Importance-weighted regression imputation

David S. Rosenberg

NYU: CDS

February 17, 2021

Contents

Importance-weighted regression imputation

Covariate shift and regression imputation

- Regression imputation had performance issues when there was
  - a misspecified model AND
  - response bias (i.e. MAR setting)
- Hypothesis: This is due to mismatch between train & target distributions.
- If we know these distributions, we can fit our imputer with importance-weighted ERM.

- Regression imputation had performance issues when there was
  - a misspecified model AND
  - response bias (i.e. MAR setting)
- Hypothesis: This is due to mismatch between train & target distributions.
- If we know these distributions, we can fit our imputer with importance-weighted ERM.

- The target distribution, sometimes just called the test distribution, is the distribution on which we'll apply our imputation function.
- In a covariate shift situation, it might be the distribution in which we deploy our prediction function.

Training distribution

- Training distribution = complete case distribution
- Complete case distribution:

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

Target distribution 1: incomplete case distribution

- Incomplete case distribution:

$$
\begin{aligned}
p(x, y \mid R=0) &= p(x, y, R=0)/\mathbb{P}(R=0) \\
&= p(y, R=0 \mid x)\, p(x)/\mathbb{P}(R=0) \\
&= p(y \mid x)\,(1-\pi(x))\, p(x)/\mathbb{P}(R=0)
\end{aligned}
$$

Importance weight 1

- **Importance weight** *(ratio of target density to training density)*:

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

- **Importance-weighted empirical risk** is

$$
\begin{aligned}
\hat{R}_{\text{iw}}(f) &= \frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i, Y_i \mid R_i = 0)}{p(X_i, Y_i \mid R_i = 1)}\ell(f(X_i), Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi(X_i))}{\pi(X_i)}\frac{\mathbb{P}(R_i = 1)}{\mathbb{P}(R_i = 0)}\ell(f(X_i), Y_i) \\
&= \frac{\mathbb{P}(R = 1)}{\mathbb{P}(R = 0)} \times \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi(X_i))}{\pi(X_i)}\ell(f(X_i), Y_i) \\
&\propto \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi(X_i))}{\pi(X_i)}\ell(f(X_i), Y_i)
\end{aligned}
$$

Importance-weighted empirical risk 1

- Importance-weighted empirical risk is

$$\begin{aligned}\hat{R}_{\text{iw}}(f) &= \frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i,Y_i \mid R_i=0)}{p(X_i,Y_i \mid R_i=1)}\ell(f(X_i),Y_i)\\ &= \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi(X_i))}{\pi(X_i)}\frac{\mathbb{P}(R_i=1)}{\mathbb{P}(R_i=0)}\ell(f(X_i),Y_i)\\ &= \frac{\mathbb{P}(R=1)}{\mathbb{P}(R=0)}\times\frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi(X_i))}{\pi(X_i)}\ell(f(X_i),Y_i)\\ &\propto \frac{1}{n}\sum_{i=1}^{n}\frac{(1-\pi(X_i))}{\pi(X_i)}\ell(f(X_i),Y_i)\end{aligned}$$

- Note that $\mathbb{P}(R_i = a)$ is just a number, the same for all i. So we just write $\mathbb{P}(R = a)$.
- This allows us to pull the ratio $\frac{\mathbb{P}(R=1)}{\mathbb{P}(R=0)}$ out of the sum.
- Note that $\frac{\mathbb{P}(R=1)}{\mathbb{P}(R=0)}$ is just a scale factor on the value of $\hat{R}_{\text{iw}}(f)$, and thus removing it has no effect on $\arg\min_f \hat{R}_{\text{iw}}(f)$.

Importance-weighted linear regression 1

- **Importance-weighted linear regression** on the **complete cases**:

$$\hat{f}_{\text{iw}-\text{linear}} = \operatorname*{arg\,min}_{\{f:f(x)=a+w^T x\}} \sum_{i=1}^{n} R_i \frac{(1-\pi(X_i))}{\pi(X_i)} (f(X_i)-Y_i)^2$$

- We'll write **impute_iw_linear** for the regression imputation estimator that uses $\hat{f}_{\text{iw}-\text{linear}}$ for imputing.

