Mis-Classified, Binary, Endogenous Regressors: Identification and Inference

Francis J. DiTraglia[1] Camilo García-Jimeno[2,3]

[1]University of Pennsylvania

[2]Emory University

[3]NBER

October 11th, 2018

Binary Regressors

- Thank you for inviting me. Joint work with Camilo Garcia-Jimeno.
- Intro. 'metrics students learn that a valid IV serves double duty: correct for endogeneity and classical measurement error
- Classical measurement error is a special case: requires true value of regressor indep. of or at least uncorrelated with measurement error
- Applied work often involves endogenous binary regressor: smoker/non-smoker or union/non-union. Binary $\implies$ non-classical error. True 0 $\implies$ can only mis-measure *upwards* as 1; true 1 $\implies$ can only mis-measure *downwards* as 0. Error *negatively correlated* with truth.
- To accommodate this, consider *non-diff* error. Say more later, but roughly non-diff means *conditionally classical*: condition on truth and controls, remaining component of error unrelated to everything else.
- Today pose simple question: binary, endog. regressor subject to non-diff. error. Can valid IV correct for *both* measurement error and endog?

What is the effect of T^*?

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$$

- y – Outcome of interest
- T^* – Unobserved, endogenous binary regressor
- T – Observed, mis-measured binary surrogate for T^*
- $\mathbf{x}$ – Exogenous covariates
- z – Discrete (typically binary) instrumental variable

What is the effect of T^*?

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$$

- y – Outcome of interest
- T^* – Unobserved, endogenous binary regressor
- T – Observed, mis-measured binary surrogate for T^*
- $\mathbf{x}$ – Exogenous covariates
- z – Discrete (typically binary) instrumental variable

- Additively separable model, want to learn the causal effect of binary regressor T^* on y. Unfortunately T^* unobserved. Observe only mis-measured binary surrogate T. Moreover, T^* is endogenous, but we have a discrete IV z.
- Additive separability is a restriction: allows very general observed heterogeneity through $\mathbf{x}$, but restricts unobserved.
- Conditionally linear: WLOG since model add. sep. & and T^* binary.
- Focus on add. sep. but also mention implications for LATE model.

Using a discrete IV to learn about $\beta(\mathbf{x})$

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$$

Contributions of This Paper

1. Show that only existing point identification result for mis-classified, endogenous T^* is incorrect.
2. Sharp identified set for β under standard assumptions.
3. Point identification of β under slightly stronger assumptions.
4. Describe problem of weak identification in mis-classification models, develop identification-robust inference for β.

Using a discrete IV to learn about $\beta(\mathbf{x})$

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon$$

Contributions of This Paper

1. Show that only existing point identification result for mis-classified, endogenous T^* is incorrect.
2. Sharp identified set for β under standard assumptions.
3. Point identification of β under slightly stronger assumptions.
4. Describe problem of weak identification in mis-classification models, develop identification-robust inference for β.

- Outline main contributions.
- Many papers consider using IV to identify effect of exog. mis-measured binary regressor, but little work on endog. case. First: show only point identification result for this case incorrect: ident. is an open question.
- Next: use standard assumptions to derive the "sharp identified set" for β. This means *fully* exploit all information in the data and our assumptions to derive tightest possible bounds for β. If bounds contain a single point, β is point identified. Otherwise partially identified.
- Novel and informative bounds for β, but not point identified. Then consider slightly stronger assumptions that allow us to exploit additional features of the data and show that these suffice to point identify β.
- Next consider inference. Show that mis-classification models, suffer from potential weak identification. Propose procedure for robust inference.
- Now a motivating example. . .

Example: Smoking and Birthweight (SNAP Trial)

Coleman et al. (N Engl J Med, 2012)

RCT with pregnant smokers in England: half given nicotine patches, the rest given placebo patches. Some given nicotine fail to quit; some given placebo quit.

- y – Birthweight
- T^* – True smoking behavior
- T – Self-reported smoking behavior
- $\mathbf{x}$ – Mother characteristics
- z – Indicator of nicotine patch

Baseline Assumptions I – Model & Instrument

Additively Separable Model

$$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0$$

Valid & Relevant Instrument: $z \in \{0, 1\}$

- $\mathbb{P}(T^* = 1|\mathbf{x}, z = 1) \neq \mathbb{P}(T^* = 1|\mathbf{x}, z = 0)$
- $\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0$
- $0 < \mathbb{P}(z = 1|\mathbf{x}) < 1$

Baseline Assumptions I – Model & Instrument

Additively Separable Model

$y = c(\mathbf{x}) + \beta(\mathbf{x})T^* + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0$

Valid & Relevant Instrument: $z \in \{0, 1\}$

- $\mathbb{P}(T^* = 1|\mathbf{x}, z = 1) \neq \mathbb{P}(T^* = 1|\mathbf{x}, z = 0)$
- $\mathbb{E}[\varepsilon|\mathbf{x}, z] = 0$
- $0 < \mathbb{P}(z = 1|\mathbf{x}) < 1$

- This slide and the next detail what I will call the "baseline" assumptions, which I will maintain through the talk.
- The first part of the baseline assumptions concern the model and instrument: all that this slide says is that if T^* were observed, then the model would be identified: usual IV relevance and validity conditions for model with T^*.

Baseline Assumptions II – Measurement Error

Notation

- $\alpha_0(\mathbf{x}, z) \equiv \mathbb{P}\left(T = 1 \middle| T^* = 0, \mathbf{x}, z\right)$
- $\alpha_1(\mathbf{x}, z) \equiv \mathbb{P}\left(T = 0 \middle| T^* = 1, \mathbf{x}, z\right)$

Mis-classification unaffected by z

$$\alpha_0(\mathbf{x}, z) = \alpha_0(\mathbf{x}), \quad \alpha_1(\mathbf{x}, z) = \alpha_1(\mathbf{x})$$

Extent of Mis-classification

$\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$ (T is positively correlated with T^*)

Non-differential Mis-classification

$$\mathbb{E}[\varepsilon | \mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon | \mathbf{x}, z, T^*]$$

Baseline Assumptions II – Measurement Error

Notation

- $\alpha_0(\mathbf{x}, z) \equiv \mathbb{P}(T = 1 | T^* = 0, \mathbf{x}, z)$
- $\alpha_1(\mathbf{x}, z) \equiv \mathbb{P}(T = 0 | T^* = 1, \mathbf{x}, z)$

Mis-classification unaffected by z

$\alpha_0(\mathbf{x}, z) = \alpha_0(\mathbf{x}), \quad \alpha_1(\mathbf{x}, z) = \alpha_1(\mathbf{x})$

Extent of Mis-classification

$\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1 \quad$ (T is positively correlated with T^*)

Non-differential Mis-classification

$\mathbb{E}[\varepsilon | \mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon | \mathbf{x}, z, T^*]$

- Second part of the baseline assumps: meas. error. . First notation: mis-class. probs. α_0 and α_1. Two errors. *Upwards*: observe $T = 1$ when truth is $T^* = 1$; occurs with prob. α_0. *Downwards*: observe $T = 0$ when truth is $T^* = 1$; occurs with prob. α_1. Convention uses subscripts to indicate the value of *truth*: α_0 is mis-classification prob. when $T^* = 0$ ($\uparrow$) and α_1 when $T^* = 1$ ($\downarrow$). So far notation: now restrictions.

- 1st: given x the mis-class. rates don't depend on z. Restrictive, but hard to do without. How reasonable? Depends on **x** and specific setting. Plausible: smoking mothers didn't know if they had the real patch.

- 2nd: $Cor(T, T^*) > 0 \iff \alpha_0 + \alpha_1 < 1$. Mild. Say more in a few slides.

- 3rd: *non-diff* assumption. Stated in terms of epsilon, but what this really requires is conditional mean of Y doesn't depend on T given $(\mathbf{x}, z, T^*)$. Plausibility depends on the situation and the controls in **x**. Example of what this rules out: "returns to lying." E.g. $Y = \log(\text{wage})$, $T^* =$ true college dummy, and $T =$self-report of college. If employers can't perfectly observe credentials, there could be a *direct* effect of T on y even after controlling for T^*. Working on a paper based on this example with Arthur Lewbel.

