Conditionally Unbiased Best Linear Predictors for Score Augmentation

Xiang Liu, Matthew S. Johnson, and Sandip Sinharay

Educational Testing Service

Abstract

The best linear predictor (BLP) has been proposed to combine different types of information in estimating true scores. The BLP is biased for individual examinees when conditioned on the true score. In this paper, we propose a conditionally unbiased BLP. Additionally, a least square method is introduced for parameter estimation.

Conditionally Unbiased Best Linear Predictors for Score Augmentation

Introduction

With the current popular trend of transition from traditional paper-pencil assessments to digital formats, an increasingly common task is to combine different sources of information in assessing some latent construct. For instance, in the context of writing assessments, in addition to the final product essays, digital platforms are now able to capture information on keystrokes during students' interaction with writing tasks. Thus, features related to writing process are now becoming available (e.g. Guo et al., 2019; Zhang et al., 2019). Furthermore, additional product features such as those related to natural language processing (NLP) could also be obtained through NLP programs (e.g. the e-rater® program; Attalí and Burstein, 2006; Burstein et al., 2004). Then a natural question is whether and how different sources of information can be combined to augment the rater scores in making inferences on examinees' writing proficiency.

There has been some previous research exploring related methods. For example, Zhang and Deane (2015) used linear regression to predict adjudicated essay scores from a large number of keystroke logging features and other product features. Sinharay et al. (2019) examined the same prediction problem utilizing machine learning methods such as the random forests and boosting. A common assumption of these methods is that essay score, the prediction target, is treated as fixed and known. However, in educational and psychological measurement, prediction targets are almost always latent and measured with error. To address this issue, the best linear predictors (BLP; e.g. Haberman et al., 2015; Yao et al., 2019) has been proposed to combine sources of information (e.g. scores from other sections) along with manifest variables such as the rater scoresand to predict some latent true score (e.g. the writing proficiency) .

The BLP minimizes the mean squared error (MSE) of prediction among all linear predictors. An important property of the BLP is that, when averaged over the population, the expected prediction is equal to the expected true score. In other words, the BLP is

unbiased. However, it exhibits shrinkage towards the population mean. Consequently, the BLP is biased conditional on the true score level. This may lead to some serious fairness concern as the BLP could favor certain groups of examinees over others.

In this paper, we introduce a conditionally unbiased best linear predictor (CUBLP) that minimizes the prediction error among all linear predictors that are unbiased at all true score levels. The proposed method is flexible and can be applied to a wide variety of problems.

Methods

Let $\boldsymbol{W} = (\boldsymbol{Y}^\top, \boldsymbol{X}^\top)^\top$ denote observed random variables from a random examinee where $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_J)^\top$ is the vector of variables measured with error[1] and $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)^\top$ is the vector of variables measured without error[2]. For simplicity, we consider an unidimensional true score S and further assume that $\boldsymbol{W}$ and S are linearly related, i.e.

$$\boldsymbol{W} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{X} \end{pmatrix} = \boldsymbol{\alpha} + \boldsymbol{\lambda} S + \boldsymbol{\epsilon}_w. \tag{1}$$

We propose a CUBLP of S in the form of $\gamma_0 + \boldsymbol{\gamma}_1^\top \boldsymbol{W}$ that minimizes

$$\text{MSE} = E\left[(S - \gamma_0 - \boldsymbol{\gamma}_1^\top \boldsymbol{W})^2\right], \tag{2}$$

subject to the constraint that it is conditionally unbiased, that is

$$E[\gamma_0 + \boldsymbol{\gamma}_1^\top \boldsymbol{W} | S] = \gamma_0 + \boldsymbol{\gamma}_1^\top E[\boldsymbol{W}|S] = S. \tag{3}$$

Under the assumption $E[\boldsymbol{\epsilon}_w | S] = 0$, we have $E[\boldsymbol{W}|S] = \boldsymbol{\lambda} S$. Combined with the observation that, in order to minimize the MSE in Equation 2, $\gamma_0 = E[S] - \boldsymbol{\gamma}_1^\top E[\boldsymbol{W}]$, it leads to

$$\begin{aligned} S &= E[S] - \boldsymbol{\gamma}_1^\top \boldsymbol{\lambda} E[S] + \boldsymbol{\gamma}_1^\top \boldsymbol{\lambda} \\ S - E[S] &= \boldsymbol{\gamma}_1^\top \boldsymbol{\lambda} (S - E[S]). \end{aligned} \tag{4}$$

[1] Commonly referred to as manifest variables of some latent trait, e.g. essay scores from raters.

