Putting Test Scores on the Right-Hand Side of Your Regression Model: What Works and What Doesn’t

J.R. Lockwood & Daniel F. McCaffrey
SREE
March 8, 2013

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. This talk has not been formally reviewed and should not be cited, quoted, reproduced, or retransmitted without permission.
Test Scores As Covariates

- In most observational studies of education data, “treatment” and “control” groups will differ on attributes related to achievement, and past achievement is typically the best control
  - Teacher value-added: students taught by each teacher are a different treatment group

- Adjusting for past achievement is tricky due to:
  - Ambiguous structural relationships among test scores
  - Large, heteroskedastic test score measurement error whose variance depends on unobserved achievement
Complexities of Test Scores Not Widely Respected in Secondary Analysis

- Gain scores or fixed effects (commonly after some form of z-scoring) are widespread approaches
  - Could correct for student heterogeneity in observational analyses, but the construction of tests raises concerns about the required assumptions

- Safer to think of scores as just correlated attributes
  - Prior achievement scores are proxy variables for unobservables related to current achievement scores
  - Motivates “kitchen-sink” type regressions or other conditional analysis requiring less stringent scaling assumptions
ANCOVA Modeling For Treatment Effects

\[ Y = \alpha + \beta' T + \gamma' Z + \lambda' \text{ACHIEVEMENT} + \epsilon \]

- \( Y \): Outcome variable
  - (e.g., target test score)

- \( T \): Treatment indicators with effects \( \beta \)
  - (e.g., teachers, interventions, policies, etc.)

- \( Z \): Error-free covariates
  - (e.g., demographics, program participation, etc.)

- \( \text{ACHIEVEMENT} \): Pre-treatment “achievement attributes”

- Hope is that \((Z, \text{ACHIEVEMENT})\) contains enough to make \( E[\epsilon|T] \approx 0 \)
Key Take-Away Points From Body of Recent Work

- References listed at end and noted along the way by [#]
- Focus on regression adjustment but similar issues apply to other observational approaches which have their own challenges
  - E.g. propensity scoring (today’s session and [2])
Point 1: We Often Do Not Really Want Test Scores as Covariates in an ANCOVA Model

- Observed test scores, $X$, are statistical estimates of achievement attributes
  - Typically from IRT model
  - $X$ are error-prone estimates of achievement attributes due to modest number of test items and their characteristics

- If treatment assignment is actually based on $X$, then they belong in the ANCOVA model

- But typical concern is that treatment assignment depends in part on unobserved underlying achievement attributes rather than (or in addition to) $X$
  - Assume this is the case for the remainder
Point 2: If Assignment Depends on Underlying Achievement Attributes, Simple Controls with $X$ Will Often Lead to Bias

- If treatment groups differ on underlying achievement attributes, then regression on $X$ will lead to bias

- In simple cases, treatment effect estimates are confounded with group means of pre-treatment achievement attributes
  
  - E.g., in teacher value-added estimation, teachers teaching higher-achieving students will have higher estimated effects than teachers teaching lower-achieving students
Point 3: If We Are Going to Ignore Test Score Measurement Error During an Analysis, Include All Available Tests as Covariates

- Make $X$ as “big” as possible - as many prior grades and subjects as feasible
  - Introduces missing data hassles but probably worth it
- Scores from the same academic subject as the outcome are best but other subjects can help too [3,4]
- With many tests, bias may be small [1]
  - Adding more tests scores can act like a measurement error correction
  - If all tests measure a low-dimensional vector of achievement attributes, then bias will go to zero as number of tests gets large
Point 4: With Few Tests, Corrections for Test Measurement Error Are Probably Essential

- If few prior test scores are available for inclusion in model, ignoring test measurement error risks huge biases for some treatments

- Test measurement error is ugly, making corrections complex
  - IRT error is heteroskedastic with variability dependent on unobserved achievement
  - Concerns about non-zero conditional mean [4] and short-term fluctuations (“bad day”)

- Developed improved method-of-moments and latent regression approaches for correcting for test measurement error in ANCOVA models [3]
Point 5: Even with Many Tests, Explicit Corrections for Test Measurement Error Can Be Helpful

- Appears to provide benefit even as number of scores included in the model grows [3]

- Can be a safeguard against treatment assignment depending on underlying achievement attributes captured in only one or a few specific prior tests

- E.g. if middle school mathematics tracking depends on specific skills measured by grade 7 mathematics test, including more prior tests may be helpful but can never proxy for an adjustment for measurement error in the grade 7 test
Point 6: Modeling Observed Nonlinear Relationships Between Current and Prior Year Tests is Not Generally Helpful [4]
Point 6: Modeling Observed Nonlinear Relationships Between Current and Prior Year Tests is Not Generally Helpful [4]

- Heteroskedastic measurement error contributes to nonlinear relationships among observed test scores

- Modeling those nonlinear relationships (e.g. through polynomial functions of prior scores in regression models) is a popular approach, but it is not a correction for measurement error, and can even make bias worse

- May lead to negligible differences compared to using linear adjustments for $X$ unless treatment groups are extremely selected

  ■ E.g. teachers teaching particularly low or high achieving classes
Future Work: Models with Nonlinear Relationships Among Underlying Achievement Attributes
Future Work: Models with Nonlinear Relationships Among Underlying Achievement Attributes

Complex:

1. Multiple, correlated latent achievement attributes
2. Each potentially nonlinearly related to outcome
3. Each measured with heteroskedastic error that may have ugly tail properties
4. Other covariates and possibly interactions

Speculate that degree of nonlinearity is not strong enough to matter much for estimated treatment effects unless treatment groups are extremely selected, in which case:

- Should we even be doing ANCOVA?
- Will test scores ever be able to balance these groups?
References


