

# Robustness to Missing Data: Breakdown Point Analysis

Daniel Ober-Reynolds

University of California, Los Angeles

*doberreynolds@gmail.com*

23 June 2023

# 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

# 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

# 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

# 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

# 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”

# 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%**



# 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%**
  - By linking data with SSA tax records, show **missing earnings data is not MAR**

# 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%**
  - By linking data with SSA tax records, show **missing earnings data is not MAR**
- ▶ Finkelstein et al. (2012), “The Oregon Health Insurance Experiment: Evidence From the First Year”

# 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%**
  - By linking data with SSA tax records, show **missing earnings data is not MAR**
- ▶ Finkelstein et al. (2012), “The Oregon Health Insurance Experiment: Evidence From the First Year”
  - Survey data shows Medicaid improved self-reported physical/mental health

# 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%**
  - By linking data with SSA tax records, show **missing earnings data is not MAR**
- ▶ Finkelstein et al. (2012), “The Oregon Health Insurance Experiment: Evidence From the First Year”
  - Survey data shows Medicaid improved self-reported physical/mental health
  - **Only 50% of survey recipients responded.**

# 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%**
  - By linking data with SSA tax records, show **missing earnings data is not MAR**
- ▶ Finkelstein et al. (2012), “The Oregon Health Insurance Experiment: Evidence From the First Year”
  - Survey data shows Medicaid improved self-reported physical/mental health
  - **Only 50% of survey recipients responded.**
  - When Lee (2009) sample selection bounds were applied, **this conclusion could no longer be supported.**

# Related literature

## ► Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ► Robustness/sensitivity analysis

- Missing data: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- Robustness/sensitivity analysis: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- Missing data: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ 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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- Robustness/sensitivity analysis: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- Missing data: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ 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)
- iv. gives results that are succinct and interpretable



# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ▶ Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ▶ Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Related literature

## ► Missing data without MAR

- Point identification: Heckman (1979), Das et al. (2003)
- Partial identification: Manski (2005), Lee (2009)
- **Robustness/sensitivity analysis**: Kline and Santos (2013)

## ► Robustness/sensitivity analysis

- **Missing data**: Kline and Santos (2013)
- Potential outcomes: Masten and Poirier (2020)
- Omitted variable bias: Diegert et al. (2022)

⇒ **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)
- iv. gives results that are succinct and interpretable

# Overview

1 Introduction

2 Setting

3 Breakdown Point Analysis

4 Estimation

5 Simulations



# Overview

1 Introduction

2 **Setting**

3 Breakdown Point Analysis

4 Estimation

5 Simulations

# Setting

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

# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .

# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .
  - $Y$  may be a vector. If present,  $X_i$  is assumed finitely supported

# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .
  - $Y$  may be a vector. If present,  $X_i$  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.).

# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .
  - $Y$  may be a vector. If present,  $X_i$  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  $(Y, X)$ .

# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .
  - $Y$  may be a vector. If present,  $X_i$  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  $g(Y, X, b) = \begin{pmatrix} Y^{(2)} \\ X \end{pmatrix} (Y^{(1)} - (Y^{(2)}, X^\top)b)$

# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .
  - $Y$  may be a vector. If present,  $X_i$  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  $g(Y, X, b) = \begin{pmatrix} Y^{(2)} \\ X \end{pmatrix} (Y^{(1)} - (Y^{(2)}, X^\top)b)$
- ▶ Conclusion to be investigated is that  $\beta$  is outside  $\mathbf{B}_0$

$$H_0 : \beta \in \mathbf{B}_0$$

vs

$$H_1 : \beta \in \mathbf{B} \setminus \mathbf{B}_0$$



# Setting

- ▶ Data is i.i.d. sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , where  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .
  - Variables of interest are  $(Y, X) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ .
  - $Y$  may be a vector. If present,  $X_i$  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  $g(Y, X, b) = \begin{pmatrix} Y^{(2)} \\ X \end{pmatrix} (Y^{(1)} - (Y^{(2)}, X^\top)b)$
- ▶ Conclusion to be investigated is that  $\beta$  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

$$(Y, X) \mid D = 1 \sim P_1, \quad (Y, X) \mid D = 0 \sim P_0,$$

$$P = p_D P_1 + (1 - p_D) P_0$$

# Setting

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

$$(Y, X) \mid D = 1 \sim P_1, \quad (Y, X) \mid D = 0 \sim P_0,$$

$$P = p_D P_1 + (1 - p_D) P_0$$

- The sample  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , identifies  $p_D$ ,  $P_1$ , and  $P_{0X} \dots$

