When large samples act small:

The importance of small-sample adjustments for cluster-robust inference in impact evaluations

James E. Pustejovsky
UT Austin
Educational Psychology Department
Quantitative Methods Program
pusto@austin.utexas.edu

Elizabeth Tipton
Columbia University
Teachers' College
Dept. of Human Development
tipton@tc.columbia.edu

July 18, 2016
American Institutes for Research
Impact Working Group Lecture Series

Regression with dependent errors

- Cluster-randomized trials
  - Angrist & Lavy (2009) studied effects of monetary incentives on passing rates for high school exit exams in Israel

- Difference-in-differences/panel data models
  - Carpenter & Dobkin (2011) examined effects of changing minimum legal drinking age on motor vehicle fatalities.
  - Effects identified by state-level changes in drinking age over time (state-by-year panel).

- Regression discontinuity designs
  - Cortez, Goodman, & Nomi (2015) evaluate effects of double-dose algebra program on students math achievement and educational attainment.
  - Lee & Card (2008) recommend clustering on unique values of the forcing variable to address specification error.

Cluster-robust variance estimation

- Method for estimating sampling variance of regression coefficients when error structure is unknown
  - Assuming that the data includes *G* independent clusters of observations.
  - White (1984); Arellano (1987); Liang & Zeger (1986)

- Valid (asymptotically consistent) when the **number of clusters** (*G*) is large.

- But can misbehave with few clusters (Cameron & Miller, 2015; Imbens & Kolesar, 2015)
  - Standard errors that are too small
  - Hypothesis tests with inflated type-I error rates
  - And it can be hard to tell if your *G* is big enough

In brief…

- McCaffrey, Bell, & Botts (2001; Bell & McCaffrey, 2002) proposed “bias-reduced linearization” variance estimator (BRL)
  - Improves bias of standard errors for small *G*
  - t-tests with Satterthwaite degrees of freedom
- Our work:
  - Extends BRL so that it works in models with fixed effects
  - Develops an F-test for multi-parameter hypothesis tests
  - Provides easy-to-use software implementation in R
- With our extensions, BRL is a general and “production-ready” approach to cluster-robust hypothesis testing.

Today

- “standard” CRVE
- Bias-reduced linearization
  - Satterthwaite t-tests
- Our extensions
  - F-tests
  - Handling fixed effects
  - Software

The model

- Suppose we have a regression model

$$\mathbf{Y}_j = \mathbf{X}_j \boldsymbol{\beta} + \mathbf{e}_j$$

  where
  - $j = 1,\ldots,G$ clusters
  - Errors have unknown variance Var($\mathbf{e}_j$)=$\boldsymbol{\Phi}_j$ for $j = 1,\ldots,G$ clusters.

- **X** might include
  - Policy indicators
  - Demographic controls
  - Fixed effects (for clusters, time periods, etc.)

- For today, I'll assume that regression is estimated by ordinary least squares.

Hypotheses

- Our goal will be to test hypotheses about elements of **β**
  - Does an intervention have non-zero effects on the outcome?

$$H_0: \ \beta_1 = 0$$

  - Do the intervention effects vary across contexts?

$$H_0: \ \beta_1 = \cdots = \beta_q = 0$$

Standard cluster-robust variance estimation

- OLS coefficient estimates have (unknown) sampling variance

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{G}\mathbf{X}_j^t\boldsymbol{\Phi}_j\mathbf{X}_j\right)\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$$

- Standard CRVE (sandwich estimator):

$$\mathbf{V}^{CR} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{G}\mathbf{X}_j^t\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j^t\mathbf{X}_j\right)\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$$

