Robustness to missing data: breakdown point analysis

Daniel Ober-Reynolds

Bates White Economic Consulting

doberreynolds@g.ucla.edu

4 February 2026

Introduction

▶ Missing data is common, as are the selection concerns it raises.

▶ Common solution: assume data are Missing (Completely) At Random.
  • Impute or ignore incomplete observations, use standard methods.
  • Convenient solution, often implausible justification.

▶ This paper proposes an interpretable measure of selection, and estimates how much selection is needed to overturn a conclusion.

Missing data

▶ Bollinger et al. (2019) "Trouble in the Tails? What We Know about Earnings Nonresponse 30 Years after Lillard, Smith, and Welch"

- CPS ASEC 2015 item and whole nonresponse rate: 43%.

- Linking data with SSA tax records shows missing earnings data is not MAR.

▶ Molina-Millán and Macours (2025) "Attrition in randomized controlled trails: Using tracking information to correct bias"

- Randomized controlled trials (RCTs) in developing countries often suffer from attrition.

- Inverse propensity weighting makes little difference to estimates, but...

- ...including subjects found after costly additional tracking can notably change estimates, suggesting data are not MAR.

Related literature

▶ Missing data without MAR
  - Point identification: Heckman (1979), Das et al. (2003), Escanciano et al. (2016)
  - Partial identification: Manski (2005), Horowitz and Manski (2006)
  - Breakdown point/sensitivity analysis: Kline and Santos (2013)

▶ Breakdown point/sensitivity analysis
  - Missing data: Kline and Santos (2013)
  - Potential outcomes: Masten and Poirier (2020), Bonvini and Kennedy (2022), Jin et al. (2022), Rosenbaum (2023), Spini (2024)
  - Omitted variable bias: Altonji et al. (2005), Oster (2019), Diegert et al. (2025)

⇒ This paper contributes a robustness exercise for missing data that
  i. allows for any number of variables to be missing,
  ii. directly uses the researcher's GMM model,
  iii. requires no additional data or modeling (no exclusion restriction), and
  iv. gives results that are succinct and interpretable.

Overview

Overview

Setting

▶ Data is i.i.d. sample $\{D_i, D_i Y_i, X_i\}_{i=1}^n$, where $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$.

- The variables of interest are $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$.
- Y may be a vector. If present, X is assumed finitely supported.
- Example: $Y_i = (Y_i^{(1)}, Y_i^{(2)}) \in \mathbb{R}^2$ collected through survey, X_i is administrative data (age, occupation, etc.).

▶ Parameter $\beta \in \mathbf{B} \subseteq \mathbb{R}^{d_b}$ is identified through moment conditions:

$$E_P[g(Y, X, b)] = 0 \text{ if and only if } b = \beta,$$

where P is the unconditional distribution of (Y, X).

- Example: OLS coefficients $g(Y, X, b) = \begin{pmatrix} Y^{(2)} \\ X \end{pmatrix} (Y^{(1)} - (Y^{(2)}, X^\mathsf{T})b)$.

▶ Conclusion to be investigated is that β is outside $\mathbf{B}_0$:

$$H_0 : \beta \in \mathbf{B}_0 \qquad \text{vs} \qquad H_1 : \beta \in \mathbf{B} \setminus \mathbf{B}_0$$

- Example: first OLS coefficient is positive. $\mathbf{B}_0 = \{b \in \mathbf{B} \; ; \; b^{(1)} \leq 0\}$.

Setting

▶ Let $p_D = P(D = 1)$, $X \mid D = 0 \sim P_{0X}$, and

$$(Y, X) \mid D = 1 \sim P_1, \qquad\qquad (Y, X) \mid D = 0 \sim P_0,$$
$$P = p_D P_1 + (1 - p_D) P_0.$$

- The sample $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ identifies p_D, P_1, and P_{0X}...
- ...but not P_0, P, or β solving $E_P[g(Y, X, \beta)] = 0$.

▶ Common solution: estimate β_1 instead, where

$$E_{P_1}[g(Y, X, \beta_1)] = 0.$$

MCAR is the assumption $P_0 = P_1$. This implies $P = P_1$ and $\beta = \beta_1$.

▶ Suppose preliminary analysis suggests $\beta_1 \in \mathbf{B} \setminus \mathbf{B}_0$, but MCAR is doubtful. MAR?

