RAND EDUCATION

Using Auxiliary Teacher Data to Improve Value-Added

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

MARCES Conference

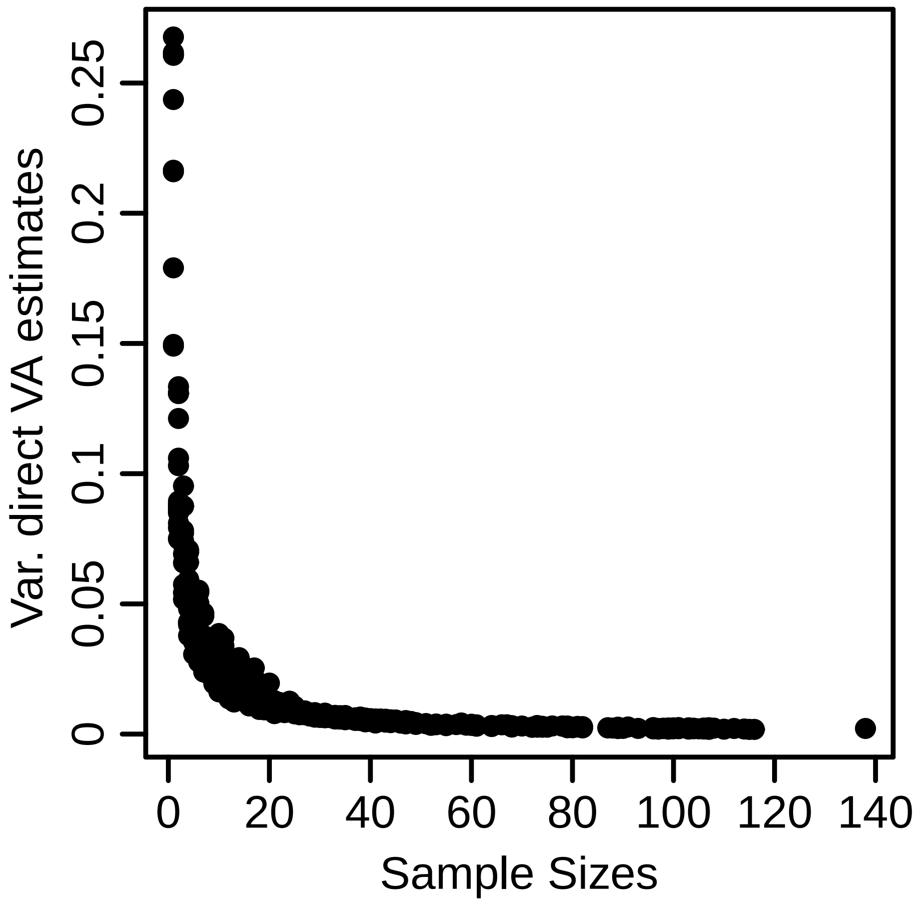
University of Maryland, October 18 - 19, 2012

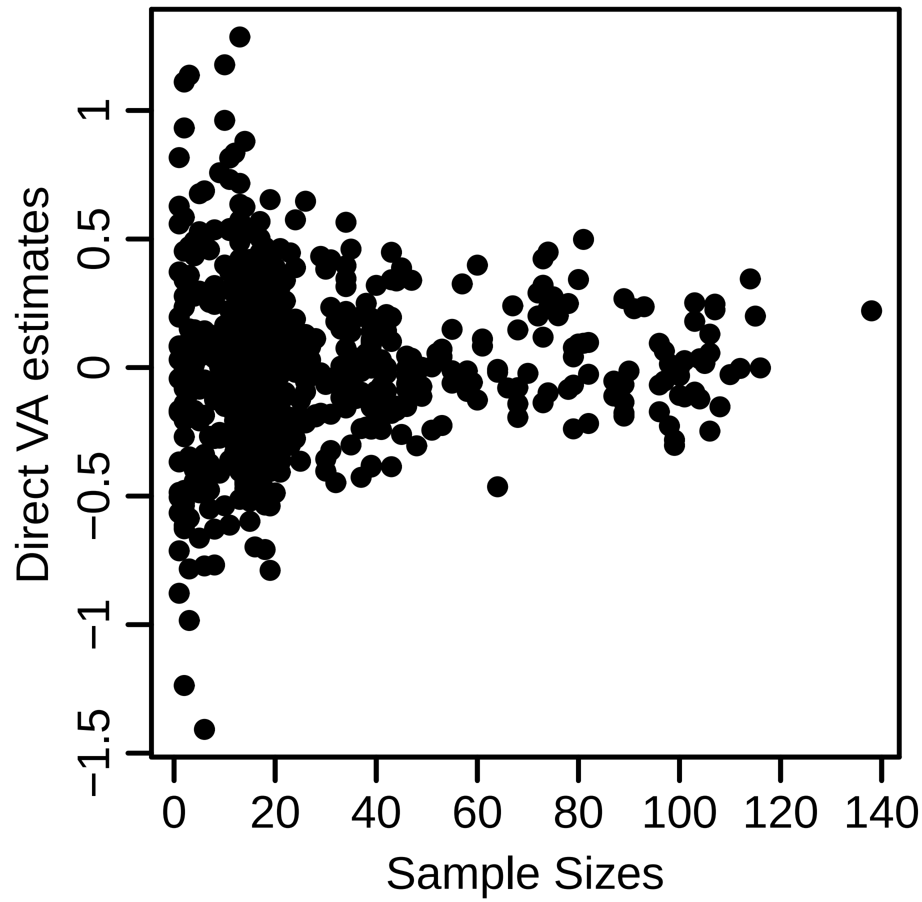
This material is based on work supported by the Department of Education Institute of Education Sciences under Grant No. R305D090011. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of these organizations.

