Value-Added Models

Matteo Paradisi
(EIEF and NBER)

Applied Micro - Lecture 7
Outline

We study two applications of fixed-effects and panel data

1. Value-added models
   • predict the impact of teachers, managers, judges

2. AKM Models
   • study the sources of wage inequality
Value-Added Metrics

- Value-added (VA) measures rate teachers based on impacts on students’ test scores
  - Long academic history: originally proposed by Hanushek (1971) and Murnane (1975)

- School districts have recently started to use VA to evaluate teachers
  - Ex: Washington DC put 50% weight on VA measures in making teacher layoff and bonus decisions under Michelle Rhee
  - Vergara v. California case on teacher tenure focused on VA measures

- More recently, VA measures developed in many other literature:
  - Health (doctors, ambulances, hospitals), Crime (police officers, judges), Neighborhoods, Public Sector Managers

- Key Challenge: Must estimate VA, leads to many econometric issues
Consider the simplest model, with random assignment.

Suppose that test scores for student $i$ with teacher $j$ are:

$$A_{ij} = \mu_j + \varepsilon_{ij}$$

Teacher-level averages are:

$$\bar{A}_j = \mu_j + \frac{\sum \varepsilon_{ij}}{N}$$

- With finite number of students, error does not disappear
- Calibration: $\sigma_\mu = 0.1$, $\sigma_\varepsilon = 0.5$, $N = 25$
  - Half of variance in $\bar{A}_j$ is from student noise.
- Intuition: Teachers with high $\bar{A}_j$ tend to be both good and lucky
Shrinkage - Intuition

\[ A_{ij} = \mu_j + \varepsilon_{ij} \]

Solution: “Shrink” estimates to account for noise.

Prediction Approach: Consider the prediction regression

\[ \bar{A}_{j,t+1} = \alpha + \beta \bar{A}_{j,t} + \varepsilon \]

• Coefficient is

\[ \beta = \frac{\text{Cov} (\bar{A}_{j,t+1}, \bar{A}_{j,t})}{\text{Var} (\bar{A}_{j,t})} = \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + \sigma^2_{\varepsilon} / N} \]

• Optimal prediction is:

\[ \hat{\mu}_j = \bar{A}_j \ast \beta = \bar{A}_j \ast \left( \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + \sigma^2_{\varepsilon} / N} \right) \]
General Empirical Framework

▶ Each teacher has only one class per year
▶ Each student attends only one class per year
▶ Teacher VA is estimated as the teacher effect $\mu_j$ from:

$$A_{ijt} = X_{ijt}\beta + \mu_j + \theta_{jt} + \epsilon_{it}$$

▶ $i$ denotes a student, $j$ a teacher and $t$ a year.
▶ $A_{ijt}$ is the student’s end-of-year test score
▶ control variables $X_{ijt}$ include student and classroom characteristics
▶ $\mu_j$ is a teacher effect (assumed to be constant over time),
▶ $\theta_{jt}$ is a classroom effect (peer effects, classroom dynamics and other effects varying from year to year)
▶ $\epsilon_{it}$ is an idiosyncratic student effect
▶ We are interested in estimating the vector of teacher effects $\mu$. 
General Empirical Framework

\[ A_{ijt} = X_{ijt}\beta + \mu_j + \theta_{jt} + \epsilon_{it} \]

- Can estimate as OLS with teacher FE
- OR, an Empirical Bayes procedure.
- **Bayesian approach intuition:** multiply a noisy estimate of teacher value added (e.g. teacher FE from OLS) by an estimate of its “reliability.”
- **Reliability:** the ratio of signal variance to signal plus noise variance
- **less reliable estimates are shrunk toward the mean of the estimates** (which is zero if the data is normalized such that teach effects have mean zero)
- This method has become particularly popular in the literature
Shrinkage: Estimation
3 Steps Procedure: Step 1

- Construct a residual $v_{ijt} = \mu_j + \theta_{jt} + \epsilon_{it}$

- Within-classroom variance in $v_{ijt}$ is used as an estimate of the variance of the student effect $\epsilon_{it}$:

  $$\hat{\sigma}_\epsilon^2 = \text{Var}(v_{ijt} - \bar{v}_{jt})$$

  where $\bar{v}_{jt}$ is the average classroom residual.