- Hope to defend $\beta \in \mathbf{B} \setminus \mathbf{B}_0$.
- So $P_0 \neq P_1$... but *how* different could they plausibly be?
- A quantitative measure of selection will allow meaningful discussion.

Quantifying selection: does (Y, X) predict missingness?

Sample is $\{D_i, D_i Y_i, X_i\}_{i=1}^n$, $p_D = P(D = 1)$,

$$(Y, X) \mid D = 1 \sim P_1, \qquad\qquad (Y, X) \mid D = 0 \sim P_0,$$
$$P = p_D P_1 + (1 - p_D) P_0.$$

▶ Selection is a greater concern when context suggests (Y, X) would predict D well.
  • Example: survey asking about arrest record, vs. survey asking about TV preferences.

▶ Illustrate this formally with densities. Let f_1, f_0 be densities of P_1, P_0 wrt P:

$$f_1(y, x) = \frac{p_D(y, x)}{p_D}, \qquad\qquad f_0(y, x) = \frac{1 - p_D(y, x)}{1 - p_D}$$

where $p_D(y, x) = P(D = 1 \mid Y = y, X = x)$.
  • Optimistic: D is independent of (Y, X).
    $\implies p_D(y, x) = p_D$, so $f_1 = f_0$ (i.e., data is MCAR).
  • Pessimistic: D is almost a function of (Y, X).
    $\implies p_D(y, x) \approx 1$ or 0; f_1 and f_0 look quite different.

Quantifying selection with squared Hellinger

▶ Measure **selection** as the **squared Hellinger distance** between P_0 and P_1:

$$H^2(P_0, P_1) = \frac{1}{2} E_P \left[(\sqrt{f_0(Y,X)} - \sqrt{f_1(Y,X)})^2 \right],$$

where $f_0(y,x)$ and $f_1(y,x)$ are densities of P_0 and P_1 wrt P.

Lemma Let $(Y, X, D) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x} \times \{0,1\}$ with $P(D = 1) \in (0,1)$. Let P_d denote the distribution of $(Y, X) \mid D = d$. Then

$$H^2(P_0, P_1) = 1 - \frac{E_P \left[\sqrt{\text{Var}(D \mid Y,X)} \right]}{\sqrt{\text{Var}(D)}}.$$

• Interpretation: expected percent standard deviation of D "explained" by (Y, X).
• Captures intuition: more predictive power, higher selection.
• Range is $[0,1]$. Equals $0 \Leftrightarrow \text{Var}(D \mid Y,X) = \text{Var}(D)$, equals $1 \Leftrightarrow \text{Var}(D \mid Y,X) = 0$.

Quantifying selection: f-divergences

▶ Squared Hellinger is intuitive, but other divergences can be used to measure selection.

▶ The results allow measuring selection with any of a broad class of f-divergences, including squared Hellinger.

▶ Measuring selection with an f-divergence effectively assumes P_0 is dominated by P_1.
  • This rules out selection mechanisms that "truncate" data; e.g. $D_i = \mathbb{1}\{Y_i \leq c\}$.

f-divergences Domination

Recap

$$(Y, X) \mid D = 1 \sim P_1, \qquad\qquad (Y, X) \mid D = 0 \sim P_0,$$
$$P = p_D P_1 + (1 - p_D) P_0.$$

▶ Setting:
- Data: $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ i.i.d., with $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$
- Model: $E_P[g(Y, X, \beta)] = 0$
- Hypothesis test: $H_0 : \beta \in \mathbf{B}_0$ vs $H_1 : \beta \in \mathbf{B} \setminus \mathbf{B}_0$
- Identified: p_D, P_1, P_{0X}. Not identified: $P = p_D P_1 + (1 - p_D) P_0$, or β
- Measure of selection: $H^2(P_0, P_1) = 1 - E_P[\sqrt{\text{Var}(D \mid Y, X)}] / \sqrt{\text{Var}(D)}$

▶ β_1 solves $E_{P_1}[g(Y, X, \beta_1)] = 0$; preliminary analysis suggests $\beta_1 \in \mathbf{B} \setminus \mathbf{B}_0$

▶ How much selection is needed to overturn the conclusion?
- Given p_D, P_1, and P_{0X}, how large must $H^2(P_0, P_1)$ be to rationalize $\beta \in \mathbf{B}_0$?

