When large samples act small:
Cluster-robust variance estimation and hypothesis testing with few clusters

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

February 19, 2016

Regression with dependent errors

- Analysis of multi-stage sample surveys
  - Blanchard & Muller (2015) use ELS:2002 to study the influence of teachers' perceptions of immigrant/language-minority students on student academic outcomes.
  - Cavanagh, Schiller, & Riegle-Crumb (2006) use Add Health to study the relationship between family structure and adolescents' academic status.
- Cluster-randomized trials
  - Burde & Linden (2012) studied effects of village-based schools in Afghanistan by randomizing 31 villages, surveying families.
- Longitudinal panel data
  - Abrevaya & Puzzello (2012) examined effects of cigarette taxes on consumption, nicotine intake, and smoking intensity using NHANES III.
  - Effects identified by state-level changes in tax rates over time. Data include 26 states.

Cluster-robust variance estimation

- A way to estimate 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) proposed “bias-reduced linearization” (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
  - Demonstrates that BRL outperforms standard CRVE across a wide range of contexts

- 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
- How to make your SEs smaller
- Further work

Errrummmm….actually….Your standard errors are too small and your p-values are all WAY too significant.

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):

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

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

Standard robust hypothesis tests

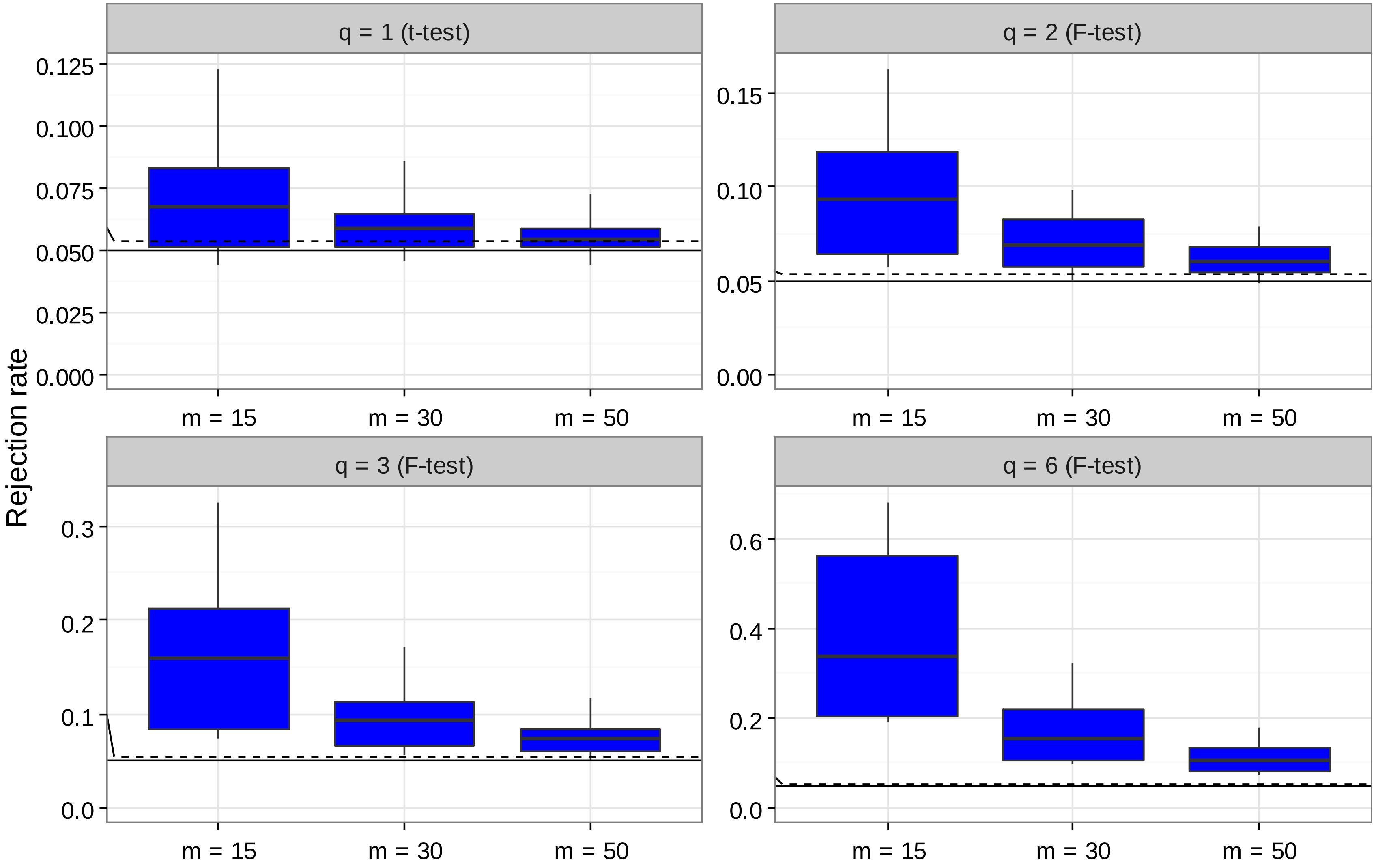
- Robust t-test (H_0: $\beta_1 = 0$)

