Correcting for selection bias and missing response

SIKS course on Causal Inference

Philip Boeken

p.a.boeken@uva.nl

[1]University of Amsterdam
The Netherlands

[2]Booking.com
The Netherlands

May 30, 2023

Contents

Selection bias:

- Regression with selection biased data: when do we have to correct?
- Causal modelling of selection mechanisms
- *Repeated regression* procedure to correct for bias.

Missing response / selective labelling:

- Re-training of predictive model used for accepting/rejecting units
- Application of the same repeated regression procedure
- Importance weighting

Selection bias

Motivating example: cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[1]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer
- Suppose we want to estimate $\mathbb{E}[Y|X]$ to predict cervical cancer in large-scale screening of the population.
- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.

[1] Available at `https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors)`.

Motivating example: cervical cancer screening

- We have data from Hospital Universitario de Caracas, Venezuela:[1]
  - X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
  - Y: Presence of cervical cancer
- Suppose we want to estimate $\mathbb{E}[Y|X]$ to predict cervical cancer in large-scale screening of the population.
- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X]$ for population screening?

(assume $\hat{\mathbb{E}}[Y|X] \approx \mathbb{E}[Y|X]$)

[1] Available at `https://archive.ics.uci.edu/ml/datasets/Cervical+cancer+(Risk+Factors)`.

Selection bias: available data

X	Y
x_1	y_1
$\vdots$	
x_m	y_m
x_{m+1}	y_{m+1}
$\vdots$	
x_n	y_n

Population $\sim \mathbb{P}(X, Y) \implies \mathbb{E}[Y|X]$

Selection bias: available data

X	Y	S
x_1	y_1	1
$\vdots$		
x_m	y_m	1
x_{m+1}	y_{m+1}	0
$\vdots$		
x_n	y_n	0

Population $\sim \mathbb{P}(X, Y) \implies \mathbb{E}[Y|X]$

Selection bias: available data

X	Y	S
x_1	y_1	1
$\vdots$		
x_m	y_m	1
x_{m+1}	y_{m+1}	0
$\vdots$		
x_n	y_n	0

Sample $\sim \mathbb{P}(X, Y|S=1) \implies \hat{\mathbb{E}}[Y|X, S=1]$

Population $\sim \mathbb{P}(X, Y) \implies \mathbb{E}[Y|X]$

Selection bias: available data

X	Y	S
x_1	y_1	1
$\vdots$		
x_m	y_m	1
x_{m+1}	y_{m+1}	0
$\vdots$		
x_n	y_n	0

Sample $\sim \mathbb{P}(X, Y | S = 1) \implies \hat{\mathbb{E}}[Y|X, S = 1]$

Population $\sim \mathbb{P}(X, Y) \implies \mathbb{E}[Y|X]$

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Selection bias: available data

X	Y	S
x_1	y_1	1
$\vdots$		
x_m	y_m	1
x_{m+1}	y_{m+1}	0
$\vdots$		
x_n	y_n	0

Sample $\sim \mathbb{P}(X, Y|S = 1) \implies \hat{\mathbb{E}}[Y|X, S = 1]$

Population $\sim \mathbb{P}(X, Y) \implies \mathbb{E}[Y|X]$

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

$\iff$ Do we have $\mathbb{E}[Y|X, S = 1] = \mathbb{E}[Y|X]$?

A taxonomy of selection mechanisms

For estimating $\mathbb{E}[Y|X]$ from $\mathbb{P}(X, Y|S = 1)$, selection is:

Ignorable[2]

$$Y \perp\!\!\!\perp S|X$$

$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, S = 1]$$

Nonignorable

$$Y \not\perp\!\!\!\perp S|X$$

$$\mathbb{E}[Y|X] \neq \mathbb{E}[Y|X, S = 1]$$

[2] Zadrozny [2004], Wei Fan et al. [2005]

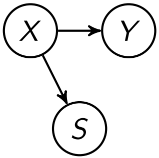
A taxonomy of selection mechanisms

For estimating $\mathbb{E}[Y|X]$ from $\mathbb{P}(X, Y|S = 1)$, selection is:

Ignorable[2]

$$Y \perp\!\!\!\perp S|X$$

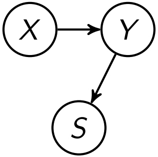
$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, S = 1]$$



Nonignorable

$$Y \not\!\perp\!\!\!\perp S|X$$

$$\mathbb{E}[Y|X] \neq \mathbb{E}[Y|X, S = 1]$$

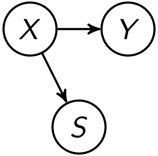


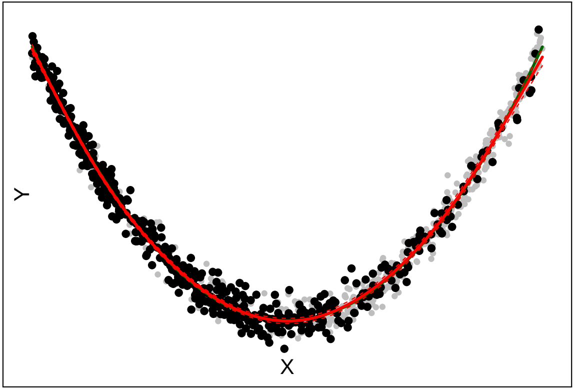
[2] Zadrozny [2004], Wei Fan et al. [2005]

Ignorable selection bias

$Y \perp\!\!\!\perp S \mid X$, hence $\mathbb{E}[Y|X, S = 1] = \mathbb{E}[Y|X]$

$$X \sim \mathcal{U}([-5, 5])$$
$$Y = X^2 + \mathcal{N}(0, 1)$$
$$S \sim \mathrm{Bernoulli}\,(\sigma(X))$$