# Setting

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

$$(Y, X) \mid D = 1 \sim P_1, \quad (Y, X) \mid D = 0 \sim P_0,$$

$$P = p_D P_1 + (1 - p_D) P_0$$

- 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  $\beta$  solving  $E_P[g(Y, X, \beta)] = 0$

# Setting

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

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

- 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  $\beta$  solving  $E_P[g(Y, X, \beta)] = 0$

- Common solution: estimate  $\beta_1$  instead

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

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

# Setting

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

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

- 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  $\beta$  solving  $E_P[g(Y, X, \beta)] = 0$

- Common solution: estimate  $\beta_1$  instead

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

MCAR is the assumption  $P_0 = P_1$ . 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?

# Setting

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

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

- 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  $\beta$  solving  $E_P[g(Y, X, \beta)] = 0$

- Common solution: estimate  $\beta_1$  instead

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

MCAR is the assumption  $P_0 = P_1$ . 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.
- Hope to defend  $\beta \in \mathbf{B} \setminus \mathbf{B}_0$

MAR?

# Setting

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

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

- 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  $\beta$  solving  $E_P[g(Y, X, \beta)] = 0$

- Common solution: estimate  $\beta_1$  instead

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

MCAR is the assumption  $P_0 = P_1$ . 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?



# Setting

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

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

- 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  $\beta$  solving  $E_P[g(Y, X, \beta)] = 0$

- Common solution: estimate  $\beta_1$  instead

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

MCAR is the assumption  $P_0 = P_1$ . 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: predictive power of $(Y, X)$

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

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

- Selection is a greater concern when context suggests  $(Y, X)$  would predict  $D$  well

# Quantifying selection: predictive power of $(Y, X)$

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

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

- ▶ 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

# Quantifying selection: predictive power of $(Y, X)$

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

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

- ▶ 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
- ▶ See this formally with densities. Let  $f_1, f_0$  be densities of  $P_1, P_0$  wrt  $P$ . Then

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

where  $p_D(y, x) = P(D = 1 \mid Y = y, X = x)$ .

# Quantifying selection: predictive power of $(Y, X)$

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

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

- ▶ 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
- ▶ See this formally with densities. Let  $f_1, f_0$  be densities of  $P_1, P_0$  wrt  $P$ . Then

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

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  $f_1 = f_0$  (data is MCAR)

# Quantifying selection: predictive power of $(Y, X)$

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

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

- ▶ 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
- ▶ See this formally with densities. Let  $f_1, f_0$  be densities of  $P_1, P_0$  wrt  $P$ . Then

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

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  $f_1 = f_0$  (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  $P$ .

# 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  $P$ .

- ▶  $f_1(y, x) = p_D(y, x)/p_D$  and  $f_0(y, x) = (1 - p_D(y, x))/p_D$  implies

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



# 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  $P$ .

- ▶  $f_1(y, x) = p_D(y, x)/p_D$  and  $f_0(y, x) = (1 - p_D(y, x))/p_D$  implies

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

- **Interpretation:** expected percent standard deviation of  $D$  “explained” by  $(Y, X)$

# 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  $P$ .

- ▶  $f_1(y, x) = p_D(y, x)/p_D$  and  $f_0(y, x) = (1 - p_D(y, x))/p_D$  implies

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

- **Interpretation**: expected percent standard deviation of  $D$  “explained” by  $(Y, X)$
- **Captures intuition**: more predictive power, higher selection

# 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  $P$ .

- ▶  $f_1(y, x) = p_D(y, x)/p_D$  and  $f_0(y, x) = (1 - p_D(y, x))/p_D$  implies

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

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

# 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  $P$ .

- ▶  $f_1(y, x) = p_D(y, x)/p_D$  and  $f_0(y, x) = (1 - p_D(y, x))/p_D$  implies

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

- **Interpretation**: expected percent standard deviation of  $D$  “explained” by  $(Y, X)$
  - **Captures intuition**: more predictive power, higher selection
  - Range is  $[0, 1]$ . Equals 0  $\Leftrightarrow \text{Var}(D | Y, X) = \text{Var}(D)$ , equals 1  $\Leftrightarrow \text{Var}(D | Y, X) = 0$
- ▶ Assumption:  $P_0$  is dominated by  $P_1$ . Domination

# 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  $P$ .

- ▶  $f_1(y, x) = p_D(y, x)/p_D$  and  $f_0(y, x) = (1 - p_D(y, x))/p_D$  implies

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

- **Interpretation**: expected percent standard deviation of  $D$  “explained” by  $(Y, X)$
  - **Captures intuition**: more predictive power, higher selection
  - Range is  $[0, 1]$ . Equals 0  $\Leftrightarrow \text{Var}(D | Y, X) = \text{Var}(D)$ , equals 1  $\Leftrightarrow \text{Var}(D | Y, X) = 0$
- ▶ Assumption:  $P_0$  is dominated by  $P_1$ . Domination
    - Rules out selection mechanisms that “truncate” data; e.g.  $D_i = \mathbb{1}\{Y_i \leq c\}$ .

