An approach for reliably leveraging machine learning predictions in subsequent statistical analyses

Based on joint work with: Kerri Lu, Tijana Zrnic, Sherrie Wang, and Stephen Bates

Kluger *et al.* "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling" (2025+) on arXiv

Lu *et al.* "Regression coefficient estimation from remote sensing maps" (2025). *Remote Sensing of Environment*

Dan Kluger

ASSA Annual Meeting

January 4, 2026

AI and ML opens doors to ask many scientific questions

AI and ML opens doors to ask many scientific questions

Predictions from ML models trained on Satellite-data
used in earth science research

AI and ML opens doors to ask many scientific questions

Predictions from ML models trained on Satellite-data
used in earth science research

Is it reliable to use predictions as if they are real data?

Prediction Errors $\sim \mathcal{N}(0,\sigma^2)$

AI and ML opens doors to ask many scientific questions

Predictions from ML models trained on Satellite-data
used in earth science research

Is it reliable to use predictions as if they are real data?

Prediction Errors $\sim \mathcal{N}(0,\sigma^2)$

Great interest in how to reliably use predictions from these complex ML models in downstream scientific analyses

AI and ML opens doors to ask many scientific questions

Predictions from ML models trained on Satellite-data
used in earth science research

Is it reliable to use predictions as if they are real data?

Prediction Errors $\sim \mathcal{N}(0,\sigma^2)$ ✘

Great interest in how to reliably use predictions from these complex ML models in downstream scientific analyses

Many fields are grappling with these questions, but will focus discussion on challenges the field of Remote Sensing is grappling with

Workflow pipelines in remote sensing

Workflow pipelines in remote sensing

Upstream Lab

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest
- Releases “dataset” of these predictions on a repository

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest
- Releases “dataset” of these predictions on a repository

Earth System Science Data

zenodo

DRYAD

scientific data

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest
- Releases “dataset” of these predictions on a repository

Downstream Lab

Earth System Science Data

zenodo

DRYAD

scientific data

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest
- Releases “dataset” of these predictions on a repository

Downstream Lab

- Downloads it

Earth System Science Data

zenodo

DRYAD

scientific data

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest
- Releases “dataset” of these predictions on a repository

Downstream Lab

- Downloads it
- Uses it in a statistical analysis as if it is real data

Earth System Science Data

zenodo

DRYAD

scientific data

Workflow pipelines in remote sensing

Upstream Lab

- Trains an ML model on satellite data to predict quantity of interest
- Releases “dataset” of these predictions on a repository

Downstream Lab

- Downloads it
- Uses it in a statistical analysis as if it is real data
- Prediction errors can bias statistical analyses…

Earth System Science Data

zenodo

DRYAD

scientific data

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either
    1. Only work for a particular model or data setup, or

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either
    1. Only work for a particular model or data setup, or
    2. Are difficult to understand or implement, or

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either
    1. Only work for a particular model or data setup, or
    2. Are difficult to understand or implement, or

<- This is an opinion :)

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either
    1. Only work for a particular model or data setup, or
    2. Are difficult to understand or implement, or
    3. Rely on distributional assumptions about prediction errors

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either
    1. Only work for a particular model or data setup, or
    2. Are difficult to understand or implement, or
    3. Rely on distributional assumptions about prediction errors
        - Assumptions on prediction errors of complex ML models difficult to reason about

Alarm bell about this issue has been raised

The Benefits and Pitfalls of Using Satellite Data for Causal Inference

Meha Jain

Remotely Incorrect? Accounting for Nonclassical Measurement Error in Satellite Data on Deforestation

Jennifer Alix-García and Daniel L. Millimet

- Recently a number of potential remedies have been proposed
- Widespread implementation of these methods remains elusive
- Many methods either
  1. Only work for a particular model or data setup, or
  2. Are difficult to understand or implement, or
  3. Rely on distributional assumptions about prediction errors
     - Assumptions on prediction errors of complex ML models difficult to reason about

Other application domains are similarly grappling with issue

Method desiderata

Method desiderata

1. No distributional assumptions about prediction errors

Method desiderata

1. No distributional assumptions about prediction errors
2. Easy to generalize to new statistical tasks and data setups

Method desiderata

1. No distributional assumptions about prediction errors
2. Easy to generalize to new statistical tasks and data setups
3. Easy for domain scientists to understand

Method desiderata

1. No distributional assumptions about prediction errors
2. Easy to generalize to new statistical tasks and data setups
3. Easy for domain scientists to understand

- We think approach originating Chen and Chen (2000) meets desiderata

Method desiderata

1. No distributional assumptions about prediction errors
2. Easy to generalize to new statistical tasks and data setups
3. Easy for domain scientists to understand

- We think approach originating Chen and Chen (2000) meets desiderata
  - more recently studied in: Tong et al. (2019), Yang and Ding (2020), Kremers (2021), Zrnic (2024), Miao and Lu (2024), and Gronsbell et al. (2024),…

Method desiderata

1. No distributional assumptions about prediction errors
2. Easy to generalize to new statistical tasks and data setups
3. Easy for domain scientists to understand

- We think approach originating Chen and Chen (2000) meets desiderata
  - more recently studied in: Tong et al. (2019), Yang and Ding (2020), Kremers (2021), Zrnic (2024), Miao and Lu (2024), and Gronsbell et al. (2024),…
  - We call it the *Predict-Then-Debias* approach (it has no consistent name)

Setting and notation

Random Vectors

Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$

Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed

Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)

Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

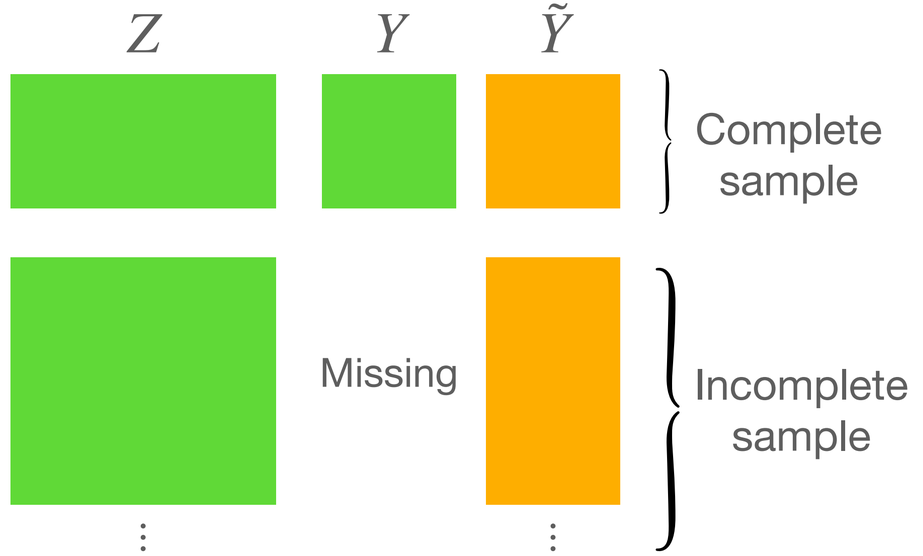
Partitioning of N samples

Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Partitioning of N samples

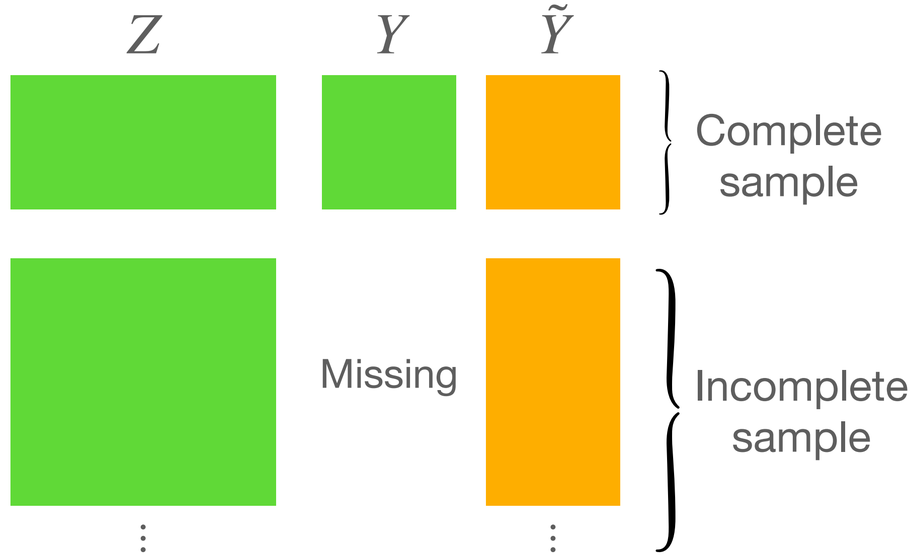


Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Partitioning of N samples



Dataset assumptions: IID and Y is missing completely at random

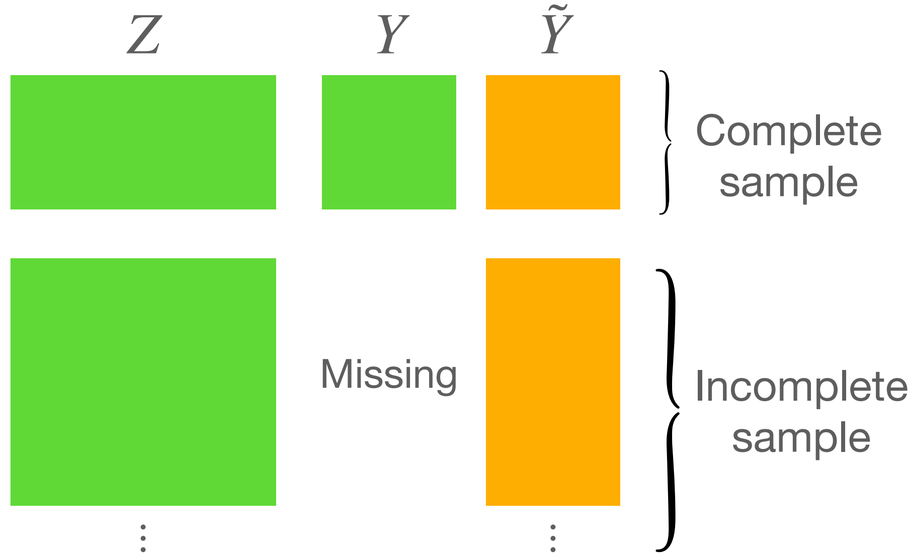
Setting and notation

Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

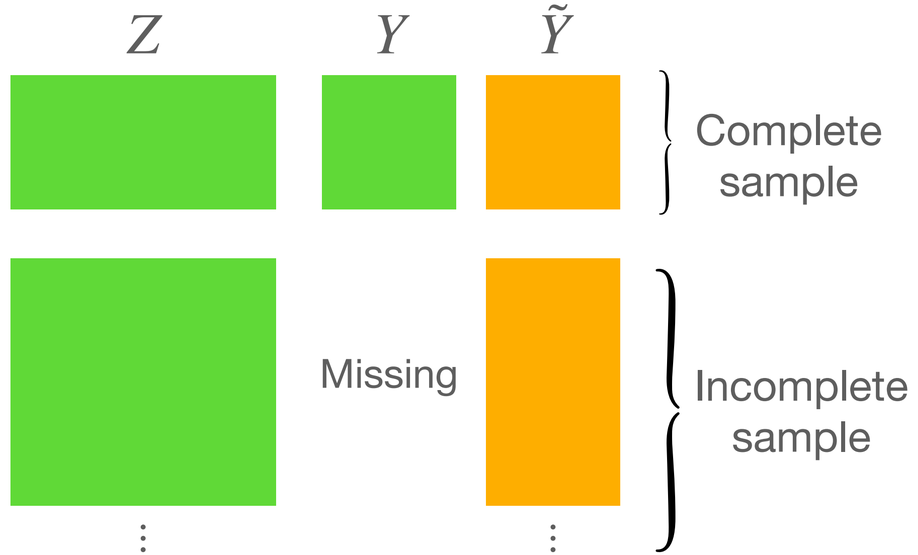
Random Vectors

- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

Random Vectors

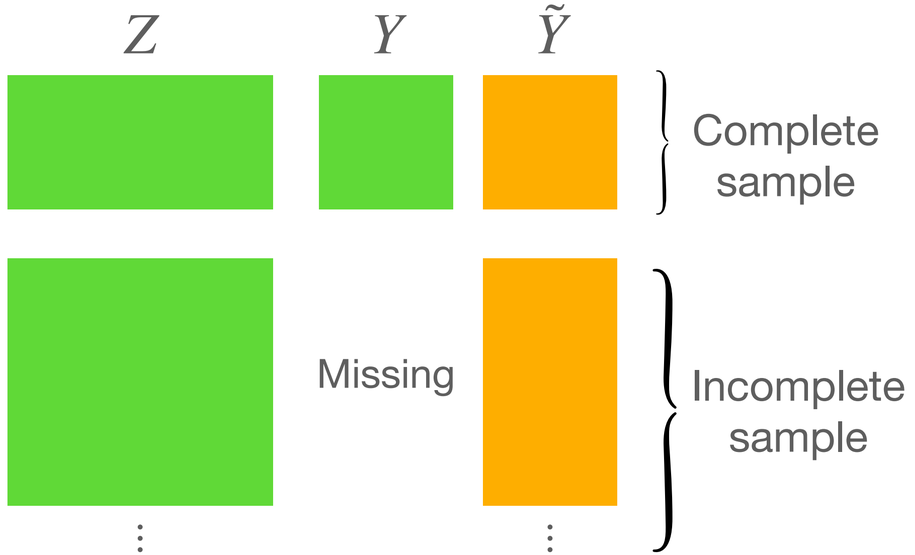
- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

- Estimate some parameter $\theta \in \mathbb{R}^d$ describing $\mathbb{P}$

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

Random Vectors

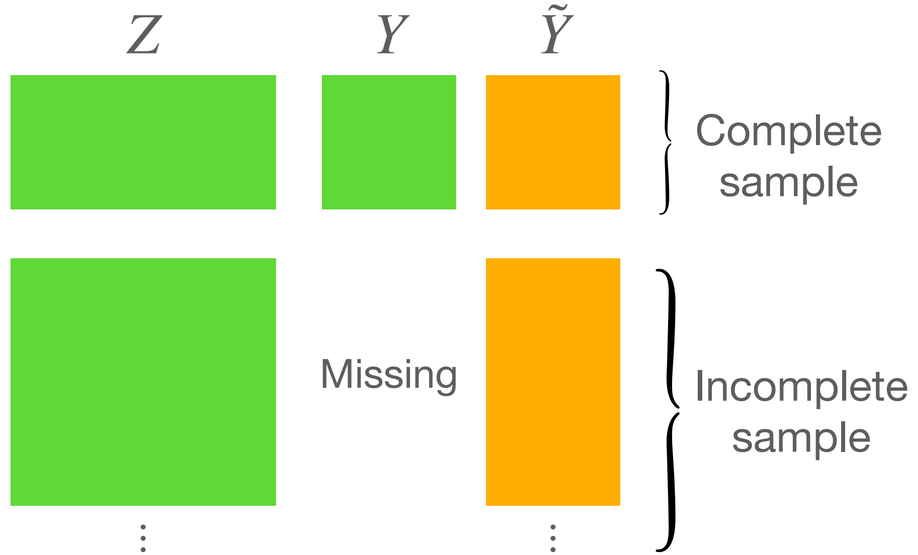
- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

- Estimate some parameter $\theta \in \mathbb{R}^d$ describing $\mathbb{P}$
  - Population means, quantiles, regression coefficients,…

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

Random Vectors

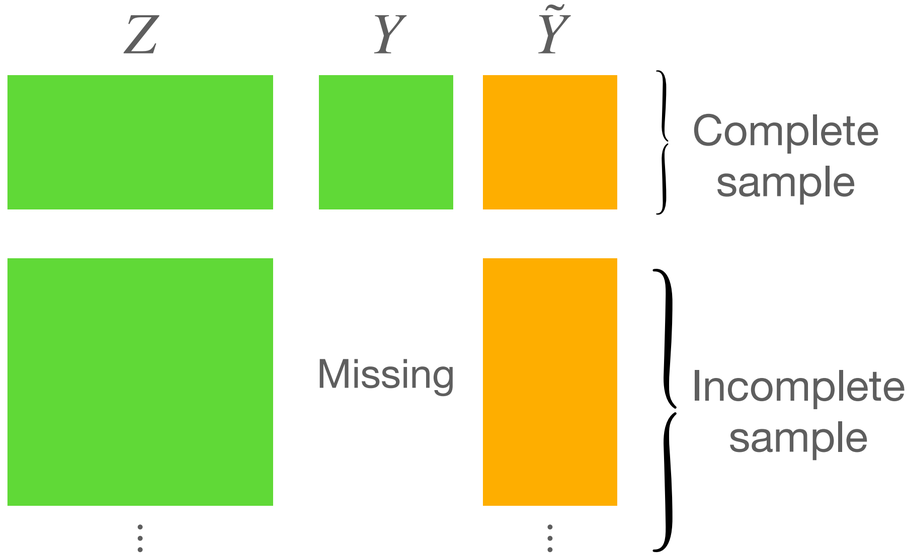
- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

- Estimate some parameter $\theta \in \mathbb{R}^d$ describing $\mathbb{P}$
  - Population means, quantiles, regression coefficients,…
  - Anything you can estimate if you had as much Y, Z data as desired

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

Random Vectors

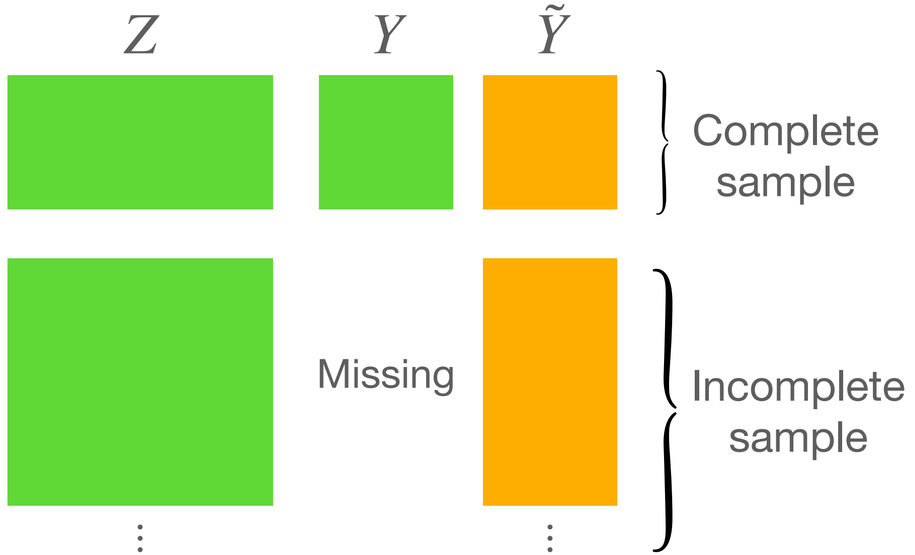
- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

- Estimate some parameter $\theta \in \mathbb{R}^d$ describing $\mathbb{P}$
  - Population means, quantiles, regression coefficients,…
  - Anything you can estimate if you had as much Y, Z data as desired
- Suppose algorithm $\mathscr{A}(\cdot)$

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

Random Vectors

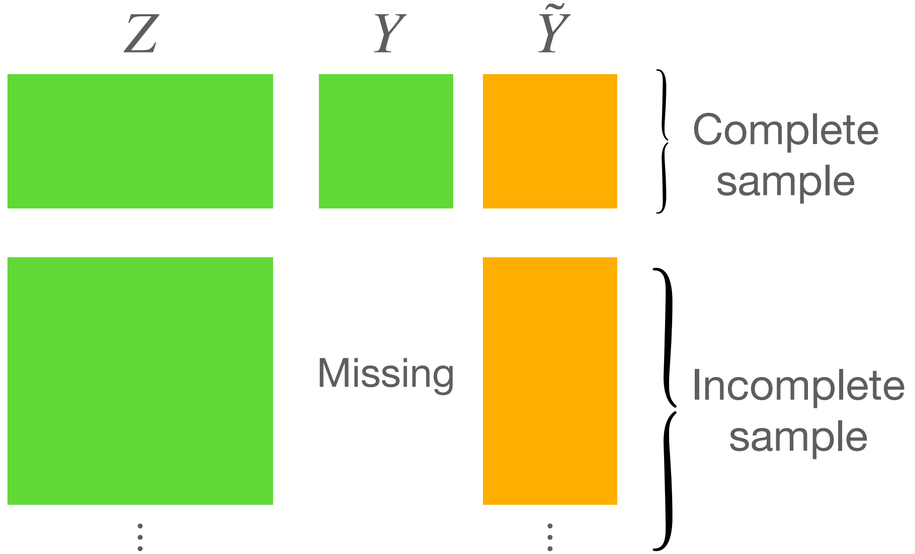
- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

- Estimate some parameter $\theta \in \mathbb{R}^d$ describing $\mathbb{P}$
  - Population means, quantiles, regression coefficients,…
  - Anything you can estimate if you had as much Y, Z data as desired
- Suppose algorithm $\mathscr{A}(\cdot)$
  - **Input:** a sample of (Y, Z) **Output:** estimate of θ

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

Setting and notation

Random Vectors

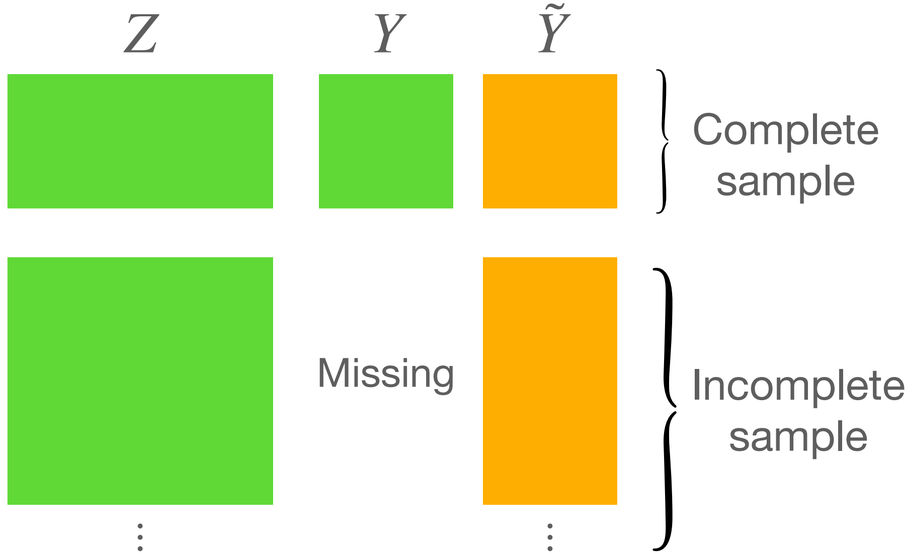
- Let $X \equiv (Y, Z) \in \mathbb{R}^p$ with $X \sim \mathbb{P}$
  - Z always observed
  - Y expensive to measure (few observations available)
- $\tilde{Y}$ is a widely available ML-based prediction of Y

Goal

- Estimate some parameter $\theta \in \mathbb{R}^d$ describing $\mathbb{P}$
  - Population means, quantiles, regression coefficients,…
  - Anything you can estimate if you had as much Y, Z data as desired
- Suppose algorithm $\mathscr{A}(\cdot)$
  - **Input:** a sample of (Y, Z) **Output:** estimate of θ