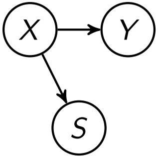


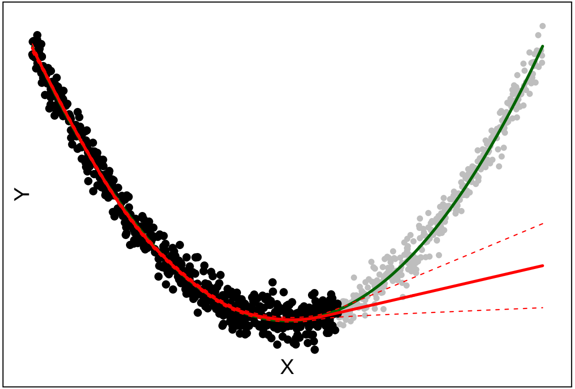


Ignorable selection bias

$Y \perp\!\!\!\perp S \mid X$, hence $\mathbb{E}[Y|X, S = 1] = \mathbb{E}[Y|X]$

$$X \sim \mathcal{U}([-5, 5])$$
$$Y = X^2 + \mathcal{N}(0, 1)$$
$$S = \mathbb{1}\{X \leq 1\}$$

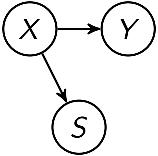


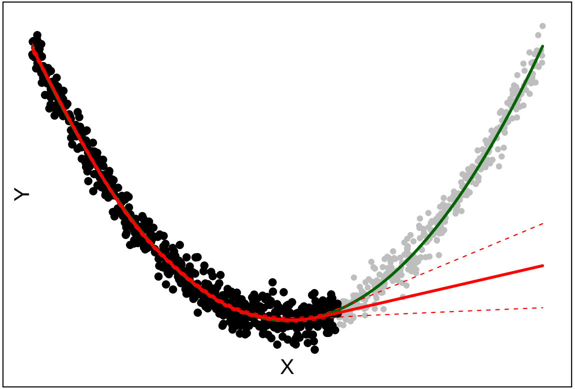


Ignorable selection bias

$Y \perp\!\!\!\perp S \mid X$, hence $\mathbb{E}[Y|X, S=1] = \mathbb{E}[Y|X]$ Positivity: $\mathrm{supp}(\mathbb{P}(X|S=1)) = \mathrm{supp}(\mathbb{P}(X))$

$$X \sim \mathcal{U}([-5,5])$$
$$Y = X^2 + \mathcal{N}(0,1)$$
$$S = \mathbb{1}\{X \leq 1\}$$

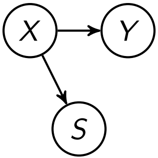


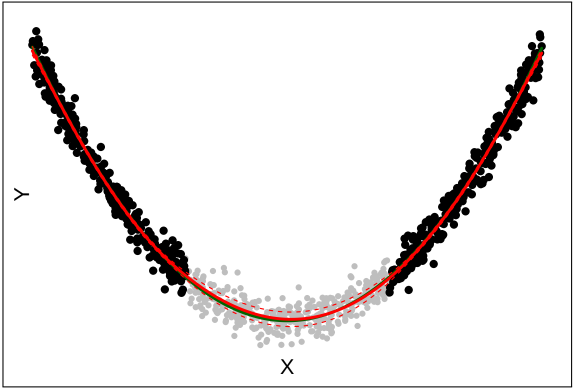


Ignorable selection bias

$Y \perp\!\!\!\perp S \mid X$, hence $\mathbb{E}[Y|X, S = 1] = \mathbb{E}[Y|X]$

$$X \sim \mathcal{U}([-5, 5])$$
$$Y = X^2 + \mathcal{N}(0, 1)$$
$$S = \mathbb{1}\{|X| > 2\}$$

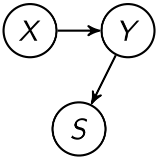


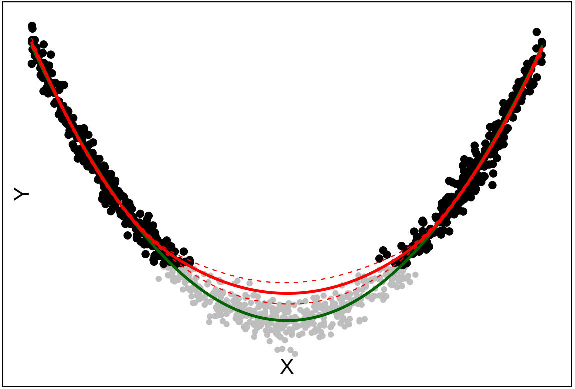


Nonignorable selection bias

$Y \not\perp\!\!\!\perp S \mid X$, hence $\mathbb{E}[Y|X, S=1] \neq \mathbb{E}[Y|X]$

$$X \sim \mathcal{U}([-5,5])$$
$$Y = X^2 + \mathcal{N}(0,1)$$
$$S = \mathbb{1}\{Y > 5\}$$



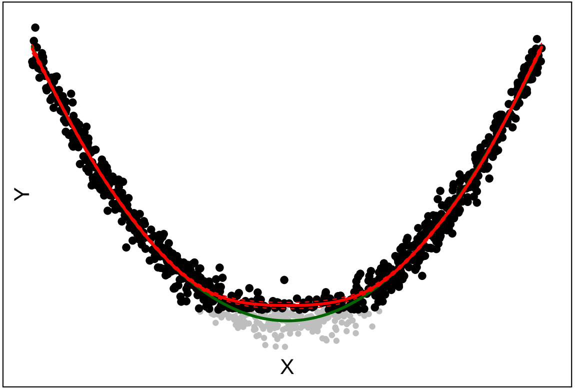


Nonignorable selection bias

$Y \not\perp\!\!\!\perp S \mid X$, hence $\mathbb{E}[Y|X, S = 1] \neq \mathbb{E}[Y|X]$

$$X \sim \mathcal{U}([-5, 5])$$
$$Y = X^2 + \mathcal{N}(0, 1)$$
$$S = \mathbb{1}\{Y > 1\}$$

$$\begin{tikzcd} X \arrow[r] & Y \arrow[dl] \\ S & \end{tikzcd}$$



Motivating example: cervical cancer screening

X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Motivating example: cervical cancer screening

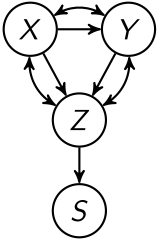
X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Answer: It depends on the selection mechanism. Suppose:



Z: Symptoms

Then $Y \not\perp\!\!\!\perp S \mid X$, so $\mathbb{E}[Y|X, S = 1] \neq \mathbb{E}[Y|X]$.

