last, I use the BOBYQA (Powell (2009)), ISRES (Runarsson and Yao (2005)), L-BFGS (Liu and Nocedal (1989)), MMA (Svanberg (2002)), and StoGO (Madsen, Zertchaninov, and Zilinskas (1998)) algorithms. They are relevant for different parts of my implementation of Bei (2024)'s procedure, as well as for obtaining maximum likelihood estimates.

I also make use of additional packages, such as Distributions (Besancon, Papamarkou, Anthoff, Arslan, Byrne, Lin, Pearson, and Schauer (2021)) and ForwardDiff (Revels, Lubin, and Papamarkou (2016)).

C Additional Topics

C.1 Choice of Topology for Model Probabilities

I elaborate here on the importance of the chosen topology when using closures in identification, which I mentioned in Section 3. I use the general setup and notation from that section, and I fix some $w \in \mathcal{W}$ and $\theta \in \Theta$, i.e. values for the conditioning variables and the parameters.

In Section 3.2, I discussed an alternative definition of the identified set for θ, based on taking a closure of model probabilities under a selected topology:

$$\left\{ \theta \in \Theta \;\middle|\; F_{Y|W=w} \in \overline{\mathcal{P}_Y^0(w,\theta)} \right\}.$$

At the time, I mentioned that a common choice for the topology, the one induced by the total

variation (TV) norm, can lead to a trivial identified set in the textbook regression model

$$Y = W'\theta + \varepsilon, \quad \mathbb{E}[\varepsilon|W] = 0,$$

although θ is consistently estimable if W has sufficient support. I now elaborate on this.

To streamline some technicalities, I assume that outcome distributions are absolutely continuous with respect to Lebesgue measure on the Borel sets of $\mathbb{R}$. I accommodate this within my framework by specifying the space of heterogeneity as the set of absolutely continuous mean-zero (conditional) distributions for the error, $\mathcal{F}_{0c}$. With admissible heterogeneity $\mathcal{P}_A^0$ as the set of finite mixtures on $\mathcal{F}_{0c}$, the set of model probabilities, $\mathcal{P}_Y^0(w,\theta)$, is a convex set. I can thus ignore the generic convexification generated by dual methods.

I now consider the TV norm closure of the set of model probabilities, $\overline{\mathcal{P}_Y^0(w,\theta)}^{TV}$. By Theorem 1 in Botosaru, Loh, and Muris (2024), it is possible to check membership in such a closed convex set, for a given outcome distribution, through the integrals of the bounded and measurable functions. By reversing my proof of Proposition 3.1 with suitable modifications, one can equivalently check non-negativity for a dual cone of bounded moment functions. However, this dual cone is actually trivial for the regression model under consideration, which implies a failure of identification.

To obtain this characterization of the dual cone, I use some ideas which are implicit in Section 5. Suppose that a measurable function $m : \mathbb{R} \to \mathbb{R}$ is bounded and that it takes negative values everywhere on a set of positive Lebesgue measure, denoted $\mathcal{S}_-$; otherwise, m has no identifying power. Without loss of generality, I assume that the set of positive measure is contained in $(-\infty, w'\theta)$. Then, to generate a contradiction for m being in the dual cone, I consider a sequence of absolutely continuous error distributions, $F_{\varepsilon,k} \in \mathcal{F}_{0c}$, which assigns an increasing amount of probability to the set $\{e - w'\theta \mid e \in \mathcal{S}_-\} \subseteq (-\infty, 0)$. As the aggregate probability on this region approaches one, increasingly large positive values of m must exist in order to maintain non-negativity of $\int m(w'\theta + e)dF_{\varepsilon,k}(e)$. However, this

eventually becomes impossible because m is bounded.

The previous paragraph shows that a bounded and measurable function belongs to the dual cone iff it is non-negative Lebesgue-almost-everywhere. As the integrals of such functions are always non-negative for any absolutely continuous outcome distribution, these test functions can never reject any θ, and there is complete non-identification. In contrast, my definitions and approach reproduce the standard moment condition, as I allow for model-integrable functions rather than just bounded ones. As discussed in Section 3, the larger class of functions generates a finer topology and thus a "tighter" sense of identification. In this specific setting, the discrepancy in the identified set is especially large.

