Regression imputation

David S. Rosenberg

NYU: CDS

February 11, 2021

Contents

Regression imputation

Recap: Missing at random (MAR) setting

- Full data: $(X_1, Y_1), \ldots, (X_n, Y_n)$
- Observed data: $(X_1, R_1, R_1 Y_1), \ldots, (X_n, R_n, R_n Y_n)$
  - where $R_1, \ldots, R_n \in \{0, 1\}$ is the response indicator.
- In the missing at random (MAR) setting, $R_i \perp\!\!\!\perp Y_i \mid X_i$
- Probability of response is given by the **propensity score function**:

$$\pi(x) = \mathbb{P}(R_i = 1 \mid X_i = x) \quad \forall i.$$

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 $(R_i = 1)$ to approximate $\mathbb{E}[Y \mid X = x]$.
- **Regression imputation estimator:** Estimate $\mathbb{E}Y$ with

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

Regression imputation

- Estimating $\mathbb{E}Y$ in the MAR setting:
- The regression imputation estimator for imputation function $\hat{f}(x)$ is

$$\hat{\mu}_{\hat{f}} = \frac{1}{n}\sum_{i=1}^{n}\left[R_i Y_i + (1 - R_i)\,\hat{f}(X_i)\right].$$

- We estimate the imputation function $\hat{f}(x)$ on the complete cases of the same data.
- Exercise: If $f(x) = \mathbb{E}[Y \mid X = x]$, show that $\mathbb{E}\hat{\mu}_f = \mathbb{E}Y$.

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

- In a learning theory context, the analogue of a model is a **hypothesis space** (of [conditional] probability distributions).
- A reasonable analogue of a well-specified model is a hypothesis space with 0 approximation error, though they're not exactly the same. Approximation error is defined in terms of an expected loss, while there is no notion of a "loss" when talking about whether a model is well specified.

Well-specified model imputation for MAR

MAR: SeaVan1 distribution

- X is drawn uniformly from $\{0, 1, 2\}$.
- $Y \mid X = x \sim \mathcal{N}(x, 1)$
- $R \mid X = x \sim \text{Bern}(\text{expit}(4 - 4x))$, where $\text{expit}(x) = 1/(1 + e^{-x})$:

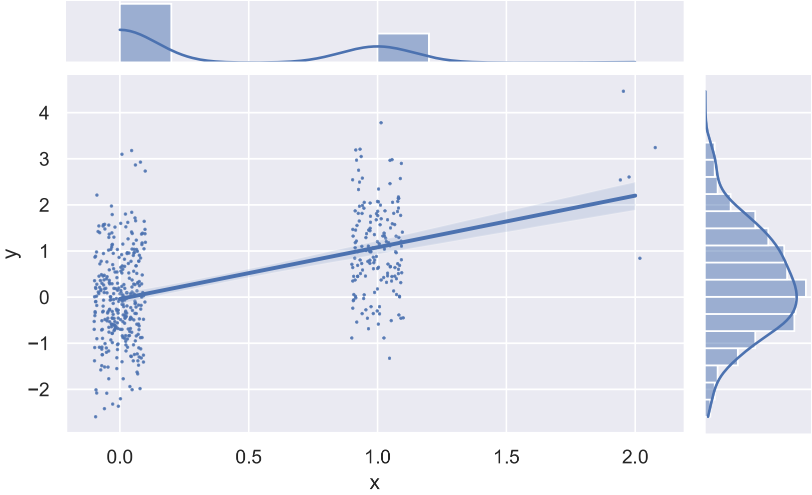
x	$\pi(x) = \mathbb{P}(R = 1 \mid X = x)$
0	.982
1	.500
2	.018

- $(X, R, Y), (X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ are i.i.d. with distribution described above.
- We'll refer to this distribution as "**SeaVan1**", based on the names of the authors who created it

- This distribution corresponds to a massive response bias: an individual with $X = 0$ is 55 times more likely to respond than an individual with $X = 2$.

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

- For impute_missing, we build a linear regression estimator $\hat{f}(x) = a + bx$ for the missing values.
- $\hat{f}(x)$ has only 2 degrees of freedom and roughly $1000/2 = 500$ observations (based on the response probabilities).
- The relatively large number of observations compared to the degrees of freedom will lead to a very low variance for $\hat{f}$. To be clear, here we're talking about the variance that's due to the randomness in the training data. This is **not** the variance that's directly reflected in the SD column of the preceding table, although it may be a significant contributor to that variance.
- The SD column is directly measuring how much $\hat{\mu}_{\hat{f}}$ varies from trial to trial. Some of that variance is coming directly from the variance in the Y_i's as they enter into the mean computation of $\hat{\mu}_{\hat{f}}$. Some of the variance comes from the variance of $\hat{f}$ and the variance of X_i's of the unobserved cases.

