Conditionally Unbiased Best Linear Predictors for Score Augmentation

Xiang Liu, Matthew S. Johnson, Sandip Sinharay

Educational Testing Service

May 18, 2022

Background

- Assessments transition from being paper-pencil based to digital.
- The Big Data movement.
- Additional information about the construct of interest comes from different sources.
- In writing assessments - scores from multiple raters, product features (e.g. NLP related ones from the e-rater®), and process features from keystroke logs, etc.
- The goal is to make inferences about the writing ability combining all these information.
- Augmenting the rater scores with additional information.
- The same can be generalized to other assessments - math, reading, speaking, etc.

Motivation

- The best linear predictors (BLP) has been proposed to combine sources of information to predict some latent true score (e.g. augmenting writing proficiency estimate with scores from other sections; Haberman et al., 2015; Yao et al., 2019).
- "Predicting" random effects vs. "estimating" fixed effects.
- The BLP minimizes the MSE of prediction; However, its exhibits shrinkage towards the population mean (Robinson, 1991).
- The BLP may biased conditional on the true score level (i.e. for individual students). It could lead to fairness concerns (e.g. favoring certain groups over others).
- We need an approach that is conditionally unbiased (or unbiased at individual level).

The model

- $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_J)^\top$ is the vector of variables measured with error (i.e. manifest variables of some latent variable of interest. e.g.essay scores from raters).
- $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)^\top$ is the vector of variables measured without error (i.e. covariates that may correlate with the latent trait, e.g. typing speed).
- $$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} = \boldsymbol{\alpha} + \boldsymbol{\lambda} S + \boldsymbol{\varepsilon}_w. \tag{1}$$

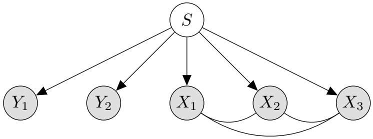
  , where S is the latent variable and $\boldsymbol{\varepsilon}_w$ denotes the residuals.
- For mathematical convenience, we center $\boldsymbol{W}$,

  $$\boldsymbol{W} = \boldsymbol{\lambda} S + \boldsymbol{\varepsilon}_w, \tag{2}$$

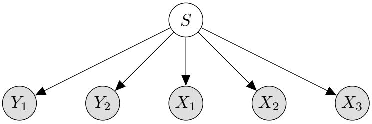
  $E(S) = 0$, $E(\boldsymbol{\varepsilon}_w) = \mathbf{0}$, and

  $$\mathrm{Cov}\begin{pmatrix} S \\ \boldsymbol{\varepsilon}_w \end{pmatrix} = \begin{bmatrix} \sigma_S^2 & 0 \\ 0 & \boldsymbol{\Sigma}_{\varepsilon_w} \end{bmatrix}.$$

Example models



(a) Correlated residuals of X



(b) Uncorrelated residuals of X

Figure: Graphical representation of some example models

- The model is flexible. Depending on the structure of $\boldsymbol{\Sigma}_{\varepsilon_w}$, it can describe a broad class of models.

BLP

- BLP, $\hat{S} = \gamma_1^\top \boldsymbol{W}$, minimizes the mean squared error for prediction[1],

$$\text{MSE} = E[(S - \gamma_1^\top \boldsymbol{W})^2] = \sigma_S^2 + \gamma_1^\top \boldsymbol{\Sigma}_{\boldsymbol{W}} \gamma_1 - 2\gamma_1^\top \boldsymbol{\lambda} \sigma_S^2 \quad (3)$$

- BLP is obtained by solving

$$\nabla \text{MSE} = \nabla(\gamma_1^\top \boldsymbol{\Sigma}_{\boldsymbol{W}} \gamma_1) - \nabla(2\gamma_1^\top \boldsymbol{\lambda} \sigma_S^2) \quad (4)$$

$$= 2\Sigma_{\boldsymbol{W}} \gamma_1 - 2\boldsymbol{\lambda} \sigma_S^2 = 0, \quad (5)$$

- The BLP coefficients are $\gamma_1 = \Sigma_{\boldsymbol{W}}^{-1} \boldsymbol{\lambda} \sigma_S^2$. By the decomposition $\Sigma_{\boldsymbol{W}} = \sigma_S^2 \boldsymbol{\lambda}\boldsymbol{\lambda}^\top + \boldsymbol{\Sigma}_{\varepsilon_w}$ and the Woodbury matrix identity, the coefficients can be alternatively expressed as

$$\gamma_1 = \frac{\Sigma_{\varepsilon_w}^{-1} \boldsymbol{\lambda}}{1/\sigma_S^2 + \boldsymbol{\lambda}^\top \Sigma_{\varepsilon_w}^{-1} \boldsymbol{\lambda}}. \quad (6)$$

[1] assumes $E[\varepsilon_w | S] = 0$

Bias

- **Averaged over the population**, the BLP is unbiased, i.e.

