Week 3 Recap

David S. Rosenberg

NYU: CDS

February 17, 2021

Contents

1. Missing data setup
2. Inverse propensity score estimators
3. Regression imputation
4. Covariate shift
5. Where are we?

Missing data setup

MAR setup

- Assume we have **covariate** X_i about each individual i.
- Also assume that X_i is **never missing**.
- Full data: $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d $p(x, y)$.
- What we actually observe:

$$(X_1, R_1, R_1 Y_1), \ldots, (X_n, R_n, R_n Y_n).$$

- **MAR assumption**: $R_i \perp\!\!\!\perp Y_i \mid X_i$ for each i
    - i.e.$p(r, y \mid x) = p(r \mid x) p(y \mid x)$

The propensity score

- Key piece in the MAR setting is the model for missingness:

$$\mathbb{P}(R=1 \mid X=x, Y=y) = \mathbb{P}(R=1 \mid X=x) = \pi(x).$$

- $\pi(x)$ is called the **propensity score**.
- If the propensity score is 0, we have a blind spot in our input space
  - can't do anything about it (at least with our estimators)

Assumption

Unless otherwise noted, we will always assume that propensity scores are strictly positive: $\pi(x) > 0$.

Inverse propensity score estimators

Inverse propensity weighted (IPW) Mean

- When[1] $\pi(x) > 0 \ \forall x \in \mathcal{X}$,
  we can define the **IPW mean estimator** for $\mathbb{E}Y$:

$$
\begin{aligned}
\hat{\mu}_{\text{ipw}} &= \frac{1}{n} \sum_{i:R_i=1} \frac{Y_i}{\pi(X_i)} \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{R_i Y_i}{\pi(X_i)}
\end{aligned}
$$

- $\hat{\mu}_{\text{ipw}}$ is **unbiased** for $\mathbb{E}Y$ (i.e. $\mathbb{E}\hat{\mu}_{\text{ipw}} = \mathbb{E}Y$.)
- $\hat{\mu}_{\text{ipw}}$ is **consistent** for $\mathbb{E}Y$ (i.e. $\hat{\mu}_{\text{ipw}} \xrightarrow{P} \mathbb{E}Y$ as $n \to \infty$)

[1] We assume here and everywhere below that $\pi(x) > 0 \ \forall x \in \mathcal{X}$.

The self-normalized IPW estimator

- If we normalize by $\sum_{i=1}^{n} W_i R_i$ instead of n, we get

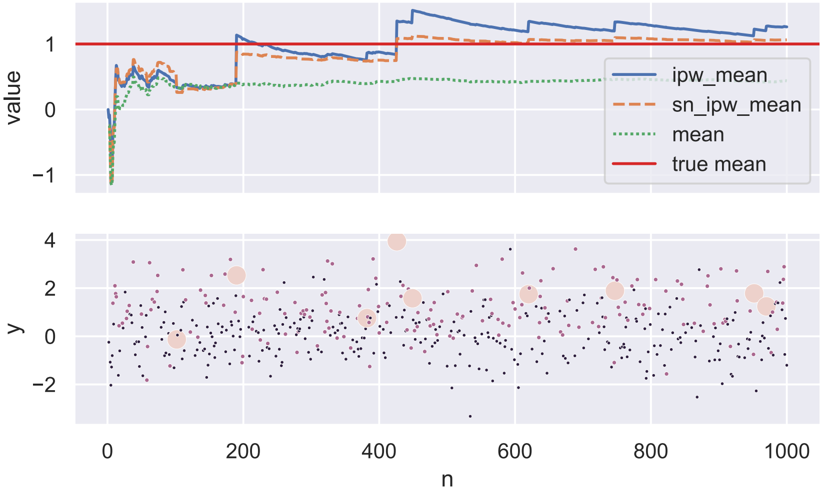
Definition (Self-normalized IPW mean)

For a dataset $(W_1, R_1, Y_1), \ldots, (W_n, R_n, Y_n)$ as described above,

$$\hat{\mu}_{\text{sn\_ipw}} = \frac{\sum_{i=1}^{n} W_i R_i Y_i}{\sum_{i=1}^{n} W_i R_i}$$

- In the MCAR case with $\pi(x) \equiv p$, $\hat{\mu}_{\text{sn\_ipw}} = \hat{\mu}_{\text{cc}}$ and seems preferable to $\hat{\mu}_{\text{ipw}}$.