Overview

Breakdown point

▶ Let $\mathbf{P}^b$ be the set of distributions Q dominated by P_1 with marginal $Q_X = P_{0X}$ and

$$0 = p_D E_{P_1}[g(Y, X, b)] + (1 - p_D) E_Q[g(Y, X, b)].$$

Say Q **rationalizes** b.

▶ The **breakdown point** is the minimum selection needed to rationalize $\beta \in \mathbf{B}_0$:

$$\delta^{BP} = \inf_{b \in \mathbf{B}_0} \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)$$

▶ Large values of δ^{BP} assuage selection concerns.
- The claim $\beta \in \mathbf{B}_0$ implies $\delta^{BP} \leq \frac{1}{2} H^2(P_0, P_1) = 1 - E_P\left[\sqrt{\text{Var}(D \mid Y, X)}\right] / \sqrt{\text{Var}(D)}$.
- If the claim (Y, X) predicts D this well is implausible, then $\beta \in \mathbf{B}_0$ is implausible.
- Context matters! Example: Survey about arrest record vs. survey about TV.

▶ δ^{BP} is point identified.
- Reporting estimates $\hat{\delta}_n^{BP}$ can facilitate selection concern discussions.
- Worries that $\hat{\delta}_n^{BP} > \delta^{BP}$ (due to sample noise) can be addressed with lower confidence intervals.
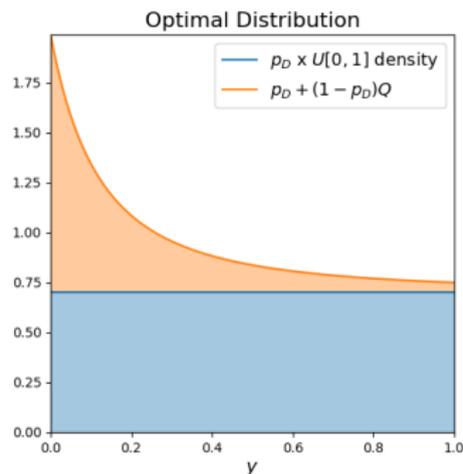
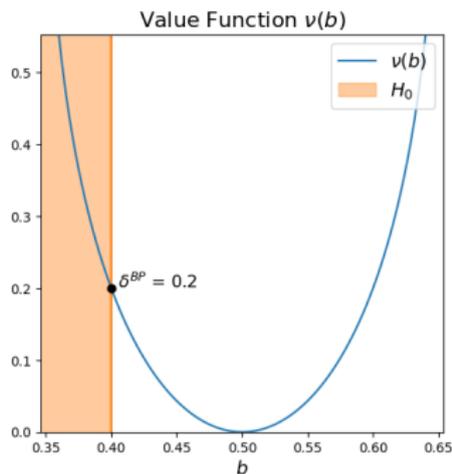Partial identification framing

Breakdown point example: expectation

$$\delta^{BP} = \inf_{b \in \mathbf{B}_0} \underbrace{\inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)}_{\nu(b)}$$

▶ **Example**: The sample is $\{D_i, D_i Y_i\}_{i=1}^n$, and $\beta = E[Y] \in \mathbb{R}$.

$$Y \mid D = 1 \sim \mathcal{U}[0,1], \qquad p_D = P(D=1) = 0.7$$

The claim to be supported is $H_1 : \beta > 0.4$.

Overview

Estimation: overview

▶ The breakdown point,

$$\delta^{BP} = \inf_{b \in \mathbf{B}_0} \underbrace{\inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)}_{\nu(b)},$$

is estimated with a two-step procedure:

1. First, $\hat{\nu}_n(b)$ estimates $\nu(b) = \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)$.
2. Second, plug-in second step $\hat{\delta}_n^{BP} = \inf_{b \in \mathbf{B}_0} \hat{\nu}_n(b)$.

▶ $\hat{\nu}_n(b)$ is based on finite dimensional, well-behaved dual problem.

▶ Second stage estimator is analyzed using the functional delta method.

Skip to Simulations

Estimation: duality

▶ The **primal problem** is

$$\nu(b) = \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1). \tag{1}$$

▶ The **dual problem** is

$$V(b) = \sup_{\lambda \in \mathbb{R}^{d_g + K}} E\left[\frac{\lambda^\intercal J(D) h(DY, X, b)}{1 - p_D} - \frac{Df^*(\lambda^\intercal h(DY, X, b))}{p_D}\right], \tag{2}$$

a finite dimensional convex optimization problem.