# Recap

► Setting:

# Recap

► Setting:

- Model:  $E_P[g(Y, X, \beta)] = 0$

# Recap

► Setting:

- Model:  $E_P[g(Y, X, \beta)] = 0$
- Hypothesis test:  $H_0 : \beta \in \mathbf{B}_0$  vs  $H_1 : \beta \in \mathbf{B} \setminus \mathbf{B}_0$



# Recap

## ► Setting:

- Model:  $E_P[g(Y, X, \beta)] = 0$
- Hypothesis test:  $H_0 : \beta \in \mathbf{B}_0$  vs  $H_1 : \beta \in \mathbf{B} \setminus \mathbf{B}_0$
- Data:  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$  i.i.d.. with  $D_i = \mathbb{1}\{Y_i \text{ is observed}\}$ .

# Recap

## ► Setting:

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

# Recap

## ► Setting:

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

# Recap

## ► Setting:

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

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

# Recap

## ► Setting:

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

►  $\beta_1$  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?

# Recap

## ► Setting:

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

## ► $\beta_1$ 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

1 Introduction

2 Setting

3 Breakdown Point Analysis

4 Estimation

5 Simulations

# 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**  $b$ .



# 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  $\beta \in \mathbf{B}_0$ :

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

# 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  $\beta \in \mathbf{B}_0$ :

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

# 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  $\beta \in \mathbf{B}_0$ :

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

# 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  $\delta^{BP}$  **assuage selection concerns**

# 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  $\delta^{BP}$  **assuage selection concerns**

- The claim  $\beta \in \mathbf{B}_0$  implies  $\delta^{BP} \leq \frac{1}{2} H^2(P_0, P_1) = 1 - E_P \left[ \sqrt{\text{Var}(D \mid Y, X)} \right] / \sqrt{\text{Var}(D)}$

# 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  $\delta^{BP}$  **assuage selection concerns**

- The claim  $\beta \in \mathbf{B}_0$  implies  $\delta^{BP} \leq \frac{1}{2} H^2(P_0, P_1) = 1 - E_P \left[ \sqrt{\text{Var}(D | Y, X)} \right] / \sqrt{\text{Var}(D)}$
- If the claim  $(Y, X)$  predicts  $D$  this well is implausible, then  $\beta \in \mathbf{B}_0$  is implausible.

# 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  $\delta^{BP}$  **assuage selection concerns**

- The claim  $\beta \in \mathbf{B}_0$  implies  $\delta^{BP} \leq \frac{1}{2} H^2(P_0, P_1) = 1 - E_P \left[ \sqrt{\text{Var}(D | Y, X)} \right] / \sqrt{\text{Var}(D)}$
- 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

# 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  $\delta^{BP}$  **assuage selection concerns**

- The claim  $\beta \in \mathbf{B}_0$  implies  $\delta^{BP} \leq \frac{1}{2} H^2(P_0, P_1) = 1 - E_P \left[ \sqrt{\text{Var}(D | Y, X)} \right] / \sqrt{\text{Var}(D)}$
- 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

- ▶  $\delta^{BP}$  is **point identified**



# 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  $\delta^{BP}$  **assuage selection concerns**

- The claim  $\beta \in \mathbf{B}_0$  implies  $\delta^{BP} \leq \frac{1}{2} H^2(P_0, P_1) = 1 - E_P \left[ \sqrt{\text{Var}(D | Y, X)} \right] / \sqrt{\text{Var}(D)}$
- 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

- ▶  $\delta^{BP}$  is **point identified**

- Reporting estimates  $\hat{\delta}_n^{BP}$  can facilitate selection concern discussions

Nominally Identified Sets

# 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  $\delta^{BP}$  **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