Partitioning of N samples

- $S^{\bullet} \cup S^{\bigcirc} = \{1, \ldots, N\}$, $S^{\bullet} \cap S^{\bigcirc} = \varnothing$
- A small complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^{\bullet}}$ of size n
- A large incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$



Dataset assumptions: IID and Y is missing completely at random

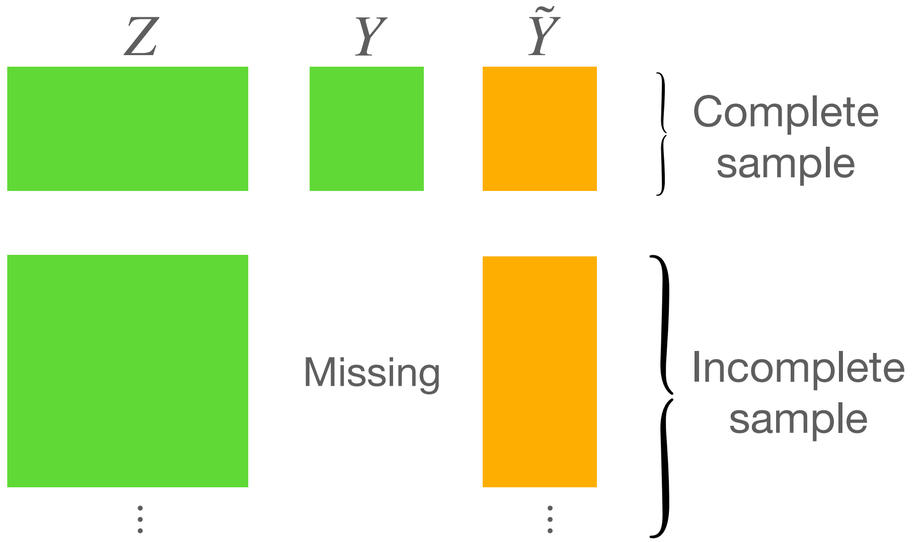
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S_{\bullet}}$

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S_{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



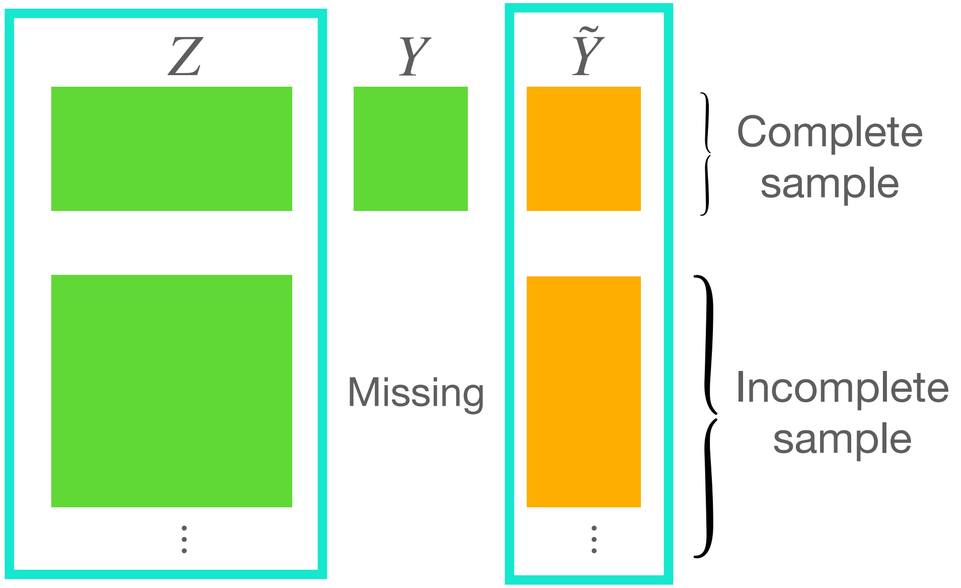
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^{N}\right)$

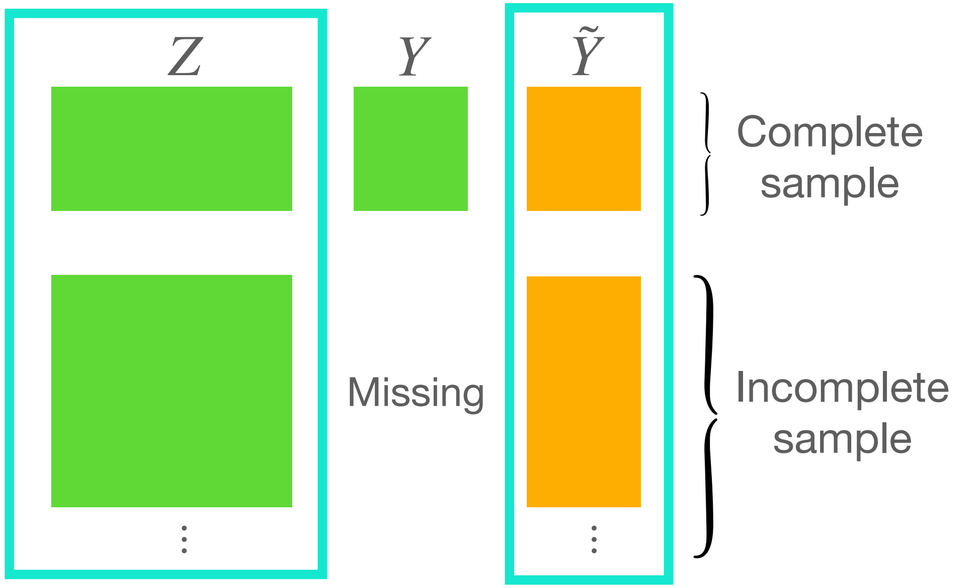
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^{N}\right)$

- Low variance: Uses all samples ✓

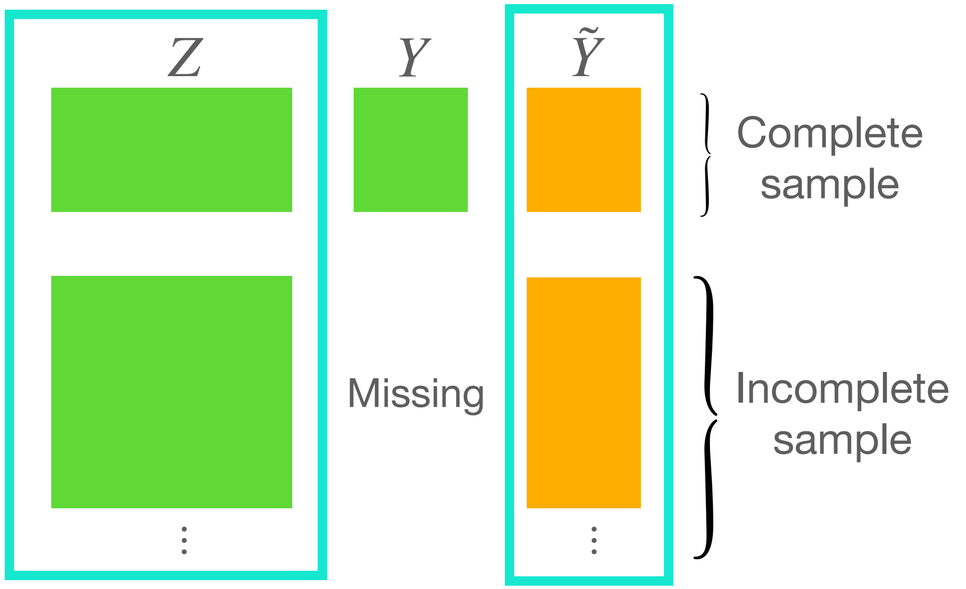
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^N\right)$

- Low variance: Uses all samples ✓
- Biased: it targets $\gamma \neq \theta$ (i.e., $\text{plim}_{N \to \infty} \hat{\gamma}^{\text{all}} = \gamma \neq \theta$)

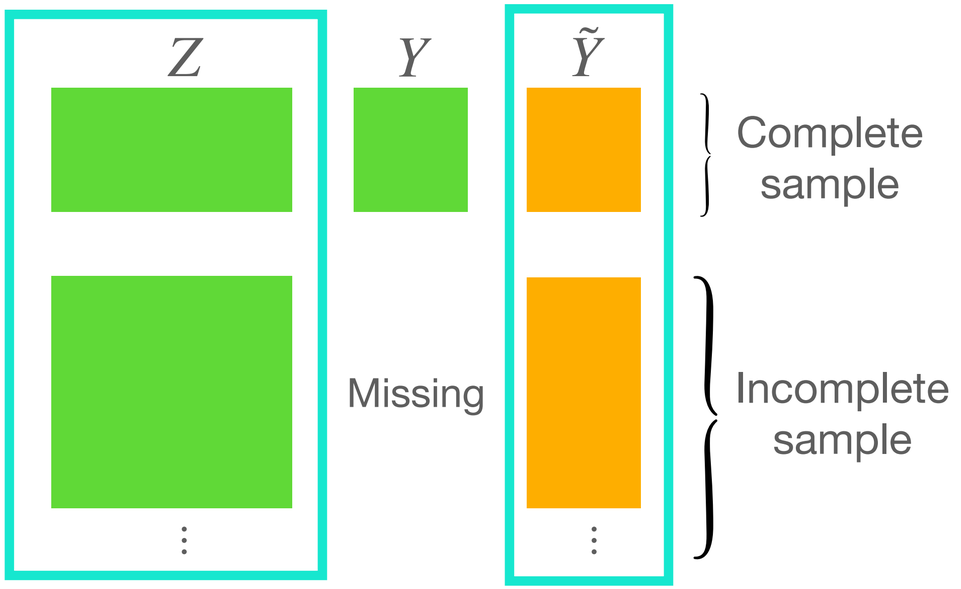
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^{N}\right)$

- Low variance: Uses all samples ✓
- Biased: it targets $\gamma \neq \theta$ (i.e., $\text{plim}_{N \to \infty} \hat{\gamma}^{\text{all}} = \gamma \neq \theta$)

γ describes joint distribution of $(\tilde{Y}, Z)$ but θ describes joint distribution of (Y, Z)

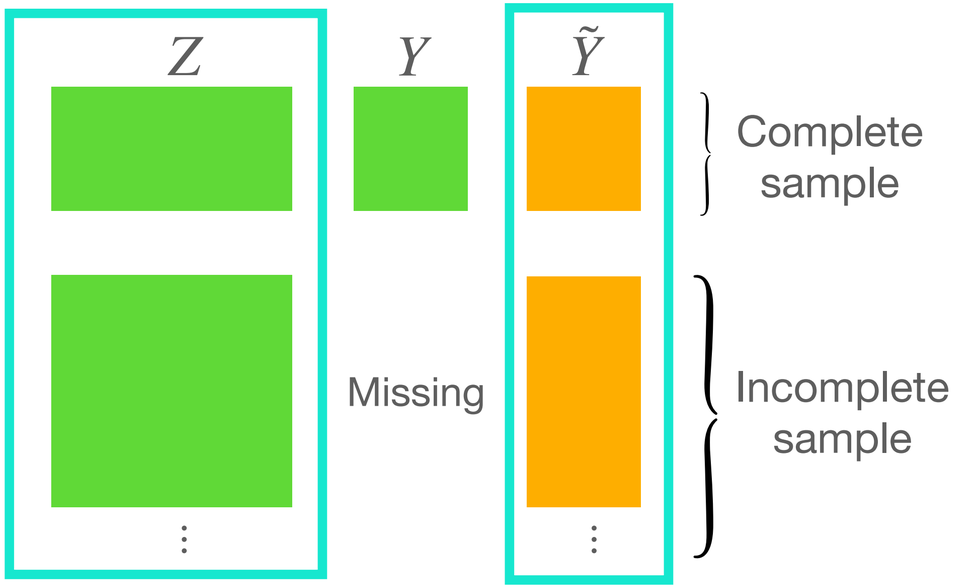
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^{N}\right)$