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

- Robust (Wald-type) F-test (H_0: **C**β = 0 for $q \times p$ matrix **C**)

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

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 effect model" assumes

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

- Doesn't this contradict goal of being robust?
- Remarkably, the working model doesn't matter much.
  - BRL greatly reduces bias even if the working model is far from the truth.

Hypothesis tests

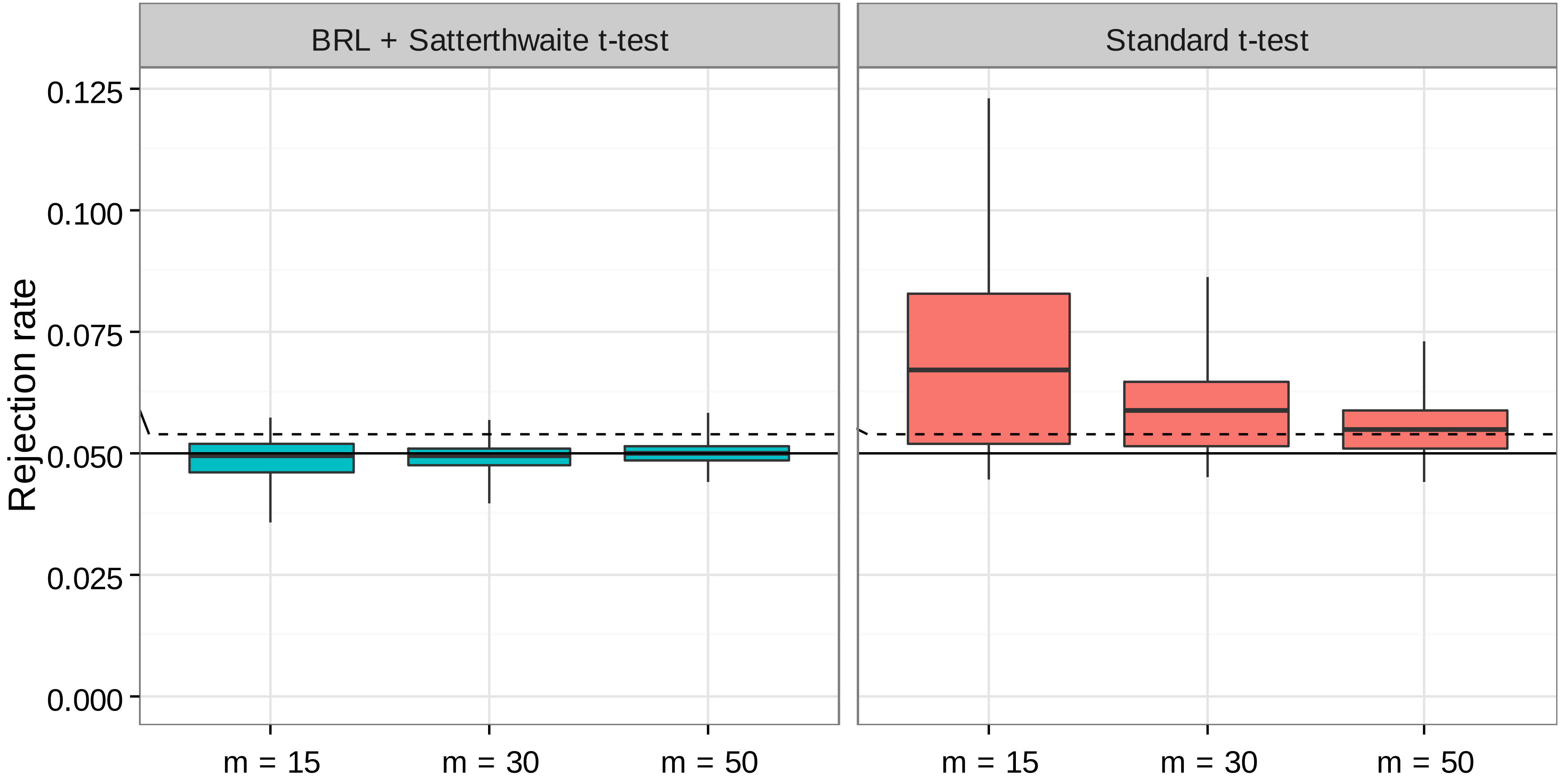
- We could use V^{BRL} in robust t and F statistics, but…
  - Bias of variance estimator is only part of the problem
  - t(G-1), F(q, G – 1) often poor approximations for reference distributions