Only Existing Result for Endogenous T^* is Incorrect

Mahajan (2006; Ecta) A2

$\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, T^*]$ + "Baseline" $\Rightarrow \beta(\mathbf{x})$ identified.

We Show:

Mahajan's assumptions imply that the instrument z is uncorrelated with T^* unless T^* is in fact *exogenous*.

Only Existing Result for Endogenous T^* is Incorrect

Mahajan (2006; Ecta) A2

$\mathbb{E}[\varepsilon|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon|\mathbf{x}, T^*]$ + "Baseline" $\Rightarrow \beta(\mathbf{x})$ identified.

We Show:

Mahajan's assumptions imply that the instrument z is uncorrelated with T^* unless T^* is in fact *exogenous*.

- The only existing result for the *endogenous* T^* appears in a paper by Mahajan. To be fair, this is *not* the main point of his paper, which primarily concerns the exogenous case. Mahajan argues that the baseline conditions plus a somewhat exotic-looking condition here implies that β is point identified. The purpose of this additional condition is to create a link with his earlier result for the exog. case.
- First contribution: we show that Mahajan's assumptions imply that z is *irrelevant*, uncorrelated with T^*, *unless* T^* is *exogenous*.
- Since Mahajan's argument for the endogenous T^* fails, identification is an open question.

"Weak" Bounds

First-Stage

$$p_k(\mathbf{x}) \equiv \mathbb{P}(T = 1|\mathbf{x}, z = k)$$

IV Estimand

$$\frac{\mathbb{E}[y|\mathbf{x}, z = 1] - \mathbb{E}[y|\mathbf{x}, z = 0]}{p_1(\mathbf{x}) - p_0(\mathbf{x})} = \frac{\beta(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}$$

Bounds for (α_0, α_1)

$\alpha_0(\mathbf{x}) \leq \min_k \{p_k(\mathbf{x})\}, \quad \alpha_1(\mathbf{x}) \leq \min_k \{1 - p_k(\mathbf{x})\}$ ▸ more

Bounds for β

$\beta(\mathbf{x})$ is between IV and Reduced form; same sign as IV. ▸ more

"Weak" Bounds

First-Stage

$$p_k(\mathbf{x}) \equiv \mathbb{P}(T = 1|\mathbf{x}, z = k)$$

IV Estimand

$$\frac{\mathbb{E}[y|\mathbf{x}, z = 1] - \mathbb{E}[y|\mathbf{x}, z = 0]}{p_1(\mathbf{x}) - p_0(\mathbf{x})} = \frac{\beta(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}$$

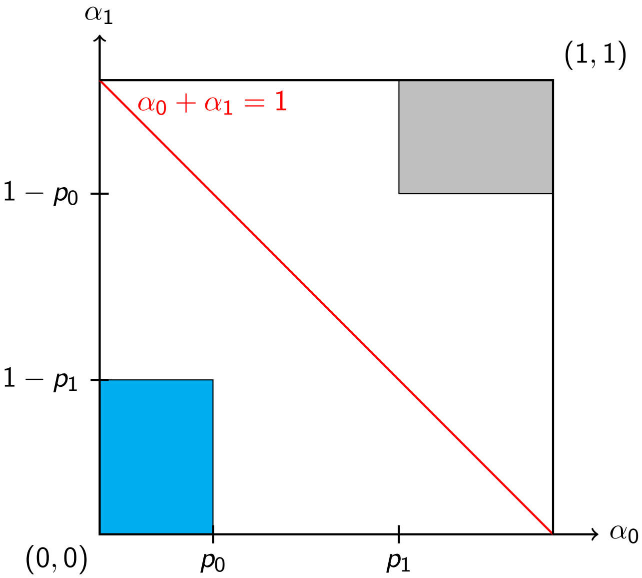
Bounds for (α_0, α_1)

$\alpha_0(\mathbf{x}) \leq \min_k \{p_k(\mathbf{x})\}, \quad \alpha_1(\mathbf{x}) \leq \min_k \{1 - p_k(\mathbf{x})\}$ more

Bounds for β

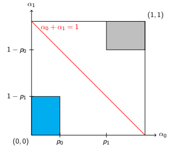
$\beta(\mathbf{x})$ is between IV and Reduced form; same sign as IV. more

- First constructive result: simple bounds for α_0, α_1 and β. Call these "weak" bounds since they don't fully exploit info. in the data & baseline assumptions.
- Before doing this, define some notation: first-stage probabilities p_k. Subscript indicates the value that z takes on: binary z gives p_0, p_1.
- Using this notation, what does IV estimate under baseline assumptions? Wald estimand *conditional* on $\mathbf{x}$. Measurement error *only* affects the denominator: instead of the first-stage probs. for *true* regressor T^* we have them for T. Simple algebra using law of total prob. and assumption that z doesn't affect mis-classification error rates shows that the constant of proportionality relating the unobserved *true* first-stage to the observed first stage is $1 - \alpha_0 - \alpha_1$. If $Cor(T^*, T) \neq 0$, denominator is non-zero. If $Cor(T^*, T) > 0$, IV has same sign as β but is *inflated*. Measurement error does *not* cause attenuation here. IV estimator corrects for endogeneity of T^* but not measurement error.
- Continuing to assume $Cor(T^*, T) > 0$, observed 1st-stage probs. bound α_0 and α_1, and we can combine these with the expression for the IV estimand see that β lies between IV and Reduced form with same sign as IV.



α_1
$(1,1)$
$\alpha_0 + \alpha_1 = 1$
$1 - p_0$
$1 - p_1$
$(0,0)$
p_0
p_1
α_0

Binary Regressors



- More about assumption $\alpha_0 + \alpha_1 < 1$. Suppress **x**. Fig. shows possible values of (α_0, α_1). Red line: $\alpha_0 + \alpha_1 = 0$ so $Cor(T, T^*) = 0$. Have to rule this out. Below red line $\alpha_0 + \alpha_1 < 1$ so $Cor(T, T^*) > 0$; above $Cor(T, T^*) < 0$. Bounds on prev. slide assume below the red line. If we relax this, still get bounds for α_0, α_1: shaded rectangles. Blue = bounds from prev slide: $\alpha_0 \leq \min\{p_0, p_1\}$ and $\alpha_1 \leq \{1 - p_0, 1 - p_1\}$. (In fig. $p_0 < p_1$). Gray means error so severe that $1 - T$ is a better predictor of T^* than T. So $\alpha_0 + \alpha_1 < 1$ just means rule out extreme error. Equiv. to assume IV and β have same sign.
- Weak bounds for $(\alpha_0, \alpha_1, \beta)$ simple and informative. Others have used related idea: Frazis & Loewenstein (2003) and Ura (forthcoming). But weak bounds don't use non-diff assump. Know that non-diff is powerful: point identifies effect of an exog T^*. Can we improve upon weak bounds for endog. T^*?
- To answer this, derive sharp identified set under baseline assumptions: new to the literature. Important even if our main concern is point identification: while we showed a flaw in Mahajan's proof, we did *not* show β not point identified.
- How to derive sharp set? Question: for what values of unknown params can we construct valid joint dist. for (y, T, T^*, z) compatible with observed joint for (y, T, z) under our assumptions? Factorize: joint for (T, T^*, z) & conditional for $y|T, T^*, z$. Turns out that weak bounds for (α_0, α_1) ensure valid joint for (T, T^*, z) so suffices to look at conditional: $y|T, T^*, z$.

Restrictions from Non-differential Mis-classification?

(Suppress **x** for simplicity)

Notation

- $r_{tk} \equiv \mathbb{P}(T^* = 1 | T = t, z = k)$
- z_k is shorthand for $z = k$

Iterated Expectations over T^*

$$\mathbb{E}(y|T=0,z_k) = (1-r_{0k})\mathbb{E}(y|T^*=0,\,T=0,z_k) + r_{0k}\mathbb{E}(y|T^*=1,\,T=0,z_k)$$

$$\mathbb{E}(y|T=1,z_k) = (1-r_{1k})\mathbb{E}(y|T^*=0,\,T=1,z_k) + r_{1k}\mathbb{E}(y|T^*=1,\,T=1,z_k)$$

Restrictions from Non-differential Mis-classification?

