Inverse Propensity Weighting

David S. Rosenberg

NYU: CDS

February 3, 2021

Contents

Recap and Warmup

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 missing at random (MAR) setting, $R_i \perp\!\!\!\perp Y_i \mid X_i$
- Probability of response is given by the **propensity score function**:

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

Exercise: $\mathbb{E}[R_i \mid X_i] = \pi(X_i)$

Show $\mathbb{E}[R_i \mid X_i] = \pi(X_i)$

- Let $f(x) = \mathbb{E}[R_i \mid X_i = x]$. (So $\mathbb{E}[R_i \mid X_i] = f(X_i)$.)
- Then

$$\begin{aligned} f(x) &= 1 \cdot \mathbb{P}(R_i = 1 \mid X_i = x) + 0 \cdot \mathbb{P}(R_i = 0 \mid X_i = x) \\ &= \pi(x) \end{aligned}$$

- So $\mathbb{E}[R_i \mid X_i] = \pi(X_i)$.

Inverse probability weighting (IPW): Introduction

Observed responses represent multiple unobserved responses

- Suppose $\mathcal{X} = \{0, 1\}$ corresponding to two types of people.
- e.g. $X_i = \mathbb{1}$ [individual i has previously responded to a survey]
- When $X_i = 1$, response probability $\pi(1) = \mathbb{P}(R_i = 1 \mid X_i = 1) = 0.2$
- When $X_i = 0$, response probability $\pi(0) = \mathbb{P}(R_i = 1 \mid X_i = 0) = 0.1$
- If we observe Y_i for an individual with $X_i = 1$,
  - That individual represents roughly 5 people from the full data.
- If we observe Y_i for an individual with $X_i = 0$,
  - That individual represents roughly 10 people from the full data.

Inverse Probability Weighted (IPW) Mean

- This motivates the **IPW mean estimator** for $\mathbb{E}Y$:

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

- We'll show that this is unbiased and asymptotically does the right thing.
- Though it has some issues, which we'll also explore

IPW: Lemmas

- In the MAR setting, where $Y_i \perp\!\!\!\perp R_i \mid X_i$, we have for a generic term in the mean:

$$
\begin{aligned}
\mathbb{E}\left[\frac{RY}{\pi(X)}\right] &= \mathbb{E}\left[\mathbb{E}\left[\frac{RY}{\pi(X)} \mid X\right]\right] \quad \text{Adam's Law} \\
&= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}\left[RY \mid X\right]\right] \quad \text{Taking out what is known} \\
&= \mathbb{E}\left[\frac{1}{\pi(X)}\mathbb{E}\left[R \mid X\right]\mathbb{E}\left[Y \mid X\right]\right] \quad \text{By MAR assumption} \\
&= \mathbb{E}\left[\mathbb{E}\left[Y \mid X\right]\right] \quad \text{since } \mathbb{E}\left[R \mid X\right] = \mathbb{P}\left(R = 1 \mid X\right) = \pi(X) \\
&= \mathbb{E}Y \quad \text{Adam's Law}
\end{aligned}
$$

IPW: Unbiased and Consistent

- The estimator $\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(X_i)}$ is **unbiased** for $\mathbb{E}Y$, since

$$\mathbb{E}\hat{\mu}_{\text{ipw}} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{R_iY_i}{\pi(X_i)}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\frac{R_iY_i}{\pi(X_i)}\right] = \mathbb{E}Y.$$

- The estimator $\hat{\mu}_{\text{ipw}}$ is "**consistent**" That is,

$$\hat{\mu}_{\text{ipw}} \xrightarrow{P} \mathbb{E}Y \text{ as } n \to \infty$$

  by the Law of Large Numbers, since $\left(\frac{R_iY_i}{p(R_i|X_i)}\right)_{i=1}^{n}$ are i.i.d. with expectation μ.
- Unbiased and consistent are nice properties to have,
  - but does NOT mean it's a great estimator....

IPW on MCAR: example and simulation

Recap: MAR, MCAR, and IPW estimators

- We started by talking about the MCAR setting
  - i.e. Response indicator R_i independent of response Y_i
- We discussed the complete case estimator for $\mathbb{E}Y$:

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

- In the MAR setting, R_i and Y_i can be dependent, so $\hat{\mu}_{\text{cc}}$ may be quite biased.
- We introduced $\hat{\mu}_{\text{ipw}}$ and showed it's **unbiased** in the MAR setting.
- But how good is $\hat{\mu}_{\text{ipw}}$?
- Let's see how it compares to $\hat{\mu}_{\text{cc}}$ in the simple MCAR setting.

IPW Mean: MCAR example

- Let's consider how
$$\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i)}$$
works in a trivial MCAR example ($R_i \perp\!\!\!\perp Y_i$).
- For $i = 1, \ldots, n$
    - Let R_i be independent of (X_i, Y_i).
    - Let $Y_i \in \{0, 1\}$ with $\mathbb{P}(Y_i = 1) = 0.75$ and
    - $\pi(x) = \mathbb{P}(R_i = 1 \mid X_i = x) \equiv 0.1$.