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., 2012).

Raw VA Estimates Based on Small Classes are Highly Variable





RAND

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$\sim$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.**

Improving the Precision

- Additional data is the best way to improve the precision of the value-added estimate
- Pool multiple years of value-added
- Use other data such as teacher qualifications or other measures of teaching

What Are We Estimating?

- Teacher’s effectiveness for a given year and with a given group of students
- Teacher’s stable level of effectiveness across several years
- The quantity of interest may depend on our purpose
- We focus on the teacher’s effectiveness for a give year

How Do We Use Other Teacher Data?

- Small-area estimation (SAE)
  - Originally used in survey data analysis
  - Goal: estimate an area-level quantity of interest
- Direct estimator for an area:
  - Estimate based on samples from the area
  - Due to small sample sizes in an area, it is imprecise
- Synthetic estimator for an area:
  - Based on pooled data from areas with similar auxiliary characteristics
  - It is precise but lacks area-level accuracy
- Composite estimator for an area:
  - A weighted average between direct and synthetic estimators
  - Weights chosen to minimize the mean squared prediction error (MSE), i.e., an optimal balance between precision and bias

Analogy between SAE and teacher VA estimate

- **True teacher VA is the area-level quantity of interest (*teacher as area*)**
- **Direct estimate is the VA estimate based on test scores of a teacher's own students (*students as survey samples*)**
- **Use VA of teachers sharing common characteristics for synthetic estimator**
  - **Assumes teachers with similar observed characteristics will have similar VA**
  - **We will use a model to determine how similar VA is for teacher with the same characteristics (*model based predictions as the mean for similar teachers*)**

A three-stage modeling approach

- **Stage 1: calculate VA estimates from an existing VAM. Treat the raw VA estimates as direct estimates**
- **Stage 2: fit an area-level SAE model by regressing teacher VA on teacher-level characteristics, from which we construct the synthetic estimate**
- **Stage 3: determine the optimal weights and combine the direct and synthetic estimates into a composite**

The Fay-Herriot model (F-H model)

- **We need a model to determine the optimal weights**
- **Let θ local value and Y equal a direct estimate**

$$\begin{aligned}\theta &= X'\beta + a \\ Y &= \theta + e = X'\beta + a + e\end{aligned}$$

  - X **are area-level auxiliary characteristics**
  - $a \sim N(0, \tau^2)$, $e \sim N(0, \sigma_i^2)$, a **and** e **are independent**
  - σ^2 **is the sampling variance of** Y
    - **E.g., estimation variance of a teacher's VA**
    - **It is assumed to be known and can vary across units**
- **The F-H model is a predictive tool, not an explanatory model**
- **The F-H is a heteroskedastic mixed model, with one measure on each subject and known (and varying) error variances**

SAE and the F-H model

- **Under F-H model:**
  - **Direct estimator Y is VA estimate**
  - **True VA is $\theta = X'\beta + a$**
  - **Synthetic estimator is $\bar{\theta} = X'\hat{\beta}$**
  - **Optimal composite estimator is**

$$\hat{\theta}(\tau) = \lambda Y + (1 - \lambda) X'\beta$$

$$\lambda = \frac{\tau^2}{\tau^2 + \sigma^2} \text{ use } \hat{\lambda} = \frac{\hat{\tau}^2}{\hat{\tau}^2 + \sigma^2} \text{ and } \hat{\beta}$$