- f^*, J and h are known functions and $p_D = P(D = 1)$.
- the expectation is wrt the distribution of (D, DY, X).
- K is the cardinality of $\mathrm{Supp}(X)$.

▶ Under regularity conditions, **strong duality** holds for $b \in B \subseteq \mathbf{B}$:

$$V(b) = \nu(b).$$

Dual problem detail   Strong duality assumptions

Estimation: estimators

▶ With strong duality, the breakdown point is $\delta^{BP} = \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$, where

$$\nu(b) = \sup_{\lambda \in \mathbb{R}^{d_g + K}} E\left[\underbrace{\frac{\lambda^\intercal J(D) h(DY, X, b)}{1 - p_D} - \frac{Df^*(\lambda^\intercal h(DY, X, b))}{p_D}}_{\equiv \varphi(D, DY, X, b, \lambda, p_D)} \right].$$

▶ Straightforward sample analogue estimators: $\hat{\delta}_n^{BP} = \inf_{b \in \mathbf{B}_0} \hat{\nu}_n(b)$, where

$$\hat{\nu}_n(b) = \sup_{\lambda \in \mathbb{R}^{d_g + K}} \frac{1}{n} \sum_{i=1}^{n} \varphi(D_i, D_i Y_i, X_i, b, \lambda, \hat{p}_{D,n}).$$

▶ $\hat{\nu}_n(b)$ is often convex, making $\hat{\delta}_n = \inf_{b \in \mathbf{B}_0} \hat{\nu}_n(b)$ straightforward to compute.

Estimation: inference

> **Theorem** Under assumptions discussed in the paper,
>
> $$\frac{\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})}{\hat{\sigma}_n} \xrightarrow{d} N(0,1),$$
>
> where $\hat{\sigma}_n^2$ is a consistent variance estimator given in the paper.

▶ A large δ^{BP} assuages selection concerns.

▶ Skeptical readers may worry $\hat{\delta}_n^{BP} > \delta^{BP}$ because of sample noise.
  • The argument is only strengthened if $\hat{\delta}_n^{BP} < \delta^{BP}$.

▶ Reporting a **lower confidence interval** $\widehat{CI}_{L,n}$ addresses this concern:

$$\lim_{n \to \infty} P\left(\underbrace{\hat{\delta}_n^{BP} - \frac{\hat{\sigma}_n}{\sqrt{n}} c_{1-\alpha}}_{\widehat{CI}_{L,n}} \leq \delta^{BP} \right) = 1 - \alpha,$$

where $c_{1-\alpha}$ is the $1-\alpha$ quantile of $N(0,1)$.

Formal assumptions

Overview

Simulations: expectation

▶ Example: The sample is $\{D_i, D_i Y_i\}_{i=1}^n$, and $\beta = E[Y] \in \mathbb{R}$.

$$Y \mid D = 1 \sim \mathcal{U}[0,1], \qquad p_D = P(D = 1) = 0.7$$

The claim to be supported is $H_1 : \beta > 0.4$.

▶ 1,000 simulations with $P(D = 1) = 0.7$, and $\delta^{BP} \approx 0.2$:

Table: Simulations, expectation

n	Bias	St. Dev.	Coverage	Ave. CI Length
1,000	0.005	0.056	98.5	0.090
3,000	0.002	0.032	96.3	0.051
5,000	0.001	0.025	95.8	0.039
10,000	0.001	0.017	95.8	0.028

Illustration

Simulations: logit

▶ Let $X \in \{0, 1\}$, $Y \in \mathbb{R}^3$. Consider the logistic model

$$P(X = 1 \mid Y) = \Lambda(Y^\mathsf{T}\beta), \qquad \Lambda(t) = \frac{\exp(t)}{1 + \exp(t)}.$$

With complete data, estimation with maximum likelihood is equivalent to nonlinear GMM:

$$E\left[(X - \Lambda(Y^\mathsf{T}\beta))\, Y\right] = 0.$$

▶ The observed data is $\{D_i, D_i Y_i, X_i\}_{i=1}^n$, and the conclusion to be investigated is

$$H_1 : \Lambda(\bar{y}^\mathsf{T}\beta) > 0.5.$$

▶ 1,000 simulations from a DGP described in the paper:

Table: Simulations, logit model

n	Bias	St. Dev.	Coverage	Ave. CI Length
1,000	0.003	0.018	94.5	0.029
3,000	-0.000	0.010	96.1	0.017
5,000	0.001	0.008	94.8	0.013
10,000	-0.000	0.005	95.9	0.009

Overview

Application: attrition in randomized controlled trials

▶ RCTs in developing countries frequently suffer from missing data due to migration.