$$\hat{\mathbf{e}}_j = \mathbf{Y}_j - \mathbf{X}_j\hat{\boldsymbol{\beta}}$$

Standard robust hypothesis tests

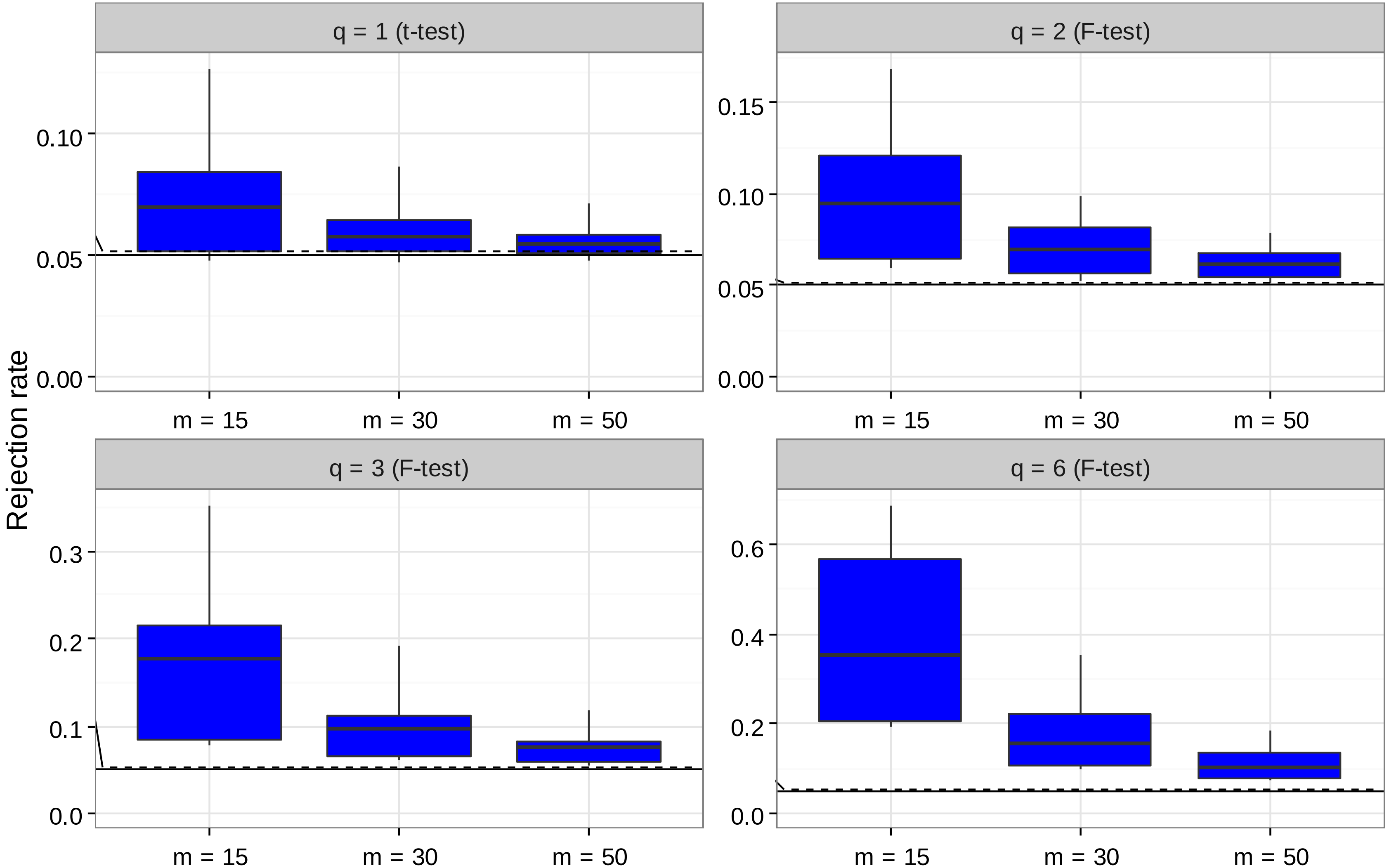
- Robust t-test (H_0: $\beta_1 = 0$)

$$t_{CR} = \hat{\beta}_1 / \sqrt{V_{11}^{CR}} \qquad t \mathrel{\dot\sim} t(G-1)$$

- Robust (Wald-type) F-test (H_0: $\mathbf{C}\beta = 0$ for $q \times p$ matrix $\mathbf{C}$)

$$F_{CR} = \frac{1}{q}\left(\mathbf{C}\hat{\boldsymbol{\beta}}\right)^t \left(\mathbf{C}\mathbf{V}^{CR}\mathbf{C}\right)^{-1}\left(\mathbf{C}\hat{\boldsymbol{\beta}}\right) \qquad F_{CR} \mathrel{\dot\sim} F\left(q, G-1\right)$$

Performance of standard tests



Bias-reduced linearization

Bias-reduced linearization

- McCaffrey, Bell, & Botts (2001) proposed a correction to $\mathbf{V}^{CR}$ based on a ***working model*** for the error covariance structure.

