A Fay-Herriot Estimator to Improve the Mean Squared Error of Teacher Value-Added Estimates

Bing Han, Daniel F. McCaffrey, J. R. Lockwood

RAND Corporation

March 1, 2012

This work is supported by grant R305D090011 from Institute of Education Sciences. Any opinions in this talk are solely of the authors and do not necessarily represent the funding agency.
Background

► Much of the research effort on value-added (VA) modeling has been devoted to reducing the biases in VA estimates, c.f. Harris & Sass, 2006; Lockwood et al., 2007, McCaffrey et al., 2009, Rothstein, 2009.

► Relatively less attention has been dedicated to the precision and efficiency in VA estimates.

► Lack of precision in some VA estimates (in particular, teachers with small classrooms) can greatly limit the utility of VA in education evaluation.

► Other sensitivity issue of VA estimates for teachers with small classrooms (Han et al., forthcoming).
Background

- Much of the research effort on value-added (VA) modeling has been devoted to reducing the biases in VA estimates, c.f. Harris & Sass, 2006; Lockwood et al., 2007, McCaffrey et al., 2009, Rothstein, 2009.
- Relatively less attention has been dedicated to the precision and efficiency in VA estimates.
- Lack of precision in some VA estimates (in particular, teachers with small classrooms) can greatly limit the utility of VA in education evaluation.
- Other sensitivity issue of VA estimates for teachers with small classrooms (Han et al., forthcoming).
Raw VA estimates based on small classes are highly variable

Consequences of imprecise VA estimates

The more variable VA estimates (most often related to teachers with small sample sizes) result in inappropriate evaluation decisions, e.g. merit-based pay and tenure decision.

▶ When the decision is based on point estimates of VA, teachers with imprecise VA estimates have an artificial advantage to be recognized due to the variability.
  ▶ E.g., a decision rule to award teachers with VA estimates greater than .5 will recognize almost exclusively teachers having fewer than 15~20 students in the previous example.

▶ When the decision is based on a statistical test or a similar measure, teachers with imprecise VA estimates have a considerable disadvantage due to the impaired power.
Consequences of imprecise VA estimates

The more variable VA estimates (most often related to teachers with small sample sizes) result in inappropriate evaluation decisions, e.g. merit-based pay and tenure decision.

- When the decision is based on point estimates of VA, teachers with imprecise VA estimates have an artificial advantage to be recognized due to the variability.
  - E.g., a decision rule to award teachers with VA estimates greater than .5 will recognize almost exclusively teachers having fewer than 15~20 students in the previous example.

- When the decision is based on a statistical test or a similar measure, teachers with imprecise VA estimates have a considerable disadvantage due to the impaired power.
Consequences of imprecise VA estimates

The more variable VA estimates (most often related to teachers with small sample sizes) result in inappropriate evaluation decisions, e.g. merit-based pay and tenure decision.

- When the decision is based on point estimates of VA, teachers with imprecise VA estimates have an artificial advantage to be recognized due to the variability.
  - E.g., a decision rule to award teachers with VA estimates greater than .5 will recognize almost exclusively teachers having fewer than 15~20 students in the previous example.

- When the decision is based on a statistical test or a similar measure, teachers with imprecise VA estimates have a considerable disadvantage due to the impaired power.
Small-area estimation (SAE)

SAE is a technique originally in the domain of survey data analysis.

- Goal: estimate an area-level quantity of interest $\theta$
- Direct estimator $\tilde{\theta}$ for an area: estimate based on samples from the area. Due to small sample sizes in an area, $\tilde{\theta}$ is imprecise.
- Synthetic estimator $\bar{\theta}$ for an area: based on pooled data from areas with similar auxiliary characteristics. $\bar{\theta}$ is precise but lacks area-level accuracy.
- Composite estimator $\hat{\theta}$ for an area: a weighted average between $\tilde{\theta}$ and $\bar{\theta}$ where weights are chosen to minimize the mean squared prediction error (MSE), i.e., an optimal estimator balancing between precision and accuracy.
Small-area estimation (SAE)

SAE is a technique originally in the domain of survey data analysis.