Misspecified model imputation for MAR

MAR: SeaVan2 distribution

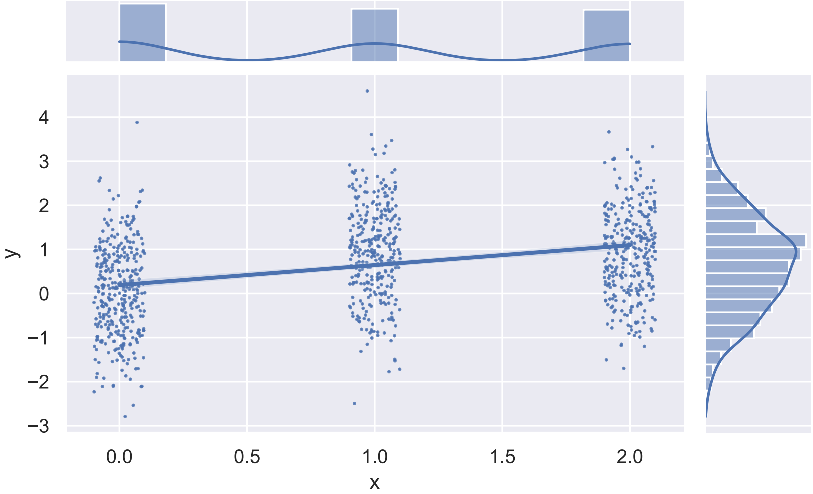
- X is drawn uniformly from $\{0, 1, 2\}$.
- $Y \mid X = x \sim \mathcal{N}(\mathbb{1}[x \geqslant 1], 1) \quad \Longleftarrow$ (THE CHANGE)
- $R \mid X = x \sim \text{Bern}(\text{expit}(4 - 4x))$, where $\text{expit}(x) = 1/(1 + e^{-x})$

x	$\mathbb{P}(R = 1 \mid X = x)$
0	.982
1	.500
2	.018

- $(X, R, Y), (X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ are i.i.d. with distribution described above.

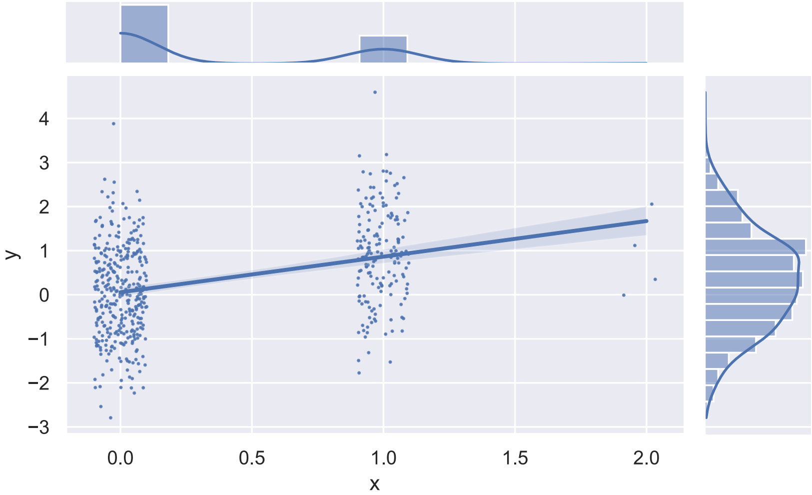
MAR: SeaVan2 distribution illustrated

- Full data for sample of size $n = 1000$



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

- The complete case mean has a large negative bias, since it has relatively more representation from $x = 1$, which has smaller y values.
- On the other hand, the impute_linear has a large positive bias, since the linear model it fits significantly overestimates $\hat{f}(2)$, imputing something close to 2 while $\mathbb{E}[Y \mid X = 2] = 1$.
- As expected, ipw_mean is unbiased and sn_ipw_mean has very small bias. The SDs and RMSEs for these two estimators follow the pattern we've seen before: self-normalized has smaller SD and improved RMSE. (Please note: it's not always true that self-normalized IPW is better than IPW, as we'll see later.)
- Comparing this example and the previous one, we see that linear imputation goes from the best estimator when the imputation model is correct to one of the worst (not much better than the naive complete case mean) when the imputation model is quite poor.
- Before you jump to any conclusions, we'll explore this phenomenon further below and see that it's not **just** that we were using a linear model to fit a nonlinear relationship. It's that misspecification **combined** with the sample bias.