  - **The composite estimator is the empirical best unbiased predictor (EBLUP)**
  - **Estimates $E(\theta | Y, X)$**

Estimation of the F-H Model

- Estimate τ^2
  - Unbiased Quadratic Estimator (UQE)
    - Method of moments
    - Expected value of mean squared residual from regression of VA on X equals τ^2 plus a known term that depends on the sampling error (σ^2) and the Xs
    - $\hat{\tau}^2$ equals mean squared residual minus the known term
    - Resolves to the standard moment estimator, in a model with just an intercept
  - REML
    - Minimize restricted log-likelihood, downweights observations with large estimation error more than UQE
- Given $\hat{\tau}^2$, estimate β by the generalized least squares

Estimation of the MSE F-H Model Estimator

- $\hat{\theta}$ **is essentially a predictor for a "mixed" effect**
- **The efficiency of a predictor is measured by its mean squared prediction error (MSE)**

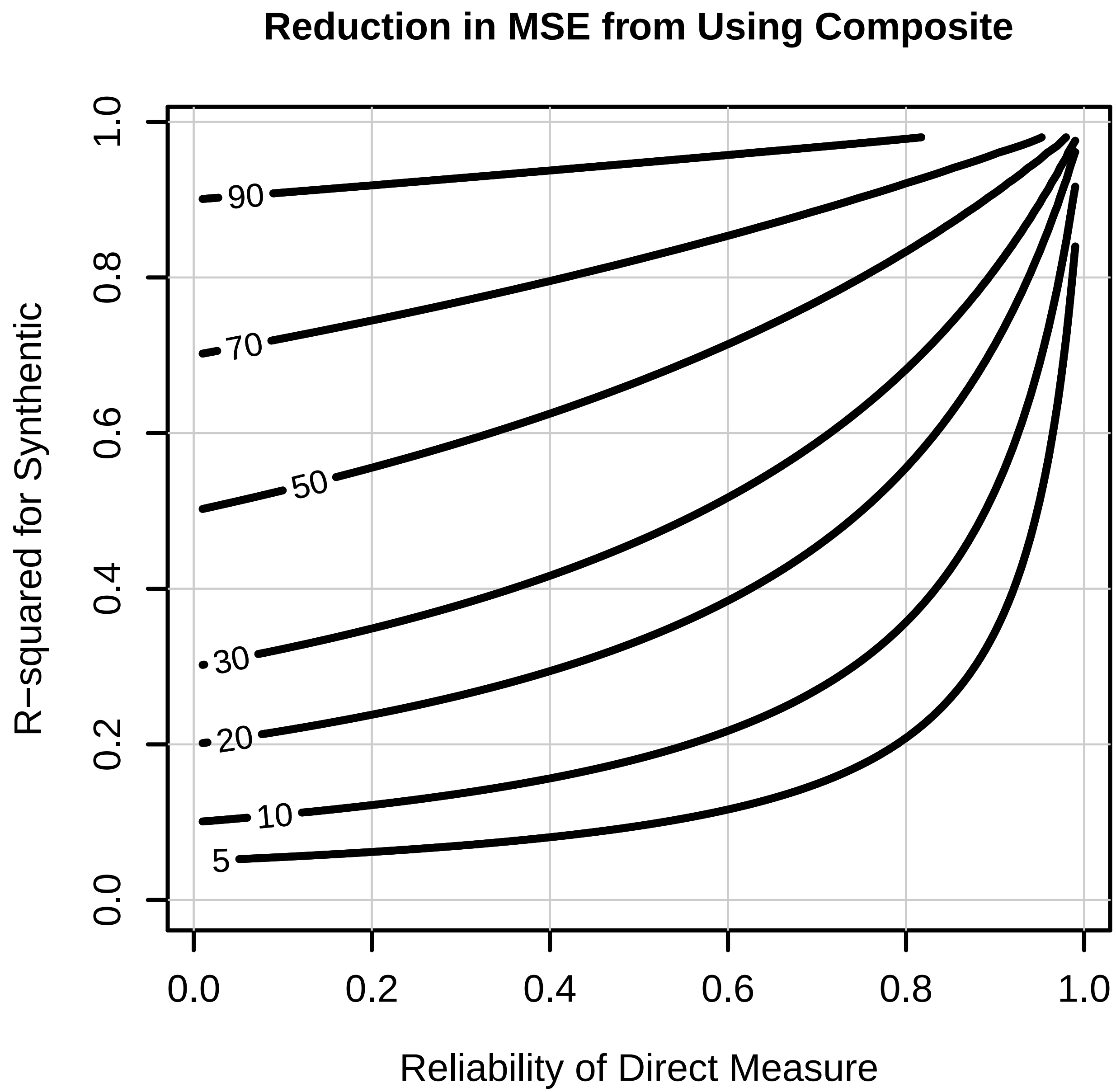
$$\begin{aligned} MSE(\hat{\theta}) &= E_Y(\hat{\theta} - \theta)^2 \\ &\approx \lambda\sigma^2 + g_2 \end{aligned}$$