Answer: No.

Motivating example: cervical cancer screening

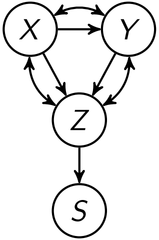
X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

- Patients in this dataset are self-selected: their own initiative caused them to be recorded in this dataset.

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Answer: It depends on the selection mechanism. Suppose:



Z: Symptoms

Then $Y \not\perp\!\!\!\perp S \mid X$, so $\mathbb{E}[Y|X, S = 1] \neq \mathbb{E}[Y|X]$.

Answer: No.

But $Y \perp\!\!\!\perp S \mid X, Z$. Can we leverage this somehow?

Repeated regression

- We have $Y \perp\!\!\!\perp S \mid X, Z$, hence we can write[3]

$$\begin{aligned}\mathbb{E}[Y|X] &= \mathbb{E}[\mathbb{E}[Y|Z,X]|X] \\ &= \mathbb{E}[\mathbb{E}[Y|X,Z,S=1]|X].\end{aligned}$$

- If we have data from $\mathbb{P}(X, Y, Z|S=1)$, then we can estimate $\hat{\mathbb{E}}[Y|X,Z,S=1]$...
- and if we additionally have data $(x,z) \sim \mathbb{P}(X,Z)$, we can
  - generate pseudo-labels $\tilde{Y} := \hat{\mathbb{E}}[Y|X=x, Z=z, S=1]$
  - and regress $\mathbb{E}[\tilde{Y}|X]$.[4]

[3] Bareinboim et al. [2014]
[4] Boeken et al. [2023]
[5] Hernán and Robins [2021], known as *standardization* or *outcome regression*

Repeated regression

- We have $Y \perp\!\!\!\perp S \mid X, Z$, hence we can write[3]

$$\begin{aligned}\mathbb{E}[Y|X] &= \mathbb{E}[\mathbb{E}[Y|Z,X]|X] \\ &= \mathbb{E}[\mathbb{E}[Y|X,Z,S=1]|X].\end{aligned}$$

- If we have data from $\mathbb{P}(X, Y, Z|S=1)$, then we can estimate $\hat{\mathbb{E}}[Y|X,Z,S=1]$...
- and if we additionally have data $(x, z) \sim \mathbb{P}(X, Z)$, we can
  - generate pseudo-labels $\tilde{Y} := \hat{\mathbb{E}}[Y|X=x, Z=z, S=1]$
  - and regress $\mathbb{E}[\tilde{Y}|X]$.[4]
- Positivity assumption: $\mathrm{supp}(\mathbb{P}(X, Z|S=1)) = \mathrm{supp}(\mathbb{P}(X, Z))$

[3] Bareinboim et al. [2014]

[4] Boeken et al. [2023]

[5] Hernán and Robins [2021], known as *standardization* or *outcome regression*

Repeated regression

- We have $Y \perp\!\!\!\perp S \mid X, Z$, hence we can write[3]

$$\begin{aligned}\mathbb{E}[Y|X] &= \mathbb{E}[\mathbb{E}[Y|Z, X]|X] \\ &= \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X].\end{aligned}$$

- If we have data from $\mathbb{P}(X, Y, Z|S = 1)$, then we can estimate $\hat{\mathbb{E}}[Y|X, Z, S = 1]$...
- and if we additionally have data $(x, z) \sim \mathbb{P}(X, Z)$, we can
  - generate pseudo-labels $\tilde{Y} := \hat{\mathbb{E}}[Y|X = x, Z = z, S = 1]$
  - and regress $\mathbb{E}[\tilde{Y}|X]$.[4]
- Positivity assumption: $\mathrm{supp}(\mathbb{P}(X, Z|S = 1)) = \mathrm{supp}(\mathbb{P}(X, Z))$

Closely related to standardization/outcome regression:[5]

$$\mathbb{E}[Y|\,\mathrm{do}(X = x)] = \mathbb{E}[\mathbb{E}[Y|X = x, Z]]$$

[3] Bareinboim et al. [2014]

[4] Boeken et al. [2023]

[5] Hernán and Robins [2021], known as *standardization* or *outcome regression*

Repeated regression

X	Z	Y	S
x_1	z_1	y_1	1
$\vdots$			
x_m	z_m	y_m	1
x_{m+1}	z_{m+1}	y_{m+1}	0
$\vdots$			
x_k	z_k	y_k	0

$\mathbb{P}(X, Y, Z | S = 1)$

X	Z
x_1	z_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_n	z_n

$\mathbb{P}(X, Z)$

Repeated regression

X	Z	Y	S
x_1	z_1	y_1	1
$\vdots$			
x_m	z_m	y_m	1
x_{m+1}	z_{m+1}	y_{m+1}	0
$\vdots$			
x_k	z_k	y_k	0

$$\mathbb{P}(X, Y, Z | S = 1)$$
$$\implies \hat{\mathbb{E}}[Y | X, Z, S = 1]$$

X	Z
x_1	z_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_n	z_n

$$\mathbb{P}(X, Z)$$

Repeated regression

X	Z	Y	S
x_1	z_1	y_1	1
$\vdots$			
x_m	z_m	y_m	1
x_{m+1}	z_{m+1}	y_{m+1}	0
$\vdots$			
x_k	z_k	y_k	0

$$\mathbb{P}(X, Y, Z | S = 1)$$
$$\implies \hat{\mathbb{E}}[Y | X, Z, S = 1]$$

X	Z	$\tilde{Y}$
x_1	z_1	$\tilde{y}_1$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
x_n	z_n	$\tilde{y}_n$

$$\mathbb{P}(X, Z, \tilde{Y})$$

Repeated regression

X	Z	Y	S
x_1	z_1	y_1	1
$\vdots$			
x_m	z_m	y_m	1
x_{m+1}	z_{m+1}	y_{m+1}	0
$\vdots$			
x_k	z_k	y_k	0

$$\mathbb{P}(X, Y, Z|S = 1)$$
$$\implies \hat{\mathbb{E}}[Y|X, Z, S = 1]$$

X	Z	$\tilde{Y}$
x_1	z_1	$\tilde{y}_1$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
x_n	z_n	$\tilde{y}_n$

$$\mathbb{P}(X, Z, \tilde{Y})$$
$$\implies \hat{\mathbb{E}}[\tilde{Y}|X]$$

Learning using privileged information

We have $Y \perp\!\!\!\perp S \mid X, Z$, so $\mathbb{E}[Y|X, Z, S = 1] = \mathbb{E}[Y|X, Z]$, so the model $\hat{\mathbb{E}}[Y|X, Z, S = 1]$ is already unbiased... Why not consider this model, instead of estimating $\mathbb{E}[Y|X]$?
In practice, measuring Z at test time might be **costly** or **unfeasible**.