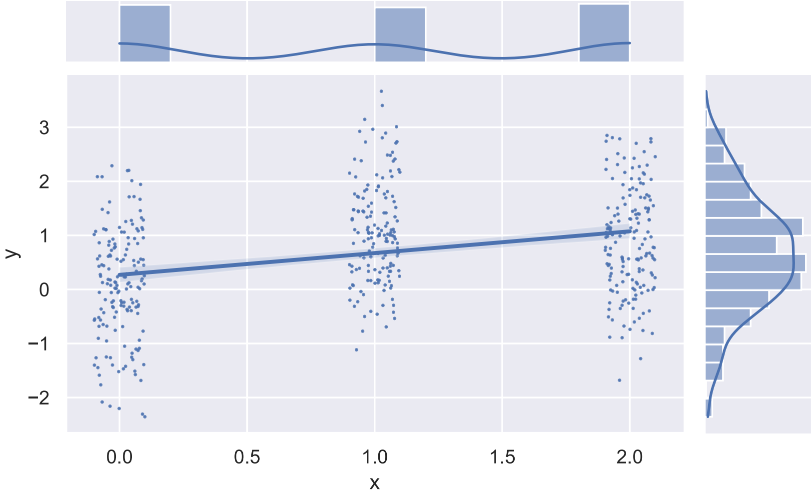
Misspecified model imputation for MCAR

SeaVan2_MCAR distribution

- X is drawn uniformly from $\{0, 1, 2\}$.
- $Y \mid X \sim \mathcal{N}(\mathbb{1}[X \geqslant 1], 1)$
- $\mathbb{P}(R = 1 \mid X = x) \equiv 0.5$.
- $(X, R, Y), (X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ are i.i.d. with distribution described above.
- Expected number of complete cases is the same for SeaVan2_MCAR and SeaVan2.
- But there is no response bias in SeaVan2_MCAR.

SeaVan2_MCAR illustrated

- Complete cases in sample size $n = 1000$



Performance on SeaVan2_MCAR

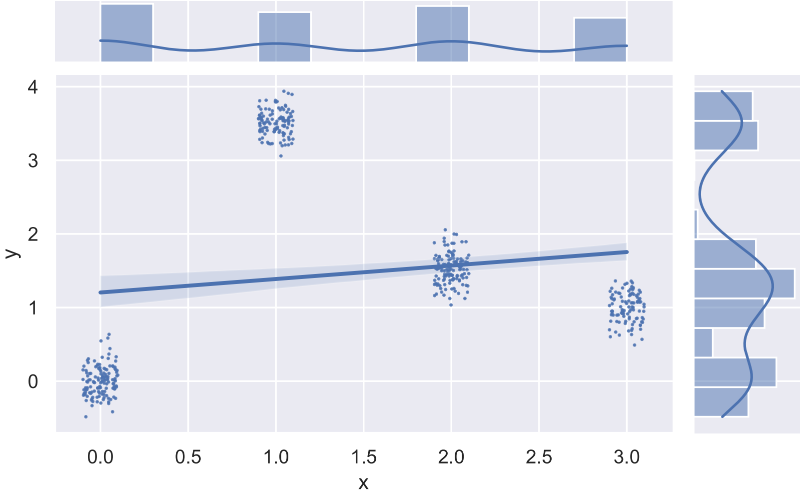
- Fit $\hat{f}(x) = a + bx$ to the complete cases.
- True mean: 0.667

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

- First note that mean and sn_ipw_mean are exactly the same, as expected when the observation probability is fixed for all x.
- Here impute_linear does quite well despite being the “wrong” model for the data. The key difference from the SeaVan2 experiment is that there is no response bias here. (i.e. here we’re in the misspecification + MCAR setting).
- Maybe you’re thinking, well, the linear fit doesn’t look too bad. Maybe that’s why it works still... That’s not quite it, as we’ll see in the next example.

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

- Note that impute_linear is about the same as everything else, despite a very poor fit.
- The key difference from SeaVan2, where impute_linear did very poorly, is that here we do not have any response bias. The model fit seems much worse here, yet still performance is fine.

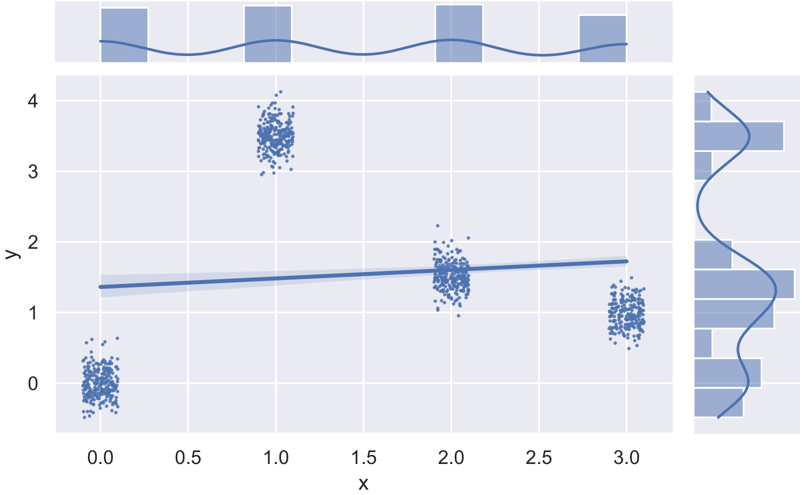
Prelude to covariate shift and importance weighting