Target distribution 2: full data

- To arrive at another common imputation function,
  - we use the full data distribution as the target distribution.
- Full data distribution:

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

- The corresponding importance weight is

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

Importance-weighted ERM 2

- The IW empirical risk with full data distribution as target is

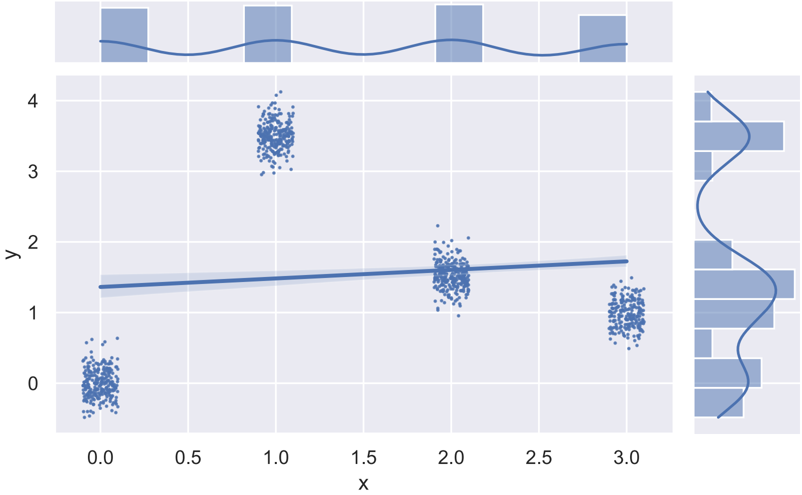
$$\begin{aligned}
\hat{R}_{\text{ipw}}(f) &= \frac{1}{n}\sum_{i=1}^{n} \frac{p(X_i, Y_i)}{p(X_i, Y_i \mid R_i = 1)} \ell(f(X_i), Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbb{P}(R_i = 1)}{\pi(X_i)} \ell(f(X_i), Y_i) \\
&= \mathbb{P}(R = 1) \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\pi(X_i)} \ell(f(X_i), Y_i) \\
&\propto \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\pi(X_i)} \ell(f(X_i), Y_i)
\end{aligned}$$

- We end up weighting by the **inverse propensity weight**.
  - We'll call this IPW-weighted linear regression.
  - We'll write **impute_ipw_linear** for the corresponding linear imputation estimator below.

Experimental results

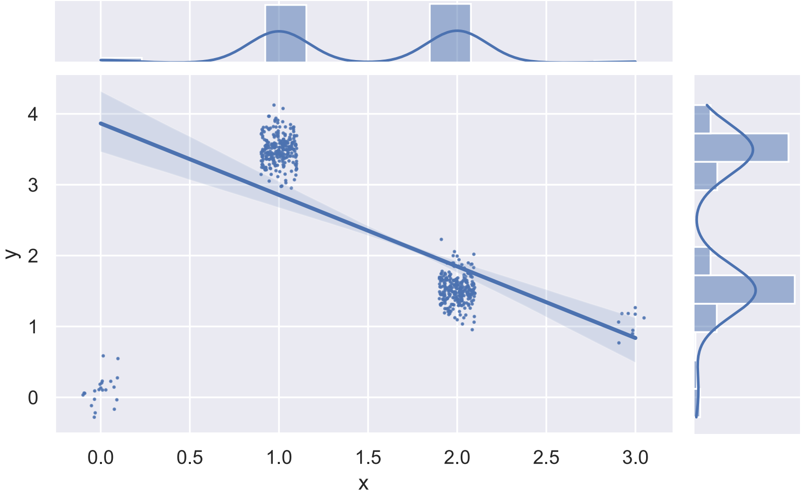
Recap: MAR_normal_nonlinear

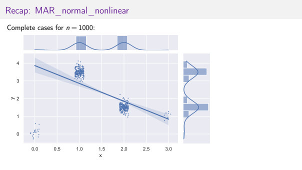
Full data for $n = 1000$:



Recap: 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.