We consider a learning paradigm called Learning Using Privileged Information (LUPI), where, at the training stage, additional information Z is provided about training example X.

The goal of the LUPI paradigm is to use privileged information to significantly increase the rate of convergence.[6]

We have just shown that privileged information can also be used to recover from selection bias.

[6] Vapnik and Vashist [2009], Vapnik and Izmailov [2015]

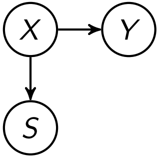
Extending the taxonomy of selection mechanisms

For estimating $\mathbb{E}[Y|X]$ from $\mathbb{P}(X, Y|S = 1)$, selection is:

Ignorable

$$Y \perp\!\!\!\perp S|X$$

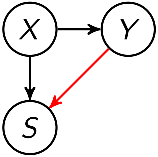
$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, S = 1]$$



Nonignorable

$$Y \not\!\perp\!\!\!\perp S|X$$

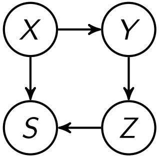
$$\mathbb{E}[Y|X] \neq \mathbb{E}[Y|X, S = 1]$$



Privilegedly ignorable[7]

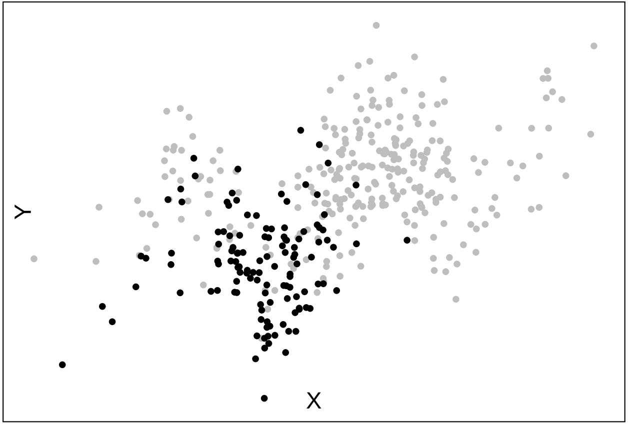
$$Y \perp\!\!\!\perp S|X, Z, \quad \mathbb{P}(X, Z)$$

$$\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X]$$

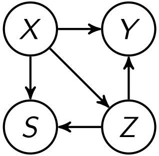


[7] Boeken et al. [2023]

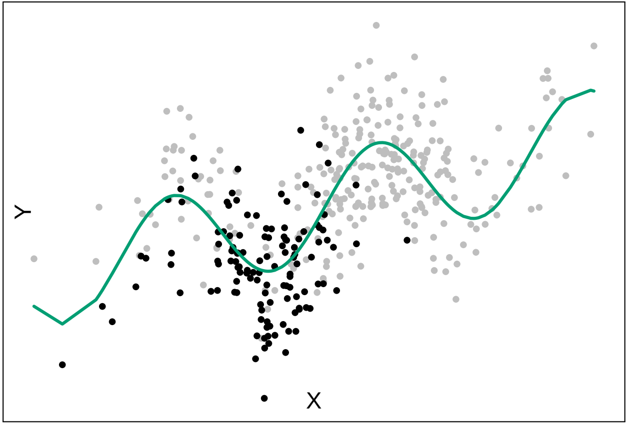
Simulated example



$$X = \varepsilon_X$$
$$Z = 3\sin(X) + \varepsilon_Z$$
$$Y = \frac{1}{2}X + Z + \varepsilon_Y$$
$$S \sim \mathrm{Bernoulli}(p_S(X, Z))$$



Simulated example



$$X = \varepsilon_X$$

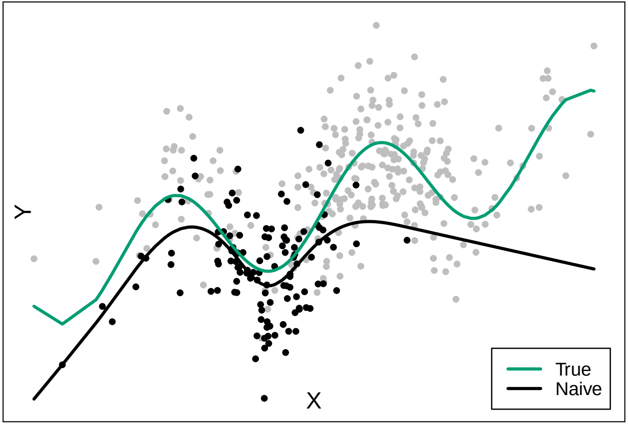
$$Z = 3\sin(X) + \varepsilon_Z$$

$$Y = \frac{1}{2}X + Z + \varepsilon_Y$$

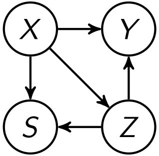
$$S \sim \mathrm{Bernoulli}(p_S(X, Z))$$

$$\begin{tikzcd} X \arrow[r] \arrow[d] \arrow[dr] & Y \\ S & Z \arrow[u] \arrow[l] \end{tikzcd}$$