[2] Covariates that may correlate with the latent trait of interest, e.g. typing speed.

It immediately follows that the constraint in Equation 3 reduces to

$$\boldsymbol{\gamma}_1^\top \boldsymbol{\lambda} = 1. \tag{5}$$

We solve the constrained minimization problem with the method of Lagrange multipliers (See Appendix A for details). The CUBLP coefficients are

$$\boldsymbol{\gamma}_1 = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}^\top \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}^{-1}\boldsymbol{\lambda}}, \tag{6}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}$ is the variance-covariance matrix of $\boldsymbol{\epsilon}_w$.

Notice that the CUBLP coefficients in Equation 6 are expressed as a function of the population parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}$. In almost all real applications, they have to be estimated. Let

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda_Y} \\ \boldsymbol{\lambda_X} \end{pmatrix},$$

where $\boldsymbol{\lambda_Y} = (\lambda_{Y_1}, \lambda_{Y_2}, \ldots, \lambda_{Y_K})^\top$ and $\boldsymbol{\lambda_X} = (\lambda_{X_1}, \lambda_{X_2}, \ldots, \lambda_{X_J})^\top$ are the loadings for $\boldsymbol{Y}$ and $\boldsymbol{X}$. In many cases, it may be desirable or necessary to assume that $\lambda_{Y_j} = \lambda_Y, \forall j$. This equal discrimination assumption is common for many measurement models. For the equal discrimination cases, assuming $\boldsymbol{W}$ are standardized and under some additional assumptions, the least square (LS) estimator for $\boldsymbol{\lambda}$ is

$$\hat{\lambda_Y} = \sqrt{\frac{\sum_{k \neq k'} r_{Y_k Y_{k'}}}{K(K-1)}} \tag{7}$$

and

$$\hat{\lambda_{X_j}} = \frac{\sum_k r_{Y_k X_j}}{K\hat{\lambda_Y}}, \tag{8}$$

where $r_{Y_k X_j}$ denotes the correlation between Y_k and X_j. For unequal discrimination cases, the LS estimator can be obtained by iterating through

$$\hat{\lambda_{Y_k}} = \frac{\sum_{k' \neq k} \lambda_{Y_{k'}} r_{Y_k Y_{k'}} + \sum_j \lambda_{X_j} r_{Y_k, X_j}}{\sum_{k' \neq k} \lambda_{Y_{k'}}^2 + \sum_j \lambda_{X_j}^2}, \tag{9}$$

and

$$\hat{\lambda_{X_j}} = \frac{\sum_k \lambda_{Y_k} r_{Y_k X_j}}{\sum_k \lambda_{Y_k}^2} \tag{10}$$

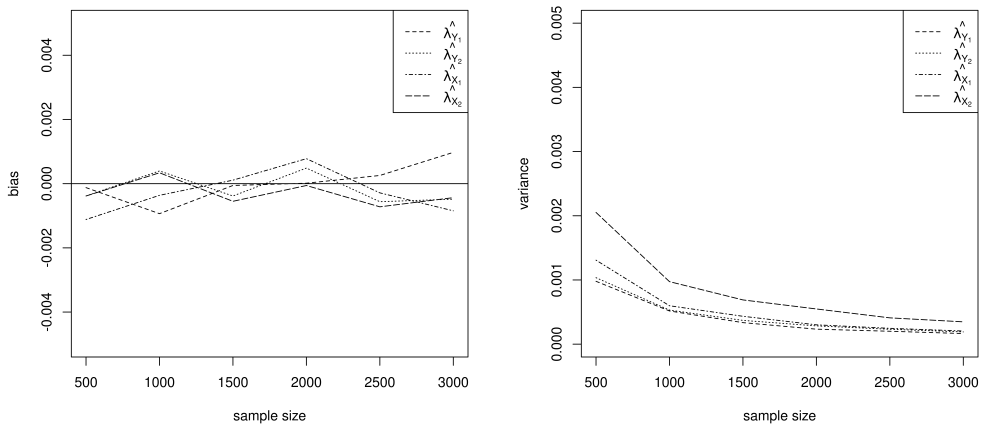
until convergence. A natural choice of an estimator for the residual covariance matrix is then

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\epsilon}_w} = \boldsymbol{\Sigma}_{\boldsymbol{W}} - \hat{\boldsymbol{\lambda}}\hat{\boldsymbol{\lambda}}^{\top}. \tag{11}$$

For a sketch of the derivations of these results, please see Appendix B.