- ▶  $\delta^{BP}$  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**

Nominally Identified Sets

## Breakdown point: uniform expectation

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

## Breakdown point: uniform 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  $\beta = E[Y] \in \mathbb{R}$ .

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

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

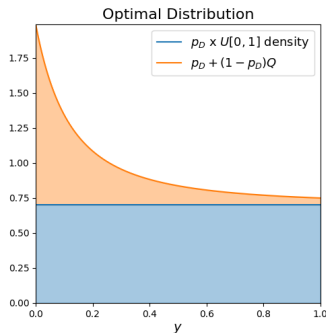
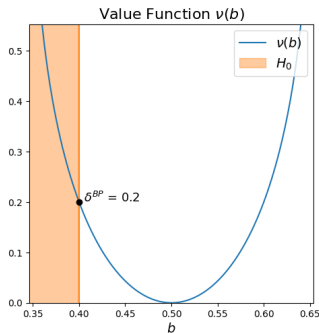
# Breakdown point: uniform expectation

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

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

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

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



# Overview

1 Introduction

2 Setting

3 Breakdown Point Analysis

4 Estimation

5 Simulations

# 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:

# 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.  $\hat{\nu}_n(b)$  estimates  $\nu(b) = \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)$



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

# 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.  $\hat{\nu}_n(b)$  estimates  $\nu(b) = \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)$
  2. Plug-in second step  $\hat{\delta}_n^{BP} = \inf_{b \in \mathbf{B}_0} \nu(b)$
- $\hat{\nu}_n(b)$  based on finite dimensional, well-behaved **dual problem**

# 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.  $\hat{\nu}_n(b)$  estimates  $\nu(b) = \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)$
  2. Plug-in second step  $\hat{\delta}_n^{BP} = \inf_{b \in \mathbf{B}_0} \nu(b)$
- ▶  $\hat{\nu}_n(b)$  based on finite dimensional, well-behaved **dual problem**
  - ▶ Second stage estimator analyzed using **functional delta method**

# 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.  $\hat{\nu}_n(b)$  estimates  $\nu(b) = \inf_{Q \in \mathbf{P}^b} H^2(Q, P_1)$
  2. Plug-in second step  $\hat{\delta}_n^{BP} = \inf_{b \in \mathbf{B}_0} \nu(b)$
- ▶  $\hat{\nu}_n(b)$  based on finite dimensional, well-behaved **dual problem**
  - ▶ Second stage estimator analyzed using **functional delta method**
  - ▶ Lower confidence intervals constructed using **bootstrap** procedure

# Duality

- The **primal problem** is

$$\nu(b) = \inf_{Q \in \mathcal{P}^b} H^2(Q, P_1) \quad (1)$$

# Duality

- ▶ The **primal problem** is

$$\nu(b) = \inf_{Q \in \mathcal{P}^b} H^2(Q, P_1) \quad (1)$$

- ▶ The **dual problem** is

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

a finite dimensional convex optimization problem.

# Duality

- The **primal problem** is

$$\nu(b) = \inf_{Q \in \mathcal{P}^b} H^2(Q, P_1) \quad (1)$$

- The **dual problem** is

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

a **finite dimensional convex optimization** problem.

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

# Duality

- The **primal problem** is

$$\nu(b) = \inf_{Q \in \mathcal{P}^b} H^2(Q, P_1) \quad (1)$$

- The **dual problem** is

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

a **finite dimensional convex optimization** problem.

- $f^*$ ,  $J$  and  $h$  are **known functions**,
  - the expectation is **wrt the distribution of  $(D, DY, X)$** , and
  - $K$  is the cardinality of  $\text{Supp}(X)$ .
- Under regularity conditions, **strong duality** holds:

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



# Duality

- The **primal problem** is

$$\nu(b) = \inf_{Q \in \mathcal{P}^b} H^2(Q, P_1) \quad (1)$$

- The **dual problem** is

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

a **finite dimensional convex optimization** problem.

- $f^*$ ,  $J$  and  $h$  are **known functions**,
  - the expectation is **wrt the distribution of  $(D, DY, X)$** , and
  - $K$  is the cardinality of  $\text{Supp}(X)$ .
- Under regularity conditions, **strong duality** holds:

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

- Assume this holds for all  $b \in B \subseteq \mathbf{B}$ , with  $\inf_{b \in B_0} \nu(b) = \inf_{b \in B \cap B_0} \nu(b)$
- $\implies$  **we can focus on the dual problem.**

# Estimators

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

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

# Estimators

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

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

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

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

# Estimators

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

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

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

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

- ▶ Under additional regularity conditions, estimators are **consistent**:

$$\hat{\nu}_n \xrightarrow{P} \nu \quad \text{in } \ell^\infty(B), \quad \hat{\delta}_n^{BP} \xrightarrow{P} \delta^{BP}$$

# Inference: asymptotic distributions

**Theorem** Under **assumptions** discussed in the paper,

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu \quad \text{in } \ell^\infty(B)$$

# Inference: asymptotic distributions

**Theorem** Under **assumptions** discussed in the paper,

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu \quad \text{in } \ell^\infty(B)$$

► Intuition: for a fixed  $b$ , view estimation as GMM:

$$\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) - \hat{\nu}_n(b) \\ \nabla_{\lambda} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) \\ D_i - \hat{\rho}_{D,n} \end{pmatrix} = 0$$

which is asymptotically linear. This linearization is shown to hold **uniformly** over  $b \in B$ .