As one explanation for this, recall that the TV norm closure is motivated by issues of statistical inference. Here, however, the relevant statistical test ultimately reduces to testing for a mean over a large set of distributions. By the classic Bahadur and Savage (1956) result, there does not exist a non-trivial test for such a hypothesis without further assumptions.

At this point, one response is to just impose the usual restrictions which are used to escape the Bahadur and Savage (1956) result, such as finiteness of higher moments or bounded support. Proceeding in this way, however, only further melds identification and inference. By extracting additional information from the model, I try to keep these issues distinct.

C.2 Necessity of Moment Equalities for Point-Identification

Here, I present one idea from my approach which is independent of particular models: It can help formalize the common wisdom that point-identification fails when no moment conditions exist. As such, my framework may also be useful for deriving negative results.

The idea goes as follows. Suppose that θ_0 is the true value in a particular model and that one views Θ_I^+ as equivalent to the identified set, as I have advanced. For simplicity, also suppose that $\mathcal{Y}$ is finite; this avoid technical complications in infinite-dimensional spaces. Endow $\mathbb{R}^{\mathcal{Y}}$ with a standard norm, ex. the supremum norm $\|\cdot\|_\infty$. Then, by the definition of

Θ_I^+, $\theta \in \Theta_I^+(F_{Y|W=w}|w)$ iff the following program has a value of zero:

$$Q(\theta) := \min_{m \in \mathcal{D}(w,\theta): \|m\|_\infty \leq 1} \sum_{y \in \mathcal{Y}} m(y) F_{Y|W=w}(\{y\}),$$

where I have made Q bounded by restricting the dual cone to the closed unit ball.

By assumption, $\theta_0 \in \Theta_I^+(F_{Y|w=w}|w)$, i.e. $Q(\theta_0) = 0$. One minimizer of the Q-program at θ_0 is thus $m = 0$. Now, suppose that there are no moment equality conditions at (w, θ_0): For all non-zero $m \in \mathcal{D}(w, \theta_0)$, there exists some $a_m \in \mathcal{A}$ such that $\int m d\mu^0_{w,a_m,\theta_0} > 0$. If $F_{Y|W=w}$ is generated by rich heterogeneity, it is then plausible that $m = 0$ is the unique minimizer at θ_0. Consequently, if the model is continuous in an appropriate sense, it seems intuitive that $m = 0$ should remain a minimizer for θ's close to θ_0, in which case point-identification fails.

Lemma C.1 below is one formalization of this outline. For any θ with $Q(\theta) < 0$, all minimizers of the Q-program must be at the boundary of the closed unit ball. Meanwhile, under the lemma's conditions, the Maximum Theorem implies that the set of minimizers, as a correspondence in θ, is upper hemicontinuous. This upper hemicontinuity implies that $Q(\theta) = 0$ for all θ in a neighborhood of θ_0; otherwise, the minimizer correspondence would "implode" from the boundary to $\{0\}$ at θ_0. Therefore, the neighborhood is also in Θ_I^+.

Lemma C.1. *Consider the setting defined by Assumptions 3.1 and 3.2, with $\mathcal{Y}$ finite and Θ a metric space. Fix $(w, \theta_0) \in \mathcal{W} \times \Theta$ and $\mu_0 \in \mathcal{M}_1(\mathcal{Y})$. Suppose that the following hold:*

- *(CS) $\theta_0 \in \Theta_I^+(\mu_0|w)$,*
- *(C) the correspondence*

$$\Gamma : \Theta \rightrightarrows \mathbb{R}^{\mathcal{Y}}, \quad \Gamma(\theta) := \{m \in \mathcal{D}(w, \theta) \mid \|m\|_\infty \leq 1\}$$

  is continuous at θ_0, and

- *(NME) $m \in \mathcal{D}(w, \theta_0)$ and $\sum_{y \in \mathcal{Y}} m(y)\mu_0(\{y\}) = 0$ jointly imply $m = 0$.*

Then $\Theta_I^+(\mu_0|w)$ contains an open neighborhood around θ_0.