IPW Mean: MCAR example

- Let's consider how
$$\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i)}$$
works in a trivial MCAR example ($R_i \perp\!\!\!\perp Y_i$).
- For $i = 1, \ldots, n$
  - Let R_i be independent of (X_i, Y_i).
  - Let $Y_i \in \{0, 1\}$ with $\mathbb{P}(Y_i = 1) = 0.75$ and
  - $\pi(x) = \mathbb{P}(R_i = 1 \mid X_i = x) \equiv 0.1$.

- In MCAR, we **could** have R_i depend on X_i. If that were the case, then X_i and Y_i would need to be independent, since for MCAR we need $R_i \perp\!\!\!\perp Y_i$.

IPW for MCAR example: $n = 1$ case

- Suppose $n = 1$:
- If $Y_1 = 0$ or $R_1 = 0$, then $\hat{\mu}_{\text{ipw}} = \frac{R_1 Y_1}{0.1} = 0$.
- If $Y_1 = 1$ and $R_1 = 1$, then $\hat{\mu}_{\text{ipw}} = \frac{R_1 Y_1}{0.1} = \frac{1}{0.1} = 10$.
- The good: $\mathbb{E}\hat{\mu}_{\text{ipw}} = 10 \cdot \mathbb{P}(Y_i = 1)\,\mathbb{P}(R_i = 1) = 10 \cdot \frac{3}{4} \cdot \frac{1}{10} = 0.75$, so unbiased.
- The bad: We can get $\hat{\mu}_{\text{ipw}} = 10$ even though $Y_i \in \{0, 1\}$.

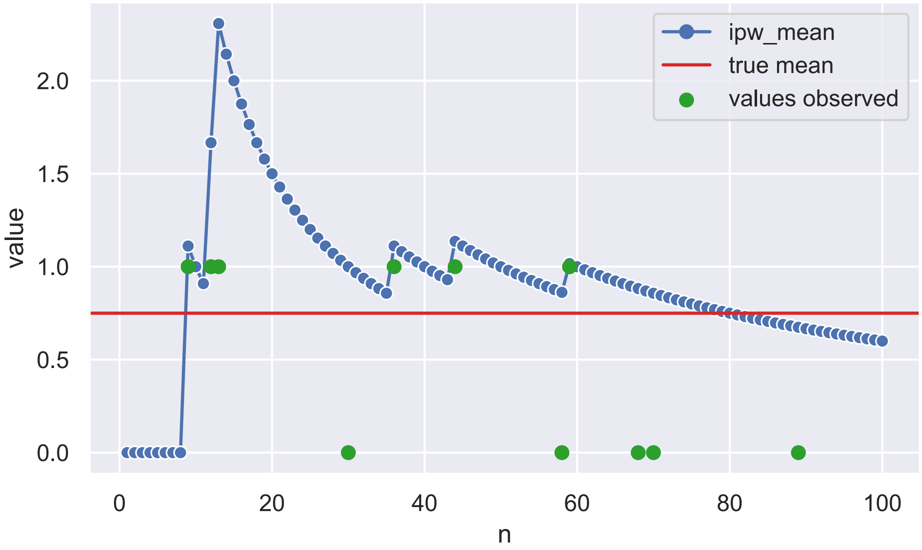
IPW for MCAR example, continued

- Now let n grow.
- Suppose $R_1 = Y_1 = 1$, but $R_2 = \cdots = R_n = 0$.
- Then

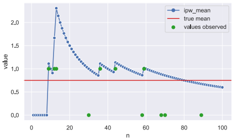
$$\hat{\mu}_{\text{ipw}} = \frac{1}{n}\sum_{i=1}^{n}\frac{R_i Y_i}{\pi(X_i)} = \frac{1}{n}\left(\frac{1\cdot 1}{0.1} + \underbrace{0 + \cdots + 0}_{(n-1)}\right) = \frac{10}{n}$$

- So the IPW estimate starts at 10, and then decreases like $O(1/n)$ towards 0.
- Our estimate keeps changing even though we don't have new observations of Y.