- Given a working model, seek a variance estimator such that

$$\mathrm{E}\left(\mathbf{V}^{BRL}\right) = \mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right)$$

- The corrected variance estimator is

$$\mathbf{V}^{BRL} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{G}\mathbf{X}_j^t\mathbf{A}_j\hat{\mathbf{e}}_j\hat{\mathbf{e}}_j^t\mathbf{A}_j^t\mathbf{X}_j\right)\left(\mathbf{X}^t\mathbf{X}\right)^{-1}$$

with adjustment matrices $\mathbf{A}_1,\ldots,\mathbf{A}_G$ chosen to satisfy BRL criterion.

Working models

- "Working independence", with $\mathbf{\Phi}_j = \mathrm{I}_j$

$$\mathbf{A}_j = \left[\mathbf{I}_j - \mathbf{X}_j \left(\mathbf{X}^t \mathbf{X} \right)^{-1} \mathbf{X}_j^t \right]^{-1/2}$$

- "Working random effects model" assumes

$$\mathbf{\Phi}_j = \rho 1_j 1_j^t + (1-\rho)\mathbf{I}_j$$

- Remarkably, the working model doesn't matter much.
  - BRL greatly reduces bias even if the working model is far from the truth.

Hypothesis tests

- We could use V^{BRL} in robust t and F statistics, but…
  - Bias of variance estimator is only part of the problem
  - t(G-1), F(q, G – 1) often poor approximations for reference distributions

- For t-tests, Bell and McCaffrey (2002) propose to use t(*v*) reference distribution, with Satterthwaite degrees of freedom

$$v = \left[\mathrm{E}\left(V_{11}^{BRL}\right)\right]^2 / \mathrm{Var}\left(V_{11}^{BRL}\right)$$

with expectation and variance estimated based on the working model.

Pustejovsky & Tipton (2016) addresses three outstanding problems with BRL

- BRL adjustments in models with lots of fixed effects
- Testing multi-parameter hypotheses
- Software availability

Handling fixed effects models

- Consider state-by-year panel data model

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + \gamma_i + \zeta_t + e_{it}$$

  - Common to treat γ_i, ζ_t as fixed effects, estimate $\boldsymbol{\beta}$ by OLS.
  - Use CRVE to allow for further correlation among errors within each state.

- BRL breaks in this model (Angrist & Pischke, 2009; Young, 2016).

- We demonstrate that the ***Moore-Penrose generalized inverse*** can be used to construct adjustment matrices that are still unbiased under the working model.

