Correcting for Selection Bias and Missing Response in Regression using Privileged Information

University of Amsterdam

Booking.com

Philip Boeken [1 2] Noud de Kroon [1] Mathijs de Jong [2] Joris M. Mooij [1] Onno Zoeter [2]

[1]University of Amsterdam [2]Booking.com

Summary

Setting: features X, target Y, missingness or selection indicator S.
Task: estimate regression model $\mathbb{E}[Y|X]$ from $\mathbb{P}(X, Y|S = 1)$.

1. *Nonignorable* missingness/selection bias: $Y \not\perp\!\!\!\perp S \mid X$ (so $\mathbb{E}[Y|X] \neq \mathbb{E}[Y|X, S = 1]$)

   Privilegedly ignorable: $Y \perp\!\!\!\perp S \mid X, Z$ with Z available during training, not at test time.
2. Three estimation procedures: (suitable for privileged information)
   a. Repeated regression;
   b. Importance weighting;
   c. A doubly robust combination of the two.
3. Empirically:
   a. All three methods can appropriately correct for bias.
   b. Repeated regression extrapolates better than importance weighting.
4. Practical challenges of evaluation:
   a. Evaluation metrics on biased data don't necessarily correlate with deployment performance.
   b. Evaluation metrics depend on auxiliary models.

(a) Nonignorable (b) Privilegedly ignorable

Settings

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

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
x_1	z_1	s_1
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$
x_n	z_n	s_n

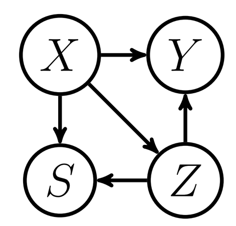
General assumptions

An underlying distribution $\mathbb{P}(X, Y, Z, S)$ such that $Y \perp\!\!\!\perp S \mid X, Z$.

Train	Test
$\mathbb{P}(X, Y, Z \mid S = 1)$	Given x, predict $\mathbb{E}[Y \mid X = x]$
$\mathbb{P}(X, Z, S)$	

Running example

$$X \sim \mathcal{N}$$
$$Z = 3\sin(X) + \mathcal{N}$$
$$Y = \frac{1}{2}X + Z + \mathcal{N}$$
$$S \sim \mathrm{Bern}(\sigma(X)\sigma(Z))$$



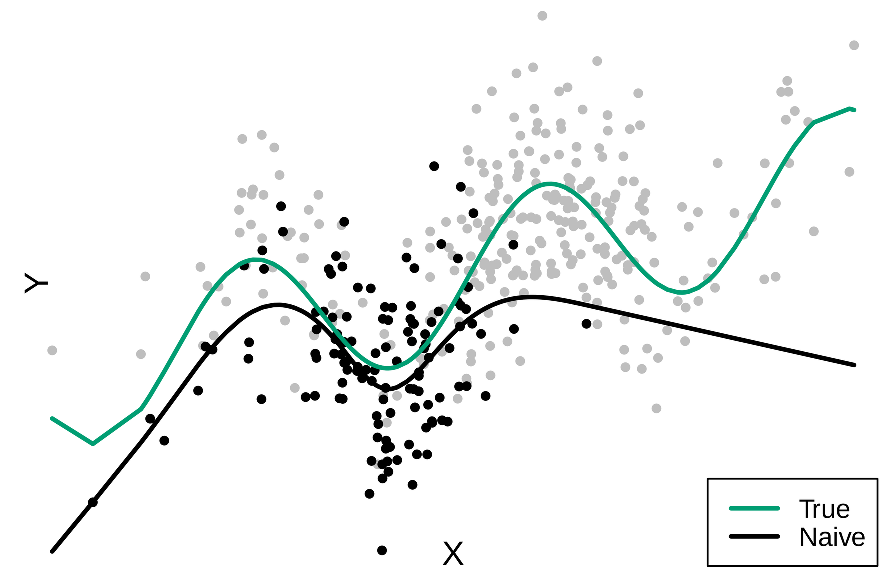


Figure 2. Black dots are observed, grey dots are missing. Task: fit the green line.