Self-normalized IPW estimator on SeaVan1



IPW vs self-normalized IPW: 5000x

- We repeat the experiment above 5000 times (1000 samples each) and get the following.
- Recall that the true mean is $\mu = 1.0$.

estimator	mean	SD	SE	bias	RMSE
mean ($\hat{\mu}_{\text{cc}}$)	0.357244	0.050305	0.000711	-0.643534	0.645497
ipw_mean ($\hat{\mu}_{\text{ipw}}$)	0.995142	0.308634	0.004365	-0.005635	0.308686
sn_ipw_mean ($\hat{\mu}_{\text{sn\_ipw}}$)	0.978119	0.197319	0.002791	-0.022659	0.198615

Regression imputation

Regression imputation: basic idea

X	R	Y
x_1	1	y_1
x_2	0	?
x_3	0	?
x_4	1	y_4
$\vdots$	$\vdots$	$\vdots$
x_n	1	y_n

$\Longrightarrow$

X	R	Y
x_1	1	y_1
x_2	0	$\hat{f}(x_2)$
x_3	0	$\hat{f}(x_3)$
x_4	1	y_4
$\vdots$	$\vdots$	$\vdots$
x_n	1	y_n

- Fit $\hat{f}(x)$ on complete cases $(R = 1)$ to approximate $\mathbb{E}[Y \mid X = x]$.
- **Regression imputation estimator:** Estimate $\mathbb{E}Y$ with

$$\frac{1}{n}\left(y_1 + \hat{f}(x_2) + \hat{f}(x_3) + y_4 + \cdots + y_n\right).$$

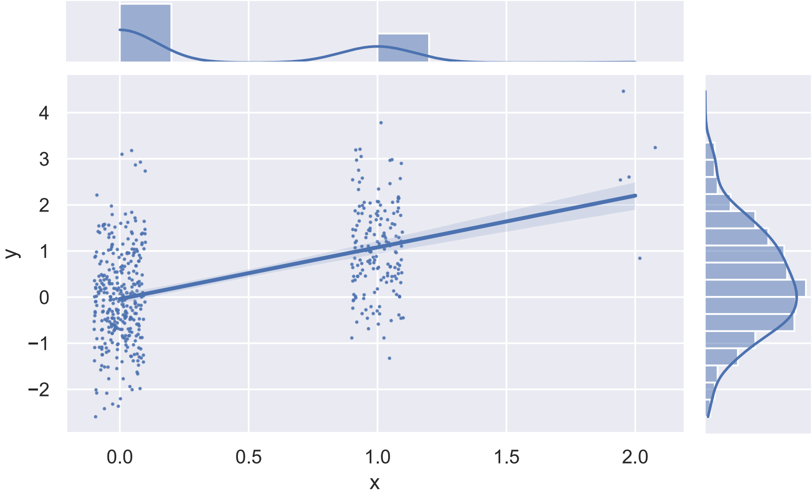
Well-specified and misspecified models

- In statistics, a **model** is a set of distributions
  - (or conditional distributions).
- A model is **well specified** if it contains the data-generating distribution.
  - Also referred to as **correctly specified.**
- If a model is not well specified, we say it's **misspecified** or **incorrectly specified**.
- We'll see that regression imputation has the following performance characteristics:

	MCAR	MAR
well specified	Good	Good
misspecified	OK/Good	**Bad**

MAR: SeaVan1 distribution illustrated

(X_i, Y_i) for which $R_i = 1$, i.e. the complete cases.



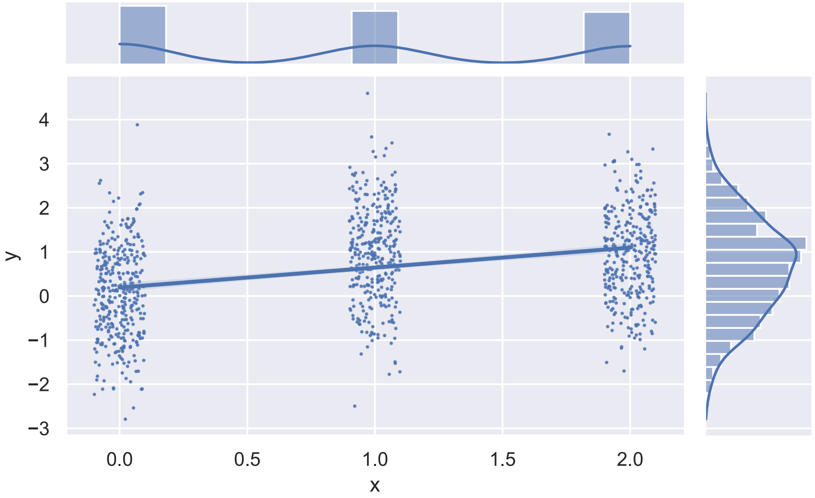
Performance on SeaVan1

- Fit $\hat{f}(x) = a + bx$ to the complete cases.
- Impute missing Y_i's with $\hat{f}(X_i)$...