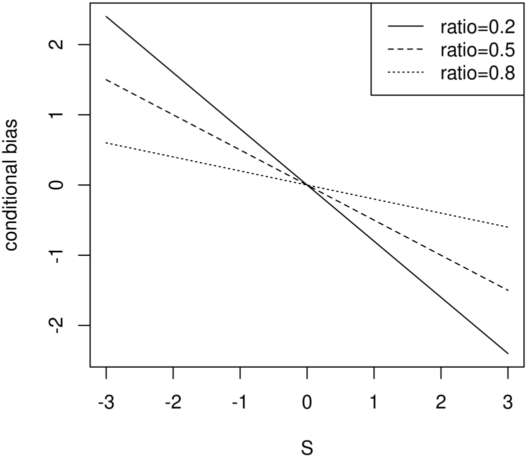
$$\begin{aligned} Bias(\hat{S}) &= E[\hat{S} - S] \\ &= E[(\sigma_S^2 \boldsymbol{\lambda}^\top \Sigma_{\boldsymbol{W}}^{-1})\boldsymbol{W} - S] \qquad (7) \\ &= \sigma_S^2 \boldsymbol{\lambda}^\top \Sigma_{\boldsymbol{W}}^{-1} E[W] - E[S] = 0. \qquad (8) \end{aligned}$$

- The BLP can be biased **for individuals** (or conditional S),

$$\begin{aligned} E[\hat{S} - S|S] &= E[\sigma_S^2 \boldsymbol{\lambda}^\top \Sigma_{\boldsymbol{W}}^{-1} \boldsymbol{W}|S] - S \\ &= E[\sigma_S^2 \boldsymbol{\lambda}^\top \Sigma_{\boldsymbol{W}}^{-1} (\boldsymbol{\lambda} S + \varepsilon_w)|S] - S \\ &= S(\sigma_S^2 \boldsymbol{\lambda}^\top \Sigma_{\boldsymbol{W}}^{-1} \boldsymbol{\lambda} - 1). \qquad (9) \end{aligned}$$

- The conditional bias depends on S and the ratio $\sigma_S^2 \boldsymbol{\lambda}^\top \Sigma_{\boldsymbol{W}}^\top \boldsymbol{\lambda}$.

Figure: Conditional bias under different covariance ratios



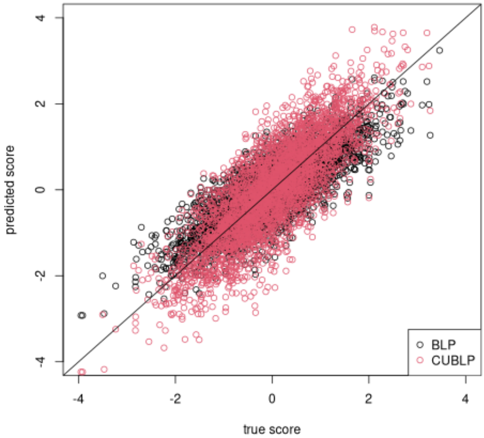
CUBLP

- The CUBLP, $\hat{S} = \gamma_1^\top \boldsymbol{W}$, minimizes the MSE $E[(S - \gamma_1^\top \boldsymbol{W})^2]$, subject to

$$E[\gamma_1^\top \boldsymbol{W}|S] = \gamma_1^\top E[\boldsymbol{W}|S] = \gamma_1^\top(\boldsymbol{\lambda} S + E[\varepsilon_w|S]) = S. \quad (10)$$

- With the assumption $E(\varepsilon_w|S) = 0$, the constraint reduces to $\gamma_1^\top \boldsymbol{\lambda} = 1$.
- This constrained optimization problem can be solved by the method of Lagrange multipliers. The Lagrange function is $\mathcal{L}(\boldsymbol{\lambda}_1, \delta) = -E[(S - \gamma_1^\top \boldsymbol{W})^2] - \delta(\gamma_1^\top \boldsymbol{\lambda} - 1)$.
- The CUBLP coefficients are obtained by solving $\nabla_{\boldsymbol{\lambda}_1,\delta}\mathcal{L}(\boldsymbol{\lambda}_1^\top, \delta) = \mathbf{0}$.
- $\boldsymbol{\lambda}_1 = \frac{\boldsymbol{\Sigma}_W^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_W^{-1}\boldsymbol{\lambda}}$ or alternatively $\boldsymbol{\lambda}_1 = \frac{\boldsymbol{\Sigma}_{\varepsilon_w}^\top \boldsymbol{\lambda}}{\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_{\varepsilon_w}^\top \boldsymbol{\lambda}}$.