(Suppress **x** for simplicity)

Notation

- $r_{tk} \equiv \mathbb{P}(T^* = 1 | T = t, z = k)$
- z_k is shorthand for $z = k$

Adding Non-differential Assumption

$$\mathbb{E}(y|T = 0, z_k) = (1 - r_{0k})\mathbb{E}(y|T^* = 0, z_k) \qquad + r_{0k}\mathbb{E}(y|T^* = 1, z_k)$$

$$\mathbb{E}(y|T = 1, z_k) = (1 - r_{1k})\mathbb{E}(y|T^* = 0, z_k) \qquad + r_{1k}\mathbb{E}(y|T^* = 1, z_k)$$

$\boxed{\text{2 equations in 2 unknowns} \Rightarrow \text{solve for } \mathbb{E}(y|T^* = t^*, z = k) \text{ given } (r_{0k}, r_{1k}).}$

Restrictions from Non-differential Mis-classification?

(Suppress **x** for simplicity)

Notation

- $r_{tk} \equiv \mathbb{P}(T^* = 1 | T = t, z = k)$
- z_k is shorthand for $z = k$

Adding Non-differential Assumption

$$\mathbb{E}(y|T=0,z_k) = (1-r_{0k})\mathbb{E}(y|T^*=0,z_k) \qquad + r_{0k}\mathbb{E}(y|T^*=1,z_k)$$
$$\mathbb{E}(y|T=1,z_k) = (1-r_{1k})\mathbb{E}(y|T^*=0,z_k) \qquad + r_{1k}\mathbb{E}(y|T^*=1,z_k)$$

$\boxed{\text{2 equations in 2 unknowns} \Rightarrow \text{solve for } \mathbb{E}(y|T^*=t^*, z=k) \text{ given } (r_{0k}, r_{1k}).}$

- Suppress **x**. Study conditional dist of $y|T, T^*, z$. Unobserved but related to dist of $y|T, z$ via a mixture model. Mixing probs are r_{tk}. These depend on (α_0, α_1) and observables only. Shorthand: z_k denotes $z = k$.
- First look at means. For each value k that the IV takes on, there are two observed means $\mathbb{E}[y|T = (0,1), z_k]$ and four unobserved means $\mathbb{E}[y|T = (0,1), T^* = (0,1), z_k]$. But the non-diff assumption restricts the four unobserved means: we can *drop* T from the conditioning set after conditioning on T^*, z. Hence, only two unknown means: color-coded to show common unknowns across equations.
- Remember: r_{tk} is known given (α_0, α_1), so we see that the non-diff. assumption lets us solve for the two unknown means at any specified pair (α_0, α_1): we simply have two linear equations in two unknowns.

Restrictions from Non-differential Mis-classification?

Mixture Representation

$$F_{tk} = (1 - r_{tk})F_{tk}^0 + r_{tk}F_{tk}^1$$

$$F_{tk} \equiv y|(T = t, z = k)$$
$$F_{tk}^{t^*} \equiv y|(T^* = t^*, T = t, z = k)$$

Restrictions

- $\mathbb{E}(y|T^*, T, z) = \mathbb{E}(y|T^*, z)$ observable given (α_0, α_1)
- r_{tk} observable given (α_0, α_1)

Question

Given (α_0, α_1) can we always find (F_{tk}^0, F_{tk}^1) to satisfy the mixture model?

Restrictions from Non-differential Mis-classification?

Mixture Representation

$$F_{tk} = (1 - r_{tk})F_{tk}^0 + r_{tk}F_{tk}^1$$

$$F_{tk} \equiv y|(T = t, z = k)$$
$$F_{tk}^{t^*} \equiv y|(T^* = t^*, T = t, z = k)$$

Restrictions

- $\mathbb{E}(y|T^*, T, z) = \mathbb{E}(y|T^*, z)$ observable given (α_0, α_1)
- r_{tk} observable given (α_0, α_1)

Question

Given (α_0, α_1) can we always find (F_{tk}^0, F_{tk}^1) to satisfy the mixture model?

- Looked at means, now look at distributions. Observe F_{tk} the distribution of $y|T, z$. This is a mixture of two unobserved distributions: F_{tk}^0 and F_{tk}^1.
- Although (F_{tk}^0, F_{tk}^1) are unobserved, they're constrained. First, they need to "integrate" to F_{tk} which is observed. Second, the mixing probability r_{tk} is a *known* function of (α_0, α_1) given observables. Third, as we saw on the preceding slide, non-differential measurement error implies that the means of F_{tk}^0 and F_{tk}^1 are *known* functions of (α_0, α_1).
- Given these constraints, can we find valid distributions (F_{tk}^0, F_{tk}^1) to satisfy the mixture representation for *any pair* (α_0, α_1)? Or are there some values for the mis-classification probabilities that are incompatible with the mixture model?

Restrictions from Non-differential Mis-classification?

Equivalent Problem

Given a specified CDF F, for what values of p and μ do there exist valid CDFs (G, H) with $F = (1-p)G + pH$ and $\mu = \text{mean}(H)$?

Necessary and Sufficient Condition if F is Continuous

$$\underline{\mu}(F,p) \leq \ \mu \ \leq \overline{\mu}(F,p)$$

$$\underline{\mu}(F,p) \equiv \int_{-\infty}^{\infty} x \left[p^{-1} f(x) \mathbf{1}\{x < F^{-1}(p)\}\right] dx = \int_{-\infty}^{\infty} x \underline{h}(x)\, dx$$

$$\overline{\mu}(F,p) \equiv \int_{-\infty}^{\infty} x \left[p^{-1} f(x) \mathbf{1}\{x > F^{-1}(1-p)\}\right] dx = \int_{-\infty}^{\infty} x \bar{h}(x)\, dx$$

Restrictions from Non-differential Mis-classification?

Equivalent Problem

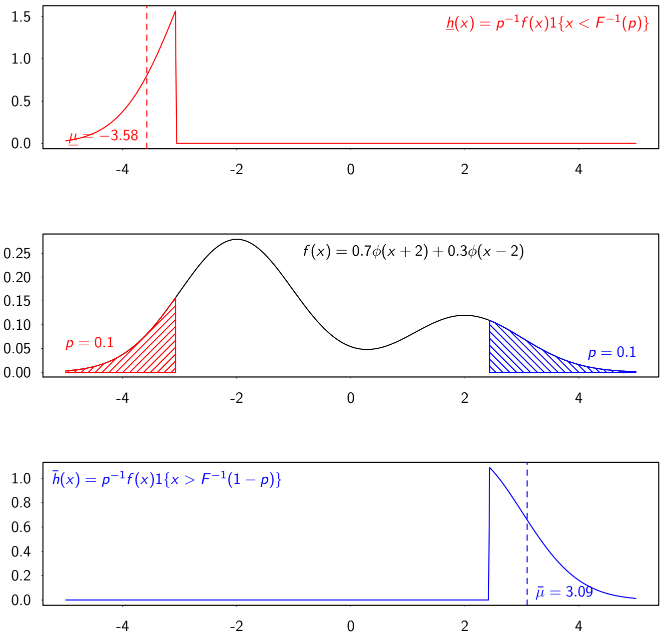
Given a specified CDF F, for what values of p and μ do there exist valid CDFs (G, H) with $F = (1-p)G + pH$ and $\mu = \text{mean}(H)$?

Necessary and Sufficient Condition if F is Continuous

$$\underline{\mu}(F,p) \leq \mu \leq \overline{\mu}(F,p)$$

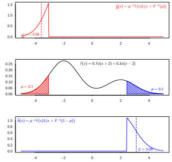
$$\underline{\mu}(F,p) \equiv \int_{-\infty}^{\infty} x\left[p^{-1}f(x)\mathbf{1}\{x < F^{-1}(p)\}\right] dx = \int_{-\infty}^{\infty} x\underline{h}(x)\, dx$$
$$\overline{\mu}(F,p) \equiv \int_{-\infty}^{\infty} x\left[p^{-1}f(x)\mathbf{1}\{x > F^{-1}(1-p)\}\right] dx = \int_{-\infty}^{\infty} x\overline{h}(x)\, dx$$

- To answer this question, we need to answer a more abstract question about mixture distributions. In particular, suppose that we observe a distribution F. Can we construct valid distributions (G, H) such that F ia s mixture of G and H in which H has mixing weight p and mean μ?
- To be clear: in this exercise F is fixed. The question is: if I postulate a mixing probability p and a mean μ for one of the mixture components, can this ever lead to a contradiction? Are we free to pick any pair (p, μ) or does the observed distribution F tie our hands?
- It turns out that if y is continuously distributed, one can derive relatively simple necessary and sufficient conditions using a first-order stochastic dominance argument.
- In particular: for any fixed (F, p) there is a lower bound $\underline{\mu}$ and an upper bound $\overline{\mu}$ within which the postulated mean μ *must* lie, for it to be possible to construct a valid mixture. These lower and upper bounds are in fact expectations taken with respect to densities constructed by *truncating* F.
- Rather than staring at these integrals, let's look at a simple example.