- Low variance: Uses all samples ✓
- Biased: it targets $\gamma \neq \theta$ (i.e., $\text{plim}_{N \to \infty} \hat{\gamma}^{\text{all}} = \gamma \neq \theta$) ✗

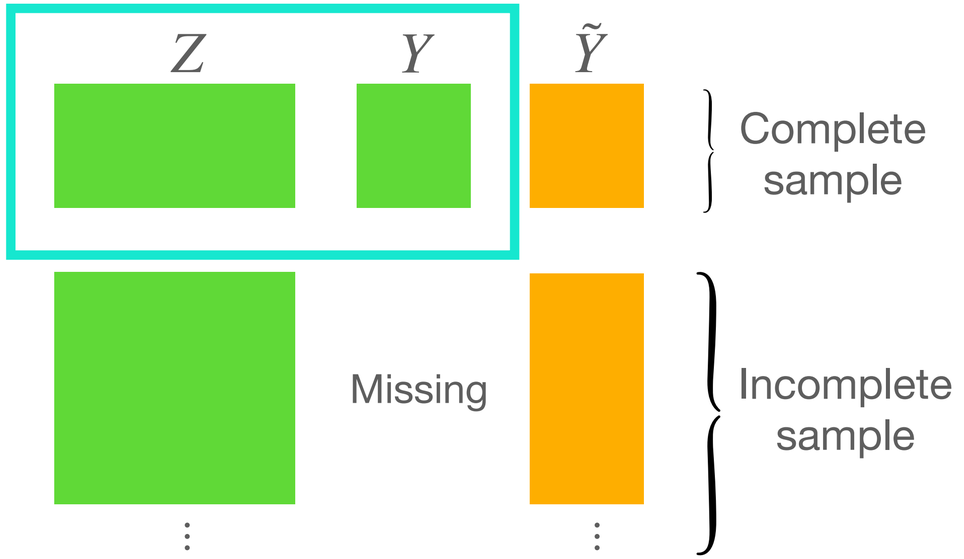
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S_\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S_\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^N\right)$

- Low variance: Uses all samples ✓
- Biased: it targets $\gamma \neq \theta$ (i.e., $\text{plim}_{N \to \infty} \hat{\gamma}^{\text{all}} = \gamma \neq \theta$) ✗

2) *Classical* Approach:

Estimate θ using $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in S_\bullet}\right)$

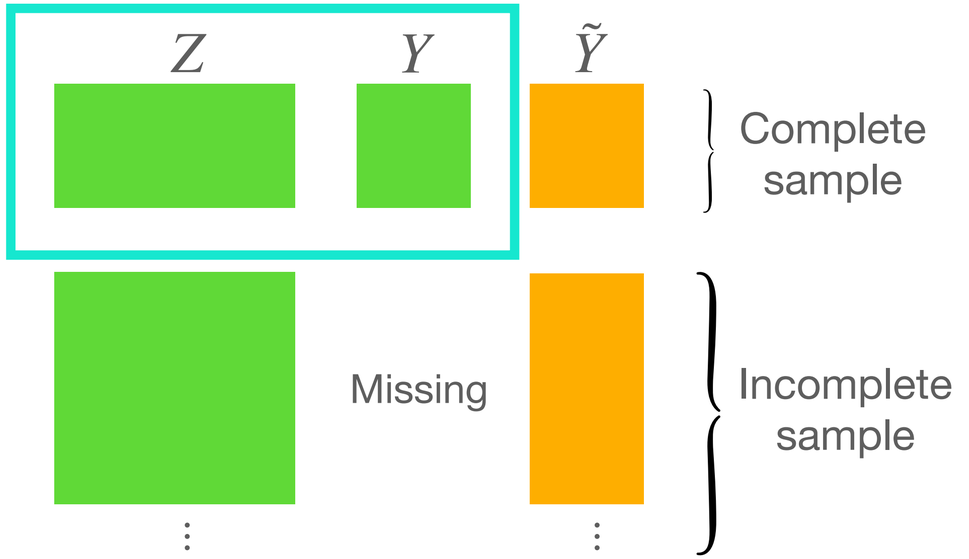
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S_\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S_\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^N\right)$

- Low variance: Uses all samples ✓
- Biased: it targets $\gamma \neq \theta$ (i.e., $\text{plim}_{N \to \infty} \hat{\gamma}^{\text{all}} = \gamma \neq \theta$) ✗

2) *Classical* Approach:

Estimate θ using $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in S_\bullet}\right)$

- High variance: uses few samples ✗

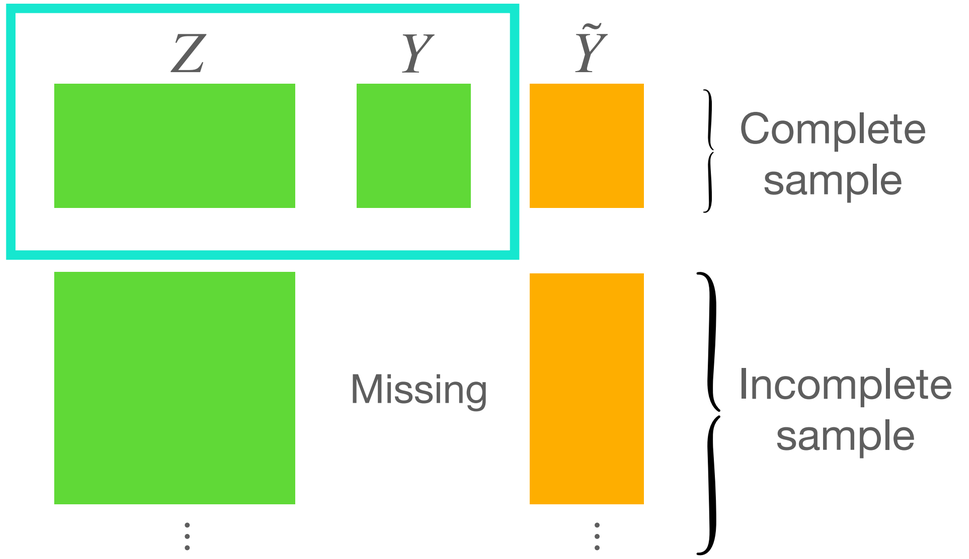
Two natural approaches

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^\circ}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



1) *Naive* Approach:

Estimate θ using $\hat{\gamma}^{\text{all}} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i=1}^N\right)$

- Low variance: Uses all samples ✓
- Biased: it targets $\gamma \neq \theta$ (i.e., $\text{plim}_{N \to \infty} \hat{\gamma}^{\text{all}} = \gamma \neq \theta$) ✗

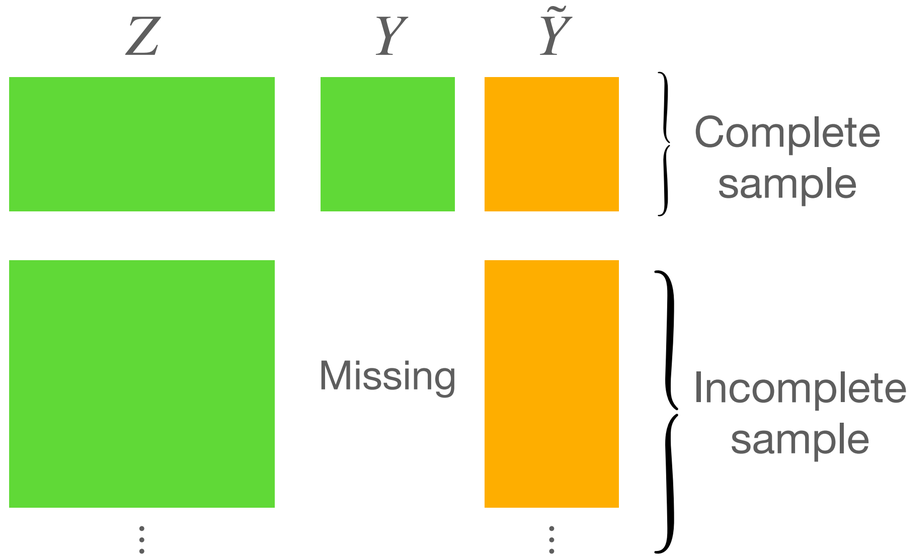
2) *Classical* Approach:

Estimate θ using $\hat{\theta}^\bullet = \mathscr{A}\left((Y_i, Z_i)_{i \in S^\bullet}\right)$

- High variance: uses few samples ✗
- Approximately unbiased: it targets θ ✓

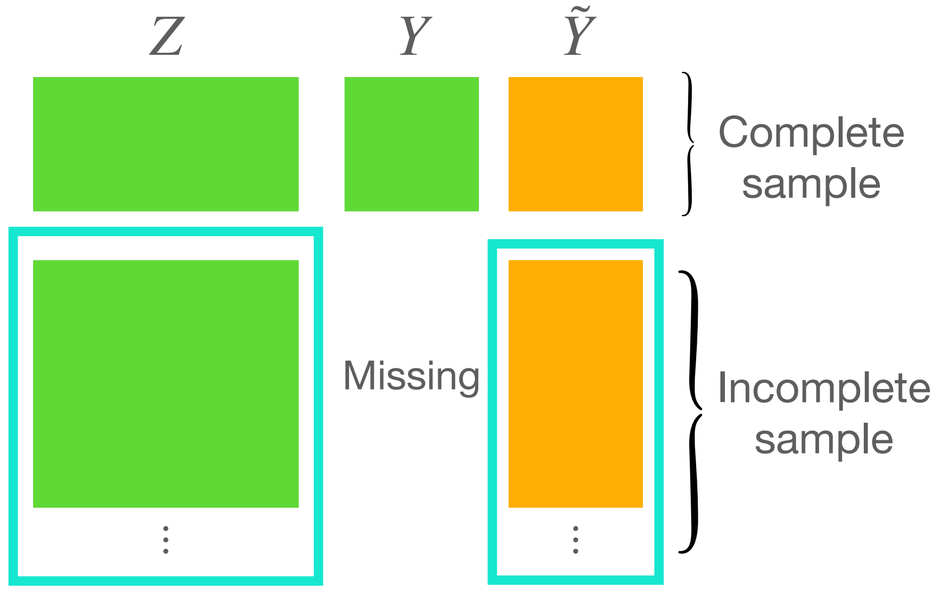
Predict-Then-Debias (PTD) approach

Predict-Then-Debias (PTD) approach



- Proceeds in 2-steps

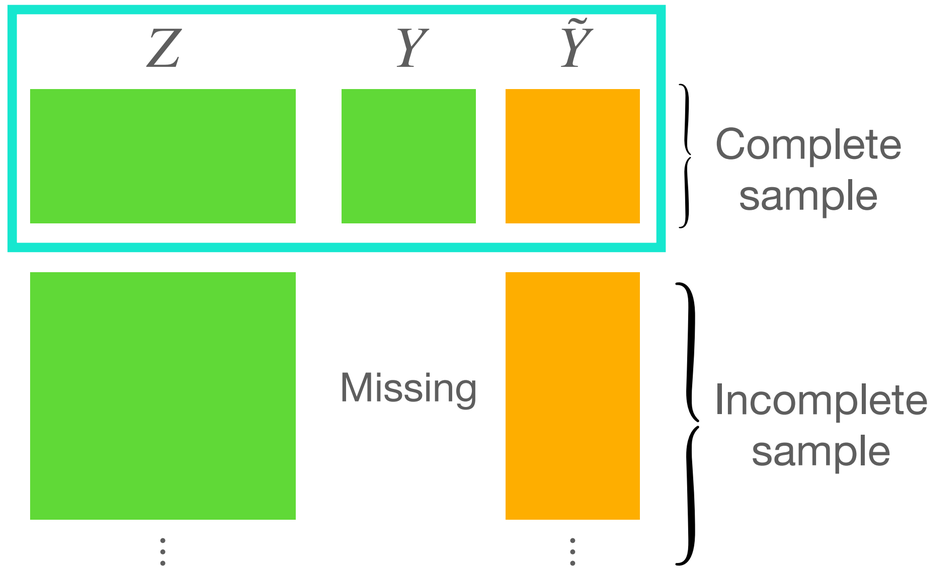
Predict-Then-Debias (PTD) approach



- Proceeds in 2-steps

    1. Start with a biased estimator based on the large, incomplete sample

Predict-Then-Debias (PTD) approach



- Proceeds in 2-steps

    1. Start with a biased estimator based on the large, incomplete sample
    2. Add a bias correction term based on the complete sample