Figure: Comparison of BLP and CUBLP



Parameter estimation

- BLP and CUBLP coefficients are expressed as functions of the population parameters $\boldsymbol{\lambda}$, and $\boldsymbol{\Sigma_W}$ (or $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}_w}$). They need to be estimated.
- Let $\boldsymbol{\lambda} = \left(\boldsymbol{\lambda}_{\boldsymbol{Y}}^\top, \boldsymbol{\lambda}_{\boldsymbol{X}}^\top\right)^\top$ In some applications, it may be desirable or necessary to assume $\lambda_{Y_j} = \lambda_Y$. We have the covariance decomposition
$$\boldsymbol{\Sigma_W} = \boldsymbol{\lambda}\boldsymbol{\lambda}^\top + \boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}, \tag{11}$$
, and
$$\boldsymbol{\Sigma_Y} = \boldsymbol{\lambda_Y}\boldsymbol{\lambda}_{\boldsymbol{Y}}^\top + \boldsymbol{\Psi_Y}, \tag{12}$$
where $\boldsymbol{\Psi_Y}$ is a diagonal matrix.
- A least square (LS) estimator is obtained by minimizing $L(\boldsymbol{\lambda}) = \sum_{k \neq k'} (\lambda_Y^2 - r_{Y_k Y_{k'}})^2 + \sum_k \sum_j (\lambda_Y \lambda_{X_j} - r_{Y_k X_j})^2$

The LS estimator

- For equal discrimination cases,

$$\hat{\lambda_Y} = \sqrt{\frac{\sum_{k \neq k'} r_{Y_k Y_{k'}}}{K(K-1)}}, \hat{\lambda_{X_j}} = \frac{\sum_k r_{Y_k X_j}}{K \hat{\lambda_Y}}. \tag{13}$$

- For unequal discrimination cases, iterate through

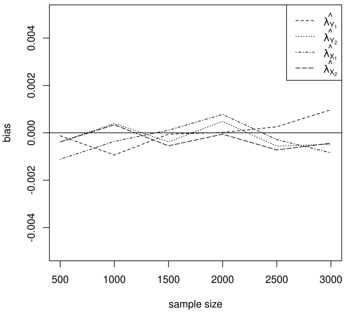
$$\hat{\lambda_{Y_k}} = \frac{\sum_{k' \neq k} \lambda_{Y_{k'}} r_{Y_k Y_{k'}} + \sum_j \lambda_{X_j} r_{Y_k, X_j}}{\sum_{k' \neq k} \lambda^2_{Y_{k'}} + \sum_j \lambda^2_{X_j}}, \tag{14}$$

and

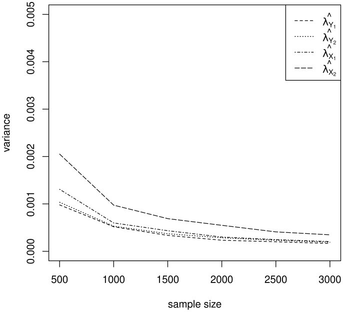
$$\hat{\lambda_{X_j}} = \frac{\sum_k \lambda_{Y_k} r_{Y_k X_j}}{\sum_k \lambda^2_{Y_k}} \tag{15}$$

- $\boldsymbol{\Sigma_W}$ can be estimated by the sample covariance matrix or its ML estimate if distribution assumption is assumed.

Figure: Behaviors of the LS estimator



(a) Bias



(b) Variance

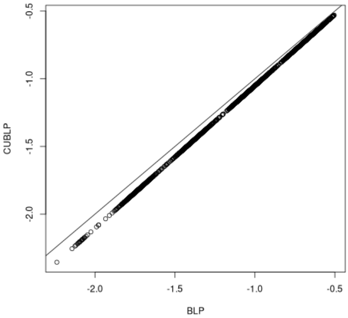
A speech scoring example

- The speaking section of an English language proficiency test.
- A listen-repeat task, The sentences vary in lengths and may situate in different scenarios such as presentations and campus tours.
- Each recorded response is scored by two different trained raters, scale from 0 to 5.
- The *SpeechRater*™of ETS provides features related to different dimensions of a speech. Composite scores on accuracy, pronunciation, fluency, and rhythm are computed.
- 7690 observations. Raters are randomly assigned.

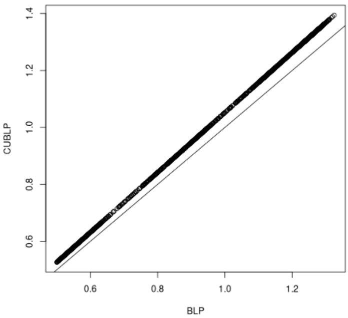
Results: estimates

	CUBLP	BLP	$\hat{\lambda}$
rating 1	0.456	0.433	0.944
rating 2	0.452	0.430	0.944
accuracy	0.144	0.137	0.839
pronunciation	0.009	0.009	0.333
fluency	0.017	0.016	0.400
rhythm	0.025	0.024	0.468

Results: BLP and CUBLP



(a) Lower performers



(b) Higher performers

Discussion

- Threats to fairness and equity.
- The algorithmic bias.
- The method is general and can be applied to a wide range of problems.
- Analyze a larger dataset.

References

Haberman, S. J., Yao, L., & Sinharay, S. (2015). Prediction of true test scores from observed item scores and ancillary data. *British Journal of Mathematical and Statistical Psychology*, *68*(2), 363–385. https://doi.org/10.1111/bmsp.12052

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects. *Statistical Science*, *6*(1), 15–32. https://doi.org/10.1214/ss/1177011926

Yao, L., Haberman, S. J., & Zhang, M. (2019). Penalized Best Linear Prediction of True Test Scores. *Psychometrika*, *84*(1), 186–211. https://doi.org/10.1007/s11336-018-9636-7

Thank you!

xliu003@ets.org