- $\lambda\sigma^2$ **is the MSE of the standard shrinkage estimator**
- g_2 **accounts for estimating** β
- **Can add terms to account for estimating** τ^2
- **Use a "plug-in" estimator of MSE,** $mse(\hat{\theta})$
- **The specific form of** $mse(\hat{\theta})$ **depends on the method fitting the F-H model.**

The Plan

- Use SAE and F - H model to improve the precision VA
- Characteristics for finding similar teachers include:
  - Teacher qualifications (education, experience)
  - Teacher personal inputs (professional development, absences)
  - Prior value-added

Will It Work?



Reduction in MSE from Using Composite
R−squared for Synthetic
Reliability of Direct Measure
0.0
0.2
0.4
0.6
0.8
1.0
90
70
50
30
20
10
5

The Challenge

- Available teacher variables other than prior VA are very weak predictors of VA
- Using many weak predictors might be inefficient because of estimation error in β
- Should we use only prior VA? What about teachers without prior VA?
- Should we use no predictors? Should we just use all the available predictors?
- Can we use the data to select variables?

Modified Plan

- Use SAE and F - H model to improve the precision VA
- Avoid overfitting
  - Variable selection
  - Model averaging

Variable Selection

- Fit models with different sets of variables
- For each model calculate a fit statistic
  - Fit statistic approximates prediction error
  - Has “penalty” term for number of parameters
- AIC and BIC are common
- cAIC generalizes AIC to hierarchical model
- Fit F- H with different sets of characteristic variables
- Select the model with the lowest cAIC

Model Averaging for the F-H Model

- Selecting a set of predictors can be inefficient because of discreteness of the choice
- Model averaging can yield more accurate predictions since we average across models rather than picking one
- Appealing for F-H since we use model to predict not make inferences about Xs
- Fit F- H with different sets of characteristic variables and generate composite estimator $\hat{\theta}_m$ for each model
- The final estimate is the weighted average of the composite estimators $\hat{\theta} = \sum_m w_m \hat{\theta}_m$
- Choose the weights to minimize MSE of model average estimator

MSE for Model Averaged F-H Model

- MSE for each model is available if the model is correct but model is not correct
- The MSE of the average depends on MSE for each model and covariance among the estimators from each model
  - If correlation among estimators is 1, then MSE is maximized and so MSE is bounded by

$$MSE(\hat{\theta}) \leq \left\{ \sum_m w_m \sqrt{MSE(\hat{\theta}_m)} \right\}^2 \qquad (1)$$

Jackknife Estimator of MSE

- Use a jackknife or leave-one-out cross-validation to estimate the MSE for model averaging for a given set of weights
- Delete observation 1 from the data
- Fit a candidate model with a given set of Xs using the dataset without observation 1
- Calculate F-H estimator for observation 1, $\hat{\theta}_{1m}$
- Calculate the prediction error for F-H estimator $Y_1 - \hat{\theta}_{1m}$
- Our goal is to estimate θ_1 not Y_1, so $Y_1 - \hat{\theta}_{1m}$ overstates the error in the model so we must adjust prediction error to account for the sampling error in Y_1
- Repeat for each of the remaining observations
- The average of squared adjusted prediction errors estimates MSE

Jackknife Estimator of Model Average Weights

- Repeat jackknife MSE algorithm for all every candidate model
- Pick a set of weights $\{w*_m\}$
- Find the model average estimator for the given set of weights, $\hat{\theta}_{1,ave} = \sum_m w*_m \hat{\theta}_{1m}$
- Calculate the prediction error for model average estimator $Y_1 - \hat{\theta}_{1,ave}$
- Repeat for each of the remaining observations
- Find weights to minimize the mean of squared adjusted prediction errors

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: 752 have direct VA estimates**

Direct VA estimates for Math Teachers

- We used the multivariate analysis of covariance (MANCOVA) method (McCaffrey et al., 2009)
- A linear model for current year math scores 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.