Recap: Performance on MAR_normal_nonlinear

- True mean: 1.50

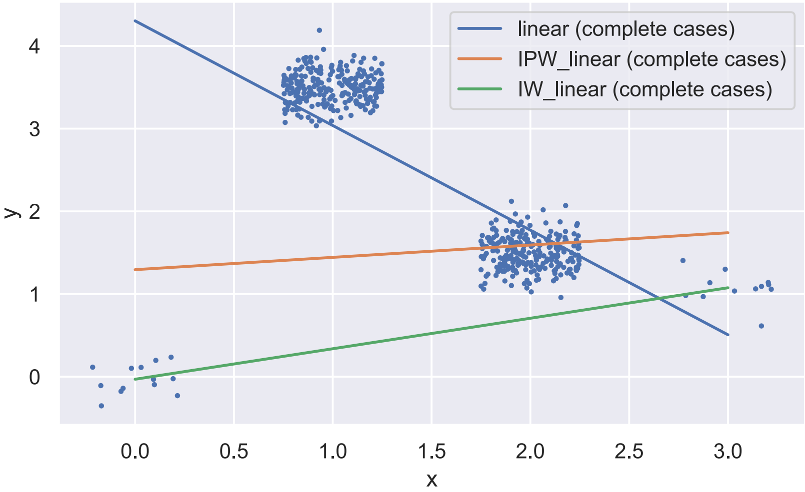
estimator	mean	SD	SE	bias	RMSE
mean	2.4075	0.0476	0.0015	0.9063	0.9075
ipw_mean	1.4985	0.0851	0.0027	-0.0027	**0.0852**
sn_ipw_mean	1.5070	0.1224	0.0039	0.0057	0.1225
impute_linear	2.4060	0.0583	0.0018	0.9048	**0.9066**

Importance-sampling imputation estimators

- Our linear model is fit to data from the complete case distribution
  - we need it to fit well on the incomplete case distribution
  - or to the full data distribution (also common)
- Two new estimators:
  - **impute_ipw_linear:** examples weighted by $\frac{1}{\pi(X_i)}$ so unbiased for full data
  - **impute_iw_linear:** examples weighted by $\frac{1-\pi(X_i)}{\pi(X_i)}$ so unbiased for incomplete data

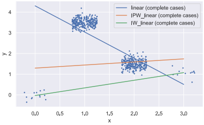
Various fits to complete cases

- Fits to the complete cases (i.e. the data we observe)



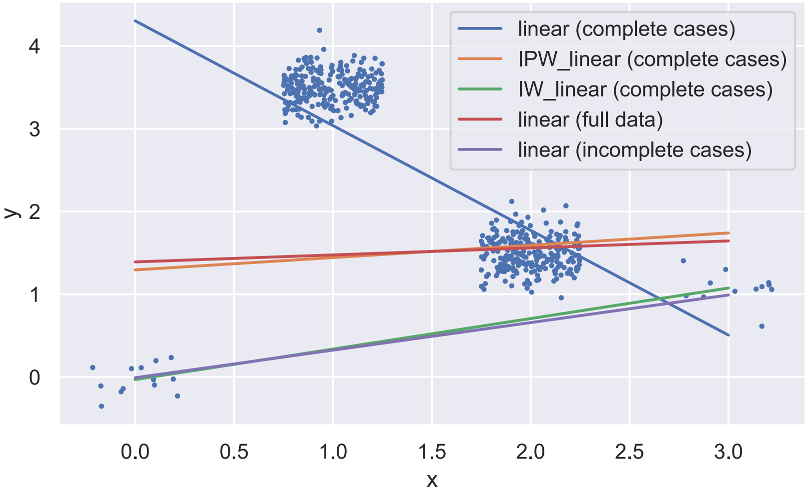
Various fits to complete cases

- Fits to the complete cases (i.e. the data we observe)

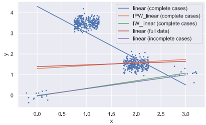


- Here we see the result of linear regression fits to the complete cases with 3 different importance weighting approaches.
- The plain "linear" fit has no importance weighting (equivalently, uniform importance weighting).
- The IPW_linear is fit with importance weights $\frac{1}{\pi(X_i)}$, and IW_linear is fit with importance weights $\frac{1-\pi(X_i)}{\pi(X_i)}$.