- Covariance between the average residual in a teacher’s class in year $t$ and $t-1$ used to estimate the variance in the teacher component (within-teacher cross-classrooms):

  $$\hat{\sigma}_\mu^2 = \text{Cov}(\bar{v}_{jt}, \bar{v}_{jt-1})$$

- Variance of the classroom component can be estimated as the remainder:

  $$\hat{\sigma}_\theta^2 = \text{Var}(v_{ijt}) - \hat{\sigma}_\mu^2 - \hat{\sigma}_\epsilon^2$$

- Notice how close this is to what we have done in RE models!
3 Steps Procedure: Step 2

- **Weighted (over time) average of the average classroom residuals** for each teacher to get a minimum variance unbiased estimate of $\mu$

- Weigh data from each classroom by its precision (the inverse of its variance), more weight on larger classrooms.

$$
\bar{v}_j = \sum_t w_{jt} \bar{v}_{jt}
$$

with

$$
w_{jt} = \frac{h_{jt}}{\sum_t h_{jt}}
$$

$$
h_{jt} = \frac{1}{\text{Var}(\bar{v}_{jt} | \mu_j)} = \frac{1}{\hat{\sigma}_\theta^2 + \frac{\hat{\sigma}_\epsilon^2}{n_{jt}}}
$$

- $n_{jt}$ is the number of students of teacher $j$ in year $t$

- estimate the “teacher fixed effect” with information from all years

- more weight to the years when in charge of a large class (which reduces noise from the student-level shocks).
3 Steps Procedure: Step 3

- Construct **Bayes estimator** of each teacher’s value-added by multiplying $\tilde{v}_j$ by an estimate of its reliability (Kane and Staiger, 2008):

$$\hat{\mu}_j = \tilde{v}_j \frac{\hat{\sigma}^2_\mu}{\text{Var}(\tilde{v}_j)}$$

with

$$\text{Var}(\tilde{v}_j) = \hat{\sigma}^2_\mu + \left( \sum_t h_{jt} \right)^{-1}$$

- **Shrinkage factor**: $\frac{\hat{\sigma}^2_\mu}{\text{Var}(\tilde{v}_j)} < 1$, the ratio of the signal variance to the total variance.

- $\text{Var}(\tilde{v}_j)$ given by the sum of signal variance and estimation error variance for $\tilde{v}_j$, which is $(\sum_t h_{jt})^{-1}$. 
Intuition for shrinkage, again

Consider experiment in which we estimate teacher impacts $j$ in year $t$ then randomly assign students to teachers in year $t + 1$

Consider a regression:

$$A_{ij(t+1)} = a + b\bar{v}_j + \eta_{ij(t+1)}$$

The coefficient in this regression is:

$$b = \frac{\text{cov}(A_{ij(t+1)} , \bar{v}_j)}{\text{var}(\bar{v}_j)}$$

$$= \frac{\text{cov}(X_{ij(t+1)} \beta + \mu_j + \theta_j(t+1) + \epsilon_i(t+1), \sum_t w_{jt}\bar{v}_{jt})}{\text{var}(\bar{v}_j)}$$

$$= \frac{\text{var}(\mu_j)}{\text{var}(\bar{v}_j)}$$

Optimal forecast of teacher $j$’s impact on future test scores is

$$\hat{\mu}_j = \bar{v}_j \frac{\sigma^2_{\mu}}{\text{var}(\bar{v}_j)} \quad (\text{we have } a = 0)$$
Shrinkage Interpretation

- **Frequentist perspective:** measurement error in $\tilde{v}_j$ makes it optimal to use a biased but more precise estimate of teacher quality to minimize the mean-square error of the forecast.

- **Bayesian perspective:** putting a normal prior with mean zero on $\mu$, the posterior mean of $\mu$ is a precision-weighted average of the sample mean $\tilde{v}_j$ and the mean of the prior (zero).
Shrinkage: Frequentist Interpretation

\[ A_{ijt} = X_{ijt} \beta + \mu_j + \theta_{jt} + \epsilon_{it} \]

- Assume \( \beta = 0 \) and \( \theta_{jt} = 0 \) \( \forall j, t \)
- we only have to estimate the vector \( \mu \)
- Want \( (A_{ij(t+1)} - \hat{\mu}_j)^2 \) to be minimal
- We can show that:

\[
\mathbb{E}[(A_{ij(t+1)} - \hat{\mu}_j)^2] = \mathbb{E}[(\mu_j - \mathbb{E}[\hat{\mu}_j])^2] + \mathbb{E}[(\hat{\mu}_j - \mathbb{E}[\hat{\mu}_j])^2] + \sigma^2 \epsilon
\]

bias \hspace{1cm} variance

- We willing to trade-off bias and precision
Bayesian Signal Extraction Approach:

- Prior: $\mu_j \sim N \left(0, \sigma^2_{\mu}\right)$
- Signal: $\bar{v}_j \sim N \left(\mu_j, \sigma^2_{\epsilon}/N\right)$

Posterior distribution:

$$\hat{\mu}_j = E \left[\mu_j \mid \bar{v}_{j,t}\right] = \bar{v}_j \ast \left(\frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + \sigma^2_{\epsilon}/N}\right)$$
Bias in VA
Sources of Bias