# Inference: asymptotic distributions

**Theorem** Under **assumptions** discussed in the paper,

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu \quad \text{in } \ell^\infty(B)$$

► Intuition: for a fixed  $b$ , view estimation as GMM:

$$\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) - \hat{\nu}_n(b) \\ \nabla_\lambda \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) \\ D_i - \hat{\rho}_{D,n} \end{pmatrix} = 0$$

which is asymptotically linear. This linearization is shown to hold **uniformly** over  $b \in B$ .

**Theorem** Suppose the same **assumptions** hold. Then  $\mathbf{m}(\nu) = \arg \min_{b \in B \cap \mathbf{B}_0} \nu(b)$  is nonempty and

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \xrightarrow{L} \inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_\nu(b)$$

# Inference: asymptotic distributions

**Theorem** Under **assumptions** discussed in the paper,

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu \quad \text{in } \ell^\infty(B)$$

- Intuition: for a fixed  $b$ , view estimation as GMM:

$$\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) - \hat{\nu}_n(b) \\ \nabla_{\lambda} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) \\ D_i - \hat{\rho}_{D,n} \end{pmatrix} = 0$$

which is asymptotically linear. This linearization is shown to hold **uniformly** over  $b \in B$ .

**Theorem** Suppose the same **assumptions** hold. Then  $\mathbf{m}(\nu) = \arg \min_{b \in B \cap \mathbf{B}_0} \nu(b)$  is nonempty and

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \xrightarrow{L} \inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_\nu(b)$$

- Follows from Hadamard directional differentiability of  $\nu \mapsto \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$  and the **functional delta method** (Fang and Santos (2019)).



# Inference: asymptotic distributions

**Theorem** Under **assumptions** discussed in the paper,

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_\nu \quad \text{in } \ell^\infty(B)$$

- Intuition: for a fixed  $b$ , view estimation as GMM:

$$\frac{1}{n} \sum_{i=1}^n \begin{pmatrix} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) - \hat{\nu}_n(b) \\ \nabla_{\lambda} \varphi(D_i, D_i Y_i, X_i, b, \hat{\lambda}_n(b), \hat{\rho}_{D,n}) \\ D_i - \hat{\rho}_{D,n} \end{pmatrix} = 0$$

which is asymptotically linear. This linearization is shown to hold **uniformly** over  $b \in B$ .

**Theorem** Suppose the same **assumptions** hold. Then  $\mathbf{m}(\nu) = \arg \min_{b \in B \cap \mathbf{B}_0} \nu(b)$  is nonempty and

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \xrightarrow{L} \inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_\nu(b)$$

- Follows from Hadamard directional differentiability of  $\nu \mapsto \inf_{b \in B \cap \mathbf{B}_0} \nu(b)$  and the **functional delta method** (Fang and Santos (2019)).
- $\mathbf{m}(\nu)$  is plausibly a singleton:  $\{b^i\}$ . If so,  $\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP})$  is **asymptotically normal**.

# Inference: lower confidence intervals

- ▶ A large  $\delta^{BP}$  assuages selection concerns

# Inference: lower confidence intervals

- ▶ A large  $\delta^{BP}$  assuages selection concerns
- ▶ Skeptical readers may worry  $\hat{\delta}_n^{BP} > \delta^{BP}$  due to sample noise

# Inference: lower confidence intervals

- ▶ A large  $\delta^{BP}$  assuages selection concerns
- ▶ Skeptical readers may worry  $\hat{\delta}_n^{BP} > \delta^{BP}$  due to sample noise
  - The argument is only strengthened if  $\hat{\delta}_n^{BP} < \delta^{BP}$

# Inference: lower confidence intervals

- ▶ A large  $\delta^{BP}$  **assuages selection concerns**
- ▶ Skeptical readers may worry  $\hat{\delta}_n^{BP} > \delta^{BP}$  due to sample noise
  - The argument is only strengthened if  $\hat{\delta}_n^{BP} < \delta^{BP}$
- ▶ Reporting a **lower confidence interval** addresses this concern:

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

# Inference: lower confidence intervals

- ▶ A large  $\delta^{BP}$  **assuages selection concerns**
- ▶ Skeptical readers may worry  $\hat{\delta}_n^{BP} > \delta^{BP}$  due to sample noise
  - The argument is only strengthened if  $\hat{\delta}_n^{BP} < \delta^{BP}$
- ▶ Reporting a **lower confidence interval** addresses this concern:

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

- ▶  $\hat{c}_{1-\alpha, n}$  is estimated with the **score bootstrap**

# Inference: lower confidence intervals

- ▶ A large  $\delta^{BP}$  **assuages selection concerns**
- ▶ Skeptical readers may worry  $\hat{\delta}_n^{BP} > \delta^{BP}$  due to sample noise
  - The argument is only strengthened if  $\hat{\delta}_n^{BP} < \delta^{BP}$
- ▶ Reporting a **lower confidence interval** addresses this concern:

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

- ▶  $\hat{c}_{1-\alpha, n}$  is estimated with the **score bootstrap**
  - Assuming  $\mathbf{m}(\nu) = \arg \min_{b \in B \cap \mathbf{B}_0} \nu(b)$  is the singleton  $\{b^i\}$ ,  $\hat{c}_{1-\alpha, n}$  is computed with a **computationally convenient procedure**

Score Bootstrap

# Overview

1 Introduction

2 Setting

3 Breakdown Point Analysis

4 Estimation

5 Simulations



# Simulations: uniform expectation

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

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

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

Illustration

# Simulations: uniform expectation

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

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

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

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

**Table:** Simulations, Squared Hellinger, Uniform, Mean

n	RMSE	Emp. Bias	Emp. CI Coverage	Ave. CI Length
1000	0.060	0.008	98.4	0.091
2000	0.040	0.005	97.6	0.063
3000	0.032	0.001	96.8	0.051
5000	0.024	0.003	96.4	0.040

Illustration

# Simulations: OLS

- ▶ Consider a linear model

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 Y_2 + \beta_3 X_2 + \varepsilon = W^T \beta + \varepsilon, \quad E[W\varepsilon] = 0$$

where  $X_1, X_2$  are discrete and  $Y_1, Y_2$  are continuous.

- ▶ The conclusion to be investigated is  $H_1 : \beta_1 > 0$ . The observed data is  $\{D_i, D_i Y_{i1}, D_i Y_{i2}, X_{i1}, X_{i2}\}_{i=1}^n$ .

# Simulations: OLS

- ▶ Consider a linear model

$$Y_1 = \beta_0 + \beta_1 X_1 + \beta_2 Y_2 + \beta_3 X_2 + \varepsilon = W^T \beta + \varepsilon, \quad E[W\varepsilon] = 0$$

where  $X_1, X_2$  are discrete and  $Y_1, Y_2$  are continuous.

- ▶ The conclusion to be investigated is  $H_1 : \beta_1 > 0$ . The observed data is  $\{D_i, D_i Y_{i1}, D_i Y_{i2}, X_{i1}, X_{i2}\}_{i=1}^n$ .
- ▶ 250 simulations from a DGP with  $P(D = 1) \approx 0.7$ , and  $\delta^{BP} \approx 0.2$ :

Table: Simulations, Squared Hellinger, OLS

n	RMSE	Emp. Bias	Emp. CI Coverage	Ave. CI Length
1000	0.043	0.009	100.0	0.078
2000	0.033	0.005	98.0	0.052
3000	0.026	0.007	98.0	0.043
5000	0.017	0.002	98.0	0.032

- ▶ Empirical coverage suggests inference is conservative.

# Conclusion

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

# 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.

# 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 breakdown point  $\delta^{BP}$  is  $\sqrt{n}$ -estimable, and lower confidence intervals can be constructed with simple bootstrap procedures.

# 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 breakdown point  $\delta^{BP}$  is  $\sqrt{n}$ -estimable, and lower confidence intervals can be constructed with simple bootstrap procedures.
- ▶ 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

- 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.
- 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 (2022). "Assessing Omitted Variable Bias when the Controls are Endogenous". In: *arXiv preprint arXiv:2206.02303*.
- Fang, Zheng and Andres Santos (2019). "Inference on directionally differentiable functions". In: *The Review of Economic Studies* 86(1), pp. 377–412.
- Finkelstein, Amy et al. (2012). "The Oregon health insurance experiment: evidence from the first year". In: *The Quarterly journal of economics* 127(3), pp. 1057–1106.
- Heckman, James J (1979). "Sample selection bias as a specification error". In: *Econometrica: Journal of the econometric society*, pp. 153–161.
- 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.
- Lee, David S (2009). "Training, wages, and sample selection: Estimating sharp bounds on treatment effects". In: *The Review of Economic Studies* 76(3), pp. 1071–1102.
- 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.