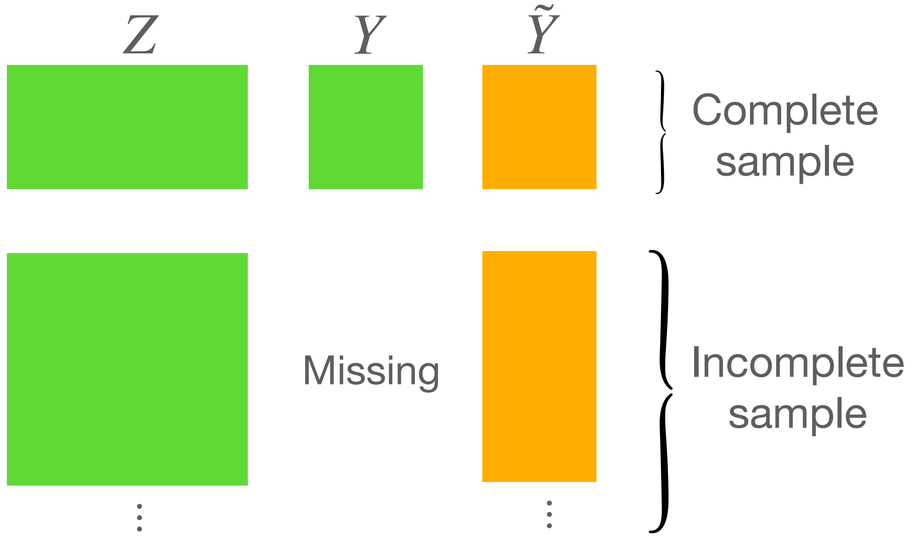
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^{\complement}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



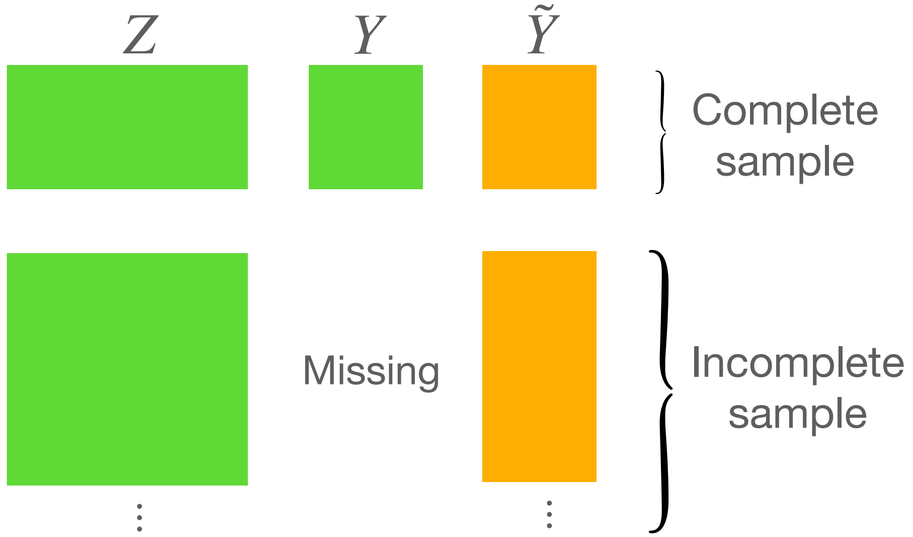
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in S}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in S^{\complement}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator

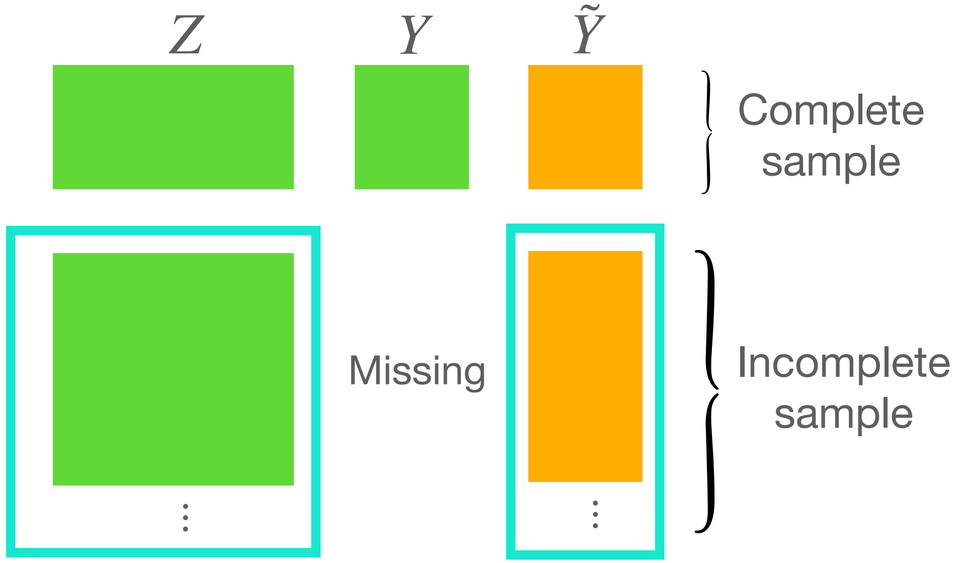
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^{\bigcirc}}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$

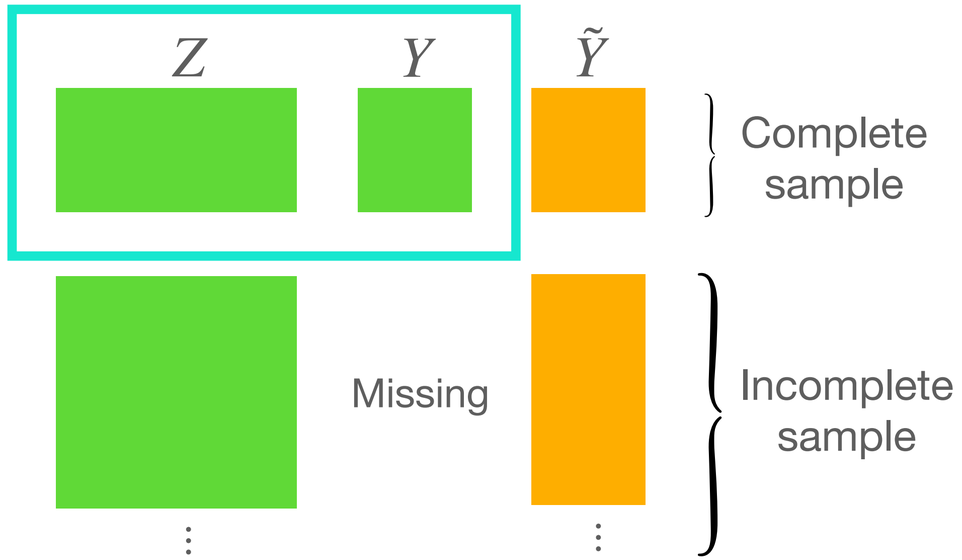
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^{\bullet}}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^{\bigcirc}}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^{\bigcirc}}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^{\bullet}}\right)$ <- Same as classical estimator

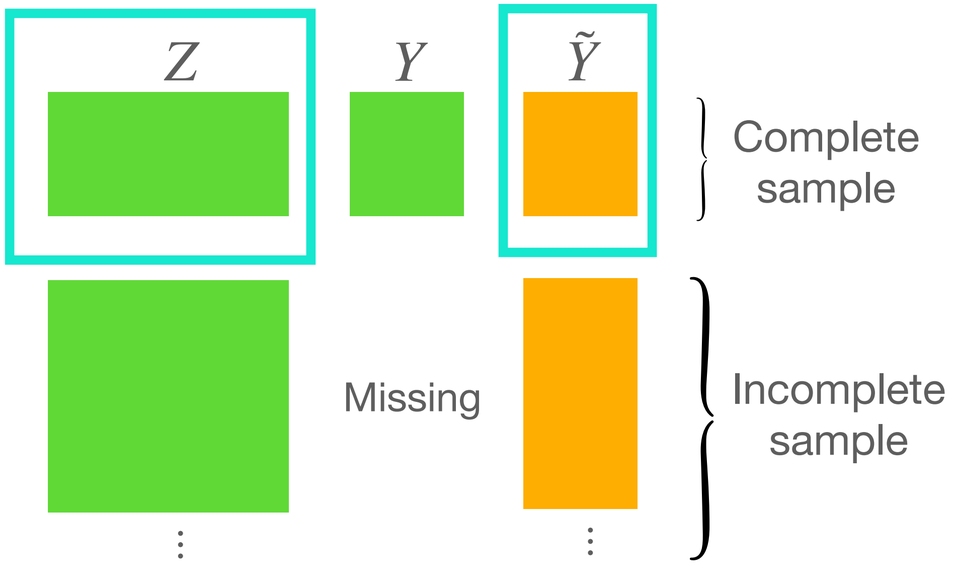
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$

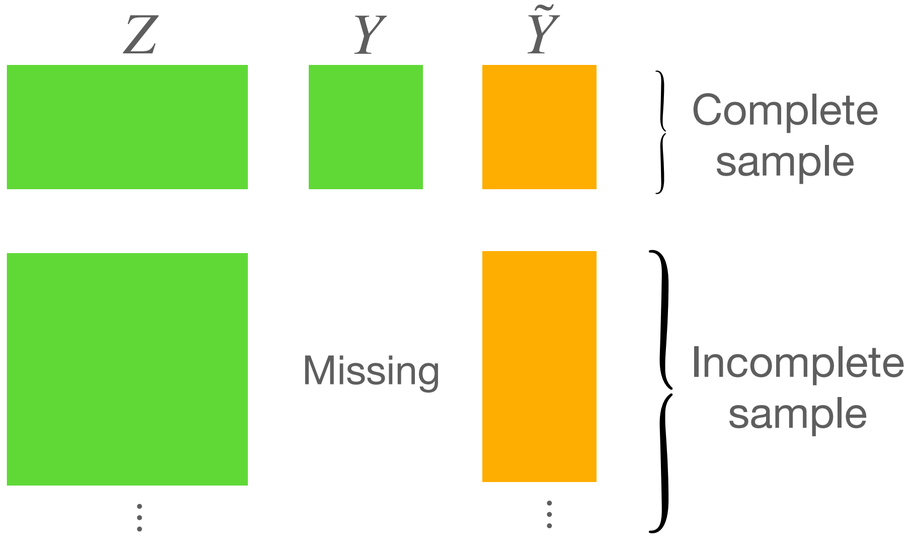
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + (\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})$