estimator	mean	SD	SE	bias	RMSE
mean ($\hat{\mu}_{\text{cc}}$)	0.3572	0.0503	0.0007	-0.6435	0.6455
ipw_mean ($\hat{\mu}_{\text{ipw}}$)	0.9951	0.3086	0.0044	-0.0056	0.3087
sn_ipw_mean ($\hat{\mu}_{\text{sn\_ipw}}$)	0.9781	0.1973	0.0028	-0.0227	0.1986
impute_linear ($\hat{\mu}_{\hat{f}}$)	0.9989	0.0777	0.0011	-0.0018	**0.0777**

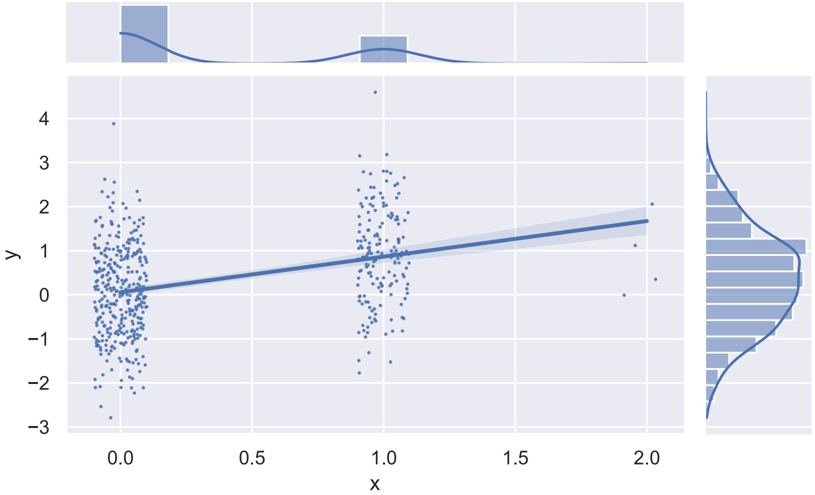
MAR: "SeaVan2" distribution illustrated

- Full data for sample of size $n = 1000$; $\mathbb{E}[Y \mid X = x] = \mathbb{1}[x \geqslant 1]$.



MAR: "SeaVan2" distribution illustrated

- Complete cases in sample of size $n = 1000$



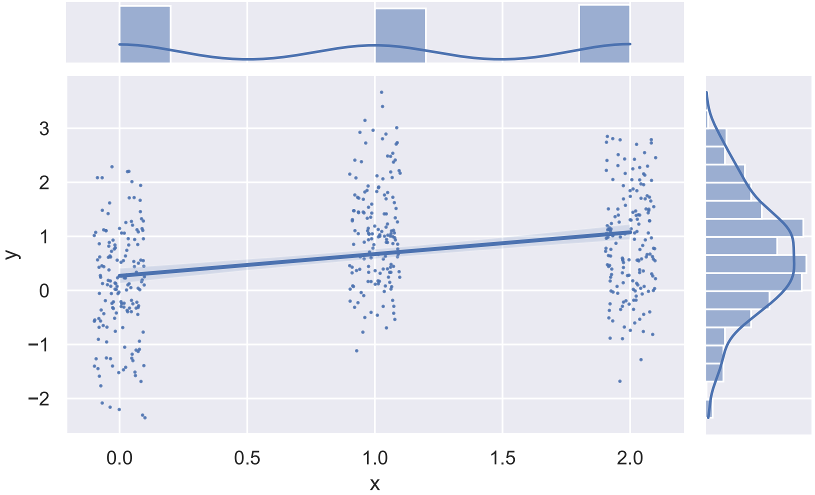
Performance on SeaVan2

- Fit $\hat{f}(x) = a + bx$ to the complete cases.

estimator	mean	SD	SE	bias	RMSE
mean $(\hat{\mu}_{cc})$	0.3453	0.0497	0.0007	-0.3221	0.3259
ipw_mean $(\hat{\mu}_{ipw})$	0.6634	0.1977	0.0028	-0.0040	0.1978
sn_ipw_mean $(\hat{\mu}_{sn\_ipw})$	0.6580	0.1462	0.0021	-0.0094	0.1465
impute_linear $(\hat{\mu}_{\hat{f}})$	0.9382	0.0793	0.0011	0.2708	**0.2821**

