Importance-weighted regression imputation

David S. Rosenberg

NYU: CDS

February 11, 2021

Contents

Importance-weighted regression imputation

Covariate shift and regression imputation

- Regression imputation had performance issues when there was
  - a misspecifed 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.

Training distribution

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

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

Target distribution 1: incomplete case distribution

- Incomplete case distribution:

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

Importance weight 1

- Importance weight:

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

Importance-weighted empirical risk 1

- So importance-weighted empirical risk is

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

Importance-weighted empirical risk 1

- So importance-weighted empirical risk is

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

- 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 $\hat{R}_{\mathrm{IW}}(f)$, and thus removing it has no effect on $\arg\min_f \hat{R}_{\mathrm{IW}}(f)$.

Importance-weighted linear regression 1

- For importance-weighted linear regression, we have

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

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

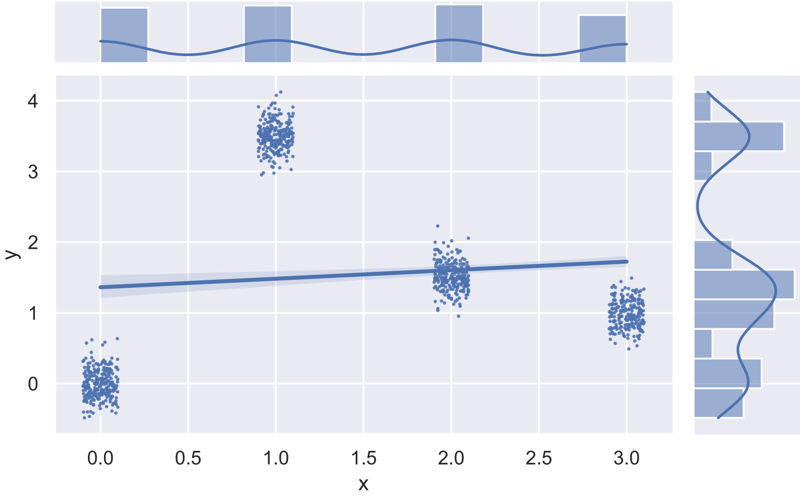
$$
\begin{aligned}
\hat{R}_{\text{IW}}(f) &= \frac{1}{n}\sum_{i=1}^{n}\frac{p(X_i, Y_i)}{p(X_i, Y_i \mid R_i = 1)}\ell(f(X_i), Y_i) \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{\mathbb{P}(R_i = 1)}{\pi(X_i)}\ell(f(X_i), Y_i) \\
&= \mathbb{P}(R = 1)\,\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\pi(X_i)}\ell(f(X_i), Y_i)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

- 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 imputation estimator below.

Experimental results

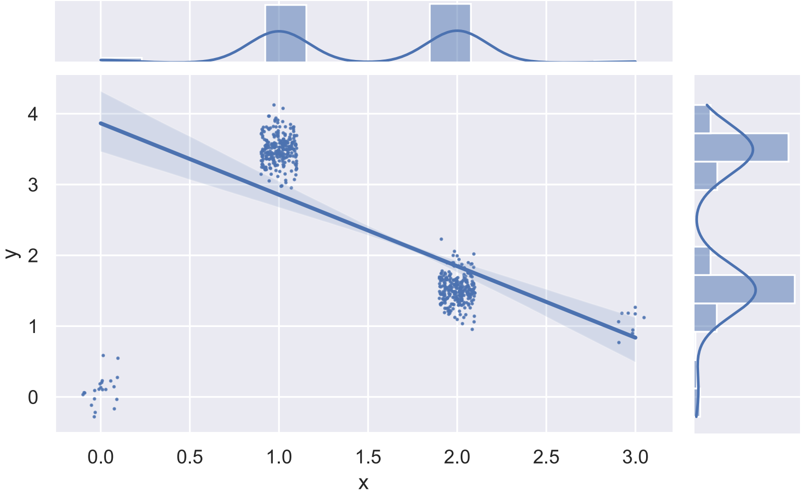
Recap: MAR_normal_nonlinear

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.

- 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 be fit to the incomplete case distribution
  - or 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_IS_linear:** examples weighted by $\frac{1-\pi(X_i)}{\pi(X_i)}$ so unbiased for incomplete data

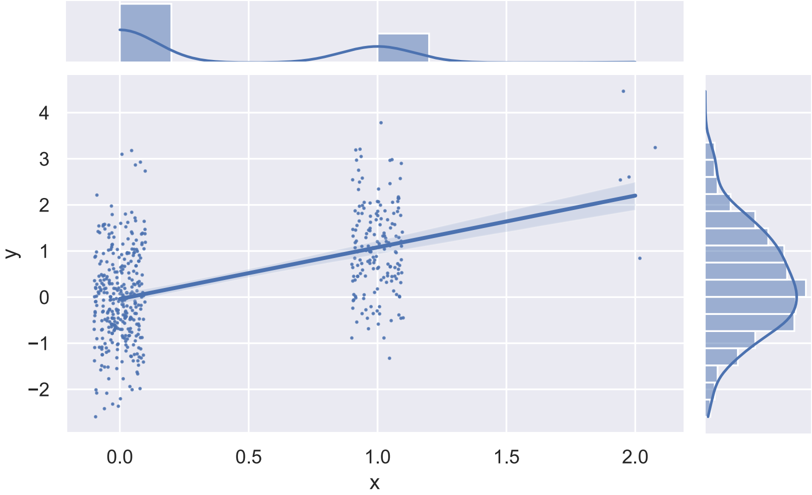
Performance on MAR_normal_nonlinear

- 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**

Recap: SeaVan1 distribution illustrated

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



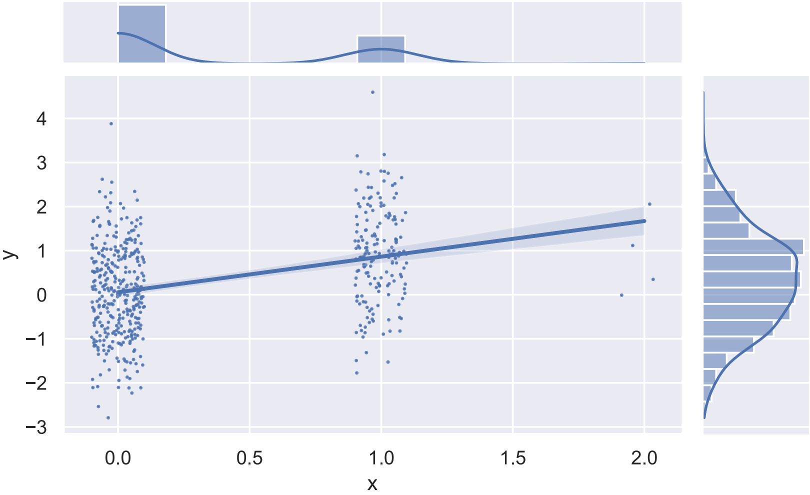
Performance on SeaVan1

- 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**

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**

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 can explore in homeworks and projects.

References

Resources

- Terminology was based on [CFV17].

References I

[CFV17] Victor Chernozhukov and Iván Fernández-Val, *Treatment effects*, Econometrics—MIT Course 14.382, Cambridge MA, 2017, MIT OpenCourseWare.