Small sample adjustments to F-tests for cluster robust standard errors

Elizabeth Tipton
Teachers College, Columbia University

January 27, 2016
Presented at NYU PRIISM

Background

The topic today is a new direction for me.

It grew out of prior work on “robust variance estimation” in meta-analysis.

I also do work developing methods for making generalizations from experiments.

Motivation

Econometric data often exhibits dependence, particularly in education contexts.

For example, nesting by:

- Schools
- Time points
- Assignment variable (in RDD)

Standard practice is **cluster robust standard errors:**

- Relies on the CLT, though the number of clusters in finite samples are often small/moderate;
- Has become recently scrutinized, e.g., Imbens & Kolesar (2015), Cameron & Miller (2015).

Overview

Joint work with James Pustejovsky (at UT-Austin).

1) Cluster robust standard errors
2) Bias reduced linearization
3) New results, with focus on F-test
4) Examples

Overview of CRVE

Model

Let's say you have a regression model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Note here that **X** might include:

- Policy variables
- Demographic controls
- Fixed effects (for clusters, for time, etc).

We can estimate **β** using OLS,

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

Hypothesis testing

You may want to test hypotheses regarding elements of $\boldsymbol{\beta}$.

For example:

1. Does *Policy A* improve student outcomes?

   H_0: $\beta_1 = 0$

   $t = b_1/se(b_1)$

Hypothesis testing

You may want to test hypotheses regarding elements of $\boldsymbol{\beta}$.

For example:

1. Does *Policy A* improve student outcomes?

   H_0: $\beta_1 = 0$

   $t = b_1/se(b_1)$

2. Do student outcomes vary *across* policies?

   H_0: $\beta_1 = \beta_2 = 0$

   $F = (\mathbf{b}_{12} - \mathbf{0})[v(\mathbf{b})_{12}]^{-1}(\mathbf{b}_{12} - \mathbf{0})/2$

Clustered standard errors

How do we estimate SE(b_1) and V($\mathbf{b}$)?

The exact variance of $\mathbf{b}$ can be written: $$V(\mathbf{b}) = (\mathbf{X'X})^{-1} \sum_{j=1}^{m} \mathbf{X}_j{'}\mathbf{\Sigma}_j\mathbf{X}_j\,(\mathbf{X'X})^{-1}$$

Assume:

- Observations across clusters are independent; and
- For clusters $j = 1\ldots m$, V($\boldsymbol{\varepsilon}_j|\mathbf{X}_j$) = $\mathbf{\Sigma}_j$.

In standard CRVE, V($\mathbf{b}$) is estimated: $$v(\mathbf{b}) = (\mathbf{X'X})^{-1} \sum_{j=1}^{m} \mathbf{X}_j{'}\mathbf{e}_j\mathbf{e}_j{'}\mathbf{X}_j\,(\mathbf{X'X})^{-1}$$

Where for clusters $j = 1\ldots m$, $\mathbf{e_j} = (\mathbf{Y}_j - \mathbf{X}_j\mathbf{b})$.

Reference distributions

Returning to the examples:

1. Under H_0, assume that

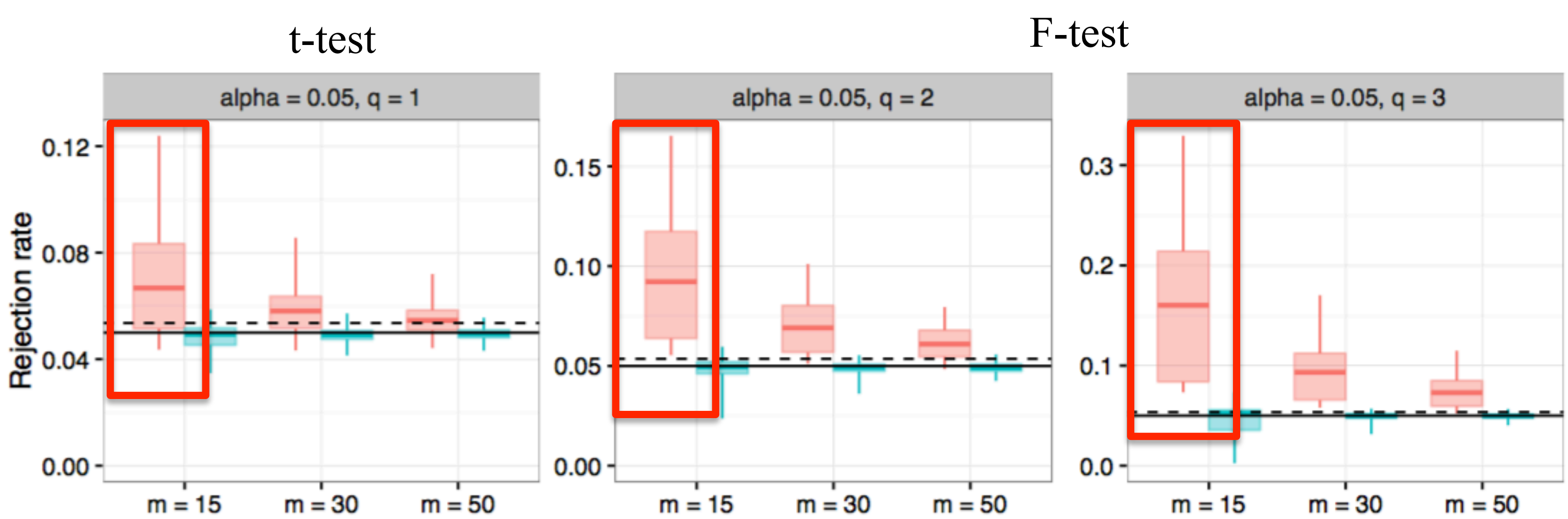
$$t \sim t(m - 1)$$

2. Under H_0, assume that

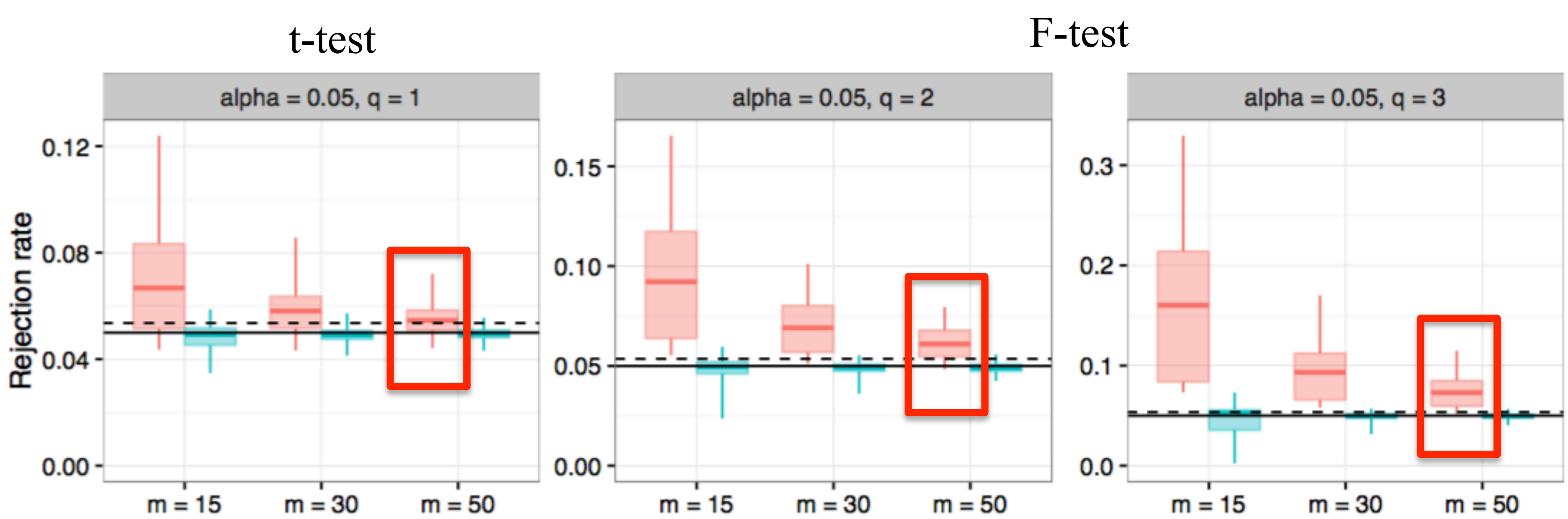
$$F \sim F(q = 2, m - 1)$$

The sample size that matters is the number of *clusters*, not the number of *observations*.