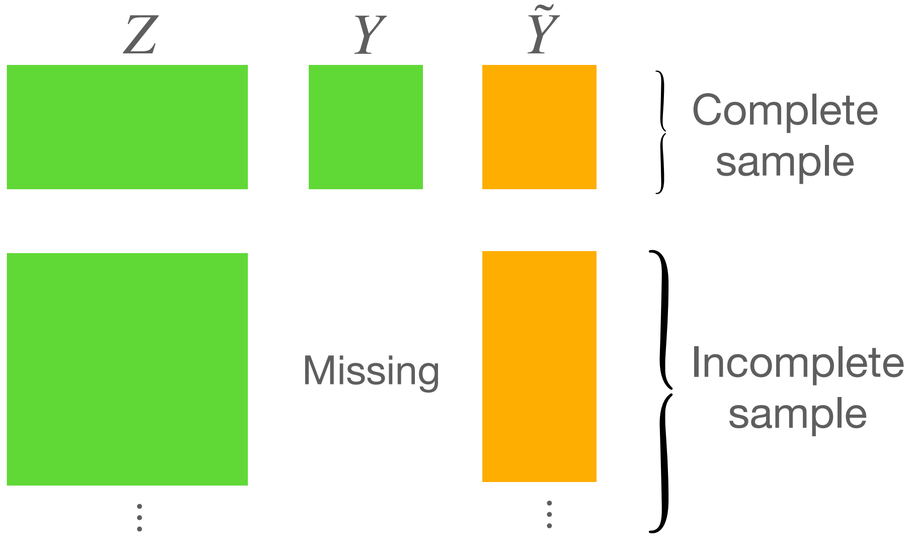
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + (\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})$

Biased, low-variance estimate

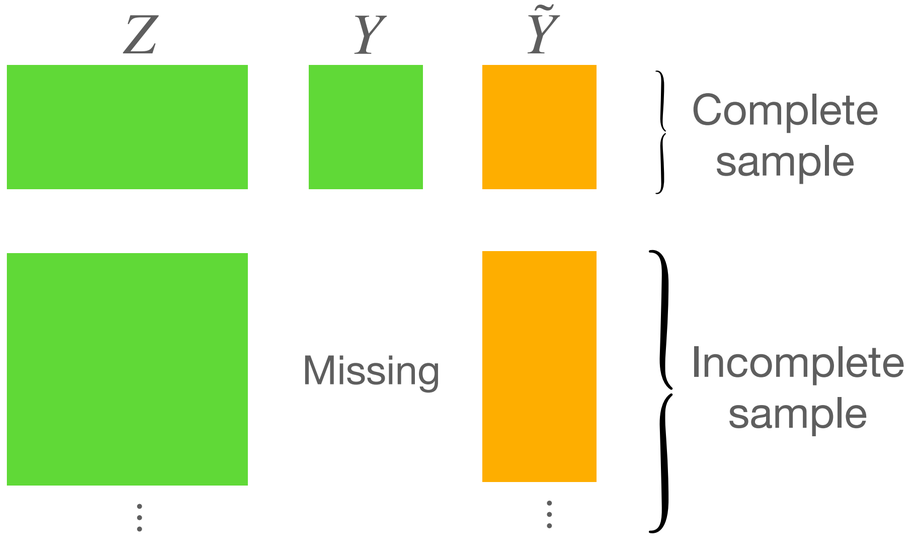
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

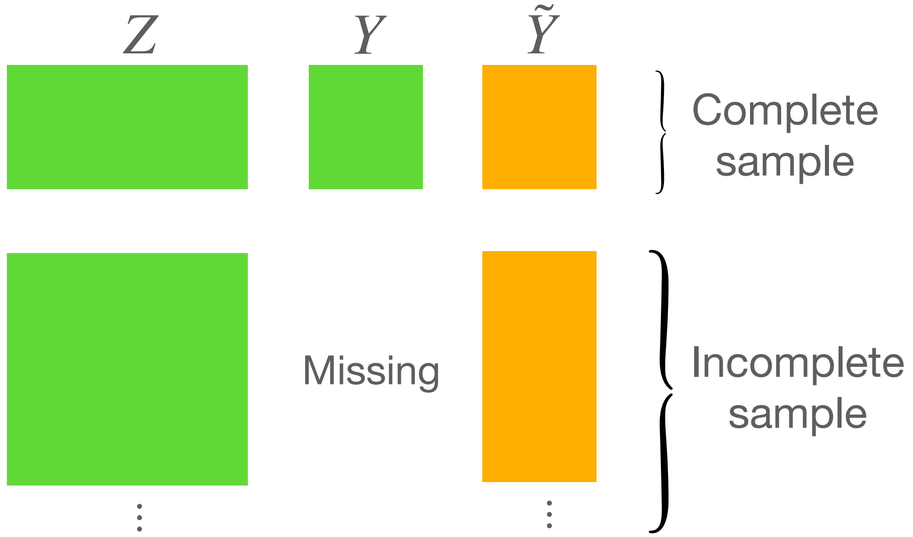
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



Properties of $\hat{\theta}^{\text{PTD}}$

- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Bias correction

Biased, low-variance estimate

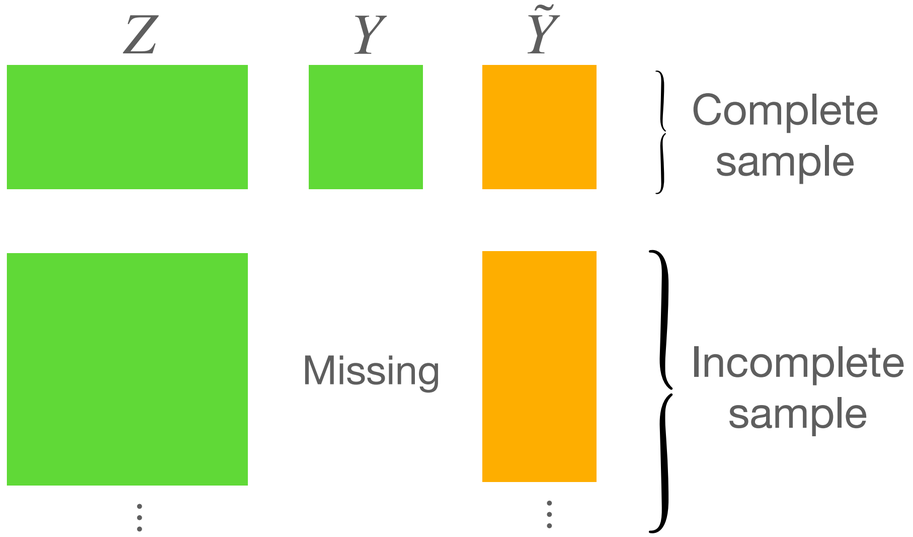
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased

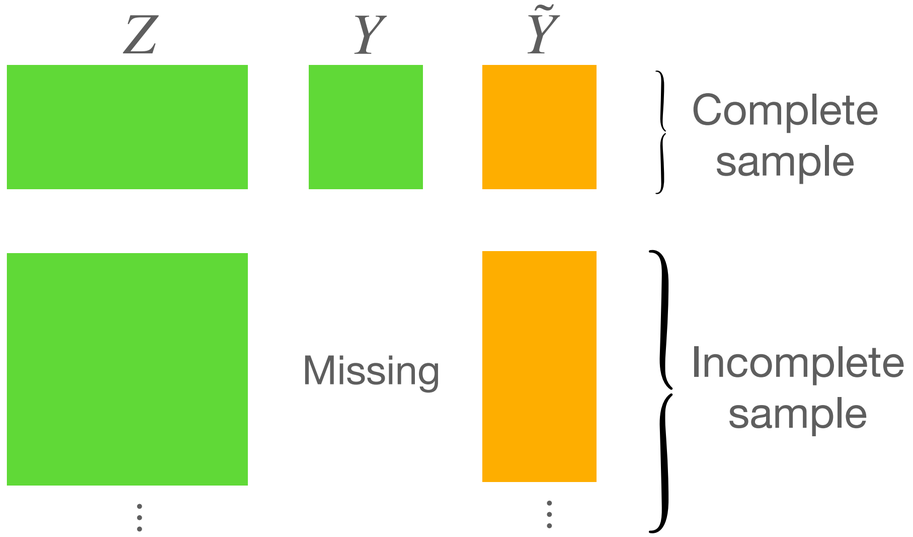
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased
  - $\hat{\theta}^{\text{PTD}}$ targets $\gamma + (\theta - \gamma) = \theta$

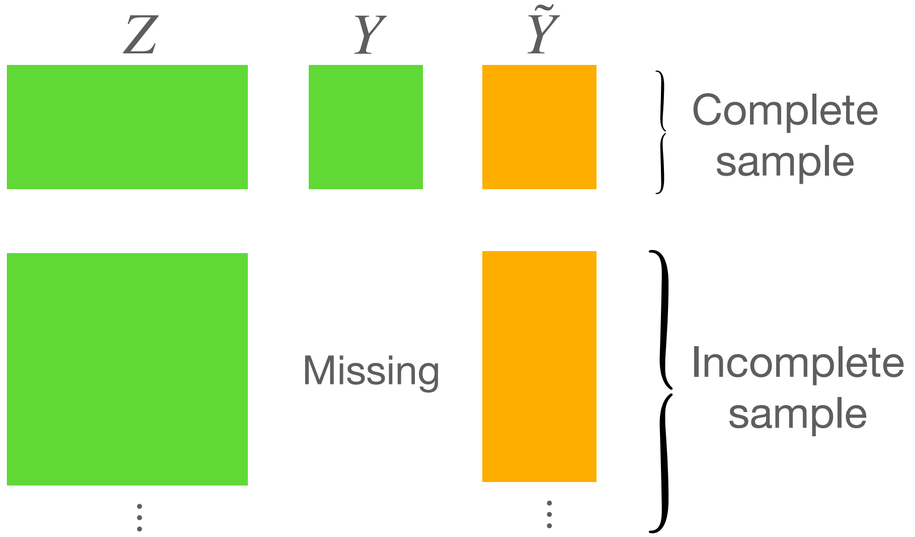
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased
  - $\hat{\theta}^{\text{PTD}}$ targets $\gamma + (\theta - \gamma) = \theta$
- Variance decomposition (under independence)

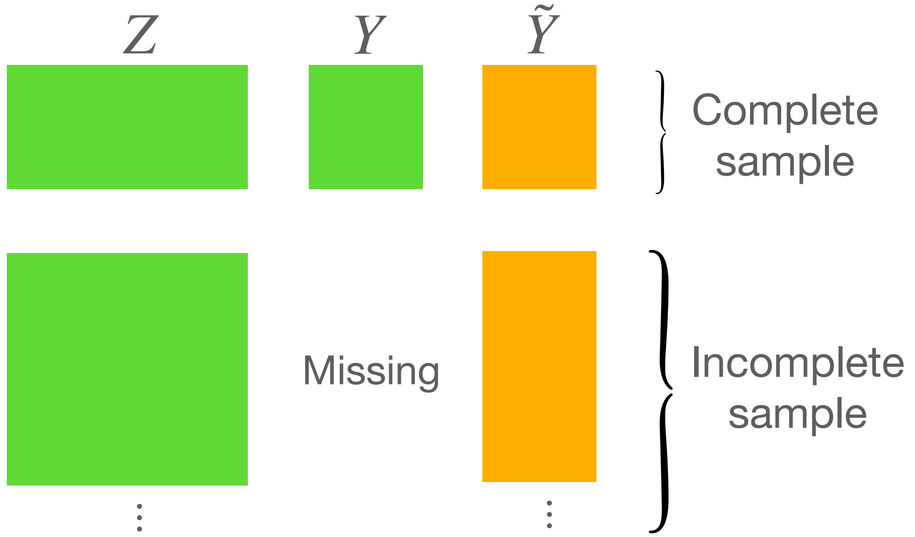
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased
  - $\hat{\theta}^{\text{PTD}}$ targets $\gamma + (\theta - \gamma) = \theta$
- Variance decomposition (under independence)
  - $\text{Var}(\hat{\theta}^{\text{PTD}}) = \text{Var}(\hat{\gamma}^{\bigcirc}) + \text{Var}(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})$