MAR_normal_nonlinear distribution

- X is drawn uniformly from $\{0, 1, 2, 3\}$.
- $\mathbb{P}(R = 1 \mid X = x) = \begin{cases} .05 & \text{for } x = 0 \\ 1.0 & \text{for } x = 1 \\ 1.0 & \text{for } x = 2 \\ .05 & \text{for } x = 3 \end{cases}$
- $\mathbb{E}[Y \mid X = x] = \begin{cases} 0 & \text{for } x = 0 \\ 3.5 & \text{for } x = 1 \\ 1.5 & \text{for } x = 2 \\ 1 & \text{for } x = 3 \end{cases}$
- $Y \mid X = x \sim \mathcal{N}(\mathbb{E}[Y \mid X = x], 0.25)$
- $(X, R, Y), (X_1, R_1, Y_1), \ldots, (X_n, R_n, Y_n)$ are i.i.d. with distribution described above.

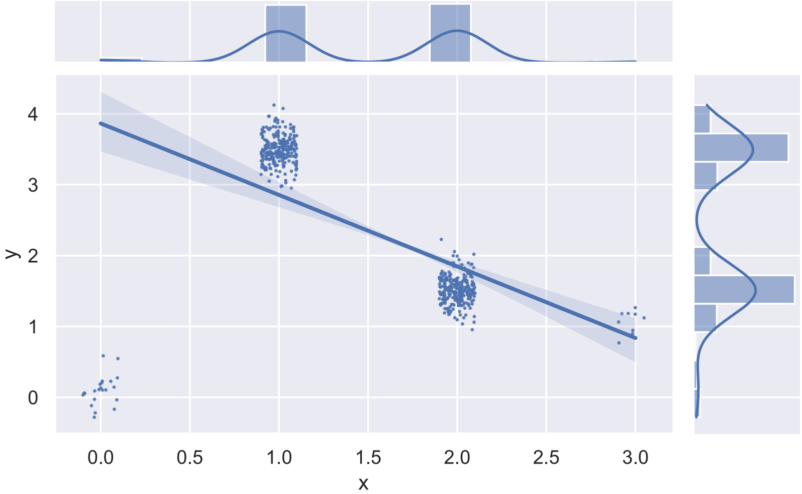
MAR_normal_nonlinear

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

- Note that impute_linear is about 10 times worse than the IPW estimators.
- Almost all the RMSE is caused by bias. This bias is dominated by the fact that the linear imputation gives $\hat{f}(0) \approx 4.5$, while $\mathbb{E}[Y \mid X = 0] = 0$. The estimate $\hat{\mu}_{\hat{f}}$ is the average of $n = 1000$ values. Approximately

$$\mathbb{P}(X = 0)\,\mathbb{P}(R = 0 \mid X = 0) = 0.25 \cdot 0.95 = 23.75\%$$

  of these values are filled in with 4.5 rather than the ideal 0, leading to a very large, positive bias for $\hat{\mu}_{\hat{f}}$.
- As a recap, if it's a poor model fit but MCAR (i.e. no response bias), performance is roughly that of the complete case mean.
- If it's a good model fit, we get good performance even in MAR (response bias) setting.
- If it's a misspecified model and MAR (i.e. response bias), then we can be in real trouble (as in this example).

What's going on?

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

Distributions of complete vs incomplete cases

- The distribution of a **complete cases** is

$$\begin{aligned}&\mathbb{P}(X=x, Y=y \mid R=1)\\ =\ &\pi(x)p(y \mid x)p(x)/\mathbb{P}(R=1)\end{aligned}$$

- The distribution of **incomplete cases** is

$$\mathbb{P}(X=x, Y=y \mid R=0) = (1-\pi(x))p(y \mid x)p(x)/\mathbb{P}(R=0).$$

- The conditional distribution of $Y \mid X$ is the same in both cases.
- The marginal distribution of X changes from $\pi(x)p(x)\frac{1}{\mathbb{P}(R=1)}$ to $(1-\pi(x))p(x)\frac{1}{\mathbb{P}(R=0)}$.
- When just the covariate distribution changes, it's called **covariate shift**.
- Next we'll discuss an approach to covariate shift called **importance weighting**.

References

Resources

- The introductions to the papers [SV18] and [KS07] discuss regression imputation, as well as the IPW and self-normalized IPW estimators (and more).

References I

[KS07] Joseph D. Y. Kang and Joseph L. Schafer, *Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data*, Statistical Science **22** (2007), no. 4, 523–539.

[SV18] Shaun R. Seaman and Stijn Vansteelandt, *Introduction to double robust methods for incomplete data*, Statistical Science **33** (2018), no. 2, 184–197.