SeaVan2_MCAR illustrated

- Complete cases in sample size $n = 1000$



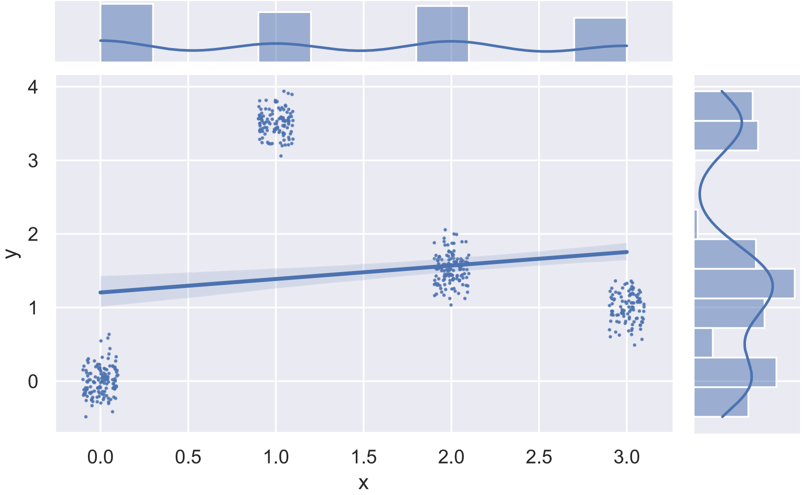
Performance on SeaVan2_MCAR

- Fit $\hat{f}(x) = a + bx$ to the complete cases.
- True mean: 0.667

estimator	mean	SD	SE	bias	RMSE
mean ($\hat{\mu}_{\text{cc}}$)	0.66724	0.05059	0.00226	0.00116	0.05061
ipw_mean ($\hat{\mu}_{\text{ipw}}$)	0.66712	0.05552	0.00248	0.00104	0.05553
sn_ipw_mean ($\hat{\mu}_{\text{sn\_ipw}}$)	0.66724	0.05059	0.00226	0.00116	0.05061
impute_linear ($\hat{\mu}_{\hat{f}}$)	0.66763	0.04953	0.00222	0.00155	**0.04955**

MCAR_normal_nonlinear

Complete cases for $\mathbb{P}(R=1 \mid X) \equiv 0.5$ and $n = 1000$:



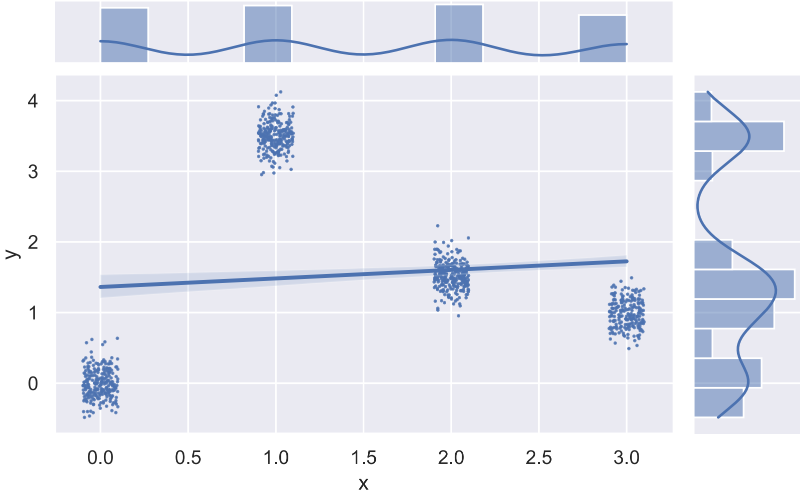
Performance on MCAR_normal_nonlinear

- True mean: 1.50

estimator	mean	SD	SE	bias	RMSE
mean	1.5021	0.0593	0.0019	0.0009	0.0593
ipw_mean	1.5014	0.0759	0.0024	0.0002	0.0759
sn_ipw_mean	1.5021	0.0593	0.0019	0.0009	0.0593
impute_linear	1.5030	0.0592	0.0019	0.0018	**0.0592**