Not so good in small samples

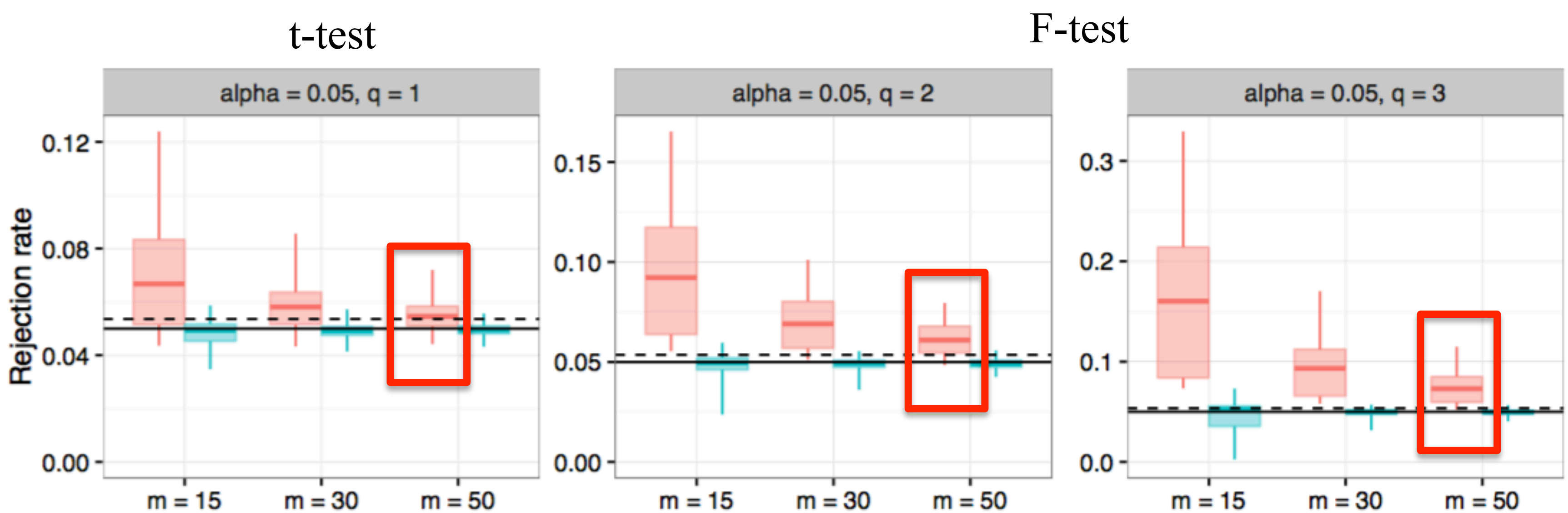


Not so good in small samples



Not so good in small samples

Not so good, even with 50 clusters!



Bias Reduced Linearization + Satterthwaite

CRVE is biased

One of the reasons for the poor performance of CRVE is that the variance estimator is biased.

To see why, note that
$$E\left(\mathbf{e}_j\mathbf{e}_j'\right) = (\mathbf{I} - \mathbf{H})_j\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{H})_j'$$

where $\mathbf{H} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X}$.

This means that:
$$E[v(\mathbf{b})] = (\mathbf{X'X})^{-1}\sum_{j=1}^{m}\mathbf{X}_j'(\mathbf{I} - \mathbf{H})_j\boldsymbol{\Sigma}(\mathbf{I} - \mathbf{H})_j'\mathbf{X}_j\ (\mathbf{X'X})^{-1}$$

$$\neq \mathrm{V}(\mathbf{b}).$$

An unbiased estimator

The goal then is to find an *adjustment matrix* $\mathbf{A}_j$:

$$v_s(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1} \sum_{j=1}^{m} \mathbf{X}_j{}'\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j{}'\mathbf{A}_j{}'\mathbf{X}_j\,(\mathbf{X}'\mathbf{X})^{-1}$$

such that: $\quad \mathrm{E}[v_s(\mathbf{b})] = \mathrm{V}(\mathbf{b})$.

$\mathbf{A}_j$ must thus be defined so:

$$\mathbf{A}_j\big[(\mathbf{I}-\mathbf{H})_j\boldsymbol{\Sigma}(\mathbf{I}-\mathbf{H})_j'\big]\mathbf{A}_j' = \boldsymbol{\Sigma}_j$$

Which means we need to know $\boldsymbol{\Sigma}_j$.