- **Goal**: estimate an area-level quantity of interest $\theta$
- **Direct estimator** $\tilde{\theta}$ for an area: estimate based on samples from the area. Due to small sample sizes in an area, $\tilde{\theta}$ is imprecise.
- **Synthetic estimator** $\bar{\theta}$ for an area: based on pooled data from areas with similar auxiliary characteristics. $\bar{\theta}$ is precise but lacks area-level accuracy.
- **Composite estimator** $\hat{\theta}$ for an area: a weighted average between $\tilde{\theta}$ and $\bar{\theta}$ where weights are chosen to minimize the mean squared prediction error (MSE), i.e., an optimal estimator balancing between precision and accuracy.
Analogy between SAE and teacher VA estimate

- True teacher VA is the area-level quantity of interest $\theta$ (teacher as area).
- Direct estimate $\tilde{\theta}$ is the VA estimate based on (test scores of) a teacher’s own students (students as survey samples).
- Teachers sharing common characteristics may have similar true VA. Such characteristics may include but not are not limited to: past VA estimates, qualification, attitude, and experience.
- Composite estimates can be constructed using the relevant teacher characteristics.
Technical difficulties in using teacher characteristics by the current VA models (VAM)

- Mixed-effect VAM:
  - Teacher effects are random effects in a large linear mixed model;
  - Possible to directly include teacher chars as additional fixed effects;
  - Composite estimator would be the EBLUP for the linear combination of teacher random effects plus fixed effects of auxiliary variables.

- Fixed-effect VAM:
  - Teacher effects are fixed effects (indicators) in a large ANCOVA model;
  - Usually impossible to utilize teacher-level characteristics due to confounding with teacher indicators.
Technical difficulties in using teacher characteristics by the current VA models (VAM)

- Mixed-effect VAM:
  - Teacher effects are random effects in a large linear mixed model;
  - Possible to directly include teacher chars as additional fixed effects;
  - Composite estimator would be the EBLUP for the linear combination of teacher random effects plus fixed effects of auxiliary variables.

- Fixed-effect VAM:
  - Teacher effects are fixed effects (indicators) in a large ANCOVA model;
  - Usually impossible to utilize teacher-level characteristics due to confounding with teacher indicators.
Technical difficulties in using teacher chars by the current VAM cont’d

- Residual-type VAM (Kane & Staiger, 2008):
  - This VAM is the same as the previous two but without teacher effects in the model (hence possibly eliminating the difference between the previous two approaches).
  - Teachers effect estimate is the summary statistic (mean/median) of residuals belonging to a teacher.
  - Difficult to utilize teacher-level characteristics due to the lack of explicit teacher effects in the models.

- Model-free VA estimates:
  - Simple gain score or similar ideas;
  - Student growth curve percentile (quantile regression) (Betebenner, 2008)
  - Difficult to utilize teacher-level characteristics due to the lack of any explicit parameterization
Technical difficulties in using teacher chars by the current VAM cont’d

- Residual-type VAM (Kane & Staiger, 2008):
  - This VAM is the same as the previous two but without teacher effects in the model (hence possibly eliminating the difference between the previous two approaches).
  - Teachers effect estimate is the summary statistic (mean/median) of residuals belonging to a teacher.
  - Difficult to utilize teacher-level characteristics due to the lack of explicit teacher effects in the models.

- Model-free VA estimates:
  - Simple gain score or similar ideas;
  - Student growth curve percentile (quantile regression) (Betebenner, 2008)
  - Difficult to utilize teacher-level characteristics due to the lack of any explicit parameterization
A two-stage modeling approach

▶ The only VAM that easily permits the inclusion of teacher characteristics is the mixed-effect approach, which is also used perhaps the least in practice.

▶ In practice, many states or districts are receiving VA estimates from third-party vendors. The technical details of VAM used by vendors may not be well disclosed or readily modifiable.

▶ We aim to develop a flexible approach that can be applied to most VAM approaches being used. We propose a two-stage procedure:

▶ Stage 1: calculate VA estimates from an existing VAM. Treat the raw VA estimates as direct estimates $\tilde{\theta}$.

▶ Stage 2: fit an area-level SAE model by regressing $\tilde{\theta}$ on teacher-level characteristics, based on which we construct the composite estimate $\hat{\theta}$. 
A two-stage modeling approach

- The only VAM that easily permits the inclusion of teacher characteristics is the mixed-effect approach, which is also used perhaps the least in practice.

- In practice, many states or districts are receiving VA estimates from third-party vendors. The technical details of VAM used by vendors may not be well disclosed or readily modifiable.

- We aim to develop a flexible approach that can be applied to most VAM approaches being used. We propose a two-stage procedure:
  
  - Stage 1: calculate VA estimates from an existing VAM. Treat the raw VA estimates as direct estimates $\tilde{\theta}$.
  
  - Stage 2: fit an area-level SAE model by regressing $\tilde{\theta}$ on teacher-level characteristics, based on which we construct the composite estimate $\hat{\theta}$. 
The Fay-Herriot model (F-H model)

The F-H model is given by (Fay & Herriot, 1979)

\[ \tilde{\theta}_i = Y_i = X_i \beta + a_i + e_i, \quad i = 1 \ldots m, \]  

(1)

where \( X_i \) are the area-level auxiliary characteristics, \( a_i \sim \text{i.i.d.} \ N(0, \tau^2) \), \( e_i \sim \text{i.i.d.} \ N(0, \sigma_i^2) \), and \( a_i \) and \( e_i \) are all indep. \( \sigma_i^2 \) is the sampling variance of \( \tilde{\theta}_i \), i.e., estimation variance of the \( i \)th raw VA estimate, which is assumed to be known.

