Inverse Propensity Weighting for evaluating employee assessments

Tania Bakhos, Árdís Elíasdóttir, Josiah Narog, John K Pate, & Rob Stewart

Global Hiring Science, Amazon

Introduction

Correlating assessment scores with performance in role (PIR) metrics provides a powerful form of validation evidence, but is complicated by the absence of PIR metrics for applicants who were not hired. Traditional range restriction perspectives state that the problem is a lack of PIR metrics for low assessment scores, and typical corrections make strong assumptions about how the relationship among incumbents extrapolates to applicants who were not hired (Bryant & Gokhale, 1972; Thorndike, 1947, 1949). If the extrapolation assumptions are strongly violated, however, traditional corrections can over- or under-estimate. This problem is particularly acute when training machine learning models to predict PIR metrics, where overfitting to observed data (incumbents with PIR measures) in a way that does not generalize to unobserved data (candidates without PIR measures) is a fundamental problem (Li et al., 2011; Strehl et al., 2010).

We propose using Inverse Propensity Weighting (IPW) as a simple and accurate method for obtaining correlation estimates that generalize to the candidate population (Lanza et al., 2013; Little, 1986; Rosenbaum, 1988; Rosenbaum & Rubin, 1983, 1984; Seaman & White, 2011; Thoemmes & Ong, 2016). A simulation study confirms that, when case-specific assumptions are violated, traditional corrections are biased and systematically over- or under-estimate the true relationship, and additional data doesn't help. IPW-based methods, however, make weaker assumptions and exhibit low bias that reduces with data volume on the same simulated data.

Method

Our simulations generate datasets with three different correlation structures, called *generative stories*, that exhibit gradations of overfitting, with examples presented in Figure 1 and details in Appendix A. The Linear Missing-At-Random (MAR) story respects the traditional linear and homoscedastic extrapolation assumptions (Pearson $R = 0.605$, in the sampled dataset of Figure 1, among hires, and 0.622 among non-hires). In the Overfit story, the correlation among incumbents (0.780) exceeds the correlation among all candidates (0.491) due to a zero correlation among a cluster of poorly-qualified candidates, a minority of incumbents but a majority of all candidates. In the Well-fit story, the correlation among incumbents is weaker (0.643) than the correlation among all candidates (0.709) due to range restriction (heteroscedastic due to the existence of two clusters). Optimization procedures prefer the Overfit story without correction.

Statistical Corrections

We consider traditional corrections designed for specific cases, all assuming linearity and homoscedasticity. Cases I and II assume direct restriction on r_i and a_i, respectively, and require

the restricted and unrestricted variance of a_i. Case V (Bryant & Gokhale, 1972) assumes indirect restriction on an unobserved variable, and requires the restricted and unrestricted variance of both a_i and r_i.

Inverse Propensity Weighting is a method for weighting samples from one probability distribution so that the weighted expected value of a function is equal to the (unweighted) expected value of that function under a different probability distribution. Pearson R is defined in terms of expected values, and we propose correcting those expectations by weighting the data we have, drawn from the distribution over incumbents $P_{\mathcal{I}}$, to compute expected values under the distribution over candidates $P_{\mathcal{C}}$ we want. See Appendix B for details.

$$\mathbb{E}_{P_{\mathcal{C}}}[f(c_i)] \quad = \mathbb{E}_{P_{\mathcal{I}}}\left[\frac{P_{\mathcal{C}}(c_i)}{P_{\mathcal{I}}(c_i)} f(c_i)\right] \quad = \mathbb{E}_{P_{\mathcal{I}}}[w_i f(c_i)]$$

Defining an incumbent to be a hired candidate, w_i depends on the overall selection rate $P_{\mathcal{C}}(\text{hired})$ and the probability with which c_i was hired:

$$\begin{aligned}\frac{P_{\mathcal{C}}(c_i)}{P_{\mathcal{I}}(c_i)} &= \frac{P_{\mathcal{C}}(c_i)}{P_{\mathcal{C}}(c_i|\text{hired})} \\ &= \frac{P_{\mathcal{C}}(c_i)}{\frac{P_{\mathcal{C}}(\text{hired}|c_i)P_{\mathcal{C}}(c_i)}{P_{\mathcal{C}}(\text{hired})}} \\ &= \frac{P_{\mathcal{C}}(\text{hired})}{P_{\mathcal{C}}(\text{hired}|c_i)} = w_i\end{aligned}$$

A *propensity model* of the probability of hire based on applicant features, such as resume text, makes propensity weights practical by substituting $\hat{Q}(\text{hired}|c_i) \approx P_{\mathcal{C}}(\text{hired}|c_i)$. Our simulations add noise to the true probability of hire to simulate imperfect propensity model output. In practice, $\hat{Q}$ should be trained and evaluated for calibration (see Appendix C). IPW additionally assumes $P_{\mathcal{C}}(\text{hired}|c_i) > 0$ for all candidates (e.g. we can only correct to basically-qualified applicants), but does not assume any specific extrapolation from incumbents to all candidates.