Simulated example



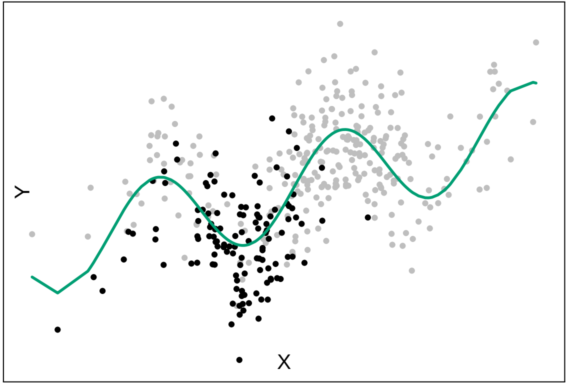
$$X = \varepsilon_X$$
$$Z = 3\sin(X) + \varepsilon_Z$$
$$Y = \frac{1}{2}X + Z + \varepsilon_Y$$
$$S \sim \mathrm{Bernoulli}(p_S(X, Z))$$



Simulated example: repeated regression

$Y \perp\!\!\!\perp S \mid X, Z$, so

$$\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X]$$



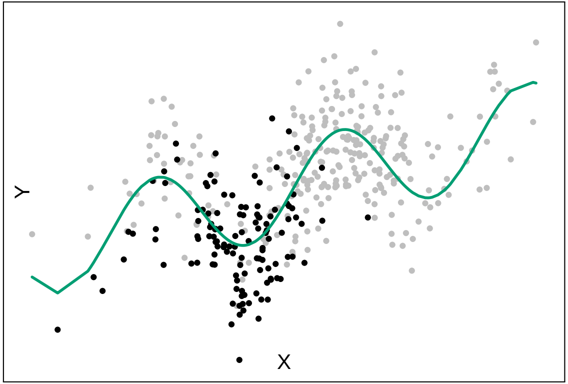
Simulated example: repeated regression

$Y \perp\!\!\!\perp S \mid X, Z$, so

$$\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X]$$

1. Estimate

$$\begin{aligned}\tilde{\mu}(x, z) &= \hat{\mathbb{E}}[Y|X = x, Z = z, S = 1] \\ &\approx \frac{1}{2}x + z\end{aligned}$$



Simulated example: repeated regression

$Y \perp\!\!\!\perp S \mid X, Z$, so

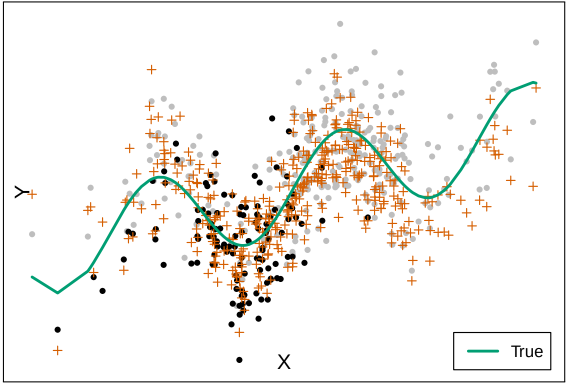
$$\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X]$$

1. Estimate

$$\tilde{\mu}(x, z) = \hat{\mathbb{E}}[Y|X = x, Z = z, S = 1] \approx \frac{1}{2}x + z$$

2. Generate pseudo-labels

$$\tilde{Y}_i = \tilde{\mu}(X_i, Z_i)$$



Simulated example: repeated regression

$Y \perp\!\!\!\perp S \mid X, Z$, so

$$\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X]$$

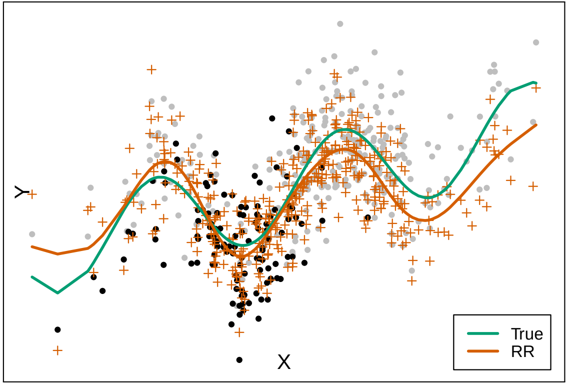
1. Estimate

$$\begin{aligned}\tilde{\mu}(x, z) &= \hat{\mathbb{E}}[Y|X = x, Z = z, S = 1] \\ &\approx \frac{1}{2}x + z\end{aligned}$$

2. Generate pseudo-labels

$$\tilde{Y}_i = \tilde{\mu}(X_i, Z_i)$$

3. Fit $\hat{\mu}(x) := \hat{\mathbb{E}}[\tilde{Y}|X]$



Motivating example: cervical cancer screening

We have data:

X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

Z: Symptoms

- Patients are self-selected, so we have data from $\mathbb{P}(X, Y, Z|S = 1)$

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Motivating example: cervical cancer screening

We have data:

X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

Z: Symptoms

- Patients are self-selected, so we have data from $\mathbb{P}(X, Y, Z | S = 1)$

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y | X, S = 1]$ for population screening?

Answer: Depends on the selection mechanism...

Motivating example: cervical cancer screening

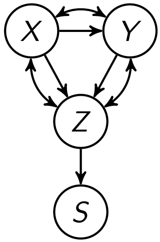
We have data:

X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

Z: Symptoms

- Patients are self-selected, so we have data from $\mathbb{P}(X, Y, Z|S = 1)$



Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Answer: Depends on the selection mechanism...

Motivating example: cervical cancer screening

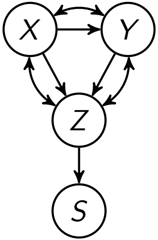
We have data:

X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)

Y: Presence of cervical cancer

Z: Symptoms

- Patients are self-selected, so we have data from $\mathbb{P}(X, Y, Z | S = 1)$

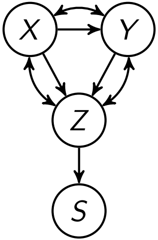


Question: Can we use the estimated model $\hat{\mathbb{E}}[Y | X, S = 1]$ for population screening?

Answer: Depends on the selection mechanism...

... so no...

Motivating example: cervical cancer screening



We have data:

- X: Demographic and medical information, available through digital medical record (age, use of contraceptives, STDs, etc.)
- Y: Presence of cervical cancer
- Z: Symptoms
- Patients are self-selected, so we have data from $\mathbb{P}(X, Y, Z | S = 1)$

Question: Can we use the estimated model $\hat{\mathbb{E}}[Y|X, S = 1]$ for population screening?