Fits on full data and incomplete cases



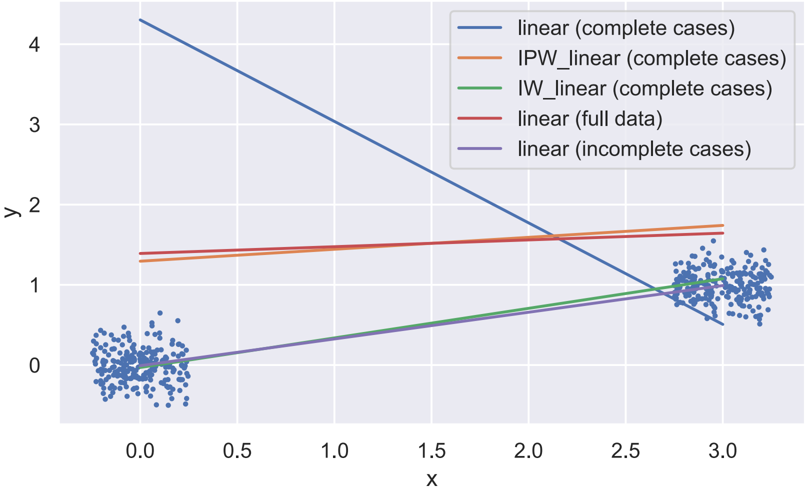
Fits on full data and incomplete cases



- Here we've added two additional fits that we would not be able to do in practice: fitting to the full data, and fitting to the incomplete cases.
- We see that the IPW_linear fit is quite close to the linear fit to the full data. This makes sense, since the importance weighting was chosen so that the objective function is an unbiased estimate for the risk with respect to the full data distribution.
- We see that the IW_linear fit is quite close to the linear fit to the incomplete cases. This makes sense, since the importance weigting was chosen so that the objective function is an unbiased estimate for the risk with respect to the incomplete case distribution.

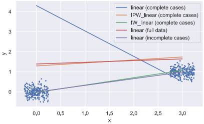
Fits overlaid on incomplete cases

- For regression imputation, we only predict on incomplete cases.



Fits overlaid on incomplete cases

- For regression imputation, we only predict on incomplete cases.



- Here the scatter plot only includes the incomplete cases (which we would not have access to in practice).
- This really highlights the effect of importance weighting for the regression imputation task.

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**
impute_ipw_linear	1.9895	0.0777	0.0025	0.4883	**0.4944**
impute_iw_linear	1.5005	0.0466	0.0015	-0.0007	**0.0466**

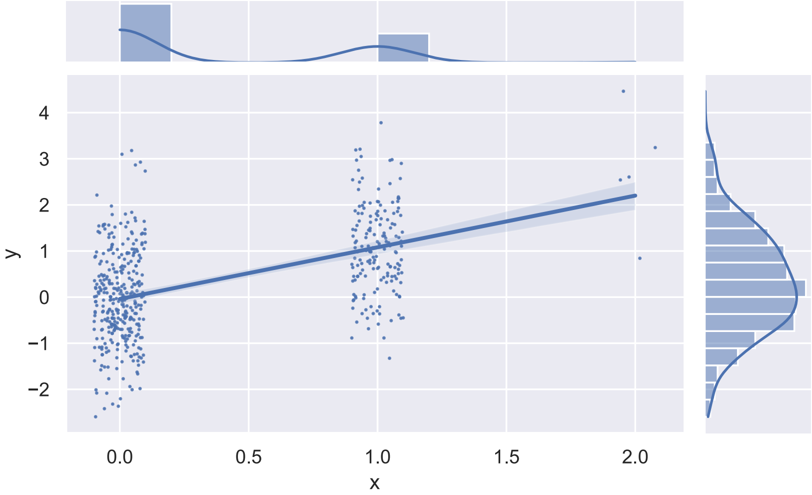
- True mean: 1.50

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**
impute_ipw_linear	1.9895	0.0777	0.0025	0.4883	**0.4944**
impute_iw_linear	1.5005	0.0466	0.0015	-0.0007	**0.0466**

- For this distribution, it's clear that importance weighting towards the incomplete case distribution is a huge win.
- Not our main point, but note also that ipw_mean performs better than sn_ipw_mean in this case. While self-normalization seems to help for most of our experiments in the MAR setting, this is not a general truth, and it's good to be reminded of that.