Small propensities increase variance by making weights arbitrarily large, so we consider variance-reducing versions. Swaminathan & Joachims (2015) proposed Self-Normalized Inverse Propensity Scores (SNIPS):

$$\widetilde{w}_i = \frac{w_i}{\frac{1}{N_{\text{Incumbents}}}\sum_{j=1}^{N_{\text{Incumbents}}} w_j}$$

where $N_{\text{Incumbents}}$ is the number of incumbents. Ionides (2008) avoided large weights, which we call Truncated IPW:

$$w_i' = \min\left(w_i, \sqrt{N_{\text{Incumbents}}}\frac{1}{N_{\text{Incumbents}}}\sum_{j=1}^{N_{\text{Incumbents}}} w_j\right)$$

Truncated SNIPS applies self-normalization to truncated weights.

Results

For each generative story, we sample 1,000 datasets containing 1,000, 10,000, and 100,000 candidates. Each story samples five things for each candidate:

1. a_i: Assessment score.
2. r_i: PIR score.
3. q_i: Probability of hire (if the hiring process repeated many times, how often would this candidate be selected).
4. h_i: Whether the candidate was hired or not (sampled using q_i).
5. $\hat{q}_i$: Estimated probability of hire (q_i with noise, used by IPW).

Figures 2, 3, and 4 plot Corrected R against Full-Sample R for each story. Ideally Corrected and Full-Sample R are the same, shown by the black $y = x$ line. Table 1 summarizes the bias of each correction (i.e. $\mathbb{E}[\text{Corrected } R - \text{Full-Sample } R]$). Traditional corrections involve the square root of a potentially-negative term, producing an error; we plot errors in red, and excluded them from the Lowess fit and Table 1. IPW and Truncated IPW systematically underestimate at high Full-Sample correlations (additional simulations showed this is due to the propensity noise), but SNIPS doesn't.

Traditional corrections overestimate in the Overfit story. While Case V performs well in the Well-fit story, Case I over-estimates, and Case II under-estimates. In all stories, SNIPS exhibits low bias, and Truncated SNIPS additionally exhibits lower variance that decreases with dataset size.

Discussion

Theory and simulation show that IPW provides more general correction for range restriction than traditional corrections. Researchers with access to large datasets of candidate features should consider using Truncated SNIPS, especially when overfitting is a risk.

Appendices

Appendix A: Generative Stories

Simple Linear MAR

Given

Dataset size N

Assessment-PIR Covariance $\sigma_{t,r}$

Propensity noise σ_q

Sample

$$t_i, r_i \sim \mathcal{N}\left([-2,0], \begin{bmatrix} 1 & \sigma_{t,r} \\ \sigma_{t,r} & 1 \end{bmatrix}\right) \quad i = 1, \ldots, N$$

$$\begin{aligned} q_i &= \text{logistic\_sigmoid}(r_i - 2) \\ h_i &\sim \text{Bernoulli}(q_i) \\ \hat{q}_i' &\sim \mathcal{N}\left(\log\frac{q_i}{1-q_i}, \sigma_q\right) \\ \hat{q}_i &= \text{logistic\_sigmoid}(\hat{q}_i') \end{aligned}$$

Overfit and Well-fit

Given

Dataset size N

Effect size μ

Strong Assessment-PIR covariance $\sigma_{t,r}^{(1)}$

Weak Assessment-PIR covariance $\sigma_{t,r}^{(0)}$

Strong proportion $\pi^{(\text{strong})}$

Strong selection rate $s^{(1)}$

Weak selection rate $s^{(0)}$

Propensity noise σ_q

Sample

$$\begin{aligned} c_i &\sim \text{Bernoulli}\left(\pi^{(\text{strong})}\right) && i = 1, \ldots, N \\ t_i, r_i &\sim \mathcal{N}\left([0, c_i\mu], \begin{bmatrix} 1 & \sigma_{t,r}^{(c_i)} \\ \sigma_{t,r}^{(c_i)} & 1 \end{bmatrix}\right) \\ q_i &\sim \text{Beta}\left(10\left(1 - s^{(c_i)}\right), 10s^{(c_i)}\right) \\ h_i &\sim \text{Bernoulli}(q_i) \\ \hat{q}_i' &\sim \mathcal{N}\left(\log\frac{q_i}{1-q_i}, \sigma_q\right) \\ \hat{q}_i &= \text{logistic\_sigmoid}(\hat{q}_i') \end{aligned}$$