▶ Molina-Millán and Macours (2025) studies RCTs where movers were carefully tracked.

  • Inverse propensity weighting with data on non-movers did not change estimates, but...

  • ...including movers often changed estimates, suggesting data are not missing at random.

▶ Breakdown point analysis on the intention to treat results of three RCTs:
  1. Barham et al. (2024),
  2. Bandiera et al. (2020), and
  3. Giacobino et al. (2024)

Application: randomized controlled trials

1. Barham et al. (2024) studies a conditional cash transfer (CCT) in Nicaragua.
   - Part of the CCT gave households with children a school attendance cash transfer.
   - Treated localities received the CCT early, at ages when boys are susceptible to dropping out.

2. Bandiera et al. (2020) studies a women's empowerment program in Uganda.
   - In treated communities, afternoon clubs taught young girls income generating activities (e.g., tailoring, agriculture) and about reproductive health.
   - Objectives included increasing economic and reproductive independence.

3. Giacobino et al. (2024) studies a scholarship for Nigerian girls to attend middle school.
   - Objectives included encouraging schooling and detering child marriage.
   - Girls in treated villages, which had no middle school, were offered a scholarship covering costs of attending (housing, food, uniforms, school supplies).

Application: estimates dropping incomplete observations

Paper	Outcome	Observations		Untreated mean	MCAR ITT estimates	
		Total (n) (1)	Complete (2)	(3)	Replication (4)	Region ind. (5)
Barham et al. (2024)	Off-farm employment	1,138	1,006	0.83	0.06 (0.02)	0.06 (0.03)
	Rank of earnings per mo. worked	1,138	1,006	497.15	41.78 (19.50)	42.87 (21.58)
	Read and write	1,138	1,007	0.87	0.05 (0.02)	0.07 (0.02)
Bandiera et al. (2020)	Income-generating activity	5,966	4,831	0.10	0.07 (0.02)	0.07 (0.01)
	Self-employed	5,966	4,831	0.06	0.06 (0.01)	0.06 (0.01)
	Wage employed	5,966	4,831	0.04	0.01 (0.01)	0.01 (0.01)
Giacobino et al. (2024)	Dropped out	1,501	1,344	0.40	-0.21 (0.05)	-0.21 (0.02)
	Married	1,501	1,344	0.14	-0.07 (0.03)	-0.07 (0.02)
	Life satisfaction	1,501	1,344	0.00	0.25 (0.11)	0.25 (0.05)

Notes: Additional results from Bandiera et al. (2020) are reported in the paper. Column (4) replicates each paper's results, including the authors' choice of additional regressors and standard errors. Column (5) uses region indicators as the additional regressors and reports HC3 standard errors in parentheses. For outcomes from Barham et al. (2024), column (3) reports the mean from subjects receiving the CCT late and estimates in column (5) make use of sampling weights from Molina-Millán & Macours (2025) rather than inverse propensity weights from Barham et al. (2024). For Bandiera et al. (2020), column (3) reports average outcomes measured at baseline.

Application: breakdown point analysis

Table: Breakdown point analyses

Paper	Outcome	$\hat{\rho}_{D,n}$	$\hat{\delta}_n^{BP}$	$\widehat{CI}_{L,n}$	Estimate of $H^2(P_{0X}, P_{1X})$
Barham et al. (2024)	Off-farm employment	0.88	0.08	0.03	0.05
	Rank of earnings per mo. worked	0.88	0.08	0.03	0.05
	Read and write	0.88	0.07	0.04	0.05
Bandiera et al. (2020)	Income-generating activity	0.82	0.05	0.04	0.03
	Self-employed	0.82	0.04	0.03	0.03
	Wage employed	0.82	0.03	0.03	0.03
Giacobino et al. (2024)	Dropped out	0.90	∞	∞	0.05
	Married	0.90	0.17	0.06	0.05
	Life satisfaction	0.90	0.26	0.09	0.05

Note: Additional results from Bandiera et al. (2020) are reported in the paper.

Conclusion

► Breakdown point analysis is a tractable approach to assessing how robust a conclusion is to relaxing common missing data assumptions.