- Estimate teacher effects from OLS regression with controls:
  \[ A_{ijt} = \alpha + \delta X_{it} + \mu_j + \epsilon_{ijt} \]
  - Controls: lagged test scores, class average (peer) test scores, demographics

- Bias Source #1: **Omitted Variable Bias**: \( \text{Cov}(\mu_j, \epsilon_{ijt}) \neq 0 \)

- Bias Source #2: **Student Clustering**: \( \text{E}(\epsilon_{ijt} | j) \neq 0 \)

- Bias Source #3: **Incorrect variance structure**
  - E.g., Individual shocks are correlated within classroom.
    - Under estimate of noise variance \( \sigma^2_\epsilon \)
    - Shrinkage coefficient \( \left( \frac{\sigma^2_\mu}{\sigma^2_\mu + \sigma^2_\epsilon / N} \right) \) is too large, predictions no longer appropriately scaled
Definitions of Bias

► "Forecast" Bias: Does $\lambda = 1$ in the regression 
\[ A_{ij,t+1} = \alpha + \lambda \hat{\mu}_j + \varepsilon \] 
when teachers are randomly assigned?

► Note that random assignment ensures \( \mathbb{E}[\varepsilon | \hat{\mu}_j] = 0 \)

Therefore
\[ \lambda = \frac{\text{Cov}(\hat{\mu}_j, A_{ij,t+1})}{\text{Var}(\hat{\mu}_j)} = \frac{\text{Cov}(\hat{\mu}_j, \mu_j)}{\text{Var}(\hat{\mu}_j)} \]

► Relevant for policy since $\lambda \cdot \Delta \hat{\mu}_j$ measures the expected gains from improving teacher quality by $\Delta \hat{\mu}_j$
Definitions of Bias

▶ "Teacher-Level" Bias: Does $E[\hat{\mu}_j] = \mu_j$?

- Relevant for policy for equity concerns; are some teachers systematically undervalued?
  - E.g., are ESL teachers rated poorly since students do poorly on tests in English?

- Not relevant for expected policy gains, conditional on forecast bias

▶ Teacher-level unbiased $\implies$ forecast unbiased, but not the reverse:

- Suppose that $E[\hat{\mu}_j] = \mu_j + \zeta_j$. Forecast bias is:

  $$\lambda = \frac{\text{Cov}(\hat{\mu}_j, \mu_j)}{\text{Var}(\hat{\mu}_j)} = \frac{\text{Var}(\mu_j) + \text{Cov}(\mu_j, \zeta_j)}{\text{Var}(\mu_j) + \text{Var}(\zeta_j) + 2\text{Cov}(\mu_j, \zeta_j)}$$

- Note that $\lambda = 1$ iff $\text{Var}(\zeta_j) = -\text{Cov}(\mu_j, \zeta_j)$
Testing for Bias

How can we test for forecast bias?

- **Kane and Staiger (2008)** takes the following approach:
  - Measure $\hat{\mu}_j$ in existing school data up to year $t$
  - Randomize students to teachers in year $t + 1$
  - Regress student performance on $\hat{\mu}_j$

Estimate three-part variance structure:

$$A_{ijt} = \alpha + \delta X_{it} + \mu_j + \phi_c + \varepsilon_{ijt}$$

- Class effect $\phi_c$ captures correlated shocks within classrooms
- Shrinkage factor is

$$\beta = \frac{\sigma^2_\mu}{\sigma^2_\mu + \sigma^2_c / C + \sigma^2_\varepsilon / N}$$
## Testing for Bias (Kane and Staiger, 2008)