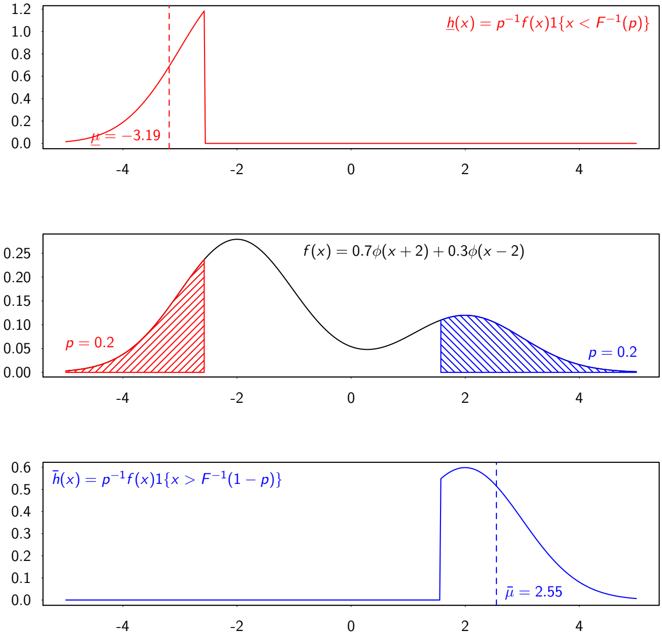


$\underline{h}(x) = p^{-1} f(x) 1\{x < F^{-1}(p)\}$
$\underline{\mu} = -3.58$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.1$
$p = 0.1$
$\bar{h}(x) = p^{-1} f(x) 1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 3.09$

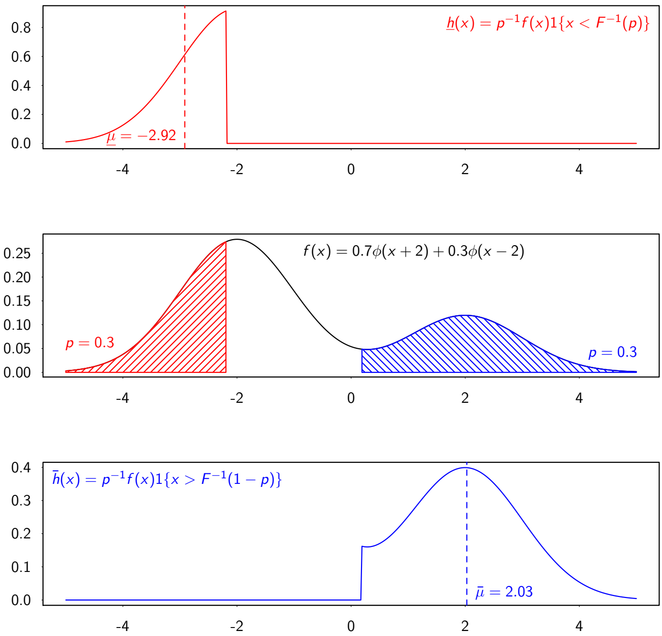
Binary Regressors



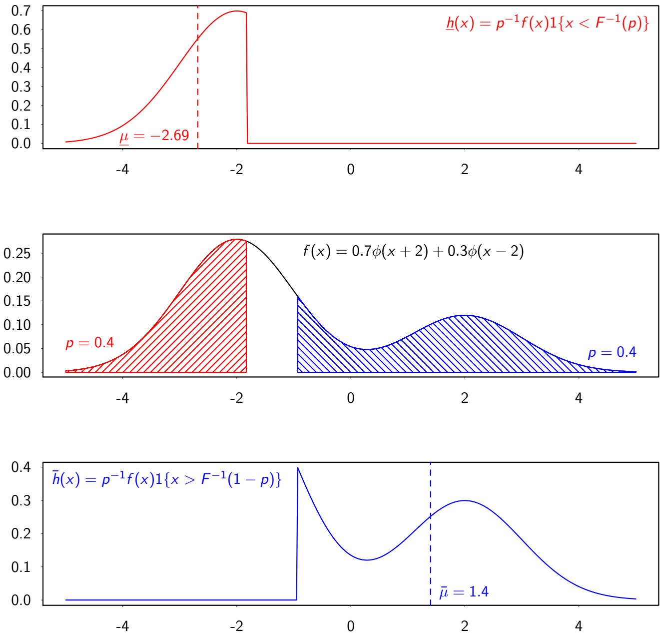
- This picture has three panels. The middle panel shows the observed distribution f. I have chosen a simple mixture of normals with variance equal to one: 70% of the weight is assigned to the one with a mean of -2 and 30% to the one with a mean of $+2$.
- The top panel depicts the "lower bound" density $\underline{h}$. This density takes its shape from the *lower tail* of f. In is simply f *truncated* to take on values below its pth quantile.
- The bottom panel depicts the "upper bound" density $\overline{h}$. This density takes its shape from the *upper tail* of f. It is simply f *trucated* to take on values above its $(1-p)$th quantile.
- For this particular choice of observed distribution f, the figure shows how a particular postulated value of p, in this instance 0.1, constrains μ: it is bounded below by $\underline{\mu} = -3.58$ and bounded above by $\overline{\mu} = 3.09$. This means that if $p = 0.1$, then μ must lie between -3.58 and 3.09 for it to be possible to construct a valid mixture that "integrates" to f. As we increase p, these bounds tighten, so we have less freedom in our choice of μ.

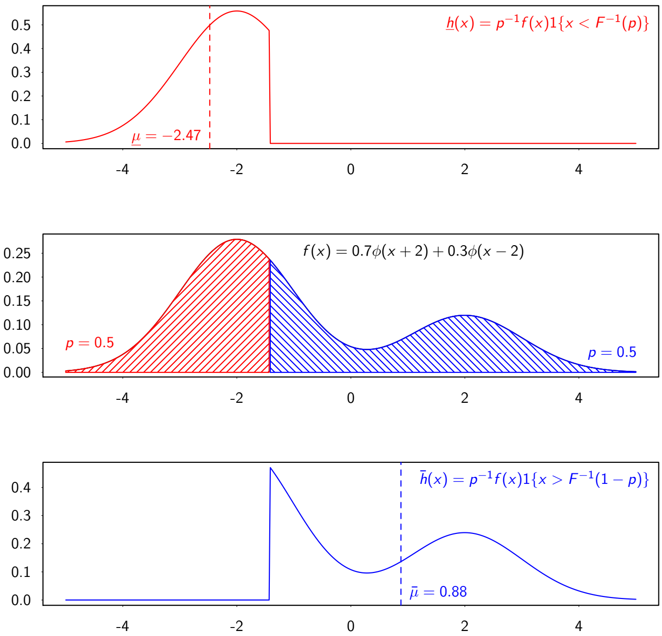


$\underline{h}(x) = p^{-1}f(x)1\{x < F^{-1}(p)\}$
$\underline{\mu} = -3.19$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.2$
$p = 0.2$
$\bar{h}(x) = p^{-1}f(x)1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 2.55$