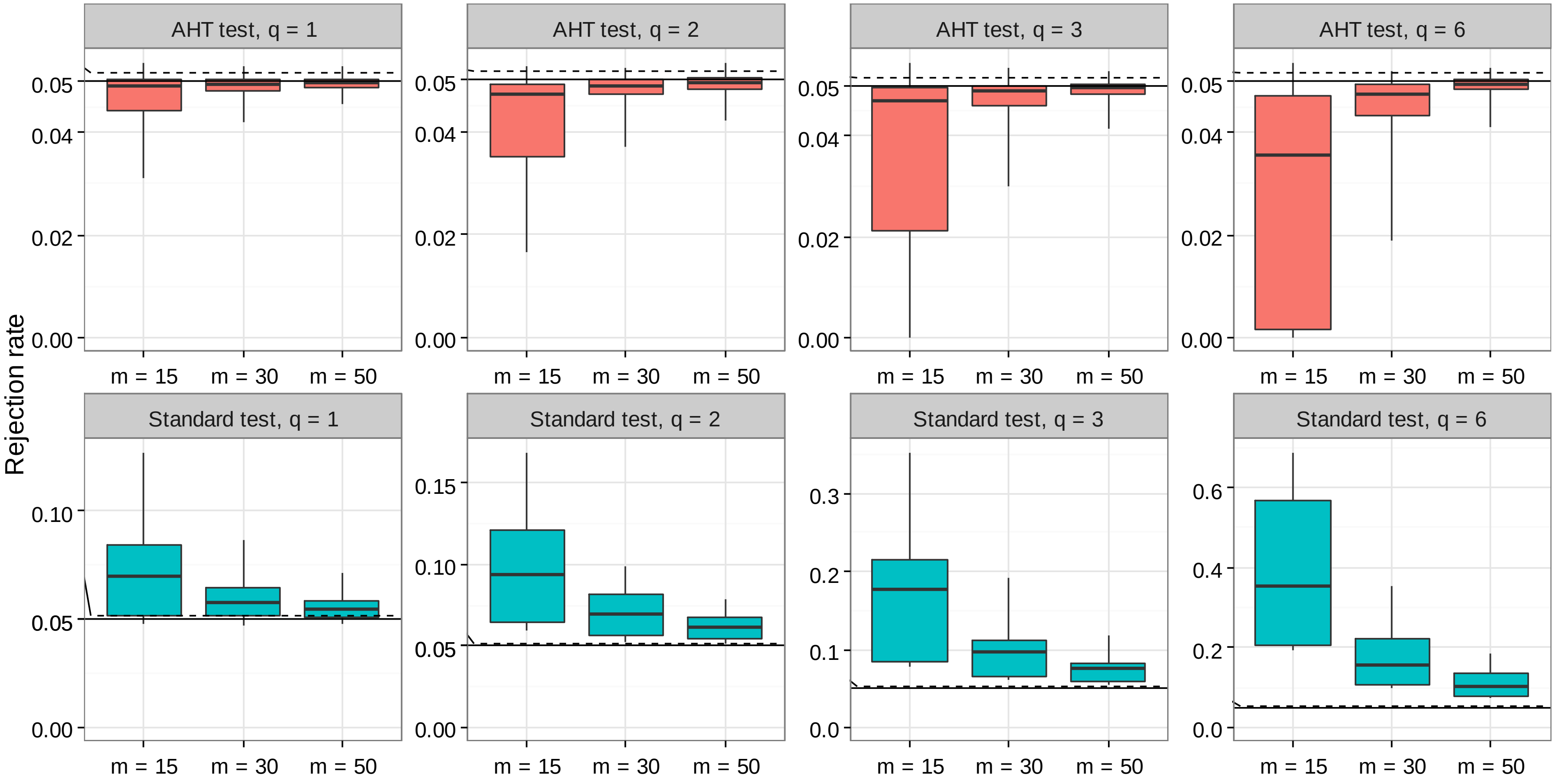
Approximate Hotelling Test

- We propose a generalization of the Satterthwaite approximation to the multi-dimensional case.
- Approximate the distribution of V^{BRL} using a Wishart distribution with degrees of freedom η and I_q scale matrix.
- Estimate η by matching mean and **total variation** of V^{BRL}.

$$F_{AHT} = \frac{\eta - q + 1}{\eta q} \left(\mathbf{C}\hat{\boldsymbol{\beta}}\right)^t \left(\mathbf{C}\mathbf{V}^{BRL}\mathbf{C}\right)^{-1} \left(\mathbf{C}\hat{\boldsymbol{\beta}}\right)$$

$$F_{AHT} \dot{\sim} F\left(q, \eta - q + 1\right)$$

AHT maintains close-to-nominal α



Software

- R package `clubSandwich`
  - https://github.com/jepusto/clubSandwich
  - Currently under active development
  - Goal is to release to CRAN by 8/1

- Works with a wide variety of fitted models
  - lm models: Ordinary/weighted least squares
  - `plm` package: Fixed-effects/random-effects panel models
  - `nlme` package: GLS and HLM models
  - Meta-analysis (`metafor` and `robumeta` packages)
  - Other packages that would be useful?

Angrist & Lavy (2009)

- Cluster-randomized trial in 40 high schools in Israel.
- Tested effects of monetary incentives on post-secondary matriculation exam (Bagrut) completion rates.
- Longitudinal data, D-in-D specification.
- Focus on effects for higher-achieving girls

Hypothesis	Test	F	df	p-value
treatment effect (q = 1)	Standard	5.746	34.00	.022
	Satterthwaite	5.169	18.13	.035
Moderation by school sector (q = 2)	Standard	3.186	34.00	.054
	AHT	1.665	7.84	.250

Carpenter & Dobkin (2011)

- Study effects of changing minimum legal drinking age on motor vehicle mortality
- State-by-year panel from FARS maintained by NHTSA.
- Difference-in-differences identification.