Appendix B: IPW Derivation

$$\begin{aligned} \mathbb{E}_{P_\mathcal{C}}[f(c_i)] &= \sum_{c_i \in \mathcal{C}} P_\mathcal{C}(c_i) f(c_i) \\ &= \sum_{c_i \in \mathcal{C}} P_\mathcal{I}(c_i) \frac{P_\mathcal{C}(c_i)}{P_\mathcal{I}(c_i)} f(c_i) \quad \text{with } P_\mathcal{I}(c_i) > 0 \\ &= \mathbb{E}_{P_\mathcal{I}}\left[\frac{P_\mathcal{C}(c_i)}{P_\mathcal{I}(c_i)} f(c_i)\right] \\ &= \mathbb{E}_{P_\mathcal{I}}[w_i f(c_i)] \end{aligned}$$

Appendix C: IPW Recipe

To use IPW, use a representative training set to prepare a calibrated propensity model $\hat{Q}$, optimizing a strictly-proper scoring rule such as binary cross-entropy loss or Brier score (e.g. Machete, 2013).

Compute weights for held-out incumbents. For Truncated SNIPS:

$$\begin{aligned}
P_{\mathcal{C}}(\text{hired}) &\approx \frac{1}{N}\sum_{i=1}^{N} h_i = \hat{P}_{\mathcal{C}}(\text{hired}) \\
w_i &= \frac{\hat{P}_C(\text{hired})}{\hat{Q}(\text{hired}|c_i)} \\
w_{\max} &= \sqrt{N}\,\frac{1}{N}\sum_{j=1}^{N} w_j \\
w_i^{\text{Truncated}} &= \min(w_i, w_{\max}) \\
w_i^{\text{TruncatedSNIPS}} &= \frac{w_i^{\text{Truncated}}}{\frac{1}{N}\sum_{j=1}^{N} w_j^{\text{Truncated}}}
\end{aligned}$$

Weight the expectations in Pearson R:

$$\begin{aligned}
\mu_a^{\mathcal{C}} &= \mathbb{E}_{P_{\mathcal{C}}}[a] = \mathbb{E}_{P_{\mathcal{J}}}[wa] \approx \frac{1}{N_{\text{Incumbents}}} \sum_{i=1}^{N_{\text{Incumbents}}} \left[w_i^{\text{TruncatedSNIPS}} a_i\right] = \hat{\mu}_a^{\mathcal{C}} \\
\mu_r^{\mathcal{C}} &\approx \frac{1}{N_{\text{Incumbents}}} \sum_{i=1}^{N_{\text{Incumbents}}} \left[w_i^{\text{TruncatedSNIPS}} r_i\right] = \hat{\mu}_r^{\mathcal{C}} \\
R_{a,r}^{\mathcal{C}} &= \frac{\text{Cov}^{\mathcal{C}}(a,r)}{\text{Var}^{\mathcal{C}}(a)\text{Var}^{\mathcal{C}}(r)} \\
&= \frac{\mathbb{E}_{P_{\mathcal{C}}}[(a-\mu_a^{\mathcal{C}})(r-\mu_r^{\mathcal{C}})]}{\sqrt{\mathbb{E}_{P_{\mathcal{C}}}\left[\left(a-\mu_a^{\mathcal{C}}\right)^2\right]\mathbb{E}_{P_{\mathcal{C}}}\left[\left(r-\mu_r^{\mathcal{C}}\right)^2\right]}} \\
&= \frac{\mathbb{E}_{P_{\mathcal{J}}}[w(a-\mu_a^{\mathcal{C}})(r-\mu_r^{\mathcal{C}})]}{\sqrt{\mathbb{E}_{P_{\mathcal{J}}}\left[w\left(a-\mu_a^{\mathcal{C}}\right)^2\right]\mathbb{E}_{P_{\mathcal{J}}}\left[w\left(r-\mu_r^{\mathcal{C}}\right)^2\right]}} \\
&\approx \frac{\sum_{i=1}^{N_{\text{Incumbents}}}\left[w_i^{\text{TruncatedSNIPS}}(a_i-\mu_a^{\mathcal{C}})(r_i-\mu_r^{\mathcal{C}})\right]}{\sqrt{\sum_{i=1}^{N_{\text{Incumbents}}}\left[w_i^{\text{TruncatedSNIPS}}\left(a_i-\hat{\mu}_a^{\mathcal{C}}\right)^2\right]\sum_{i=1}^{N_{\text{Incumbents}}}\left[w_i^{\text{TruncatedSNIPS}}\left(r_i-\hat{\mu}_r^{\mathcal{C}}\right)^2\right]}}
\end{aligned}$$