Planned simulations and data analysis

We plan to conduct simulation studies to examine the properties of the proposed CUBLP and the LS estimator of the parameters. preliminary results suggest that the proposed LS estimator has good statistical properties (See Figure 1). Further simulation



(a) *Bias* (b) *Variance*

Figure 1

Bias and variance of the LS estimator

studies would investigate the prediction performance of the CUBLP under different scenarios, e.g. different sample sizes, high/low correlations between the manifest variables $\boldsymbol{Y}$ and the true score S. Prediction bias and variance would be examined. We will also apply the proposed method to a real dataset from a pilot study of an English language test where examinees were asked to respond to several speech tasks. Each response was scored

by two human raters. And NLP features would be used to augment the human rater scores in predicting the speaking proficiency.

Practical Implications

The proposed CULBP method utilizes multiple sources of information about the examinees to improve measurement precision. At the same time, it does it in an equitable fashion. This is particularly important when we develop methods that can leverage the opportunities that the current trend of Big Data brings.

References

Attalí, Y., & Bursteín, J. (2006). Automated essay scoring with e-rater® V.2. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The criterion online writing service. *AI Magazine*, *25*(3), 27–36.

Guo, H., Zhang, M., Deane, P., & Bennett, R. E. (2019). Writing Process Differences in Subgroups Reflected in Keystroke Logs. *Journal of Educational and Behavioral Statistics*, *44*(5), 571–596. https://doi.org/10.3102/1076998619856590

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

Sinharay, S., Zhang, M., & Deane, P. (2019). Prediction of Essay Scores From Writing Process and Product Features Using Data Mining Methods. *Applied Measurement in Education*, *32*(2), 116–137. https://doi.org/10.1080/08957347.2019.1577245

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

Zhang, M., Bennett, R. E., Deane, P., & van Rijn, P. W. (2019). Are There Gender Differences in How Students Write Their Essays? An Analysis of Writing Processes. *Educational Measurement: Issues and Practice*, *38*(2), 14–26. https://doi.org/10.1111/emip.12249

Zhang, M., & Deane, P. (2015). Process Features in Writing: Internal Structure and Incremental Value Over Product Features. *ETS Research Report Series*, *2015*(2), 1–12. https://doi.org/10.1002/ets2.12075

Appendix A

Constrained Minimization of the MSE

To find the optimal value for γ we use a Lagrange multiplier approach where we optimize

$$\frac{1}{2}MSE - \delta(\boldsymbol{\gamma}_1^\top \boldsymbol{\lambda} - 1)$$

The gradient of this function is

$$-E\left[(\boldsymbol{W} - E[\boldsymbol{W}])(S - E[S] - (\boldsymbol{W} - E[\boldsymbol{W}])^\top \boldsymbol{\gamma}_1)\right] - \delta\boldsymbol{\lambda}$$

Because $\boldsymbol{W} = \boldsymbol{\lambda}S + \boldsymbol{\epsilon}_w$ and $\boldsymbol{\lambda}^\top\boldsymbol{\gamma} = 1$ by constraint, this gradient becomes

$$E\left[(\boldsymbol{W} - E[\boldsymbol{W}])\boldsymbol{\epsilon}_w^\top \boldsymbol{\gamma}_1\right] - \delta\boldsymbol{\lambda}$$

$$E\left[(\boldsymbol{\lambda}(S - E[S]) + \boldsymbol{\epsilon}_w)\boldsymbol{\epsilon}_w^\top \boldsymbol{\gamma}_1\right] - \delta\boldsymbol{\lambda}$$

Therefore, the solution satisfies

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}\boldsymbol{\gamma}_1 &= \delta\boldsymbol{\lambda} \\ \boldsymbol{\gamma}_1 &= \delta\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}^{-1}\boldsymbol{\lambda} \end{aligned}$$

where δ is a constant to ensure the constraint is met. Therefore the CUBLP coefficients are

$$\boldsymbol{\gamma}_1 = \frac{\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}^{-1}\boldsymbol{\lambda}}{\boldsymbol{\lambda}^\top\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w}^{-1}\boldsymbol{\lambda}} \tag{12}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}_w} = \mathrm{Cov}(\boldsymbol{\epsilon}_w)$ is the covariance of the error terms $\boldsymbol{\epsilon}_w = \boldsymbol{W} - \boldsymbol{\lambda}S$.