Recap: 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.

estimator	mean	SD	SE	bias	RMSE
mean	0.3564	0.0515	0.0016	-0.6431	0.6452
ipw_mean	1.0127	0.2968	0.0094	0.0132	0.2971
sn_ipw_mean	0.9906	0.1890	0.0060	-0.0089	0.1892
impute_linear	1.0022	0.0781	0.0025	0.0027	**0.0782**
impute_ipw_linear	1.0039	0.1439	0.0046	0.0044	**0.1440**
impute_iw_linear	1.0047	0.1529	0.0048	0.0052	**0.1530**

Performance on SeaVan1

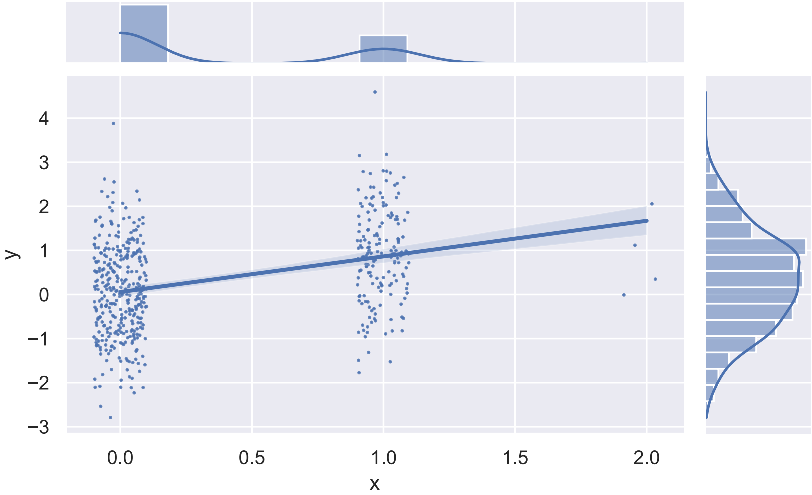
- Fit $\hat{f}(x) = a + bx$ to the complete cases.

estimator	mean	SD	SE	bias	RMSE
mean	0.3564	0.0515	0.0016	-0.6431	0.6452
ipw_mean	1.0127	0.2968	0.0094	0.0132	0.2971
sn_ipw_mean	0.9906	0.1890	0.0060	-0.0089	0.1892
impute_linear	1.0022	0.0781	0.0025	0.0027	**0.0782**
impute_ipw_linear	1.0039	0.1439	0.0046	0.0044	**0.1440**
impute_iw_linear	1.0047	0.1529	0.0048	0.0052	**0.1530**

- Here the model is well-specified, so the linear fit converges to the right thing, even with response bias (i.e. even when our covariate distribution changes between complete and incomplete cases).
- The additional variance caused by importance weighting ends up hurting our performance, compared to the unweighted regression imputation.

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	0.3425	0.0493	0.0007	-0.3244	0.3282
ipw_mean	0.6655	0.1939	0.0027	-0.0014	0.1939
sn_ipw_mean	0.6594	0.1446	0.0020	-0.0075	0.1448
impute_linear	0.9364	0.0792	0.0011	0.2695	**0.2809**
impute_ipw_linear	0.6750	0.1503	0.0021	0.0081	**0.1505**
impute_iw_linear	0.6677	0.1561	0.0022	0.0008	**0.1561**

Performance on SeaVan2

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

estimator	mean	SD	SE	bias	RMSE
mean	0.3425	0.0493	0.0007	-0.3244	0.3282
ipw_mean	0.6655	0.1939	0.0027	-0.0014	0.1939
sn_ipw_mean	0.6594	0.1446	0.0020	-0.0075	0.1448
impute_linear	0.9364	0.0792	0.0011	0.2695	**0.2809**
impute_ipw_linear	0.6750	0.1503	0.0021	0.0081	**0.1505**
impute_iw_linear	0.6677	0.1561	0.0022	0.0008	**0.1561**

- Here we have model misspecification and response bias.
- In each trial, we have very few complete cases with $x = 2.0$, yet when we do observe them they're upweighted very highly (20x for the impute_ipw_linear imputer and .95/.05=19x for the impute_iw_linear). This manages to improve the fit substantially over the unweighted regression imputation (impute_linear).

Caveat on results

- The importance-sampled regression imputation estimators seem promising.
- The estimators rely on knowing the importance weights $p(x)/q(x)$.
- Performance may be significantly worse when we use estimates $\hat{p}(x)/\hat{q}(x)$.
- This is something we will explore further in homework and potentially in your projects.