Auxiliary Variables

- Years of experience
- Absences
- Total professional development hours and math professional development hours
- Master's degree or more
- Master's degree GPA (equal to zero if no degree), $n = 283$ (38%) missing
- Bachelor's degree GPA $n = 435$ (58%) missing
- Quality of bachelor's institution, 3 levels, $n = 424$ (56%) missing
- Bachelor's degree major in education, $n = 412$ (55%) missing
- Prior year VA, $n = 212$ (28%) missing

Missing Data Patterns

- 16 missing data patterns in the observed data
- Made bachelor's variables have consistent set of observed or missing values
  - All four variables have $n = 445$ teachers (59%) with "missing" data
- Results in 8 missing data patterns
- Six teachers are missing only Master's GPA and prior VA, mean imputed Master's GPA
- Results in 7 missing data patterns

Candidate Models

- Seven blocks of variables: 1) experience 2) absences, 3) total professional development hours and math professional development hours, 4) Master's Plus, 5) Master's GPA, 6) Bachelor's GPA, quality of Bachelor's institution, Bachelor's degree major in education, 7) prior value-added
- Considered the 128 models including or excluding each block

Missing Data and Model Selection or Averaging

Approach 1

- **Stratify the sample by missing data pattern**
- **Within each stratum use available candidate models**
- **Pick the model with lowest cAIC**
- **Estimate weights for model averaging using jackknife**

Missing Data and Model Selection or Averaging

Approach 2

- **For each potential model find all teachers with the necessary data to fit the model**
- **Fit the model and estimate prediction error using the jackknife procedure**
- **Estimate weights for model averaging using jackknife using the entire group**
- **For each pattern of observed data, identify all teachers with the observed pattern**
- **Assign these weights to all teachers with the observed data pattern**

Limited Gains from the Auxiliary Data

Strata	Direct	Intercept	cAIC	MA (1)	MA (2)
1	39	13	13	17	17
2	16	13	12	12	12
3	11	7	6	8	8
4	38	22	22	18	18
5	47	22	22	21	21
6	17	13	12	11	12
7	11	8	7	8	8
All	23	14	13	13	14

Reductions in MSE from Model Averaging

- Smallest average MSE comes from model averaging within strata defined by pattern of observed data (Approach 1)
- But method does not lead to smaller MSE in each stratum
  - Error in direct estimates differs across strata
  - Standard errors of direct estimates may be biased low because they do not account for classroom to classroom variance

JK Estimates of MSE

Strata	Direct	Intercept	cAIC	JK Min	MA (1)	MA (2)	All
1	39	21	21	21	21	21	22
2	16	9	9	9	9	9	9
3	11	6	6	5	6	6	7
4	38	24	24	24	24	24	28
5	47	25	25	25	25	25	28
6	17	11	11	10	11	10	11
7	11	6	6	6	5	5	6
All	23	13	13	13	13	13	14

MSE Relative to Intercept

Strata	cAIC	JK Min	MA (1)	MA (2)	All
1	102	99	99	100	105
2	103	99	99	99	103
3	107	98	100	100	124
4	100	100	100	100	118
5	100	100	100	100	109
6	96	95	95	95	96
7	104	99	94	94	106
All	101	99	98	98	108

MSE Relative to Intercept, Low Reliability VA

Strata	cAIC	JK Min	MA (1)	MA (2)	All
1	103	99	99	100	105
2	86	85	85	89	85
3	110	96	102	102	134
4	100	100	100	100	122
5	100	100	100	100	109
6	92	91	91	91	91
7	106	95	88	88	109
All	97	95	95	95	107

Estimating Stable Value-Added

- We estimated the teachers VA in 2008
- Stable VA may also be of interest
- Reliability of estimates of stable VA is lower because of annual variation
- Could follow same approach to estimate stable component, will obtain greater reductions in MSE

Conclusions

- Small area estimation approaches provide means for using auxiliary to reduce MSE of VA
- Gains for current year VA are likely to be small given available auxiliary data
- Gains for stable VA will be greater
- Using a large group of weak predictors was inefficient because of estimation error
- Variable selection and model averaging both can be used to choose model
  - Jackknife model averaging was less sensitive to model assumptions than using cAIC for variable selection