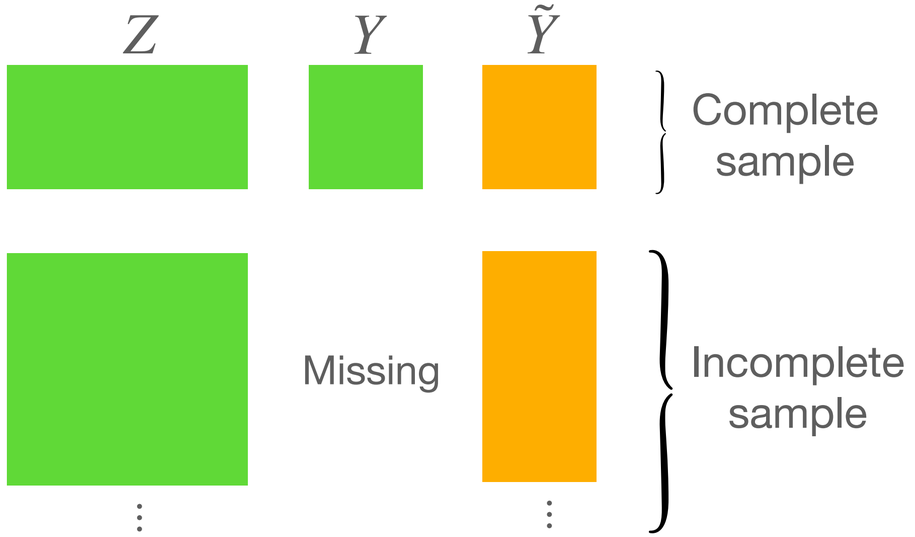
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Bias correction

Biased, low-variance estimate

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased
  - $\hat{\theta}^{\text{PTD}}$ targets $\gamma + (\theta - \gamma) = \theta$
- Variance decomposition (under independence)
  - $\text{Var}(\hat{\theta}^{\text{PTD}}) = \text{Var}(\hat{\gamma}^{\bigcirc}) + \text{Var}(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})$

Small for large N

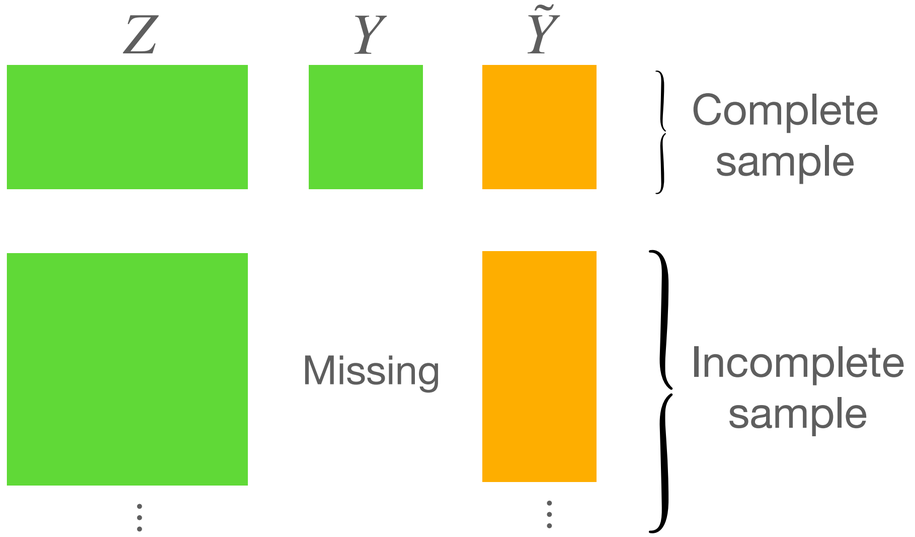
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased
  - $\hat{\theta}^{\text{PTD}}$ targets $\gamma + (\theta - \gamma) = \theta$
- Variance decomposition (under independence)
  - $\text{Var}(\hat{\theta}^{\text{PTD}}) = \text{Var}(\hat{\gamma}^{\bigcirc}) + \text{Var}(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})$

Small for large N

Small if $\tilde{Y} \approx Y$

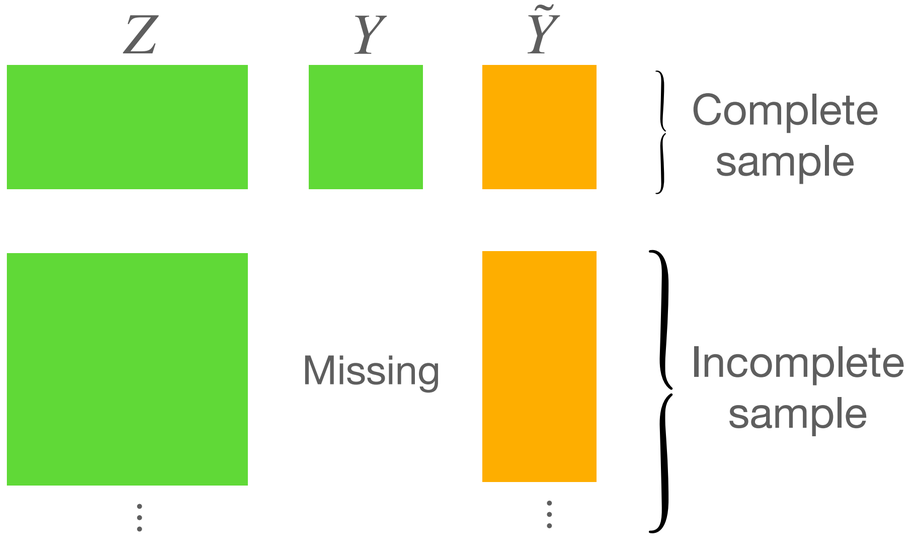
The Predict-Then-Debias estimator (untuned version)

Recall setting:

a complete sample $(Y_i, \tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}$.

an incomplete sample $(\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}$

an algorithm $\mathscr{A}(\cdot)$ that estimates θ using an input sample



- Constructing the Predict-Then-Debias estimator
  - $\hat{\gamma}^{\bigcirc} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bigcirc}\right)$ <- Similar to naive estimator $\hat{\gamma}^{\text{all}}$
  - $\hat{\theta}^{\bullet} = \mathscr{A}\left((Y_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$ <- Same as classical estimator
  - $\hat{\gamma}^{\bullet} = \mathscr{A}\left((\tilde{Y}_i, Z_i)_{i \in \mathcal{S}^\bullet}\right)$
  - $\hat{\theta}^{\text{PTD}} = \hat{\gamma}^{\bigcirc} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})}$

Biased, low-variance estimate

Bias correction

Properties of $\hat{\theta}^{\text{PTD}}$

- Approximately unbiased
  - $\hat{\theta}^{\text{PTD}}$ targets $\gamma + (\theta - \gamma) = \theta$
- Variance decomposition (under independence)
  - $\text{Var}(\hat{\theta}^{\text{PTD}}) = \text{Var}(\hat{\gamma}^{\bigcirc}) + \text{Var}(\hat{\theta}^{\bullet} - \hat{\gamma}^{\bullet})$

Small for large N

Small if $\tilde{Y} \approx Y$

If $\tilde{Y} - Y$ is large may have large variance…

Tuning to reduce variance

Tuning to reduce variance

- For each tuning matrix tuning matrix $\Omega \in \mathbb{R}^{d \times d}$, consider $\hat{\theta}^{\text{PTD},\Omega} = \Omega\hat{\gamma}^{\bigcirc} + (\hat{\theta}^{\bullet} - \Omega\hat{\gamma}^{\bullet})$

Tuning to reduce variance

- For each tuning matrix tuning matrix $\Omega \in \mathbb{R}^{d \times d}$, consider $\hat{\theta}^{\text{PTD},\Omega} = \Omega\hat{\gamma}^{\bigcirc} + (\hat{\theta}^{\bullet} - \Omega\hat{\gamma}^{\bullet})$
- "best" tuning matrix simultaneously minimizes $\text{Var}(\hat{\theta}_j^{\text{PTD},\Omega})$ for each $j = 1, \ldots, d$
  - $\Omega_{\text{opt}} = \text{Cov}(\hat{\theta}^{\bullet}, \hat{\gamma}^{\bullet})[\text{Var}(\hat{\gamma}^{\bigcirc}) + \text{Var}(\hat{\gamma}^{\bullet})]^{-1}$

Tuning to reduce variance

- For each tuning matrix tuning matrix $\Omega \in \mathbb{R}^{d \times d}$, consider $\hat{\theta}^{\text{PTD},\Omega} = \Omega\hat{\gamma}^{\bigcirc} + (\hat{\theta}^{\bullet} - \Omega\hat{\gamma}^{\bullet})$
- "best" tuning matrix simultaneously minimizes $\text{Var}(\hat{\theta}_j^{\text{PTD},\Omega})$ for each $j = 1, \ldots, d$
  - $\Omega_{\text{opt}} = \text{Cov}(\hat{\theta}^{\bullet}, \hat{\gamma}^{\bullet})[\text{Var}(\hat{\gamma}^{\bigcirc}) + \text{Var}(\hat{\gamma}^{\bullet})]^{-1}$
- In practice Ω_{opt} is not known precisely
  - Let $\hat{\Omega}$ be a data based estimate of Ω_{opt}
  - Proposed Predict-Then-Debias Estimator: $\hat{\theta}^{\text{PTD},\hat{\Omega}} = \hat{\Omega}\hat{\gamma}^{\bigcirc} + (\hat{\theta}^{\bullet} - \hat{\Omega}\hat{\gamma}^{\bullet})$

Constructing Confidence Intervals

Constructing Confidence Intervals

- Show $\sqrt{N}(\hat{\theta}^{\text{PTD},\hat{\Omega}} - \theta) \xrightarrow{d} \mathcal{N}(0,\Sigma)$ (Chen & Chen 2000, and others)

Constructing Confidence Intervals

- Show $\sqrt{N}(\hat{\theta}^{\text{PTD},\hat{\Omega}} - \theta) \xrightarrow{d} \mathcal{N}(0,\Sigma)$ (Chen & Chen 2000, and others)
- But requires additional mathematical calculations to generalize to new estimands.

Constructing Confidence Intervals

- Show $\sqrt{N}(\hat{\theta}^{\text{PTD},\hat{\Omega}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ (Chen & Chen 2000, and others)
- But requires additional mathematical calculations to generalize to new estimands.

Algorithm: The Predict-Then-Debias Bootstrap (Kluger et al., 2025)

1. **For** $b = 1, \ldots, B$
2. Sample $i_1, \ldots, i_N \sim_{iid} \text{Unif}(\{1, \ldots, N\})$
3. Let $\mathcal{K}^{\bullet,*} = \{k \in \{1, \ldots, N\} \ : \ i_k \in S^{\bullet}\}$
4. Let $\mathcal{K}^{\bigcirc,*} = \{k \in \{1, \ldots, N\} \ : \ i_k \in S^{\bigcirc}\}$
5. Set $\hat{\theta}^{\bullet,(b)} = \mathcal{A}\left((Y_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bullet,*}}\right)$
6. Set $\hat{\gamma}^{\bullet,(b)} = \mathcal{A}\left((\tilde{Y}_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bullet,*}}\right)$
7. Set $\hat{\gamma}^{\bigcirc,(b)} = \mathcal{A}\left((\tilde{Y}_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bigcirc,*}}\right)$
8. **End For**
9. Select tuning matrix $\hat{\Omega}$
10. Set $\hat{\theta}^{\text{PTD},\hat{\Omega},(b)} = \hat{\Omega}\hat{\gamma}^{\bigcirc,(b)} + (\hat{\theta}^{\bullet,(b)} - \hat{\Omega}\hat{\gamma}^{\bullet,(b)})$ for $b = 1, \ldots, B$
11. **Return** $\mathcal{C}_j = \left(\text{Quantile}_{\alpha/2}(\{\hat{\theta}_j^{\text{PTD},\hat{\Omega},(b)}\}_{b=1}^{B}), \text{Quantile}_{1-\alpha/2}(\{\hat{\theta}_j^{\text{PTD},\hat{\Omega},(b)}\}_{b=1}^{B})\right)$ for $j = 1, \ldots, d$

Constructing Confidence Intervals

- Show $\sqrt{N}(\hat{\theta}^{\text{PTD},\hat{\Omega}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ (Chen & Chen 2000, and others)
- But requires additional mathematical calculations to generalize to new estimands.