References

Bryant, N. D., & Gokhale, S. (1972). Correcting correlations for restrictions in range due to selection on an unmeasured variable. *Educational and Psychological Measurement*, *32*(2), 305–310.

Ionides, E. L. (2008). Truncated importance sampling. *Journal of Computational and Graphical Statistics*, *17*(2), 295–311. https://doi.org/10.1198/106186008X320456

Lanza, S. T., Moore, J. E., & Butera, N. M. (2013). Drawing causal inferences using propensity scores: A practical guide for community psychologists. *American Journal of Community Psychology*, *52*(3-4).

Li, L., Chu, W., Langford, J., & Wang, X. (2011). Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, 297–306.

Little, R. J. A. (1986). Survey nonresponse adjustments for estimates of means. *International Statistical Review*, *54*(2), 139–157.

Machete, R. L. (2013). Contrasting probabilistic scoring rules. *Journal of Statistical Planning and Inference*, *143*(10).

Rosenbaum, P. R. (1988). Model-based direct adjustment. *Journal of the American Statistical Association*, *82*(398), 387–394.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516–524.

Seaman, S. R., & White, I. R. (2011). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3).

Strehl, A., Langford, J., & Kakade, S. (2010). Learning from logged implicit exploration data. *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, 2217–2225.

Swaminathan, A., & Joachims, T. (2015). The self-normalized estimator for counterfactual learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 28). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2015/file/39027dfad5138c9ca0c474d71db915c3-Paper.pdf

Thoemmes, F., & Ong, A. D. (2016). A primer on inverse probability of treatment weighting and marginal structural models. *Emerging Adulthood*, *4*(1).

Thorndike, R. L. (1947). *Research problems and techniques* (Report No. 3). U.S. Government Printing Office.

Thorndike, R. L. (1949). *Personnel selection; test and measurement techniques*. Wiley.

Figure 1: Three generative story datasets.