► For the conclusion $\beta \in \mathbf{B} \setminus \mathbf{B}_0$, the claim $\beta \in \mathbf{B}_0$ implies

$$\delta^{BP} \leq 1 - \frac{E_P[\sqrt{\text{Var}(D \mid Y, X)}]}{\sqrt{\text{Var}(D)}}.$$

If it is implausible (Y, X) predicts D this well, then $\beta \in \mathbf{B}_0$ is similarly implausible.

► The estimator of δ^{BP} is $\sqrt{n}$-consistent and asymptotically normal.

► Reporting $\hat{\delta}_n^{BP}$ and the lower confidence interval $\widehat{CI}_{L,n}$ is a succinct summary of a conclusion's robustness.

References I

Altonji, Joseph G, Todd E Elder, and Christopher R Taber (2005). "Selection on observed and unobserved variables: Assessing the effectiveness of Catholic schools". In: *Journal of political economy* 113(1), pp. 151–184.

Bandiera, Oriana et al. (2020). "Women's empowerment in action: evidence from a randomized control trial in Africa". In: *American Economic Journal: Applied Economics* 12(1), pp. 210–259.

Barham, Tania, Karen Macours, and John A Maluccio (2024). "Experimental evidence from a conditional cash transfer program: schooling, learning, fertility, and labor market outcomes after 10 years". In: *Journal of the European Economic Association* 22(4), pp. 1844–1883.

Bollinger, Christopher R et al. (2019). "Trouble in the tails? What we know about earnings nonresponse 30 years after Lillard, Smith, and Welch". In: *Journal of Political Economy* 127(5), pp. 2143–2185.

Bonvini, Matteo and Edward H Kennedy (2022). "Sensitivity analysis via the proportion of unmeasured confounding". In: *Journal of the American Statistical Association* 117(539), pp. 1540–1550.

Das, Mitali, Whitney K Newey, and Francis Vella (2003). "Nonparametric estimation of sample selection models". In: *The Review of Economic Studies* 70(1), pp. 33–58.

Diegert, Paul, Matthew A Masten, and Alexandre Poirier (2025). "Assessing Omitted Variable Bias when the Controls are Endogenous". In: *arXiv preprint arXiv:2206.02303*.

Escanciano, Juan Carlos, David Jacho-Chávez, and Arthur Lewbel (2016). "Identification and estimation of semiparametric two-step models". In: *Quantitative Economics* 7(2), pp. 561–589.

Giacobino, Hélène et al. (2024). "Schoolgirls, not brides: Education as a shield against child marriage". In: *American Economic Journal: Applied Economics* 16(4), pp. 109–143.

Heckman, James J (1979). "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society*, pp. 153–161.

Horowitz, Joel L and Charles F Manski (2006). "Identification and estimation of statistical functionals using incomplete data". In: *Journal of Econometrics* 132(2), pp. 445–459.

Jin, Ying, Zhimei Ren, and Zhengyuan Zhou (2022). "Sensitivity analysis under the f-sensitivity models: a distributional robustness perspective". In: *arXiv preprint arXiv:2203.04373*.

References II

Kline, Patrick and Andres Santos (2013). "Sensitivity to missing data assumptions: Theory and an evaluation of the US wage structure". In: *Quantitative Economics* 4(2), pp. 231–267.

Manski, Charles F (2005). "Partial identification with missing data: concepts and findings". In: *International Journal of Approximate Reasoning* 39(2-3), pp. 151–165.

Masten, Matthew A and Alexandre Poirier (2020). "Inference on breakdown frontiers". In: *Quantitative Economics* 11(1), pp. 41–111.

Molina-Millán, Teresa and Karen Macours (2025). "Attrition in Randomized Controlled Trials: Using Tracking Information to Correct Bias". In: *Economic Development and Cultural Change* 73(2), pp. 811–834.

Oster, Emily (2019). "Unobservable selection and coefficient stability: Theory and evidence". In: *Journal of Business & Economic Statistics* 37(2), pp. 187–204.

Rosenbaum, Paul R (2023). "Sensitivity analyses informed by tests for bias in observational studies". In: *Biometrics* 79(1), pp. 475–487.

Spini, Pietro Emilio (2024). "Robustness, heterogeneous treatment effects and covariate shifts". In: *arXiv preprint arXiv:2112.09259*.