The F-H model is a predictive tool, rather than an explanatory model.

The F-H model is a special heteroskedastic mixed-effect model, which has a single measurement on each subject and known (and varying) error variances.
The Fay-Herriot model (F-H model)

The F-H model is given by (Fay & Herriot, 1979)

\[ \tilde{\theta}_i = Y_i = X_i \beta + a_i + e_i, \quad i = 1 \ldots m, \]  

where \( X_i \) are the area-level auxiliary characteristics, \( a_i \sim \text{i.i.d.} \, N(0, \tau^2) \), \( e_i \sim \text{i.i.d.} \, N(0, \sigma^2_i) \), and \( a_i \) and \( e_i \) are all indep. \( \sigma^2_i \) is the sampling variance of \( \tilde{\theta}_i \), i.e., estimation variance of the \( i \)th raw VA estimate, which is assumed to be known.

The F-H model is a predictive tool, rather than an explanatory model.

The F-H model is a special heteroskedastic mixed-effect model, which has a single measurement on each subject and known (and varying) error variances.
SAE under the F-H model

Under (1),

▶ the true VA is \( \theta_i = X_i \beta + a_i \);

▶ the synthetic estimator is \( \bar{\theta}_i = X_i \hat{\beta} \): high precision but less accurate;

▶ the sensible **composite estimator** \( \hat{\theta} \) is the empirical best unbiased predictor (EBLUP),

\[
\hat{\theta}_i(\hat{\tau}) = \hat{Y}_i(\hat{\tau}) = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma_i^2} Y_i + \frac{\sigma_i^2}{\hat{\tau}^2 + \sigma_i^2} X'_i \hat{\beta}.
\] (2)
Estimation of the F-H model

- Estimate the parameters:
  - Estimating $\tau^2$ (Prasad-Rao estimator, Fay-Herriot estimator, REML, MLE, and other methods)
  - Given $\hat{\tau}^2$, estimate $\beta$ by the generalized least squares method.

- Efficiency of the composite VA estimator $\hat{\theta}$
  - Note that $\hat{\theta}$ is essentially a predictor for a mixed effect.
  - The efficiency of a predictor is usually measured by its mean squared prediction error (MSE)

$$MSE(\hat{\theta}) = E_Y (\hat{\theta} - \theta)^2.$$ (3)

- $MSE(\hat{\theta})$ itself needs to be estimated. The estimator is hereafter denoted by $mse(\hat{\theta})$, which quantifies the level of improvement of $\hat{\theta}$ compared to $\bar{\theta}$
- There is a rich collection of work on the efficient $mse(\hat{\theta})$ (e.g., Lahiri & Rao, 1995; Das et al., 2004). The specific form of $mse(\hat{\theta})$ depends on the method fitting the F-H model.
Estimation of the F-H model

- Estimate the parameters:
  - Estimating $\tau^2$ (Prasad-Rao estimator, Fay-Herriot estimator, REML, MLE, and other methods)
  - Given $\hat{\tau}^2$, estimate $\beta$ by the generalized least squares method.

- Efficiency of the composite VA estimator $\hat{\theta}$
  - Note that $\hat{\theta}$ is essentially a predictor for a mixed effect.
  - The efficiency of a predictor is usually measured by its mean squared prediction error (MSE)
    
    $$MSE(\hat{\theta}) = E_Y(\hat{\theta} - \theta)^2.$$  

- $MSE(\hat{\theta})$ itself needs to be estimated. The estimator is hereafter denoted by $mse(\hat{\theta})$, which quantifies the level of improvement of $\hat{\theta}$ compared to $\bar{\theta}$

- There is a rich collection of work on the efficient $mse(\hat{\theta})$ (e.g., Lahiri & Rao, 1995; Das et al., 2004). The specific form of $mse(\hat{\theta})$ depends on the method fitting the F-H model.
Variable selection issue in F-H model

- The utility of the composite estimator depends on the strength of auxiliary teacher characteristics for predicting teacher effects.
- There is a wide list of available teacher-level covariates, including education, experience, historic performance measures, professional development, etc.
- Many auxiliary variables are only weakly associated with the direct VA estimates (Harris and Sass, 2006).
- What variables should we use in the F-H model? In the case of all weak predictors, should we just use an intercept-only F-H model?
- In practice many auxiliary variables have missing information.
Existing model selection approaches for F-H model