Appendix B

Parameter estimation

Following the results in the previous section, the BLUP estimator is a function of the factor loadings $\boldsymbol{\lambda}$. However, this loading vector is generally unknown and has to be estimated. Here we consider the single score case. Let

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda_Y} \\ \boldsymbol{\lambda_X} \end{pmatrix}$$

, where $\boldsymbol{\lambda_Y} = (\lambda_{Y_1}, \lambda_{Y_2}, \ldots, \lambda_{Y_K})'$ and $\boldsymbol{\lambda_X} = (\lambda_{X_1}, \lambda_{X_2}, \ldots, \lambda_{X_J})'$. Assume all observed variables are standardized and $E(S) = \mu_S = 0$ and $\sigma_S^2 = 1$. It leads to

$$\boldsymbol{\Sigma_W} = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Sigma_\epsilon}. \tag{13}$$

We further assume that the covariance of $\boldsymbol{Y}$ is fully explained by S. That is

$$\boldsymbol{\Sigma_Y} = \boldsymbol{\lambda_Y}\boldsymbol{\lambda_Y}' + \boldsymbol{\Psi_Y}, \tag{14}$$

where $\boldsymbol{\Psi_Y}$ is a diagonal matrix with residual variances of $\boldsymbol{Y}$ on the main diagonal. This is a common assumption for many popular measurement models. Now, we impose some additional restrictions on the residual covariance matrix $\boldsymbol{\Sigma_\epsilon}$. $\boldsymbol{Y}$ is assumed to be uncorrelated with $\boldsymbol{X}$ given S. Equivalently, $\mathrm{Cov}(\boldsymbol{\epsilon_Y}, \boldsymbol{\epsilon_X}) = \mathbf{0}$.

In many cases, it may be desirable or necessary to assume that $\lambda_{Y_j} = \lambda_Y$, $\forall j$. This equal discrimination assumption also simplifies the estimation of $\boldsymbol{\lambda}$. Consider a quadratic loss function,

$$L(\boldsymbol{\lambda}) = \sum_{k \neq k'} (\lambda_Y^2 - r_{Y_k Y_{k'}})^2 + \sum_k \sum_j (\lambda_Y \lambda_{X_j} - r_{Y_k X_j})^2. \tag{15}$$

Taking the gradient to minimize the loss function,

$$\nabla L = \begin{pmatrix} 4\lambda_Y \sum_{k \neq k'} (\lambda_Y^2 - r_{Y_k Y_{k'}}) + 2K\lambda_Y \sum_j \lambda_{X_j}^2 - 2\sum_j \lambda_{X_j} \sum_k r_{Y_k X_j} \\ \vdots \\ 2K\lambda_Y^2 \lambda_{X_j} - 2\lambda_Y \sum_k r_{Y_k X_j} \\ \vdots \end{pmatrix} = \mathbf{0}. \tag{16}$$

Solving Equation 16 leads to the unweighted least square estimator of the loadings,

$$\hat{\lambda_Y} = \sqrt{\frac{\sum_{k \neq k'} r_{Y_k Y_{k'}}}{K(K-1)}} \tag{17}$$

and

$$\hat{\lambda_{X_j}} = \frac{\sum_k r_{Y_k X_j}}{K\hat{\lambda_Y}}. \tag{18}$$

The solutions under unequal discriminations are

$$\hat{\lambda_{Y_k}} = \frac{\sum_{k' \neq k} \lambda_{Y_{k'}} r_{Y_k Y_{k'}} + \sum_j \lambda_{X_j} r_{Y_k, X_j}}{\sum_{k' \neq k} \lambda^2_{Y_{k'}} + \sum_j \lambda^2_{X_j}}, \tag{19}$$

and

$$\hat{\lambda_{X_j}} = \frac{\sum_k \lambda_{Y_k} r_{Y_k X_j}}{\sum_k \lambda^2_{Y_k}} \tag{20}$$