Assumption (CS) states that the model is correctly specified, (C) is a model continuity condition, and (NME) says that there are no moment equality conditions at θ_0, with μ_0 spanning heterogeneity in some way.

More generally, I view the dual cone approach as a potentially unified way of understanding existing partial identification results. For example, see Corollary 3 in Bonhomme, Dano, and Graham (2023), which concerns a fixed effects panel logit model with a predetermined covariate. There are no moment equality conditions in that setting. To obtain a failure of point-identification, the authors assume "interior" heterogeneity, alongside model parameters being at a "regular point." I interpret these as concrete primitive conditions which can give rise to counterparts of (NME) and (C) from Lemma C.1.

C.2.1 Proof of Lemma C.1

Proof. Fix $(w, \theta_0) \in \mathcal{W} \times \Theta$ and $\mu_0 \in \mathcal{M}_1(\mathcal{Y})$. Suppose that the conditions in the lemma hold, ex. $\mathcal{Y}$ is finite.

I define the function $Q : \Theta \to \mathbb{R}$ by

$$Q(\theta) = \min_{m \in \Gamma(w,\theta)} \sum_{y \in \mathcal{Y}} m(y)\mu_0(\{y\}).$$

The minimum is achieved because $\Gamma(w, \theta)$ is compact-valued and non-empty (it always contains zero), while the objective is continuous in m due to the finiteness of $\mathcal{Y}$. Based on this, I further define $\varphi : \Theta \rightrightarrows \mathbb{R}^{\mathcal{Y}}$ as the minimizer correspondence for Q, i.e. $\varphi(\theta)$ is the set of minimizers at θ. Then,

$$\theta \in \Theta_I^+(\mu_0|w) \quad \Leftrightarrow \quad Q(\theta) = 0 \quad \Leftrightarrow \quad 0 \in \varphi(\theta),$$

by the definition of Θ_I^+.

By assumption (C), Γ is continuous at θ_0, in addition to being compact-valued for all θ. Meanwhile, the objective, as mentioned, is continuous in m, with no dependence on θ. Together, these allow for an application of the Maximum Theorem, which yields that φ is upper hemicontinuous at θ_0.

This upper hemicontinuity is sufficient to establish the existence of the open neighborhood described in the lemma. Consider the open ball of radius $1/2$ around 0 in $\mathbb{R}^{\mathcal{Y}}$ endowed with the supremum norm, denoted $B_{\mathbb{R}^{\mathcal{Y}},1/2}(0)$. By the upper hemicontinuity of φ at θ_0 and the implication, from (CS) and (NME), that $\varphi(\theta_0) = \{0\}$, there exists $\epsilon > 0$ such that

$$\varphi(B_{\Theta,\epsilon}(\theta_0)) \subseteq B_{\mathbb{R}^{\mathcal{Y}},1/2}(0),$$

where $B_{\Theta,\epsilon}(\theta_0)$ is the open ball of radius ϵ around θ_0 in Θ.

Now, suppose toward a contradiction that there exists $\theta \in B_{\Theta,\epsilon}(\theta_0)$ with $0 \notin \varphi(\theta)$; that is, θ can be rejected from Θ_I^+. An immediate implication is that $Q(\theta) < 0$. Moreover, by the properties of the program which defines Q, it is evident that every element of $\varphi(\theta)$ must lie on the boundary of the closed unit ball in $\mathbb{R}^{\mathcal{Y}}$ (see the definition of Γ), because any minimizer must be scaled as much as possible. However, this is at odds with the previous display, which states that elements of the φ-image of $B_{\Theta,\epsilon}(\theta_0)$ have supremum norm bounded by $1/2$.

Therefore, $0 \in \varphi(\theta)$ for all $\theta \in B_{\Theta,\epsilon}(\theta_0)$, which means that $B_{\Theta,\epsilon}(\theta_0) \subseteq \Theta_I^+(\mu_0|w)$. $\square$