- 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 moments estimated based on the working model.

BRL + Satterthwaite t-tests work well



Outstanding problems with BRL

1. How do you do test multi-parameter hypotheses?
2. BRL adjustment matrices are sometimes undefined in models with lots of fixed effects.
3. In models with fixed effects, BRL adjustments depends on how you calculate the coefficient estimates.

Our work

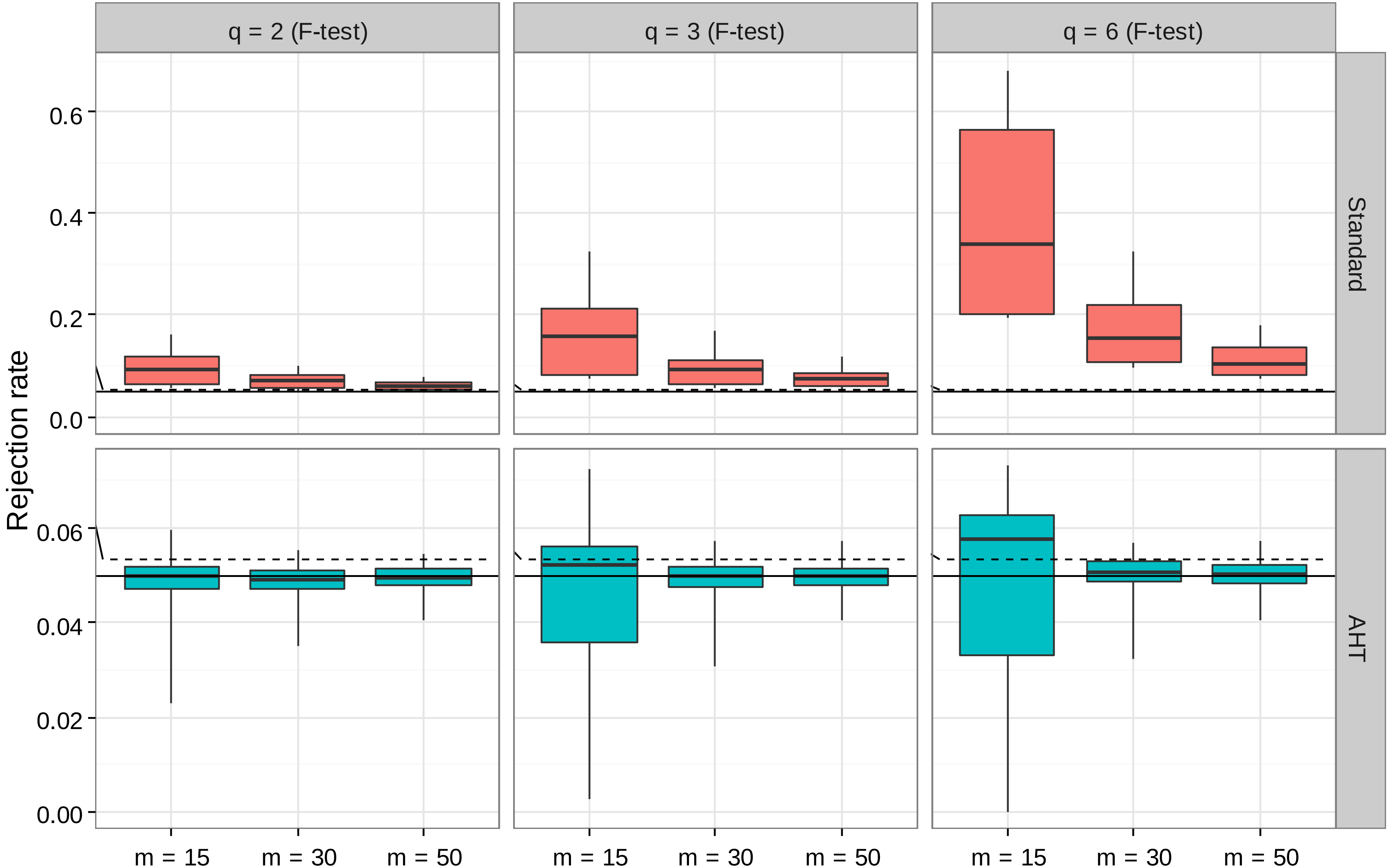
Approximate Hotelling Test

- 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}.

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

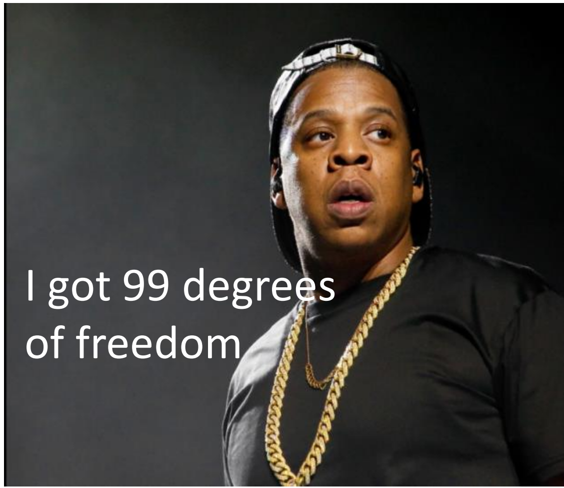
$$F_{AHT} \dot{\sim} F\left(q, \eta - q + 1\right)$$

AHT maintains close-to-nominal α



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

- 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 down in this model (Angrist & Pischke, 2009).
  - Adjustment matrices are not calculable because of rank-deficiency.

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

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.

But does this matter in practice?

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	24.74	.006
Hausman test of endogeneity (q = 2)	Standard	2.930	49.00	.063
	AHT	2.489	8.69	.140

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, diff-in-diff 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	15.86	.037
Moderation by school sector (q = 2)	Standard	3.186	34.00	.054
	AHT	0.091	3.19	.915

How to make your SEs smaller

Hierarchical linear modeling

- Develop "working" hierarchical models.
- Use estimated error structures for weighted least squares (WLS) estimation.
- Use BRL standard errors + AHT degrees of freedom
  - Based on the same working model as for WLS.
  - Adjustment matrices get a little more complicated, but it all works.

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***.

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?)

Thank you

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