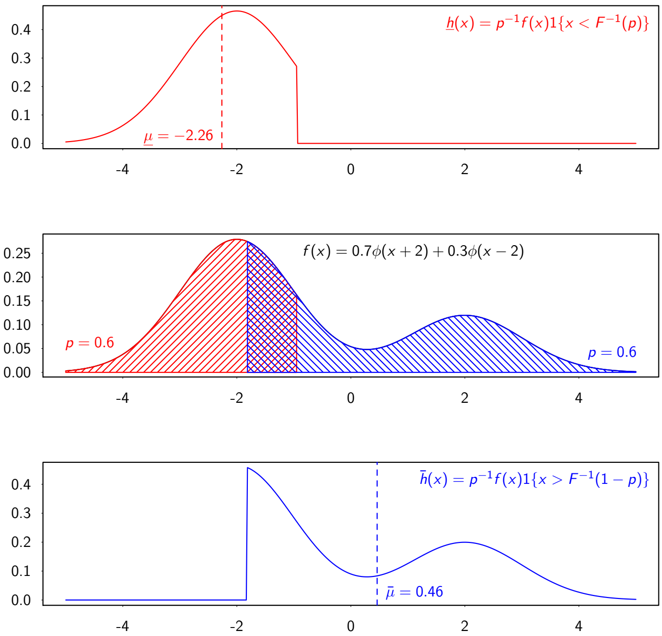


$\underline{h}(x) = p^{-1}f(x)1\{x < F^{-1}(p)\}$
$\underline{\mu} = -2.92$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.3$
$p = 0.3$
$\bar{h}(x) = p^{-1}f(x)1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 2.03$

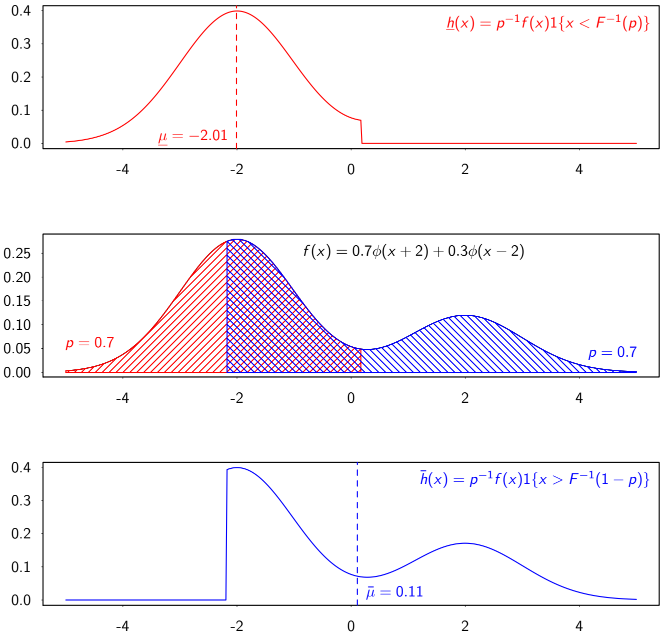

$\underline{h}(x) = p^{-1}f(x)1\{x < F^{-1}(p)\}$
$\underline{\mu} = -2.69$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.4$
$p = 0.4$
$\bar{h}(x) = p^{-1}f(x)1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 1.4$



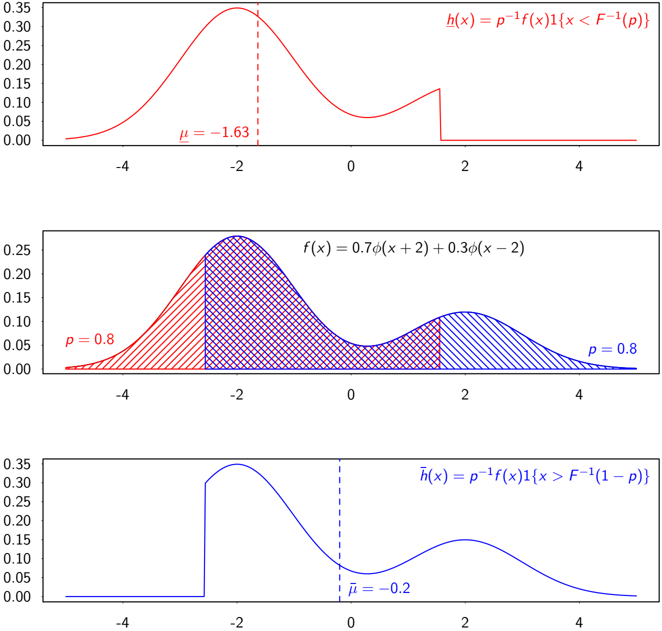
$\underline{h}(x) = p^{-1}f(x)1\{x < F^{-1}(p)\}$
$\underline{\mu} = -2.47$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.5$
$p = 0.5$
$\bar{h}(x) = p^{-1}f(x)1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 0.88$

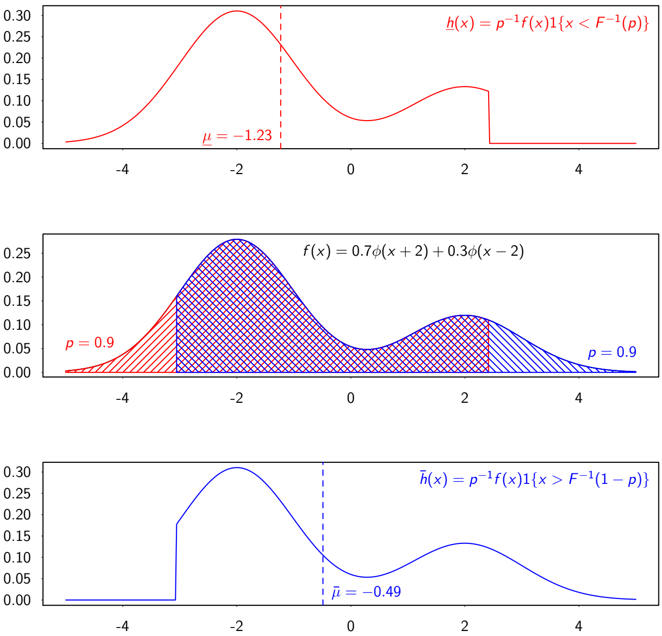


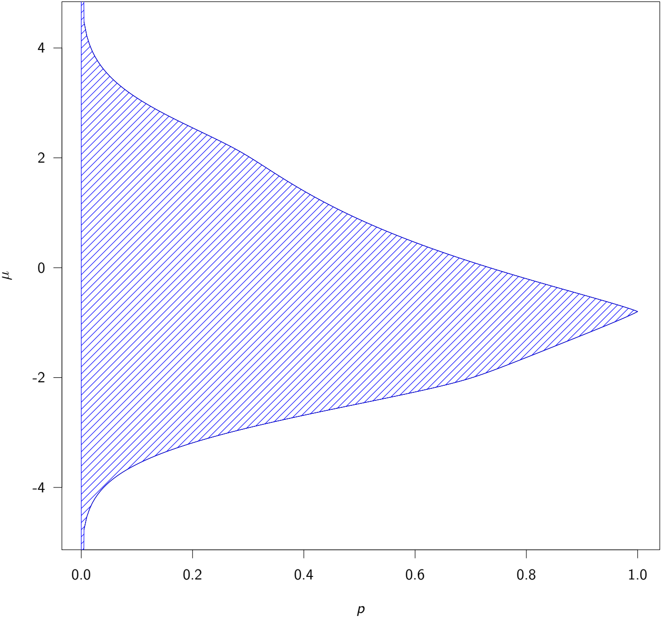
$\underline{h}(x) = p^{-1} f(x) 1\{x < F^{-1}(p)\}$
$\underline{\mu} = -2.26$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.6$
$p = 0.6$
$\bar{h}(x) = p^{-1} f(x) 1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 0.46$



$\underline{h}(x) = p^{-1} f(x) 1\{x < F^{-1}(p)\}$
$\underline{\mu} = -2.01$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.7$
$p = 0.7$
$\bar{h}(x) = p^{-1} f(x) 1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = 0.11$

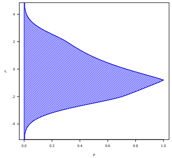

$\underline{h}(x) = p^{-1}f(x)1\{x < F^{-1}(p)\}$
$\underline{\mu} = -1.63$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.8$
$p = 0.8$
$\bar{h}(x) = p^{-1}f(x)1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = -0.2$


$\underline{h}(x) = p^{-1} f(x) 1\{x < F^{-1}(p)\}$
$\underline{\mu} = -1.23$
$f(x) = 0.7\phi(x+2) + 0.3\phi(x-2)$
$p = 0.9$
$p = 0.9$
$\bar{h}(x) = p^{-1} f(x) 1\{x > F^{-1}(1-p)\}$
$\bar{\mu} = -0.49$



4
2
0
-2
-4
μ
0.0
0.2
0.4
0.6
0.8
1.0
p

Binary Regressors



- For this particular choice of f, a mixture of normals, the blue shaded region shows all pairs (p, μ) that are compatible with the mixture.
- If $p = 0$, μ is unconstrained. This makes sense: in this case H can have any mean because it contributes nothing to the mixture that generates F.
- If $p = 1$ then μ must *equal* the mean of the observed distribution F, here -0.8, since this is a degenerate mixture where $F = H$.
- Relation to original problem? We observe dist of $y|T, z$ which is related to the unobserved dist of $y|T, T^*, z$ via a mixture model. Mixing prob. depends only on observables and (α_0, α_1); same for means of mixture components. Hence, some values of (α_0, α_1) are incompatible with the mixture model. This in restricts β since IV is $\beta/(1 - \alpha_0 - \alpha_1)$. Joint restrictions for all (t, k) so the book-keeping is complicated, but intuition is same as in simple mixture of normals example.