Figure 2: Linear MAR

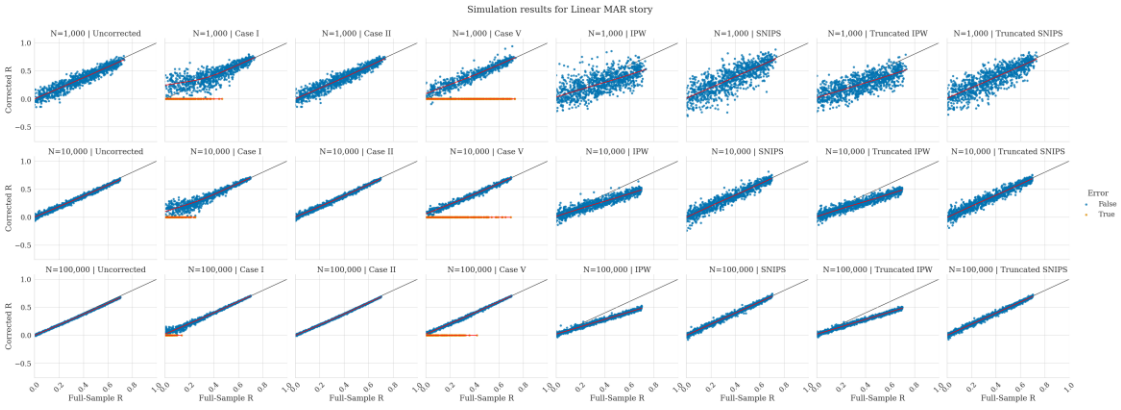

Figure 3: Overfit

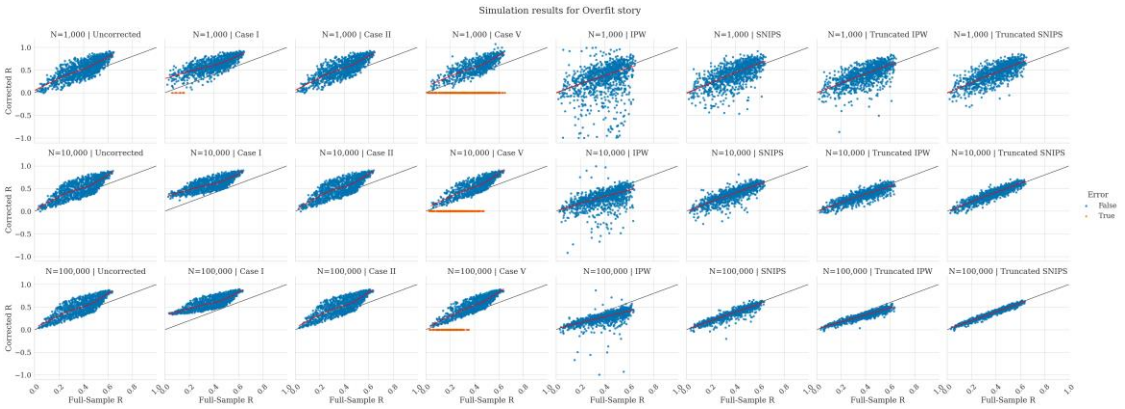

Figure 4: Well-fit

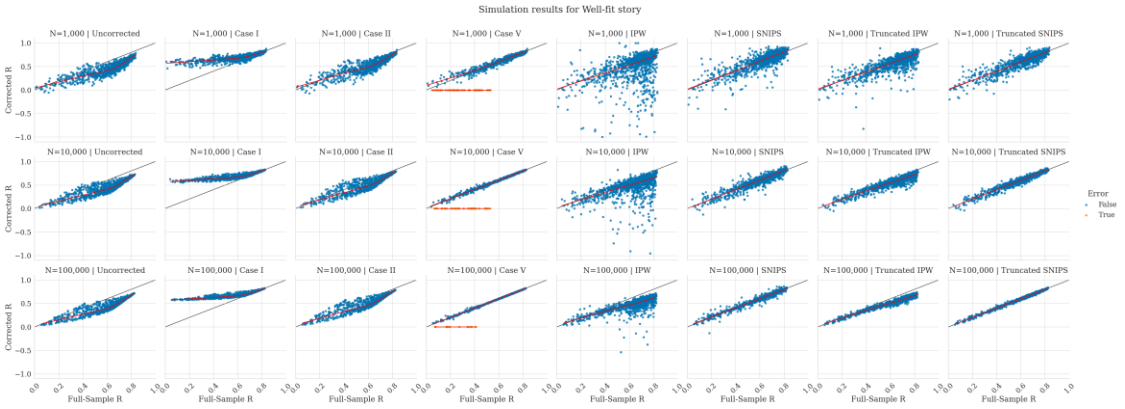

Table 1: Simulation Biases.

Correction			**Linear MAR**						**Well-fit**
	1000	10000	100000	1000	10000	100000	1000	10000	100000
Uncorrected	-0.021	-0.020	-0.020	0.129	0.130	0.122	-0.149	-0.145	-0.151
Case I	0.047	0.014	0.004	0.216	0.225	0.216	0.080	0.089	0.083
Case II	-0.015	-0.015	-0.016	0.150	0.151	0.143	-0.080	-0.076	-0.081
Case V	0.037	0.014	0.003	0.169	0.151	0.146	0.002	0.001	0.000
IPW	-0.080	-0.096	-0.101	-0.091	-0.103	-0.128	-0.160	-0.157	-0.157
SNIPS	-0.009	-0.007	-0.004	0.017	-0.007	-0.022	0.001	-0.002	-0.000
Truncated IPW	-0.098	-0.099	-0.101	-0.002	-0.029	-0.075	-0.055	-0.064	-0.098
Truncated SNIPS	-0.013	-0.008	-0.004	0.037	0.009	-0.008	-0.006	-0.001	-0.000

Table 2: Hyperparameters for the stories

	Linear MAR Story	**Overfitting Story**	**Well-fit Story**
$\sigma_{m,r}$	$\sim$ Uniform$(0,0.7)$	–	–
μ	–	$\sim$ Uniform$(0,2)$	$\sim$ Uniform$(0,5)$
$\sigma_{t,r}^{(1)}$	–	$\sim$ Uniform$(0,1.0)$	$\sim$ Uniform$(0,0.6)$
$\sigma_{t,r}^{(0)}$	–	0	$\sim$ Uniform$\left(0,\sigma_{t,r}^{(1)}\right)$
$\pi^{(\text{strong})}$	–	0.3	0.5
$s^{(1)}$	–	0.5	0.5
$s^{(0)}$	–	0.05	0.05
σ_q	1	1	1