Repeated regression

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

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

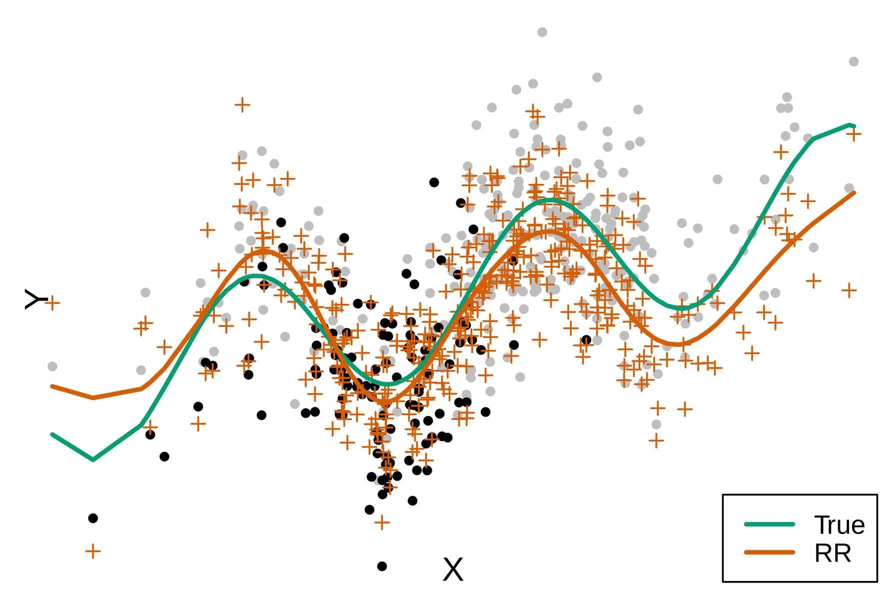


Figure 3. Orange crosses: imputed values. Orange line: fitted repeated regression model.

Importance weighting

Assuming e.g. $\mathbb{E}[Y|X] = g(X; \beta)$, we have

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

so estimate

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

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

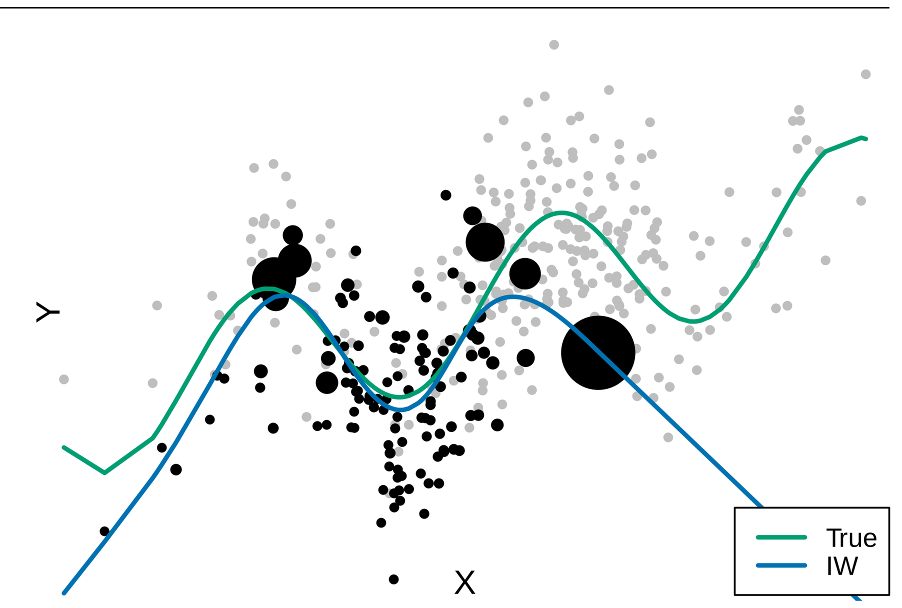


Figure 4. Dot size represents the associated weight. Blue line: fitted weighted regression model.