Sharp Identified Set under Baseline Assumptions

Theorem

(i) If $\mathbb{E}[y|\mathbf{x}, T = 0, z = k] \neq \mathbb{E}[y|\mathbf{x}, T = 1, z = k]$ for some k, non-differential assump. strictly improves upon weak bounds.

(ii) Under the baseline assumptions, β is not point identified, regardless of how many (discrete) values z takes on.

Corollary

Bounds for α_0, α_1, and β remain valid in a LATE model. They may not be sharp, however, sharp, since they do not incorporate the testable implications of the LATE assumptions.

Sharp Identified Set under Baseline Assumptions

Theorem

(i) If $\mathbb{E}[y|\mathbf{x}, T=0, z=k] \neq \mathbb{E}[y|\mathbf{x}, T=1, z=k]$ for some k, non-differential assump. strictly improves upon weak bounds.

(ii) Under the baseline assumptions, β is not point identified, regardless of how many (discrete) values z takes on.

Corollary

Bounds for α_0, α_1, and β remain valid in a LATE model. They may not be sharp, however, sharp, since they do not incorporate the testable implications of the LATE assumptions.

- Second main contribution: sharp identified set for $(\alpha_0, \alpha_1, \beta)$ under the baseline assumptions. The description of the set is fairly complicated, so I'm not going to show it on the slide. But the form that this set takes leads to two important results. First, the non-differential measurement error assumption *generically* improves upon the weak bounds. Second, under the baseline assumptions β is *never* point identified, regardless of how many different (discrete) values z takes.
- Some intuition: the true β always lies within the identified set by definition. It turns out that $\alpha_0 = \alpha_1 = 0$ implies that the mixing probabilities r_{tk} are all either zero or one. But in this case the mixtures are trivial, so we can simply set $F = H$. Hence, the IV estimand always lies in the sharp identified set.
- Corollary: everything I've said so far concerns an additively separable model. But in fact, bounds we derive under the baseline assumptions remain valid if we re-state our assumptions so that they involve a LATE model. These bounds may not be sharp in a LATE model, however, because the LATE assumptions themselves have testable implications. We don't impose these since we're mainly interested in the additively separable case.
- What now? Sharp bounds quite informative in practice, but they do not point identify β. Baseline assumptions aren't enough. Are there slightly stronger but still plausible assumptions that allow us to point identify β? Yes!

Point Identification: 1st Ingredient

Reparameterization

$$\theta_1(\mathbf{x}) = \beta(\mathbf{x}) / \left[1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]$$

$$\theta_2(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^2 \left[1 + \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right]$$

$$\theta_3(\mathbf{x}) = \left[\theta_1(\mathbf{x})\right]^3 \left[\left\{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\right\}^2 + 6\alpha_0(\mathbf{x})\left\{1 - \alpha_1(\mathbf{x})\right\}\right]$$

Lemma

Baseline Assumptions $\implies$ $\mathrm{Cov}(y, z|\mathbf{x}) = \theta_1(\mathbf{x})\mathrm{Cov}(z, T|\mathbf{x})$.

Point Identification: 1st Ingredient

Reparameterization

$\theta_1(\mathbf{x}) = \beta(\mathbf{x}) / [1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]$
$\theta_2(\mathbf{x}) = [\theta_1(\mathbf{x})]^2 [1 + \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]$
$\theta_3(\mathbf{x}) = [\theta_1(\mathbf{x})]^3 \left[\{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})\}^2 + 6\alpha_0(\mathbf{x})\{1 - \alpha_1(\mathbf{x})\}\right]$

Lemma

Baseline Assumptions $\implies$ $\text{Cov}(y, z|\mathbf{x}) = \theta_1(\mathbf{x})\text{Cov}(z, T|\mathbf{x})$.

- Re-parameterize: "reduced form" parameters $(\theta_1, \theta_2, \theta_3)$ are functions of "structural parameters" $(\alpha_0, \alpha_1, \beta)$. θ_1 is the IV estimand; (θ_2, θ_3) less intuitive: "correct" parameterization *after* finishing proof, then re-write!
- Notice that $\beta = 0$ iff $\theta_1 = \theta_2 = \theta_3 = 0$. Important later for inference.
- Identification argument: three lemmas to obtain equations that point identify reduced form parameters $(\theta_1, \theta_2, \theta_3)$. Then show that we can invert the mapping from structural to reduced form.
- First lemma identifies θ_1. Already showed this: tells us what IV identifies.

Point Identification: 2nd Ingredient

Assumption (II)

$\mathbb{E}[\varepsilon^2|\mathbf{x}, z] = \mathbb{E}[\varepsilon^2|\mathbf{x}]$

Lemma

(Baseline) + (II) $\implies$

$$\mathrm{Cov}(y^2, z|\mathbf{x}) = 2\mathrm{Cov}(yT, z|\mathbf{x})\theta_1(\mathbf{x}) - \mathrm{Cov}(T, z|\mathbf{x})\theta_2(\mathbf{x})$$

Corollary

(Baseline) + (II) + $[\beta(\mathbf{x}) \neq 0] \implies [\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})]$ is identified.

Point Identification: 2nd Ingredient

Assumption (II)

$\mathbb{E}[\varepsilon^2|\mathbf{x}, z] = \mathbb{E}[\varepsilon^2|\mathbf{x}]$

Lemma

(Baseline) + (II) $\implies$

$\text{Cov}(y^2, z|\mathbf{x}) = 2\text{Cov}(yT, z|\mathbf{x})\theta_1(\mathbf{x}) - \text{Cov}(T, z|\mathbf{x})\theta_2(\mathbf{x})$

Corollary

(Baseline) + (II) + $[\beta(\mathbf{x}) \neq 0] \implies [\alpha_1(\mathbf{x}) - \alpha_0(\mathbf{x})]$ is identified.

-
- Notice that the corollary implies that β is point identified if mis-classification is one-sided, as it might well be in the smoking example.

Point Identification: 3rd Ingredient

Assumption (III)

(i) $\mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*, T] = \mathbb{E}[\varepsilon^2|\mathbf{x}, z, T^*]$

(ii) $\mathbb{E}[\varepsilon^3|\mathbf{x}, z] = \mathbb{E}[\varepsilon^3|\mathbf{x}]$

Lemma

(Baseline) + (II) + (III) $\implies$

$$\mathrm{Cov}(y^3, z|\mathbf{x}) = 3\mathrm{Cov}(y^2 T, z|\mathbf{x})\theta_1(\mathbf{x}) - 3\mathrm{Cov}(yT, z|\mathbf{x})\theta_2(\mathbf{x}) + \mathrm{Cov}(T, z|\mathbf{x})\theta_3(\mathbf{x})$$

Point Identification Result

Theorem

(Baseline) + (II) + (III) $\implies \beta(\mathbf{x})$ is point identified. If $\beta(\mathbf{x}) \neq 0$, then $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are likewise point identified.