IPW for MCAR example visualized

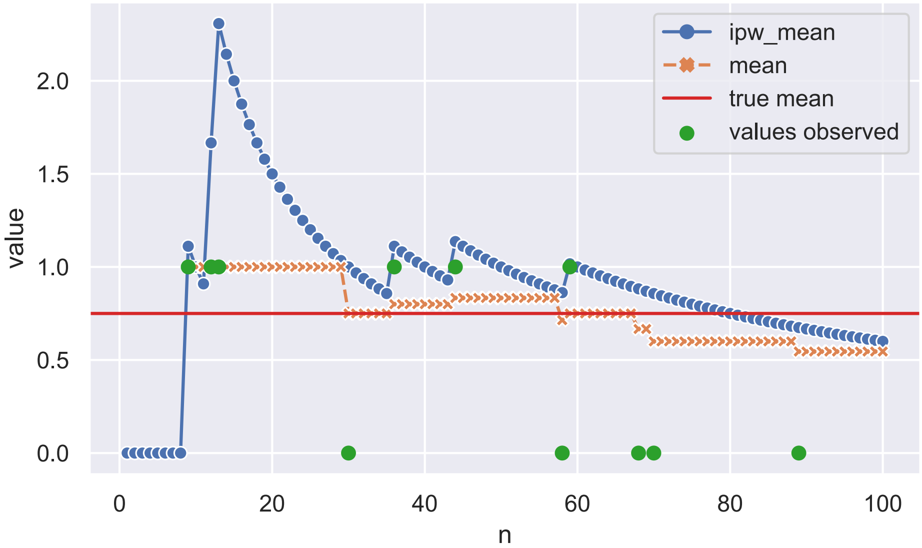


IPW for MCAR example visualized



- The green dots represent observed values of Y_i.
- We can see that we had no observations of Y until $Y_9 = 1$.
- The horizontal red line shows the true mean of the Y_i's.
- The blue dots show $\hat{\mu}_{\text{ipw}}$ as n increases.
- Note the large jumps in $\hat{\mu}_{\text{ipw}}$ whenever we get an observation with $Y_i = 1$. This is because each observed Y_i is scaled up by an inverse propensity weight of 10.
- Also note that between observations with $Y_i = 1$, ipw_mean decays like $1/n$ towards 0.

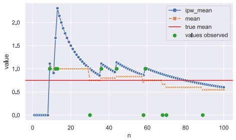
IPW vs complete case mean



2021-02-01

DS-GA 3001: Tools and Techniques for ML
└IPW on MCAR: example and simulation
└IPW vs complete case mean

IPW vs complete case mean



- We've added in the orange line here, which represent the mean of the complete cases.
- Note that the orange line is constant between observations, and jumps whenever we get a new observation.

IPW vs complete case mean, 5000x

- We repeat the experiment[1] above 5000 times ($n = 100$ samples in each) and get the following.
- Recall that the true mean is $\mu = 0.75$.

estimator	mean	SD	SE	bias	RMSE
mean	0.751654	0.143943	0.002036	0.001654	0.143953
ipw_mean	0.752540	0.262224	0.003708	0.002540	0.262237

[1] In the very rare event that the sample in a particular repetition has 0 complete cases, we take $\hat{\mu}_{cc} = 0$.

└─IPW vs complete case mean, 5000x

IPW vs complete case mean, 5000x

- We repeat the experiment[1] above 5000 times ($n = 100$ samples in each) and get the following.
- Recall that the true mean is $\mu = 0.75$.

estimator	mean	SD	SE	bias	RMSE
mean	0.751654	0.143943	0.002036	0.001654	0.143953
ipw_mean	0.752540	0.262224	0.003708	0.002540	0.262237

[1] In the very rare event that the sample in a particular repetition has 0 complete cases, we take $\hat{\mu}_{cc} = 0$.

To generate this table, we repeated the following 5000 times:

1. Get a sample of size $n = 100$ from the distribution described previously.
2. Evaluate the estimators on this sample.

- The table then displays various summary statistics of the performance of these estimators across the 5000 repetitions.
- Mean and SD are just the mean and SD of the estimate across the repeats.
- The SE is the uncertainty of the mean as an estimator $\mathbb{E}\hat{\mu}_{ipw}$ and $\mathbb{E}\hat{\mu}_{cc}$ (depending on the row) – so SE = SD / sqrt(num_repeats=5000).
- "Bias" is computed as the difference between the mean of the estimators and the true mean μ .
- The SE is also the uncertainty of this bias.
- For each repeat we compute the RMS difference between the estimate and μ.

IPW vs Complete case mean, 5000x (continued)

estimator	mean	SD	SE	bias	RMSE
mean	0.751654	0.143943	0.002036	0.001654	0.143953
ipw_mean	0.752540	0.262224	0.003708	0.002540	0.262237

- Both estimators have
    - bias within 1 SE of 0 (i.e. indistinguishable from 0).
    - SD is orders of magnitude larger than bias (and essentially equals the RMS error).
- **Complete case mean is better than ipw_mean in RMSE for this MCAR distribution.**

IPW vs Complete case mean, 5000x (continued)

estimator	mean	SD	SE	bias	RMSE
mean	0.751654	0.143943	0.002036	0.001654	0.143953
ipw_mean	0.752540	0.262224	0.003708	0.002540	0.262237

- Both estimators have
  - bias within 1 SE of 0 (i.e. indistinguishable from 0).
  - SD is orders of magnitude larger than bias (and essentially equals the RMS error).
- **Complete case mean is better than ipw_mean in RMSE for this MCAR distribution.**

- The bias of mean (the mean of the observations) and ipw_mean are both within 1 SE of 0, so we can't conclude from this data that there's bias.
- Indeed, we know theoretically that the IPW estimator is unbiased.
- In both cases, the bias is orders of magnitude smaller than the SD, which is what drives the RMSE.
- For this dataset, it seems like the complete case mean is the clear winner, between the two.
- Recall that $\text{RMSE}^2 = \text{SD}^2 + \text{bias}^2$.
- This explains why the contribution of the bias to the RMSE is even less than one might expect from looking at the relative magnitudes of SD and bias.