Answer: Depends on the selection mechanism...

... so no...

...but if we additionally have data from $\mathbb{P}(X, Z)$, then we can estimate $\mathbb{E}[Y|X]$ with repeated regression!

Summary (selection bias)

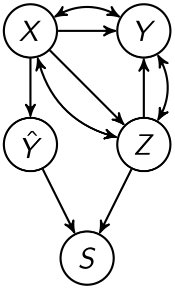
When estimating $\mathbb{E}[Y|X]$ from a dataset with selection bias:

- It is generally not testable whether we have to correct for bias
- Can motivate this by modelling the causal graph of the DGP
- Ignorable: $Y \perp\!\!\!\perp S \mid X$, then no correction is necessary.
  Watch out for positivity violations!
- Nonignorable: $Y \not\perp\!\!\!\perp S \mid X$, naive regression is biased, but
  - Privilegedly ignorable: $Y \perp\!\!\!\perp S \mid X, Z$ and unbiased sample $\mathbb{P}(X, Z)$, then we can apply the repeated regression procedure

Missing response (selective labelling)

Example: selective labelling (the bank loan problem)

- X: digital data in loan application
- Y: default
- $\hat{Y}$: estimated probability of default
- Z: information from interview
- S: issue of the loan



Goal: re-train $\hat{Y} = \hat{\mathbb{E}}[Y|X]$

Available data: missing response

X	Z	S	Y
x_1	z_1	1	y_1
$\vdots$			
x_m	z_m	1	y_m
x_{m+1}	z_{m+1}	0	y_{m+1}
$\vdots$			
x_n	z_n	0	y_n

$\mathbb{P}(X, Y, Z | S = 1)$

$\mathbb{P}(X, Z, S)$

Available data: missing reponse vs. selection bias

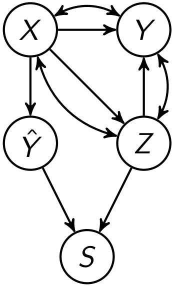
X	Z	S	Y
x_1	z_1	1	y_1
$\vdots$			
x_m	z_m	1	y_m
x_{m+1}	z_{m+1}	0	y_{m+1}
$\vdots$			
x_n	z_n	0	y_n

X	Z	S	Y
x_1	z_1	1	y_1
$\vdots$			
x_m	z_m	1	y_m
x_{m+1}	z_{m+1}	0	y_{m+1}
$\vdots$			
x_n	z_n	0	y_n

X	Z
x_1	z_1
$\vdots$	$\vdots$
$\vdots$	$\vdots$
$\vdots$	$\vdots$
x_n	z_n

Example: selective labelling (automated hiring)

- X: job application (cv, letter)
- Y: successful hire (binary)
- $\hat{Y}$: estimated probability of success
- Z: psychological test
- S: hire



Goal: re-train $\hat{Y} = \hat{\mathbb{E}}[Y|X]$

We have $Y \not\perp\!\!\!\perp S \mid X$, so $\mathbb{E}[Y|X, S = 1] \neq \mathbb{E}[Y|X]$

Exercise (work in pairs)

Hypothesize a setting in which we have

- covariates X
- target variable Y
- prediction $\hat{Y} = \hat{\mathbb{E}}[Y|X]$
- privileged information Z
- selection indicator S;

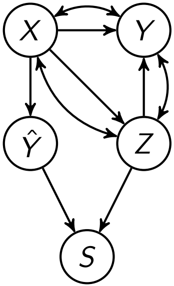
draw a causal graph G of this setting, and check that it satisfies

- $X \not\perp_G Y$
- $Y \not\perp_G S \mid X$
- $Y \perp_G S \mid X, Z.$

BREAK

Example: selective labelling (automated hiring)

- X: job application (cv, letter)
- Y: successful hire (binary)
- $\hat{Y}$: estimated probability of success
- Z: psychological test
- S: hire



Goal: re-train $\hat{Y} = \hat{\mathbb{E}}[Y|X]$

We have $Y \not\perp\!\!\!\perp S \mid X$, so $\mathbb{E}[Y|X, S=1] \neq \mathbb{E}[Y|X]$, but $Y \perp\!\!\!\perp S \mid X, Z$, so...

Repeated regression for missing response

X	Z	S	Y
x_1	z_1	1	y_1
$\vdots$			
x_m	z_m	1	y_m
x_{m+1}	z_{m+1}	0	y_{m+1}
$\vdots$			
x_n	z_n	0	y_n

$\mathbb{P}(X, Y, Z | S = 1)$

$\mathbb{P}(X, Z, S)$

Repeated regression for missing response

X	Z	S	Y
x_1	z_1	1	y_1
$\vdots$			
x_m	z_m	1	y_m
x_{m+1}	z_{m+1}	0	y_{m+1}
$\vdots$			
x_n	z_n	0	y_n

$\mathbb{P}(X, Z, S)$

$\mathbb{P}(X, Y, Z | S = 1)$

$\implies \hat{\mathbb{E}}[Y | X, Z, S = 1]$

Repeated regression for missing response

X	Z	S	Y	$\tilde{Y}$
x_1	z_1	1	y_1	$\tilde{y}_1$
$\vdots$				
x_m	z_m	1	y_m	$\tilde{y}_m$
x_{m+1}	z_{m+1}	0	y_{m+1}	$\tilde{y}_{m+1}$
$\vdots$				
x_n	z_n	0	y_n	$\tilde{y}_n$

$\mathbb{P}(X, Z, \tilde{Y}, S)$

$\mathbb{P}(X, Y, Z | S = 1)$

$\Longrightarrow \hat{\mathbb{E}}[Y | X, Z, S = 1]$

Repeated regression for missing response

X	Z	S	Y	$\tilde{Y}$
x_1	z_1	1	y_1	$\tilde{y}_1$
$\vdots$				
x_m	z_m	1	y_m	$\tilde{y}_m$
x_{m+1}	z_{m+1}	0	y_{m+1}	$\tilde{y}_{m+1}$
$\vdots$				
x_n	z_n	0	y_n	$\tilde{y}_n$

$$\mathbb{P}(X, Y, Z | S = 1)$$
$$\implies \hat{\mathbb{E}}[Y | X, Z, S = 1]$$

$$\mathbb{P}(X, Z, \tilde{Y}, S)$$
$$\implies \mathbb{E}[\tilde{Y} | X]$$

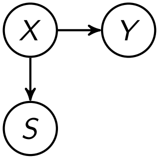
A taxonomy of missingness mechanisms[10]