Missing (completely) at random

▶ With i.i.d. sample $\{D_i, D_i Y_i, X_i\}_{i=1}^n$, where $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$

$$(Y, X) \mid D = 1 \sim P_1, \qquad\qquad (Y, X) \mid D = 0 \sim P_0,$$
$$P = p_D P_1 + (1 - p_D)P_0,$$

two common assumptions restore point identification of P:

▶ **Missing completely at random** (MCAR) assumes $P_0 = P_1$.
   - Partially testable: do distributions of X match? $P_{0X} = P_{1X}$?
   - Justifies dropping observations where $D_i = 0$.

▶ **Missing at random** (MAR) assumes $Y \mid X = x, D = 0$ follows the same distribution as $Y \mid X = x, D = 1$.
   - Not testable.
   - Justifies imputating $Y \mid X = x, D = 0$ based on distribution of $Y \mid X = x, D = 1$.

▶ Preliminary analysis may be based on either assumption.

Setting

Assumption: P_0 is dominated by P_1

▶ **Assumption:** P_0 is dominated by P_1, i.e. $P_0 \ll P_1$.
  - For any set A with $P_1((X, Y) \in A) = 0$, then $P_0((X, Y) \in A) = 0$.
  - Simplifies analysis considerably; set of possible P_0 characterized by densities wrt P_1.
  - Effectively assumed when selection is measured by an f-divergence.

▶ Some support assumption is typically necessary for an interesting exercise.
  - Example: $\beta = E[Y]$. P_1 and P_0 given by

$$P_1(Y = -1) = 0.5 \qquad P_1(Y = 1) = 0.5$$
$$P_0(Y = -1) = 0.5 \qquad P_0(Y = 1) = 0.5 - \alpha \qquad P_0(Y = y) = \alpha$$

  Then

$$H^2(P_0, P_1) = (\sqrt{0.5 - \alpha} - \sqrt{0.5} + \sqrt{\alpha})^2$$

  can be made **arbitrarily close to zero** by choice of $\alpha > 0$. For any $\alpha > 0$,

$$E_P[Y] = (1 - p_D)E_{P_0}[Y] = (1 - p_D)\alpha(y - 1)$$

  can be made **any number** by choice of $y \in \mathbb{R}$.

`f-divergence` `Squared Hellinger`

Breakdown Point through Partial Identification

▶ Breakdown point analysis can be framed as an exercise in partial identification, as in Kline and Santos (2013), Masten and Poirier (2020), and Diegert et al. (2025).

▶ In this framing, consider assumptions of the form $H^2(P_0, P_1) \leq \delta$ for some $\delta > 0$.

▶ The *nominal* identified set $\mathbf{B}_{ID}(\delta)$ for β grows with δ. As long as $\mathbf{B}_{ID}(\delta) \subseteq \mathbf{B} \setminus \mathbf{B}_0$, it is clear the conclusion holds.

▶ The **breakdown point** δ^{BP} can then be defined as either:
  1. the largest δ for which $\mathbf{B}_{ID}(\delta) \subseteq \mathbf{B} \setminus \mathbf{B}_0$, or
  2. the smallest δ for which $\mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \varnothing$.

Breakdown Point

Other selection measures: f-divergences

▶ Given a convex function $f : \mathbb{R} \to [0, \infty]$ satisfying $f(t) = \infty$ for $t < 0$ and taking a unique minimum of $f(1) = 0$, the corresponding f-**divergence** is the function given by

$$d_f(Q\|P) = \begin{cases} \int f\left(\frac{dQ}{dP}\right) dP & \text{if } Q \ll P \\ \infty & \text{otherwise} \end{cases} \tag{3}$$

▶ Many popular divergences can be written as f-divergences (when $Q \ll P$):

Name	Common formula	$f(t)$ when $t \geq 0$
Squared Hellinger	$H^2(Q, P) = \frac{1}{2} \int \left(\sqrt{\frac{dQ}{dP}(z)} - 1 \right)^2 dP(z)$	$f(t) = \frac{1}{2}(\sqrt{t} - 1)^2$
Kullback-Leibler (KL)	$KL(Q\|P) = \int \log\left(\frac{dQ}{dP}(z)\right) dQ(z)$	$f(t) = t\log(t) - t + 1$
"Reverse" KL	$KL(P\|Q) = \int \log\left(\frac{dP}{dQ}(z)\right) dP(z)$	$f(x) = -\log(t) + t - 1$
Cressie-Read	–	$f_\gamma(t) = \frac{t^\gamma - \gamma t + \gamma - 1}{\gamma(\gamma - 1)}, \ \gamma < 1$

Table: Common f-divergences

▶ Results in the paper allow any f-divergence (satisfying certain regularity conditions) to be used to measure selection.