Explicit Solution

$$\beta(\mathbf{x}) = \text{sign}\left[\theta_1(\mathbf{x})\right] \sqrt{3\left[\theta_2(\mathbf{x})/\theta_1(\mathbf{x})\right]^2 - 2\left[\theta_3(\mathbf{x})/\theta_1(\mathbf{x})\right]}$$

Sufficient for (II) and (III)

(a) T is conditionally independent of (ε, z) given $(T^*, \mathbf{x})$

(b) z is conditionally independent of ε given $\mathbf{x}$

Point Identification Result

Theorem

(Baseline) + (II) + (III) $\implies \beta(\mathbf{x})$ is point identified. If $\beta(\mathbf{x}) \neq 0$, then $\alpha_0(\mathbf{x})$ and $\alpha_1(\mathbf{x})$ are likewise point identified.

Explicit Solution

$$\beta(\mathbf{x}) = \text{sign}\left[\theta_1(\mathbf{x})\right]\sqrt{3\left[\theta_2(\mathbf{x})/\theta_1(\mathbf{x})\right]^2 - 2\left[\theta_3(\mathbf{x})/\theta_1(\mathbf{x})\right]}$$

Sufficient for (II) and (III)

(a) T is conditionally independent of (ε, z) given $(T^*, \mathbf{x})$

(b) z is conditionally independent of ε given $\mathbf{x}$

Comment on the sufficient conditions: say that we really think these are what people have in mind in a natural experiment setting. Explain about reporting results in both logs and levels.

Inference for a Mis-classified Regressor

Weak Identification

- β small $\Rightarrow$ moment equalities uninformative about (α_0, α_1) more
- (α_0, α_1) could be on the boundary of the parameter space
- Also true of existing estimators that assume T^* exogenous

Our Approach

- Sharp identified set yields *inequality* moment restrictions that remain informative even if $\beta \approx 0$. more
- Identification-robust inference with equality and inequality MCs.

Inference with Moment Equalities and Inequalities

Moment Conditions

$$\mathbb{E}\left[m_j(\mathbf{w}_i, \vartheta_0)\right] \geq 0, \quad j = 1, \cdots, J$$

$$\mathbb{E}\left[m_j(\mathbf{w}_i, \vartheta_0)\right] = 0, \quad j = J+1, \cdots, J+K$$

Test Statistic

$$T_n(\vartheta) = \sum_{j=1}^{J} \left[\frac{\sqrt{n}\ \bar{m}_{n,j}(\vartheta)}{\widehat{\sigma}_{n,j}(\vartheta)}\right]_-^2 + \sum_{j=J+1}^{J+K} \left[\frac{\sqrt{n}\ \bar{m}_{n,j}(\vartheta)}{\widehat{\sigma}_{n,j}(\vartheta)}\right]^2$$

Critical Value

- $\sqrt{n}\,\bar{m}_n(\vartheta_0) \rightarrow_d$ normal limit with covariance matrix $\Sigma(\vartheta_0)$
- Use this to bootstrap the limit dist. of $T_n(\vartheta)$ under $H_0\colon \vartheta = \vartheta_0$

Inference with Moment Equalities and Inequalities

Moment Conditions

$\mathbb{E}\left[m_j(\mathbf{w}_i, \vartheta_0)\right] \geq 0, \quad j = 1, \cdots, J$

$\mathbb{E}\left[m_j(\mathbf{w}_i, \vartheta_0)\right] = 0, \quad j = J+1, \cdots, J+K$

Test Statistic

$$T_n(\vartheta) = \sum_{j=1}^{J} \left[\frac{\sqrt{n}\, \bar{m}_{n,j}(\vartheta)}{\widehat{\sigma}_{n,j}(\vartheta)}\right]_-^2 + \sum_{j=J+1}^{J+K} \left[\frac{\sqrt{n}\, \bar{m}_{n,j}(\vartheta)}{\widehat{\sigma}_{n,j}(\vartheta)}\right]^2$$

Critical Value

- $\sqrt{n}\, \bar{m}_n(\vartheta_0) \to_d$ normal limit with covariance matrix $\Sigma(\vartheta_0)$
- Use this to bootstrap the limit dist. of $T_n(\vartheta)$ under $H_0\colon \vartheta = \vartheta_0$

Explain about the meaning of the m-var, the sigma-hat and the “minus” subscript

Generalized Moment Selection

Andrews & Soares (2010)

- Inequalities that don't bind reduce power of test, so eliminate those that are "far from binding" before calculating critical value:

$$\boxed{\text{Drop inequality } j \text{ if } \frac{\sqrt{n}\,\bar{m}_{n,j}(\vartheta_0)}{\widehat{\sigma}_{n,j}(\vartheta_0)} > \sqrt{\log n}}$$

- Uniformly valid test of $H_0 \colon \vartheta = \vartheta_0$ even if ϑ_0 is not point identified.
- Not asymptotically conservative.

Problem

Joint test for the whole parameter vector but we're only interested in β. Projection is conservative and computationally intensive.

Generalized Moment Selection

Andrews & Soares (2010)

- Inequalities that don't bind reduce power of test, so eliminate those that are "far from binding" before calculating critical value:
  $\boxed{\text{Drop inequality } j \text{ if } \frac{\sqrt{n}\,\bar{m}_{n,j}(\vartheta_0)}{\widehat{\sigma}_{n,j}(\vartheta_0)} > \sqrt{\log n}}$
- Uniformly valid test of $H_0: \vartheta = \vartheta_0$ even if ϑ_0 is not point identified.
- Not asymptotically conservative.

Problem

Joint test for the whole parameter vector but we're only interested in β. Projection is conservative and computationally intensive.

Explain what not asymptotically conservative means. Explain what projection is and why it's conservative and computationally intensive.

Our Solution: Bonferroni-Based Inference

Special Structure

- β only enters MCs through $\theta_1 = \beta/(1 - \alpha_0 - \alpha_1)$
- Strong instrument $\Rightarrow$ inference for θ_1 is standard.
- Nuisance pars $\boldsymbol{\gamma}$ strongly identified under null for (α_0, α_1)

Procedure

1. Concentrate out $(\theta_1, \boldsymbol{\gamma}) \Rightarrow$ joint GMS test for (α_0, α_1)
2. Invert test $\Rightarrow (1 - \delta_1) \times 100\%$ confidence set for (α_0, α_1)
3. Project $\Rightarrow$ CI for $(1 - \alpha_0 - \alpha_1)$
4. Construct standard $(1 - \delta_2) \times 100\%$ IV CI for θ_1
5. Bonferroni $\Rightarrow (1 - \delta_1 - \delta_2) \times 100\%$ CI for β

Our Solution: Bonferroni-Based Inference

Special Structure

- β only enters MCs through $\theta_1 = \beta/(1 - \alpha_0 - \alpha_1)$
- Strong instrument $\Rightarrow$ inference for θ_1 is standard.
- Nuisance pars γ strongly identified under null for (α_0, α_1)

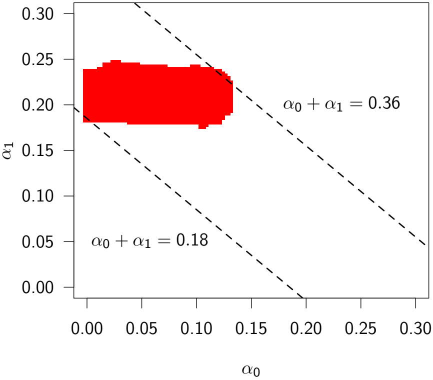
Procedure

1. Concentrate out $(\theta_1, \gamma) \Rightarrow$ joint GMS test for (α_0, α_1)
2. Invert test $\Rightarrow (1 - \delta_1) \times 100\%$ confidence set for (α_0, α_1)
3. Project $\Rightarrow$ CI for $(1 - \alpha_0 - \alpha_1)$
4. Construct standard $(1 - \delta_2) \times 100\%$ IV CI for θ_1
5. Bonferroni $\Rightarrow (1 - \delta_1 - \delta_2) \times 100\%$ CI for β

Explain that the procedure works well in simulations etc. Possibly add link to simulation here.