For estimating $\mathbb{E}[Y|X]$ from $\mathbb{P}(X, Y|S = 1)$, selection is:

Ignorable (MAR)[8]

$$Y \perp\!\!\!\perp S | X$$

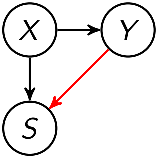
$$\mathbb{E}[Y|X] = \mathbb{E}[Y|X, S = 1]$$



Nonignorable (MNAR)

$$Y \not\perp\!\!\!\perp S | X$$

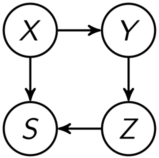
$$\mathbb{E}[Y|X] \neq \mathbb{E}[Y|X, S = 1]$$



Privilegedly ignorable (PMAR)[9]

$$Y \perp\!\!\!\perp S | X, Z$$

$$\mathbb{E}[Y|X] = \mathbb{E}[\mathbb{E}[Y|X, Z, S = 1]|X]$$



[9] Robins and Rotnitzky [1995]
[9] Boeken et al. [2023]
[10] Rubin [1976]

Weighted regression

Empirical risk minimization:
Assuming e.g. a parametric model $\mathbb{E}[Y|X] = g(X; \beta)$, given data $(X_1, Y_1), ..., (X_n, Y_n) \sim \mathbb{P}(X, Y)$ estimate

$$\hat{\beta} := \arg\min_{\beta} \hat{\mathbb{E}}[\ell(X, Y)] = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} \ell(g(X_i; \beta), Y_i)$$

and use $\hat{\mathbb{E}}[Y|X = x] = g(x; \hat{\beta})$.

Weighted regression[11]

Assuming PMAR we have $Y \perp\!\!\!\perp S \mid X, Z$, so

[11] Horvitz and Thompson [1952], inverse probability weighting, inverse propensity weighting

Weighted regression[11]

Assuming PMAR we have $Y \perp\!\!\!\perp S \mid X, Z$, so

$$\begin{aligned}
\mathbb{E}[\ell(X,Y)] &= \int \ell(x,y)p(x,y,z)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \int \ell(x,y)\frac{p(x,y,z)}{p(x,y,z|S=1)}p(x,y,z|S=1)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \int \ell(x,y)\frac{p(S=1)}{p(S=1|x,y,z)}p(x,y,z|S=1)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \int \ell(x,y)\frac{p(S=1)}{p(S=1|x,z)}p(x,y,z|S=1)\mathrm{d}x\mathrm{d}y\mathrm{d}z \\
&= \mathbb{E}[w(X,Z)\ell(x,y)|S=1]
\end{aligned}$$

$$w(X,Z) = \mathbb{P}(S=1)/\mathbb{P}(S=1|X,Z)$$

Weighted regression[11]

Assuming PMAR we have $Y \perp\!\!\!\perp S \mid X, Z$, so

$$\mathbb{E}[\ell(X, Y)] = \mathbb{E}[w(X, Z)\ell(x, y)|S = 1]$$
$$w(X, Z) = \mathbb{P}(S = 1)/\mathbb{P}(S = 1|X, Z)$$

[11] Horvitz and Thompson [1952], inverse probability weighting, inverse propensity weighting

Weighted regression[11]

Assuming PMAR we have $Y \perp\!\!\!\perp S \mid X, Z$, so

$$\mathbb{E}[\ell(X, Y)] = \mathbb{E}[w(X, Z)\ell(x, y)|S = 1]$$
$$w(X, Z) = \mathbb{P}(S = 1)/\mathbb{P}(S = 1|X, Z)$$

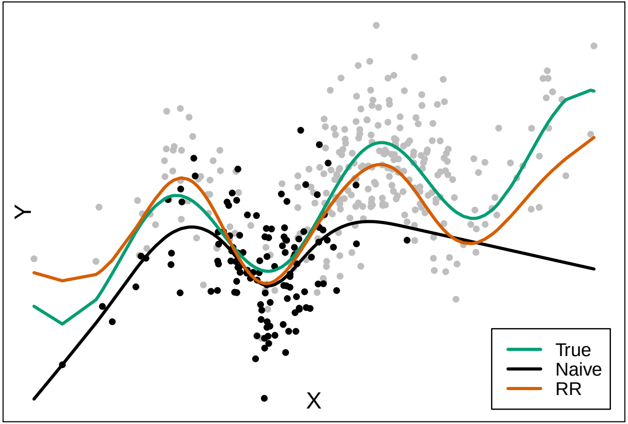
Given data $(X_1, Y_1, Z_1), \dots, (X_n, Y_n, Z_n) \sim \mathbb{P}(X, Y, Z|S = 1)$ estimate

$$\hat{\beta} := \arg\min_{\beta} \sum_{i=1}^{n} w(X_i, Z_i)\ell(g(X_i; \beta), Y_i)$$

and use $\hat{\mathbb{E}}[Y|X = x] = g(x; \hat{\beta})$.