- Ad hoc
  - Validation sample (Fay & Herriot, 1979)
  - Combining expert opinions, residual diagnostics, classic Akaike information criterion (AIC), and hypothesis tests (Small Area Income and Poverty Estimation project, Panel of NRC, 2000)

- Other methods: model averaging (Longford, 2005), adaptive fence method (Jiang et al., 2008)
- Classic AIC and conditional AIC (Vaida and Blanchard, 2005)
Existing model selection approaches for F-H model

- Ad hoc
  - Validation sample (Fay & Herriot, 1979)
  - Combining expert opinions, residual diagnostics, classic Akaike information criterion (AIC), and hypothesis tests (Small Area Income and Poverty Estimation project, Panel of NRC, 2000)

- Other methods: model averaging (Longford, 2005), adaptive fence method (Jiang et al., 2008)

- Classic AIC and conditional AIC (Vaida and Blanchard, 2005)
Conditional AIC is a suitable tool for the F-H model

- The Fay-Herriot model is a prediction tool and the interpretation of model parameters is usually not of main research interest.

- A selection criterion for Fay-Herriot models should be based on the prediction performance.

- Akaike information, i.e., the expected Kullback-Leibler divergence between the working model and the true model, is particularly suitable for selecting F-H model.

- Classic AIC estimates the prediction performance of the synthetic estimators (of a candidate model).

- Conditional AIC estimates the prediction performance of the composite estimators (of a candidate model).

- However, the existing form of conditional AIC does not directly apply to the heteroskedastic F-H model. Han (2012) derived a series of theoretical results, which provide closed-form conditional AIC for the F-H model fitted by several popular estimators (Prasad-Rao, REML, and MLE).
Conditional AIC is a suitable tool for the F-H model

- The Fay-Herriot model is a prediction tool and the interpretation of model parameters is usually not of main research interest.
- A selection criterion for Fay-Herriot models should be based on the prediction performance.
- Akaike information, i.e., the expected Kullback-Leibler divergence between the working model and the true model, is particularly suitable for selecting F-H model.
- Classic AIC estimates the prediction performance of the synthetic estimators (of a candidate model).
- Conditional AIC estimates the prediction performance of the composite estimators (of a candidate model).
- However, the existing form of conditional AIC does not directly apply to the heteroskedastic F-H model. Han (2012) derived a series of theoretical results, which provide closed-form conditional AIC for the F-H model fitted by several popular estimators (Prasad-Rao, REML, and MLE).
Case study

- Data source: from a large urban school district
- Sample: roughly 27,000 elementary and middle school students; excluded special education and alternative schools but included magnet schools. Grades 4 to 8.
- Student demographics: 50% African-American, 36% white, 11% Hispanic, 3% Asian or other ethnic groups.
- Subject: math
- Test scores (outcome of VAM): from spring of 2008 and prior achievement scores from 2007, 2006 and 2005. (4th grade had only one year of prior testing and 5th grade had only two years of prior testing). Ranks of scale scores within grade transformed by the inverse cdf of normal distribution.
- Teachers: 725 have direct VA estimates
Direct VA estimates for math teachers

- We used the multivariate analysis of covariance (MANCOVA) method (McCaffrey et al., 2009). This is a fixed-effect approach which cannot utilize teacher-level auxiliary variables due to confounding.
- A linear model with effects for individual teachers and student prior achievement test scores and demographic variables.
- The teacher effects are parameterized to sum to zero within grade level (4 to 8).
- A pattern mixture approach for missing information in student variables.
Composite VA estimates for math teachers

- Synthetic estimates: a preliminary exploratory analysis suggests four potentially useful predictors.
  - Teaching experience (categorized as 0, 1, 2, 2+ years)
  - Qualification (BA degree)
  - Days absent
  - Prior year VA estimate

- Pattern mixture approach for the two categorical variables (experience and qualification)

- 508 math teachers have sufficient auxiliary variables.

- Conduct model selection for 16 candidate F-H models by conditional AIC

- Quantify the improvement of composite estimator by \( mse(\hat{\theta}) \).
Summary of conditional AIC (cAIC, smaller is better).

Methods make a sizable difference (REML outperforms UQE).
Prior year VA has the most notable impact of cAIC.

BA degree also reduces cAIC.
Estimated MSE of the composite estimator based on the chosen model versus variance of direct value-added estimates.
Summary

- Teacher-level auxiliary variables contain useful information to boost the efficiency in VA estimates.
- The proposed 2-stage VA estimation using the F-H model in the second stage is a practically useful approach.
- We addressed several technical difficulties in the second stage, in particular, conditional AIC for variable selection.
- Future work: missing data, robust estimation, finer parametrization.
Major references