# Missing (completely) at random

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

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

two common assumptions restore point identification of  $P$

- ▶ **Missing completely at random (MCAR)** assumes  $P_0 = P_1$ 
  - 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 imputing  $Y \mid X = x, D = 0$  based on distribution of  $Y \mid X = x, D = 1$
- ▶ Preliminary analysis may be based on **either assumption**.

# 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  $P_1$
  - Allows squared Hellinger to be written as an  $f$ -divergence
- ▶ Some support assumption is typically necessary for an interesting exercise.
  - **Example:**  $\beta = E[Y]$ .  $P_1$  and  $P_0$  given by

$$\begin{array}{lll} P_1(Y = -1) = 0.5 & P_1(Y = 1) = 0.5 & \\ P_0(Y = -1) = 0.5 & P_0(Y = 1) = 0.5 - \alpha & P_0(Y = y) = \alpha \end{array}$$

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  $y \in \mathbb{R}$ .

## Other selection measures: $f$ -divergences

- Given a convex function  $f : \mathbb{R} \rightarrow [0, \infty]$  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

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

- 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

# 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. (2022).
- ▶ 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  $\beta$  grows with  $\delta$ . As long as  $\mathbf{B}_{ID}(\delta) \subseteq \mathbf{B} \setminus \mathbf{B}_0$ , it is clear the conclusion holds.
- ▶ The **breakdown point**  $\delta^{BP}$  can then be defined as either:
  1. the largest  $\delta$  for which  $\mathbf{B}_{ID}(\delta) \subseteq \mathbf{B} \setminus \mathbf{B}_0$ , or
  2. the smallest  $\delta$  for which  $\mathbf{B}_{ID}(\delta) \cap \mathbf{B}_0 \neq \emptyset$

Breakdown Point

# Dual problem (detailed)

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

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

where

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

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

and  $\{x_1, \dots, x_K\}$  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.

# Formal 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

- (i)  $p_D = P(D = 1) \in (0, 1)$
- (ii)  $X \mid D = 1$  and  $X \mid D = 0$  have the same finite support  $\{x_1, \dots, x_K\}$
- (iii)  $E[\sup_{b \in \mathbf{B}} \|g(Y, X, b)\| \mid D = 1] < \infty$

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

- (i) there exists  $Q^b \in \mathbf{P}^b$  such that  $0 < \frac{\partial Q^b}{\partial P_1}(y, x) < \infty$ , almost-surely  $P_1$ .
- (ii)  $\lambda(b)$  solving the dual problem is in the interior of  $\{\lambda ; E[|f^*(\lambda^\top h(Y, X, b))| \mid D = 1] < \infty$ .

# Formal assumptions: consistency

## Assumption 3 (Consistency)

- (i)  $B$  is compact
- (ii)  $g(y, x, b)$  is continuous in  $b$  for all  $(y, x)$
- (iii) For each  $b \in B$ ,  $\{h_j(y, x, b)\}_{j=1}^{d_g+K}$  are linearly independent in the sense that for any  $\lambda \neq 0 \in \mathbb{R}^{d_g+K}$ ,
$$P(\lambda^\top h(Y, X, b) \neq 0 \mid D = 1) > 0$$
- (iv) For each  $b \in B$ , there exists a closed convex  $\bar{\Lambda}^b$  with  $\lambda(b) \in \text{int}(\bar{\Lambda}^b)$  such that  $\bar{\Lambda}^B = \{(b, \lambda) ; b \in B, \lambda \in \bar{\Lambda}^b\}$  is compact, and for some open  $\mathcal{N} \subset \mathbb{R}$  containing  $p_D$ ,

$$E \left[ \sup_{p \in \mathcal{N}} \sup_{(b, \lambda) \in \bar{\Lambda}^B} |\varphi(D, DY, X, b, \lambda, p)| \right] < \infty,$$

$$E \left[ \sup_{(b, \lambda) \in \bar{\Lambda}^B} \|\nabla_\lambda \varphi(D, DY, X, b, \lambda, p_D)\| \right] < \infty, \quad E \left[ \sup_{(b, \lambda) \in \bar{\Lambda}^B} \|\nabla_\lambda^2 \varphi(D, DY, X, b, \lambda, p_D)\| \right] < \infty$$

If assumptions 1, 2, and 3 hold, then  $\hat{\nu}_n \xrightarrow{P} \nu$  in  $\ell^\infty(B)$  and  $\hat{\delta}_n^{BP} \xrightarrow{P} \delta^{BP}$ .