Algorithm: The Predict-Then-Debias Bootstrap (Kluger et al., 2025)

1. **For** $b = 1, \ldots, B$
2. Sample $i_1, \ldots, i_N \sim_{iid} \text{Unif}(\{1, \ldots, N\})$
3. Let $\mathcal{K}^{\bullet,*} = \{k \in \{1, \ldots, N\} \ : \ i_k \in S^{\bullet}\}$
4. Let $\mathcal{K}^{\bigcirc,*} = \{k \in \{1, \ldots, N\} \ : \ i_k \in S^{\bigcirc}\}$
5. Set $\hat{\theta}^{\bullet,(b)} = \mathcal{A}\left((Y_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bullet,*}}\right)$
6. Set $\hat{\gamma}^{\bullet,(b)} = \mathcal{A}\left((\tilde{Y}_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bullet,*}}\right)$
7. Set $\hat{\gamma}^{\bigcirc,(b)} = \mathcal{A}\left((\tilde{Y}_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bigcirc,*}}\right)$
8. **End For**
9. Select tuning matrix $\hat{\Omega}$
10. Set $\hat{\theta}^{\text{PTD},\hat{\Omega},(b)} = \hat{\Omega}\hat{\gamma}^{\bigcirc,(b)} + (\hat{\theta}^{\bullet,(b)} - \hat{\Omega}\hat{\gamma}^{\bullet,(b)})$ for $b = 1, \ldots, B$
11. **Return** $\mathcal{C}_j = \left(\text{Quantile}_{\alpha/2}(\{\hat{\theta}_j^{\text{PTD},\hat{\Omega},(b)}\}_{b=1}^{B}), \text{Quantile}_{1-\alpha/2}(\{\hat{\theta}_j^{\text{PTD},\hat{\Omega},(b)}\}_{b=1}^{B})\right)$ for $j = 1, \ldots, d$

- Percentile bootstrap

Constructing Confidence Intervals

- Show $\sqrt{N}(\hat{\theta}^{\text{PTD},\hat{\Omega}} - \theta) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ (Chen & Chen 2000, and others)
- But requires additional mathematical calculations to generalize to new estimands.

Algorithm: The Predict-Then-Debias Bootstrap (Kluger et al., 2025)

1. **For** $b = 1, \ldots, B$
2. Sample $i_1, \ldots, i_N \sim_{iid} \text{Unif}(\{1, \ldots, N\})$
3. Let $\mathcal{K}^{\bullet,*} = \{k \in \{1, \ldots, N\} \ : \ i_k \in S^{\bullet}\}$
4. Let $\mathcal{K}^{\bigcirc,*} = \{k \in \{1, \ldots, N\} \ : \ i_k \in S^{\bigcirc}\}$
5. Set $\hat{\theta}^{\bullet,(b)} = \mathcal{A}\left((Y_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bullet,*}}\right)$
6. Set $\hat{\gamma}^{\bullet,(b)} = \mathcal{A}\left((\tilde{Y}_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bullet,*}}\right)$
7. Set $\hat{\gamma}^{\bigcirc,(b)} = \mathcal{A}\left((\tilde{Y}_{i_k}, Z_{i_k})_{k \in \mathcal{K}^{\bigcirc,*}}\right)$
8. **End For**
9. Select tuning matrix $\hat{\Omega}$
10. Set $\hat{\theta}^{\text{PTD},\hat{\Omega},(b)} = \hat{\Omega}\hat{\gamma}^{\bigcirc,(b)} + (\hat{\theta}^{\bullet,(b)} - \hat{\Omega}\hat{\gamma}^{\bullet,(b)})$ for $b = 1, \ldots, B$
11. **Return** $\mathcal{C}_j = \left(\text{Quantile}_{\alpha/2}(\{\hat{\theta}_j^{\text{PTD},\hat{\Omega},(b)}\}_{b=1}^{B}), \text{Quantile}_{1-\alpha/2}(\{\hat{\theta}_j^{\text{PTD},\hat{\Omega},(b)}\}_{b=1}^{B})\right)$ for $j = 1, \ldots, d$

- Percentile bootstrap
- General and flexible (algorithmically agnostic to choice of estimand θ and estimation approach $\mathcal{A}$)

Theoretical guarantees (informally stated)

Theoretical guarantees (informally stated)

Assumptions

1. $(Y_i, \tilde{Y}_i, Z_i)_{i=1}^N$ is IID, with Y_i missing (unobserved) completely at random

2. $\mathscr{A}(\cdot)$ gives a consistent and sufficiently smooth estimate of θ using data as inputs

Theorem (Kluger et al., 2025) **:** Under Assumptions 1 and 2, then the Predict-Then-Debias bootstrap returns asymptotically valid confidence intervals:

$$\lim_{n,N,B \to \infty} \mathbb{P}(\theta_j \in \mathscr{C}_j) = 1 - \alpha$$

Theoretical guarantees (informally stated)

Assumptions

1. $(Y_i, \tilde{Y}_i, Z_i)_{i=1}^N$ is IID, with Y_i missing (unobserved) completely at random
2. $\mathscr{A}(\cdot)$ gives a consistent and sufficiently smooth estimate of θ using data as inputs

Theorem (Kluger et al., 2025) **:** Under Assumptions 1 and 2, then the Predict-Then-Debias bootstrap returns asymptotically valid confidence intervals:

$$\lim_{n,N,B \to \infty} \mathbb{P}(\theta_j \in \mathscr{C}_j) = 1 - \alpha$$

See paper for

- More details on Assumption 2 and generalizations beyond Assumption 1
- Real data-based simulations that validate these theoretical guarantees

Generalizations and other findings

See paper for:

- Generalizations when data comes from a weighted, clustered or stratified sampling scheme
- Computational speedups for settings where N is very large
- Data-based experiments demonstrating Predict-Then-Debias approach leads to confidence intervals that are up to 4 times narrower than those from classical approach

Kluger *et al.* "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling" (2025+) on arXiv

Questions

Email: dkluger@mit.edu

- For more information see:

Kluger *et al.* "Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling" (2025+) on arXiv

Lu *et al.* "Regression coefficient estimation from remote sensing maps" (2025). *Remote Sensing of Environment*

Kluger *et al.* "*A Preview of the Predict-Then-Debias Bootstrap*" (2026). To appear in *AEA Papers & Proceedings*

R Package and Python module on Github

- Thank you to Matthew Gordon (session chair), Ed Rubin (discussant) and organizers
- One line summary of method

$$\hat{\theta}^{\text{PTD}} = \underbrace{\Omega\hat{\gamma}^{\bigcirc}}_{\text{Biased, low variance term}} + \underbrace{(\hat{\theta}^{\bullet} - \Omega\hat{\gamma}^{\bullet})}_{\text{Bias correction}} \Bigg\} \text{ Apply percentile bootstrap to this estimator}$$

References

Alix-Garcia, Jennifer and Daniel L Millimet (2023). “Remotely incorrect? Accounting for nonclassical measurement error in satellite data on deforestation”. In: Journal of the Association of Environmental and Resource Economists 10.5, pp. 1335–1367.

Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., and Zrnic, T. (2023a). Prediction-powered inference. Science, 382(6671):669–674.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition. Chapman and Hall/CRC, London, 2nd edition.

Chen, Y.-H. and Chen, H. (2000). A unified approach to regression analysis under doublesampling designs. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 62(3):449–460.

Deines, J.M., Wang S., and Lobell D.B. Satellites reveal a small positive yield effect from conservation tillage across the US Corn Belt. Environmental Research Letters, 14(12), 2019

Deines, J.M., Patel R., Liang S.Z., Dado W., and Lobell D.B. A million kernels of truth: Insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. Remote Sensing of Environment, 253:112174, 2021

Gronsbell, J., Gao, J., Shi, Y., McCaw, Z. R., and Cheng, D. (2024). Another look at inference after prediction. arXiv preprint arXiv:2411.19908.

Jain, Meha (2020). “The benefits and pitfalls of using satellite data for causal inference”. In: Review of Environmental Economics and Policy.

Kluger, Dan M. et al. (2025). “Prediction-Powered Inference with Imputed Covariates and Nonuniform Sampling”. In: arXiv:2501.18577 [stat.ME]. https://doi.org/10.48550/arXiv.2501.18577.

Kosorok, M. R. (2008). Introduction to Empirical Processes and Semiparametric Inference. Springer Series in Statistics. Springer.

Kremers, W. K. (2021). A general, simple, robust method to account for measurement error when analyzing data with an internal validation subsample. arXiv preprint arXiv:2106.14063

Little, R. J. A. and Rubin, D. B. (2019). Statistical Analysis with Missing Data: Third Edition. Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ, 3rd edition.

Lu K., Kluger D.M., Bates S., Wang S. (2025). “Regression Coefficient Estimating from Remote Sensing Maps”. In: arXiv:2407.13659 [stat.AP]. https://arxiv.org/pdf/2407.13659v5

Miao, J. and Lu, Q. (2024). Task-agnostic machine-learning-assisted inference. In 38th Conference on

Proctor, J., Carleton, T., and Sum, S. (2023). Parameter recovery using remotely sensed variables. Technical report, National Bureau of Economic Research.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association, 89(427):846–866.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., and Hsiang, S. (2021a). A generalizable and accessible approach to machine learning with global satellite imagery. Nature Communications, 12(1):4392.

Rolf, E., Proctor, J., Carleton, T., Bolliger, I., Shankar, V., Ishihara, M., Recht, B., and Hsiang, S. (2021b). A generalizable and accessible approach to machine learning with global satellite imagery. https://www.codeocean.com/capsule/6456296/tree/v2.

Tong, J., Huang, J., Chubak, J., Wang, X., Moore, J. H., Hubbard, R. A., and Chen, Y. (2019). An augmented estimation procedure for ehr-based association studies accounting for differential misclassification. Journal of the American Medical Informatics Association, 27(2):244–253.

van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Yang, S. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. Journal of the American Statistical Association, 115(531):1540–1554. PMID: 33088006.

Zhou, Q., et al. (2022). Recent Rapid Increase of Cover Crop Adoption Across the U.S. Midwest Detected by Fusing Multi-Source Satellite Data. Geophysical Research Letters, 49(22).

Zrnic, T. (2024). A note on the prediction-powered bootstrap. arXiv preprint arXiv:2405.18379.

Appendix slides

Overview of Predict-Then-Debias Bootstrap

- Guarantees of lower variance than simply ignoring ML data
- Few line characterization

$$\hat{\theta}^{\text{PTD},\hat{\Omega}} = \underbrace{\hat{\Omega}\hat{\gamma}^{\bigcirc}}_{\text{Biased, low variance term}} + \underbrace{(\hat{\theta}^{\bullet} - \hat{\Omega}\hat{\gamma}^{\bullet})}_{\text{Bias correction}}$$

  - CIs from bootstrap have theoretical guarantees
- Construction allows for natural generalizations to new settings:
  - Works for a variety of estimators $\mathscr{A}(\cdot)$ and estimands θ without additional modifications
  - If the labelling scheme is known, bootstrap scheme can be modified accordingly
    - E.g., if the labels collected according to weighted, clustered, or stratified sampling
  - More extensions?

Takeaways

- Many fields are grappling with questions of how to reliably use machine learning predictions in downstream statistical analyses
- Methods exist but are not widespread in some fields
- Predict-Then-Debias Bootstrap is an option
  - Requires a small complete sample
  - Construction also allows for natural generalizations to new settings
  - R and Python Packages online!