<table>
<thead>
<tr>
<th>Specification Used for Non-experimental Teacher Effect</th>
<th>Test Score First Year</th>
<th>Test Score Second Year</th>
<th>Test Score Third Year</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>R2</td>
<td>Coefficient</td>
</tr>
<tr>
<td>Math Levels with...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Controls</td>
<td>0.511***</td>
<td>0.185</td>
<td>0.282**</td>
</tr>
<tr>
<td>(0.108)</td>
<td></td>
<td></td>
<td>(0.107)</td>
</tr>
<tr>
<td>Student/Peer Controls (incl. prior scores)</td>
<td>0.852***</td>
<td>0.210</td>
<td>0.359*</td>
</tr>
<tr>
<td>(0.177)</td>
<td></td>
<td></td>
<td>(0.172)</td>
</tr>
<tr>
<td>Student/Peer Controls (incl. prior scores) &amp; School F.E.</td>
<td>0.905***</td>
<td>0.226</td>
<td>0.390*</td>
</tr>
<tr>
<td>(0.180)</td>
<td></td>
<td></td>
<td>(0.176)</td>
</tr>
<tr>
<td>Student Fixed Effects</td>
<td>1.859***</td>
<td>0.153</td>
<td>0.822</td>
</tr>
<tr>
<td>(0.470)</td>
<td></td>
<td></td>
<td>(0.445)</td>
</tr>
<tr>
<td>Math Gains with...</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No Controls</td>
<td>0.794***</td>
<td>0.162</td>
<td>0.342</td>
</tr>
<tr>
<td>(0.201)</td>
<td></td>
<td></td>
<td>(0.185)</td>
</tr>
<tr>
<td>Student/Peer Controls</td>
<td>0.828***</td>
<td>0.171</td>
<td>0.356</td>
</tr>
<tr>
<td>(0.207)</td>
<td></td>
<td></td>
<td>(0.191)</td>
</tr>
<tr>
<td>Student/Peer Controls &amp; School F.E.</td>
<td>0.865***</td>
<td>0.177</td>
<td>0.382</td>
</tr>
<tr>
<td>(0.213)</td>
<td></td>
<td></td>
<td>(0.200)</td>
</tr>
</tbody>
</table>
Testing for Bias (Kane and Staiger, 2008)

- Estimates cannot reject no bias ($\beta = 1$) for models with controls, but estimates are consistently low
  - Underlying problem: Very costly to randomize students to classrooms (just 78 classrooms)
  - Since replicated at slightly larger scale (1000 teachers) in 6 cities (MET Study, Kane et al. 2013), similar findings

- Two possibilities:
  - Controls fail to account for student differences
  - Variance model not sufficient: Does VA drift?

- Chetty Friedman and Rockoff (2014a) addresses these issues using quasi-experimental variation in New York City test scores
Value Added with Drift (Chetty et al, 2014a)

Allow for drift in VA, three-step procedure to estimate VA ($\hat{\mu}_{j,t+1}$)

1. Form residual test scores $A_{is}$, controlling for observables $X_{is}$
   - Regress raw test scores $A^*_{is}$ on observable student characteristics $X_{is}$, including prior test scores $A^*_{i,s-1}$

2. Regress mean class-level test score residuals in year $t$ on class-level test score residuals in years $0$ to $t - 1$:
   \[
   \bar{A}_{jt} = a + \psi_{t-1}\bar{A}_{j,t-1} + \ldots + \psi_0\bar{A}_{j0} + \varepsilon_{jt}
   \]

3. Use estimated coefficients $\psi_1, \ldots, \psi_{t-1}$ to predict VA in year $t$ based on mean test score residuals in years $1$ to $t$ for each teacher $j$:
   \[
   \hat{\mu}_{j,t} = \sum_{s=1}^{t-1} \psi_s \bar{A}_{js}
   \]
Value Added with Drift (Chetty et al, 2014a)

Two special cases:

1. Forecast VA in year using data from only year $t - s$:

$$\hat{\mu}_{jt} = r_s \bar{A}_{j,t-s}$$

where $r_s = \text{Corr}(\bar{A}_t, \bar{A}_{t-s})$ is autocorrelation at lag $s$

2. Without drift, put equal weight on all prior scores:

$$\hat{\mu}_{jt} = \bar{A}_j^{-t} \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + (\sigma^2_\theta + \sigma^2_\epsilon / n) / T}$$

- Bayesian interpretation: shrinkage based on signal-noise ratio (Kane and Staiger 2008)
Testing for Bias (Chetty et al., )
VA Impact on Long-term Outcomes

- Do teachers who raise test scores also improve students’ long-run outcomes?

- Regress long-term outcomes on teacher-level VA estimates
  - Then validate using cross-cohort switchers design

- Interpretation of these reduced-form coefficients (Todd and Wolpin 2003):
  - Impact of having better teacher, as measured by VA, for single year during grades 4-8 on earnings
  - Includes benefit of better teachers, peers, etc. in later grades via tracking
Open Questions on VA

Two key issues remain to be resolved before one can determine optimal way to use VA for policy

1. Gains may be eroded when VA is actually used (Lucas critique, Campbell’s “Law”)
   - Using VA in high-stakes evaluation could lead to teaching to the test or cheating

2. Need to compare VA to other metrics
   - Classroom observation, principal/peer evaluation, non-cognitive assessments (Jackson 2016)
   - What is the optimal weight on each measure to predict earnings impacts?

Teacher switching methodology can be used for both purposes
Many important questions remain

- Are some teachers better at teaching some types of students?
- Could resort teachers and students instead of hiring new teachers

Complementarities in production across teachers

- Are effects of having a good teacher in grade $g$ and grade $g + 1$ super-additive?
- Which grades are most important?