Example (sim data: $\beta = 1, \alpha_0 = 0.1, \alpha_1 = 0.2, n = 5000$)

97.5% GMS Confidence Region for (α_0, α_1)



Bonferroni Interval

1. 97.5% CI for $(1 - \alpha_0 - \alpha_1) = (0.64, 0.82)$
2. 97.5% CI for $\theta_1 = (1.20, 1.47)$
3. $>$ 95% CI for β:
   $(0.64 \times 1.20, 0.82 \times 1.47) = (0.77, 1.21)$

Comparisons

- (0.88, 1.04) for IV if T^* were observed
- (1.22,1.45) for naive IV interval using T

Conclusion

This Paper

- Partial and point identification results for effect of binary, endogenous regressor using a valid instrument.
- Identification-robust inference in models with mis-classification

Related Work

- Relaxing Instrument Validity: "A Framework for Eliticing, Incorporating, and Disciplining Identification Beliefs in Linear Models" (with Camilo Garcia-Jimeno)
- Relaxing Non-differential Measurement Error: "Estimating the Returns to Lying" (with Arthur Lewbel)

Simple Bounds for Mis-classification from First-stage

Unobserved	Observed
$p_k^*(\mathbf{x}) \equiv \mathbb{P}(T^* = 1 \mid \mathbf{x}, z = k)$	$p_k(\mathbf{x}) \equiv \mathbb{P}(T = 1 \mid \mathbf{x}, z = k)$

Relationship

$$p_k^*(\mathbf{x}) = \frac{p_k(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}, \quad k = 0, 1$$

z does not affect (α_0, α_1); denominator $\neq 0$

Bounds for Mis-classification

$$\alpha_0(\mathbf{x}) \leq p_k(\mathbf{x}) \leq 1 - \alpha_1(\mathbf{x}), \quad k = 0, 1$$

$\alpha_0(\mathbf{x}) + \alpha_1(\mathbf{x}) < 1$

back

What does IV estimate under mis-classification?

Unobserved

$$\beta(\mathbf{x}) = \frac{\mathbb{E}[y|\mathbf{x}, z=1] - \mathbb{E}[y|\mathbf{x}, z=0]}{p_1^*(\mathbf{x}) - p_0^*(\mathbf{x})}$$

Wald (Observed)

$$\frac{\mathbb{E}[y|\mathbf{x}, z=1] - \mathbb{E}[y|\mathbf{x}, z=0]}{p_1(\mathbf{x}) - p_0(\mathbf{x})} = \beta(\mathbf{x})\left[\frac{p_1^*(\mathbf{x}) - p_0^*(\mathbf{x})}{p_1(\mathbf{x}) - p_0(\mathbf{x})}\right] = \frac{\beta(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}$$

$$\boxed{p_1^*(\mathbf{x}) - p_0^*(\mathbf{x}) = \frac{p_1(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0 - \alpha_1(\mathbf{x})} - \frac{p_0(\mathbf{x}) - \alpha_0(\mathbf{x})}{1 - \alpha_0 - \alpha_1(\mathbf{x})} = \frac{p_1(\mathbf{x}) - p_0(\mathbf{x})}{1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})}}$$

back

Partial Identification Bounds for $\beta(\mathbf{x})$

$$\boxed{\beta(\mathbf{x}) = [1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x})]\left[\frac{\mathbb{E}[y|\mathbf{x}, z=1] - \mathbb{E}[y|\mathbf{x}, z=0]}{p_1(\mathbf{x}) - p_0(\mathbf{x})}\right]}$$

$$\boxed{0 \leq \alpha_0 \leq \min_k\{p_k(\mathbf{x})\}, \quad 0 \leq \alpha_1 \leq \min_k\{1 - p_k(\mathbf{x})\}}$$

No Mis-classification

$$\alpha_0(\mathbf{x}) = \alpha_1(\mathbf{x}) = 0 \implies \beta(\mathbf{x}) = \text{Wald}$$

Maximum Mis-classification

$$\alpha_0(\mathbf{x}) = p_{\min}(\mathbf{x}),\ \alpha_1(\mathbf{x}) = 1 - p_{\max}(\mathbf{x})$$

$$\implies 1 - \alpha_0(\mathbf{x}) - \alpha_1(\mathbf{x}) = p_{\max}(\mathbf{x}) - p_{\min}(\mathbf{x}) = |p_1(\mathbf{x}) - p_0(\mathbf{x})|$$

$$\implies \beta(\mathbf{x}) = \text{sign}\{p_1(\mathbf{x}) - p_0(\mathbf{x})\} \times (\text{Reduced Form})$$

back

Just-Identified System of Moment Equalities

Suppress dependence on $\mathbf{x}$...

$$\mathbb{E}\left[\{\boldsymbol{\Psi}(\boldsymbol{\theta})\mathbf{w}_i - \boldsymbol{\kappa}\} \otimes \begin{pmatrix} 1 \\ z \end{pmatrix}\right] = \mathbf{0}$$

$$\boldsymbol{\Psi}(\boldsymbol{\theta}) \equiv \begin{bmatrix} -\theta_1 & 1 & 0 & 0 & 0 & 0 \\ \theta_2 & 0 & -2\theta_1 & 1 & 0 & 0 \\ -\theta_3 & 0 & 3\theta_2 & 0 & -3\theta_1 & 1 \end{bmatrix}$$

$$\mathbf{w}_i = (T_i, y_i, y_i T_i, y_i^2, y_i^2 T_i, y_i^3)' \qquad \theta_1 = \beta/(1-\alpha_0-\alpha_1)$$

$$\boldsymbol{\kappa} = (\kappa_1, \kappa_2, \kappa_3)' \qquad \theta_2 = \theta_1^2(1+\alpha_0-\alpha_1)$$

$$\theta_3 = \theta_1^3\left[(1-\alpha_0-\alpha_1)^2 + 6\alpha_0(1-\alpha_1)\right]$$

back

Moment Inequalities I – First-stage Probabilities

$\alpha_0 \leq p_k \leq 1 - \alpha_1$ becomes $\mathbb{E}\left[m(\mathbf{w}_i, \boldsymbol{\vartheta})\right] \geq \mathbf{0}$ for all k where

$$m(\mathbf{w}_i, \boldsymbol{\vartheta}) \equiv \left[\begin{array}{l} \mathbf{1}(z_i = k)(T - \alpha_0) \\ \mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \end{array} \right]$$

Moment Inequalities II – Non-differential Assumption

For all k, we have $\mathbb{E}[m(\mathbf{w}_i, \vartheta, \mathbf{q}_k)] \geq 0$ where

$$m(\mathbf{w}_i, \vartheta, \mathbf{q}_k) \equiv \begin{bmatrix} y_i \mathbf{1}(z_i = k)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{0k})(1 - T_i)\left(\frac{1-\alpha_0-\alpha_1}{\alpha_1}\right)\right\} \\ -y_i \mathbf{1}(z_i = k)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{0k})(1 - T_i)\left(\frac{1-\alpha_0-\alpha_1}{\alpha_1}\right)\right\} \\ y_i \mathbf{1}(z_i = k)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i \leq \underline{q}_{1k}) T_i \left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1}\right)\right\} \\ -y_i \mathbf{1}(z_i = k)\left\{(T_i - \alpha_0) - \mathbf{1}(y_i > \overline{q}_{1k}) T_i \left(\frac{1-\alpha_0-\alpha_1}{1-\alpha_1}\right)\right\} \end{bmatrix}$$

and $\mathbf{q}_k \equiv (\underline{q}_{0k}, \overline{q}_{0k}, \underline{q}_{1k}, \overline{q}_{1k})'$ defined by $\mathbb{E}[h(\mathbf{w}_i, \vartheta, \mathbf{q}_k)] = 0$ with

$$h(\mathbf{w}_i, \vartheta, \mathbf{q}_k) = \begin{bmatrix} \mathbf{1}(y_i \leq \underline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \overline{q}_{0k})\mathbf{1}(z_i = k)(1 - T_i) - \left(\frac{1-\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \\ \mathbf{1}(y_i \leq \underline{q}_{1k})\mathbf{1}(z_i = k) T_i - \left(\frac{1-\alpha_1}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(T_i - \alpha_0) \\ \mathbf{1}(y_i \leq \overline{q}_{1k})\mathbf{1}(z_i = k) T_i - \left(\frac{\alpha_0}{1-\alpha_0-\alpha_1}\right)\mathbf{1}(z_i = k)(1 - T_i - \alpha_1) \end{bmatrix}$$

back