[11] Horvitz and Thompson [1952], inverse probability weighting, inverse propensity weighting

Simulated example



$$X = \varepsilon_X$$

$$Z = 3\sin(X) + \varepsilon_Z$$

$$Y = \frac{1}{2}X + Z + \varepsilon_Y$$

$$S \sim \mathrm{Bernoulli}(p_S(X, Z))$$

$$\begin{tikzcd}
X \arrow[r] \arrow[d] \arrow[dr] & Y \\
S & Z \arrow[l] \arrow[u]
\end{tikzcd}$$

Simulated example: weighted regression

$Y \perp\!\!\!\perp S \mid X, Z$, so

$$\mathbb{E}[\ell(X,Y)] = \mathbb{E}[w(X,Z)\ell(X,Y)|S=1]$$
$$w(X,Z) = \mathbb{P}(S=1)/\mathbb{P}(S=1|X,Z)$$

Simulated example: weighted regression

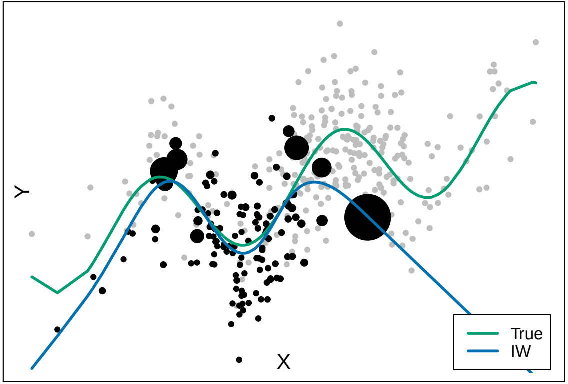
$Y \perp\!\!\!\perp S \,|\, X, Z$, so

$$\mathbb{E}[\ell(X, Y)] = \mathbb{E}[w(X, Z)\ell(X, Y)|S = 1]$$
$$w(X, Z) = \mathbb{P}(S = 1)/\mathbb{P}(S = 1|X, Z)$$

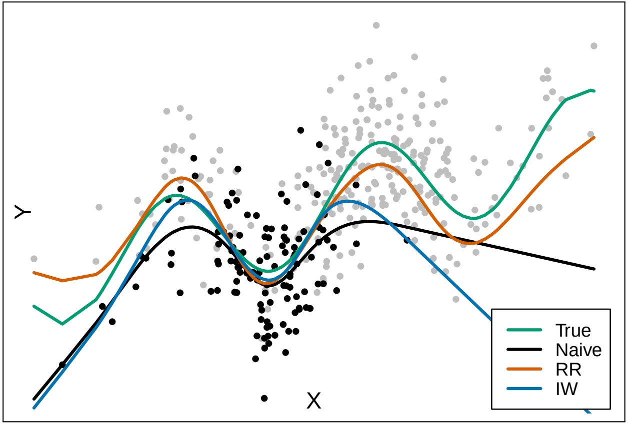
Assuming e.g. a parametric model $\mathbb{E}[Y|X] = g(X; \beta)$, estimate

$$\hat{\beta} := \arg\min_{\beta} \sum_{i=1}^{n} w(X_i, Z_i)\ell(g(X_i; \beta), Y_i)$$

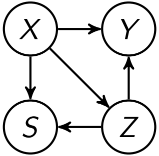
and use $\hat{\mathbb{E}}[Y|X = x] = g(x; \hat{\beta})$.



Simulated example: comparing methods



$$X = \varepsilon_X$$
$$Z = 3\sin(X) + \varepsilon_Z$$
$$Y = \frac{1}{2}X + Z + \varepsilon_Y$$
$$S \sim \mathrm{Bernoulli}(p_S(X, Z))$$



Summary

When estimating $\mathbb{E}[Y|X]$ from a dataset with selection bias:

- It is generally not testable whether we have to correct for bias
- Can motivate this by modelling the causal graph of the DGP
- Ignorable: $Y \perp\!\!\!\perp S \mid X$, then no correction is necessary.
  Watch out for positivity violations!
- Nonignorable: $Y \not\!\perp\!\!\!\perp S \mid X$, naive regression is biased, but
  - Privilegedly ignorable: $Y \perp\!\!\!\perp S \mid X, Z$ and unbiased sample $\mathbb{P}(X, Z)$, then we can apply the repeated regression procedure

Missingness response / selective labelling:

- Example: prediction models used for selective labelling
- Same characterisation of the regression problem under different missingness mechanisms
- Characterisation not testable, but can be motivated with causal model.
- Repeated regression can also be applied for re-training
- Importance weighting as an alternative estimation method

Takeaways

Before deploying an ML model, pay attention to any mismatch between your train and test set.

Causal modelling is a convenient tool for characterising such differences!

Repeated regression and importance weighting can be used for estimating a regression model from biased data.

References I

E. Bareinboim, J. Tian, and J. Pearl. Recovering from Selection Bias in Causal and Statistical Inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, page 12, 2014.

P. Boeken, N. de Kroon, M. de Jong, J. M. Mooij, and O. Zoeter. Correcting for Selection Bias and Missing Response in Regression using Privileged Information. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence (To appear)*. PMLR, 2023.

M. Hernán and J. M. Robins. *Causal Inference*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Chapman & Hall/CRC, Boca Raton, 2021. ISBN 978-1-4200-7616-5.

D. G. Horvitz and D. J. Thompson. A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47(260):663–685, 1952. ISSN 0162-1459. doi: 10.2307/2280784. URL `https://www.jstor.org/stable/2280784`.

J. M. Robins and A. Rotnitzky. Semiparametric Efficiency in Multivariate Regression Models with Missing Data. *Journal of the American Statistical Association*, 90(429):122–129, 1995. ISSN 0162-1459. doi: 10.2307/2291135. URL `https://www.jstor.org/stable/2291135`.

References II

D. B. Rubin. Inference and Missing Data. *Biometrika*, 63(3):581–592, 1976. ISSN 0006-3444. doi: 10.2307/2335739. URL https://www.jstor.org/stable/2335739.

V. Vapnik and R. Izmailov. Learning Using Privileged Information: Similarity Control and Knowledge Transfer. *Journal of Machine Learning Research*, 16(61):2023–2049, 2015. ISSN 1533-7928. URL http://jmlr.org/papers/v16/vapnik15b.html.

V. Vapnik and A. Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5):544–557, July 2009. ISSN 0893-6080. doi: 10.1016/j.neunet.2009.06.042. URL https://www.sciencedirect.com/science/article/pii/S0893608009001130.

Wei Fan, I. Davidson, B. Zadrozny, and Philip S. Yu. An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 605–608, Houston, TX, USA, 2005. IEEE. ISBN 978-0-7695-2278-4. doi: 10.1109/ICDM.2005.24. URL http://ieeexplore.ieee.org/document/1565737/.

B. Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114, 2004.