Squared Hellinger

Dual problem (detailed)

▶ The **dual problem** using squared Hellinger is

$$V(b) = \sup_{\lambda \in \mathbb{R}^{d_g+K}} E\left[\frac{\lambda^\intercal J(D)h(DY, X, b)}{1 - p_D} - \frac{Df^*(\lambda^\intercal h(DY, X, b))}{p_D}\right]$$

where

$$J(D) = \begin{bmatrix} -DI_{d_g} & 0 \\ 0 & (1-D)I_K \end{bmatrix}, \qquad h(DY, X, b) = \begin{pmatrix} g(DY, X, b) \\ \mathbb{1}\{X = x_1\} \\ \vdots \\ \mathbb{1}\{X = x_K\} \end{pmatrix},$$

$$f^*(r) = \begin{cases} \frac{1}{2}\left(\frac{1}{1-2r} - 1\right) & \text{if } r < 1/2 \\ \infty & \text{o.w.} \end{cases}$$

and $\{x_1, \ldots, x_K\}$ is the support of X.

▶ $f^*(r) = \sup_{t \in \mathbb{R}}\{rt - f(t)\}$ is the **convex conjugate** of $f(t)$, the function defining the f-divergence used to measure selection.

Duality f-divergences

Assumptions: setting and strong duality

Assumption 1 *(Setting)* $\{(D_i, D_i Y_i, X_i)\}_{i=1}^n$ is an i.i.d. sample from a distribution satisfying

1. $p_D = P(D = 1) \in (0, 1)$,
2. $X \mid D = 1$ and $X \mid D = 0$ have the same finite support $\{x_1, \ldots, x_K\}$,
3. $E\left[\sup_{b \in B} \|g(Z, b)\| \mid D = 1\right] < \infty$, where $Z = (Y, X)$,
4. $P_0 \ll P_1$, and
5. $f : \mathbb{R} \to [0, \infty]$ is closed, proper, strictly convex, essentially smooth, takes its unique minimum of $f(t) = 0$ at $t = 1$, and satisfies $f(t) = \infty$ for all $t < 0$. The interior of $\operatorname{dom}(f) \equiv \{t \in \mathbb{R} \; : \; f(t) < \infty\}$, denoted (ℓ, u), satisfies $\ell < 1 < u$, and f is twice continuously differentiable on (ℓ, u).

Assumption 2 *(Strong duality)* $B \subseteq \mathbf{B}$ is convex, compact, and satisfies $\inf_{b \in \mathbf{B}_0} \nu(b) = \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$. Furthermore, for each $b \in B$,

1. there exists $Q^b \in \mathbf{P}^b$ such that $\ell < \frac{\partial Q^b}{\partial P_1}(z) < u$, almost surely P_1, and
2. $\lambda(b)$ solving the dual problem is in the interior of $\{\lambda \; : \; E[|f^*(\lambda^\intercal h(Z, b))| \mid D = 1] < \infty\}$.

Estimation: inference

Assumption: estimation

Assumption 3 *(Estimation)* Suppose that

1. $\mathbf{B}_0$ is closed,
2. $\min_{b \in B \cap \mathbf{B}_0} \nu(b)$ has a unique solution,
3. the matrix $E[h(Y, X, b)h(Y, X, b)^{\mathsf{T}} \mid D = 1]$ is nonsingular for each $b \in B$,
4. $g(y, x, b)$ is continuously differentiable with respect to b for each (y, x), and
5. there exists a convex, compact set Θ^B containing $\mathrm{Gr}(\theta_0)^{\eta}$ for some $\eta > 0$ satisfying

$$E\left[\sup_{(b,\theta)\in\Theta^B}\|\phi(D, DY, X, b, \theta)\|^2\right] < \infty, \text{ and } E\left[\left(\sup_{(b,\theta)\in\Theta^B}\|\nabla_{(b,\theta)}\phi(D, DY, X, b, \theta)\|\right)^2\right] < \infty.$$

Theorem Under assumptions 1, 2, and 3,

$$\frac{\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})}{\hat{\sigma}_n} \xrightarrow{d} N(0, 1),$$

where $\hat{\sigma}_n^2$ is a consistent variance estimator given in the paper.

Estimation: inference