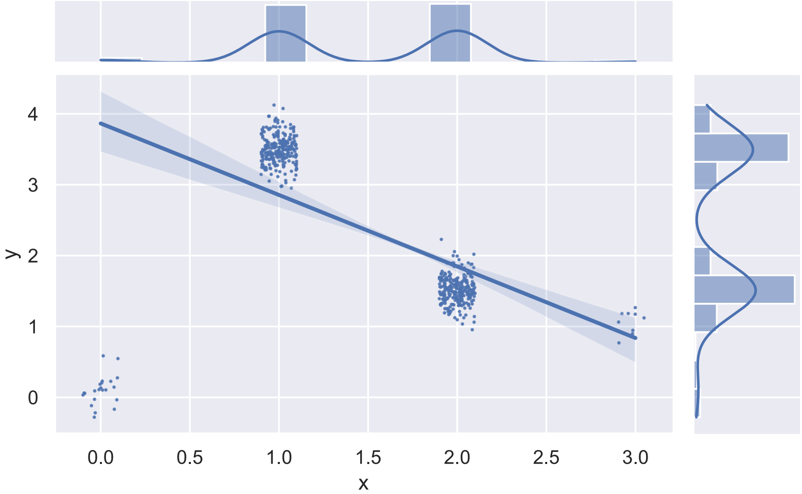
MAR_normal_nonlinear

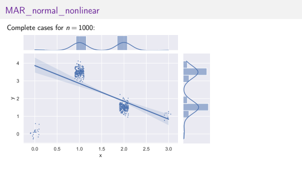
Full data for $n = 1000$:



MAR_normal_nonlinear

Complete cases for $n = 1000$:





Note that the linear fit is completely off from the fit to the full data (preceding slide) because of the sample bias.

Performance on MAR_normal_nonlinear

- True mean: 1.50

estimator	mean	SD	SE	bias	RMSE
mean	2.4075	0.0476	0.0015	0.9063	0.9075
ipw_mean	1.4985	0.0851	0.0027	-0.0027	0.0852
sn_ipw_mean	1.5070	0.1224	0.0039	0.0057	0.1225
impute_linear	2.4060	0.0583	0.0018	0.9048	**0.9066**

What's going on?

- The best linear fit to the complete cases is
    - COMPLETELY DIFFERENT from the best linear fit to full data.
- Essential issue: model is fit to the **complete cases**,
    - but applied on **incomplete cases.**
- Complete cases and incomplete cases have different distributions!

Covariate shift

Covariate shift

- Goal: Find f minimizing risk $R(f) = \mathbb{E}\ell(f(X), Y)$ where

$$(X, Y) \sim p(x, y) = p(x)p(y \mid x).$$

- Standard: $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ is i.i.d. from

$$p(x, y) = p(x)p(y \mid x).$$

- **Covariate shift**: $\mathcal{D}_n = ((X_1, Y_1), \dots, (X_n, Y_n))$ is i.i.d. from

$$q(x, y) = q(x)p(y \mid x).$$

- The covariate distribution has changed, but
  - the conditional distribution $p(y \mid x)$ is the same in both cases.

Covariate shift: the issue

- Under covariate shift,

$$\mathbb{E}_{(X_i,Y_i)\sim q(x,y)}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(f(X_i),Y_i)\right] \neq \mathbb{E}_{(X,Y)\sim p(x,y)}\ell(f(X),Y).$$

- i.e the empirical risk is a **biased** estimator for risk.
- Naive empirical risk minimization is optimizing the wrong thing.
- Can we get an unbiased estimate of risk with $\mathcal{D}_n \sim q(x,y)$?
- **Importance weighting** is one approach to this problem.

Importance weighting for covariate shift

- $\mathcal{D}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ is i.i.d. from

$$q(x, y) = q(x)p(y \mid x).$$

- Then the **importance-weighted empirical risk** is

$$\begin{aligned}\hat{R}_{\text{iw}}(f) &= \frac{1}{n}\sum_{i=1}^{n}\frac{p(x)p(y \mid x)}{q(x)p(y \mid x)}\ell(f(X_i), Y_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{p(x)}{q(x)}\ell(f(X_i), Y_i).\end{aligned}$$

- Note that $\mathbb{E}_{\mathcal{D}_n \sim q(x,y)}\hat{R}_{\text{iw}}(f) = \mathbb{E}_{(X,Y)\sim p(x,y)}\ell(f(X), Y)$.
- So the **importance-weighted empirical risk** is unbiased.

Where are we?

Techniques and applications so far

	Techniques	Applications
So far	Inverse propensity weighting (IPW)	Missing data / response bias
	Self-normalization	
	Regression imputation	
	Importance sampling / weighting	Covariate shift
This week	Control variates	Average treatment effect estimation
	Doubly robust estimators	Conditional ATE estimation
Next few weeks	Policy gradient	Bandit optimization
	Thompson sampling	Offline bandit optimization
	REINFORCE	Reinforcement learning