# Formal assumptions: inference

Let  $\theta(b) = (\nu(b), \lambda(b), p_D)$ ,  $\theta = (\nu, \lambda, p)$ ,

$$\phi(D, DY, X, b, \theta) = \phi(D, DY, X, b, \nu, \lambda, p) = \begin{pmatrix} \varphi(D, DY, X, b, \lambda, p) - \nu \\ \nabla_{\lambda} \varphi(D, DY, X, b, \lambda, p) \\ D - p \end{pmatrix},$$

$\Theta^b = \left\{ \theta = (\nu, \lambda, p) ; \nu \in [0, \bar{\nu}], \lambda \in \bar{\Lambda}^b, p \in [\underline{p}, \bar{p}] \right\}$ , and  $\theta^B = \left\{ (b, \theta) ; b \in B, \theta \in \Theta^b \right\}$ .

**Assumption 4** (Inference) Suppose that

- (i)  $B_0$  is closed
- (ii)  $B$  is convex
- (iii)  $g(z, b)$  is continuously differentiable with respect to  $b$
- (iv)  $\hat{\theta}_n(b) = (\hat{\nu}_n(b), \hat{\lambda}_n(b), \hat{p}_{D,n}) \in \Theta^b$  for each  $b$
- (v) There exists  $F(d, dy, x)$  such that

$$\sup_{b \in B} \sup_{\theta \in \Theta^b} \|\nabla_{(b, \theta)} \phi(d, dy, x, b, \theta)\| \leq F(d, dy, x)$$

and  $E[F(D, DY, X)^2] < \infty$ .

If assumptions 1, 2, 3, and 4 hold, then

$$\sqrt{n}(\hat{\nu}_n - \nu) \xrightarrow{L} \mathbb{G}_{\nu} \text{ in } \ell^{\infty}(B),$$

and

$$\sqrt{n}(\hat{\delta}_n^{BP} - \delta^{BP}) \xrightarrow{L} \inf_{b \in \mathbf{m}(\nu)} \mathbb{G}_{\nu}(b) \text{ in } \mathbb{R}$$

# Score bootstrap

- ▶ Let  $\{W_i\}_{i=1}^n$  be i.i.d. scalars, independent of  $\{D_i, D_i Y_i, X_i\}_{i=1}^n$ , satisfying
  - (i)  $E[W] = 0$ ,
  - (ii)  $E[W^2] = 1$ , and
  - (iii)  $E[|W|^{2+a}] < \infty$  for some  $a > 0$ .
- ▶ Let  $\hat{\Phi}_n(b) = \frac{1}{n} \sum_{i=1}^n \nabla_{\theta} \phi(D_i, D_i Y_i, X_i, b, \hat{\theta}_n(b))$ ,

$$\hat{G}_n^*(b) = \hat{\Phi}_n(b)^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n W_i \phi(D_i, D_i Y_i, X_i, b, \hat{\theta}_n(b))$$

and  $\hat{G}_n^*(b, 1)$  be the first coordinate of the vector  $\hat{G}_n^*(b)$ .

## Bootstrap procedure

1. Compute  $\hat{b}_n^i = \arg \min_{b \in B \cap \mathbf{B}_0} \hat{\nu}_n(b)$ ,
2. Generate  $N$  bootstrap samples  $\{W_i\}_{i=1}^n$  from a distribution with moments described above, and compute  $\hat{G}_n^*(\hat{b}_n^i, 1)$  for each of the  $N$  bootstrap samples,
3. Let  $\hat{c}_{1-\alpha, n}$  be the  $1 - \alpha$  quantile of  $\{\hat{G}_{n, k}^*(\hat{b}_n^i, 1)\}_{k=1}^N$ .

If assumptions 1, 2, 3, and 4 hold, and  $\mathbf{m}(\nu) = \arg \min_{b \in B \cap \mathbf{B}_0} \nu(b)$  is the singleton  $\{b^i\}$ , then

$$\lim_{n \rightarrow \infty} P\left(\hat{\delta}_n^{BP} - \frac{1}{\sqrt{n}} \hat{c}_{1-\alpha, n} \leq \delta^{BP}\right) = 1 - \alpha.$$

Inference: lower confidence intervals