Double robust regression

1. Estimate $\hat{\mu}_{RR}(x)$ with repeated regression.
2. Calculate residuals $R = Y - \hat{\mu}_{RR}(X)$.
3. Estimate $\hat{r}_{IW}(x) \approx \mathbb{E}[R|X = x]$ with importance weighting.
4. Construct
$$\hat{\mu}_{DR}(x) := \hat{\mu}_{RR}(x) + \hat{r}_{IW}(x).$$

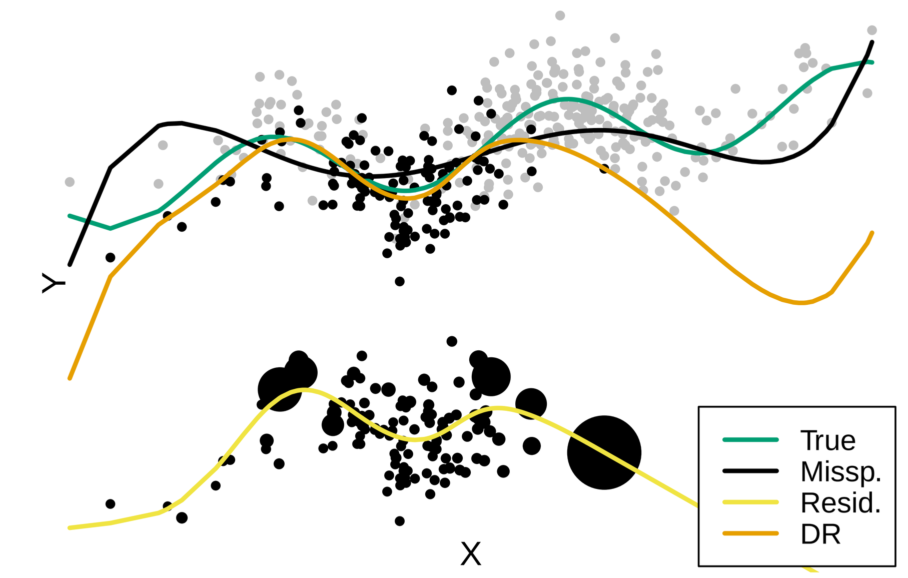


Figure 5. Double robust regression.

Simulations

- Iterate all ADMGs with variables X, Y, Z, S such that
$$X \not\perp\!\!\!\perp Y$$
$$Y \not\perp\!\!\!\perp S \mid X$$
$$Y \perp\!\!\!\perp S \mid X, Z$$
- Simulate data according to additive noise structural equation model $V = f(\mathrm{Pa}(V)) + \varepsilon$, with
$$f \sim \mathcal{GP} \qquad \varepsilon \sim \mathcal{GP}(\mathrm{Unif}[0, 1]) \qquad S \sim \mathrm{Bern}\left(\prod_{v \in \mathrm{Pa}(S)} \sigma(v)\right)$$

	MSE	naive MSE	pseudo-label MSE	weighted MSE
Naive	3.11 (20.6)	0.85 (0.5)	2.81 (18.0)	0.90 (0.6)
RR	**2.01** (2.6)	1.26 (1.4)	**0.60** (0.9)	1.20 (1.3)
IW	4.18 (23.2)	0.86 (0.5)	3.83 (20.7)	0.91 (0.6)
DR	4.51 (45.1)	**0.81** (0.4)	4.47 (49.0)	**0.79** (0.4)

Table 1. Means (standard deviations) over 27.500 simulated datasets.

	MSE	MSE-interp.	MSE-extrap.
Naive	3.31 (8.6)	1.28 (0.7)	5.51 (17.5)
RR	**2.13** (2.7)	1.46 (1.8)	**2.89** (4.2)
IW	5.82 (16.4)	1.27 (0.6)	10.81 (32.6)
DR	6.93 (72.3)	**1.24** (0.6)	12.34 (94.9)

Table 2. Interpolation and extrapolation results, on graphs where S and X are adjacent.