Hypothesis	Test	F	df	p-value
Policy effect (q = 1)	Standard	9.660	49.00	.003
	Satterthwaite	9.116	24.58	.006
Hausman test of endogeneity (q = 2)	Standard	2.930	49.00	.063
	AHT	2.560	11.91	.119

Conclusions

- Standard tests based on CRVE do not perform well with few or even a moderate number of clusters.
- It can be difficult to tell whether you have enough clusters to trust standard methods because it depends on
  - The hypothesis being tested.
  - The structure of the covariates in the model.
- Satterthwaite t-test/AHT F-test perform well across a broad range of applications. We recommend that they be ***used by default***.

Thank you

- pusto@austin.utexas.edu
- http://jepusto.github.io/
- Working paper available at http://arxiv.org/abs/1601.01981

References

- Angrist, J. D., & Lavy, V. (2009). The effects of high stakes high school achievement awards : Evidence from a randomized trial. *American Economic Review*, *99*(4), 1384–1414. doi:10.1257/aer.99.4.1384
- Angrist, J. D., & Pischke, J. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton, NJ: Princeton University Press.
- Arellano, M. (1987). Computing robust standard errors for within-groups estimators. *Oxford Bulletin of Economics and Statistics*, *49*(4), 431–434.
- Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–181.
- Cameron, A. C., & Miller, D. L. (2015). *A practitioner's guide to cluster-robust inference*.
- Carpenter, C., & Dobkin, C. (2011). The minimum legal drinking age and public health. *Journal of Economic Perspectives*, *25*(2), 133–156. doi:10.1257/jep.25.2.133
- Cortes, K. E., Goodman, J. S., & Nomi, T. (2015). Intensive math instruction and educational attainment: Long-run impacts of double-dose algebra. *Journal of Human Resources, 50*(1), 108–158. doi:10.3386/w20211
- Imbens, G. W., & Kolesar, M. (2015). *Robust standard errors in small samples: Some practical advice*. Retrieved from https://www.princeton.edu/~mkolesar/papers/small-robust.pdf
- Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674. doi:10.1016/j.jeconom.2007.05.003
- Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.
- McCaffrey, D. F., Bell, R. M., & Botts, C. H. (2001). Generalizations of biased reduced linearization. In *Proceedings of the Annual Meeting of the American Statistical Association*.
- Young, A. (2016). *Improved, nearly exact, statistical inference with robust and clustered covariance matrices using effective degrees of freedom corrections*.

Future work

- Compare BRL + AHT to other recent proposals
  - Cluster-wild bootstrap (Webb & MacKinnon, 2013)
  - Re-weighted, containment t-test (Imbragimov & Muller, 2015)

- Application to more complex models
  - Instrumental variables
  - Cross-classified/multiple-membership models

- Software
  - clubSandwich R package under active development (https://github.com/jepusto/clubSandwich)
  - Need to implement in Stata (Wanna help?)

Degrees of freedom (η)

- For single-dimensional tests, $\eta = v$ (Satterthwaite df).
- Degrees of freedom are diagnostic.
  - large η indicates large effective sample size
  - small η (i.e., much less than $G - 1$) indicates that you've got small-sample problems.
- Degrees of freedom capture the influence of covariates on the distribution of $\mathbf{V}^{BRL}$
  - Unbalanced covariates
  - Skewed/leveraged covariates
  - Unequal cluster sizes

Handling fixed effects models

- Two ways to calculate OLS estimates in fixed effects models:
  - Use dummy variables, estimate the full regression.
  - Absorb the fixed effects, estimate only the remaining coefficents.
- BRL gives different results depending on which design matrix you use to calculate $\mathbf{A}_1,..,\mathbf{A}_G$.
- We identify conditions where it is okay to use the absorbed design matrix to calculate $\mathbf{A}_1,..,\mathbf{A}_G$.
  - With OLS estimation, it's okay if you are using a working identity model.
  - Absorb the within-cluster fixed effects only.