BRL

Bell & McCaffrey using a "working" model for $\mathbf{\Sigma}_j$.

For example, they propose setting $\mathbf{\Sigma}_j = \mathbf{I}_j$.

This seems contradictory:

- The goal is an estimator that *does not* require specification of the dependence structure,
- Yet the estimator *requires* a dependence structure to be specified.

Yet, simulation results consistently show:

- that the BRL approach reduces bias,
- even when the working model is far from the truth.

But that's not all

We can then use this BRL estimator $v_s(\mathbf{b})$ in:

- t-tests
- F-tests

So the problem is solved?

- Bias isn't the only problem.
- The sampling distribution is also problematic.

Distributional problems

Bell & McCaffrey show that in small samples,

$$t \not\sim t(m-1).$$

Instead,

$$t \sim t(\nu)$$

where the degrees of freedom ν can be estimated using a Satterthwaite approximation.

Degrees of freedom (ν)

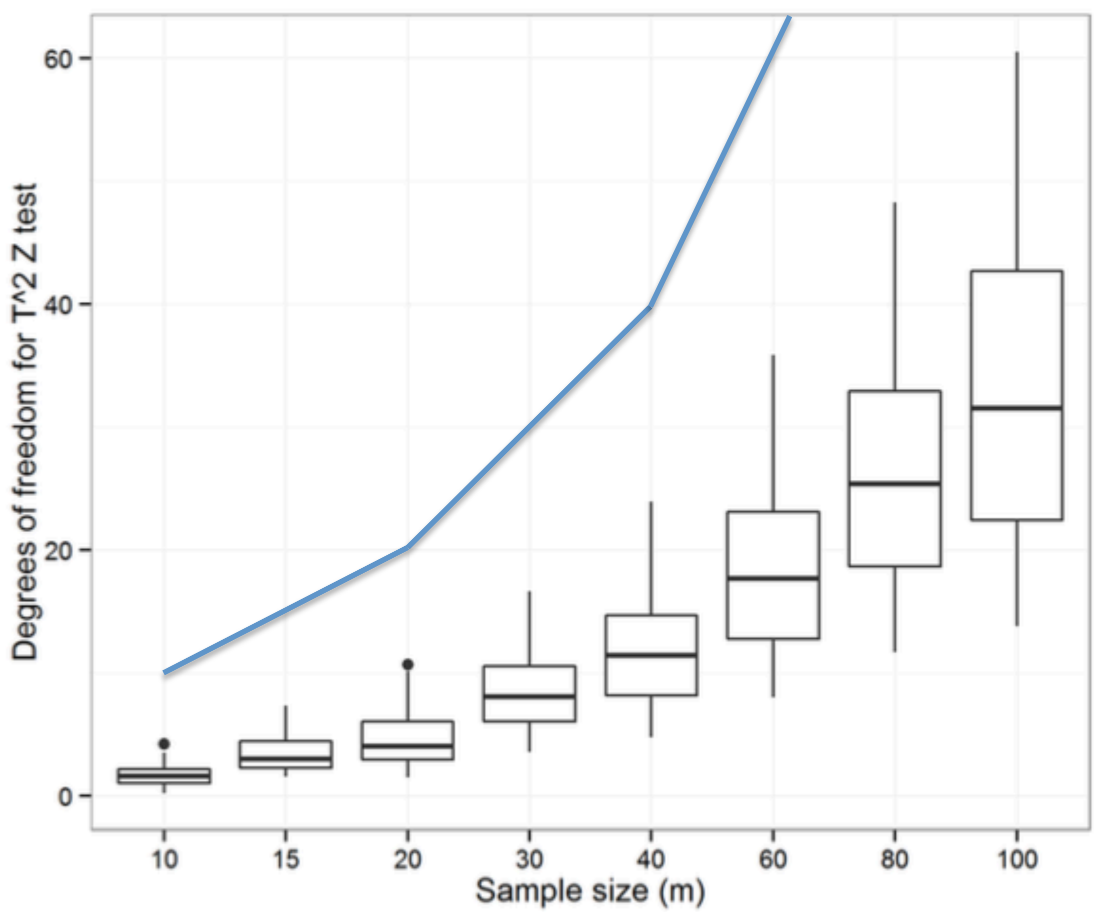
These estimated degrees of freedom depend not only on the number of clusters (m) but on *features of the covariate*.

For example, imagine a model with a single covariate, a policy indicator.

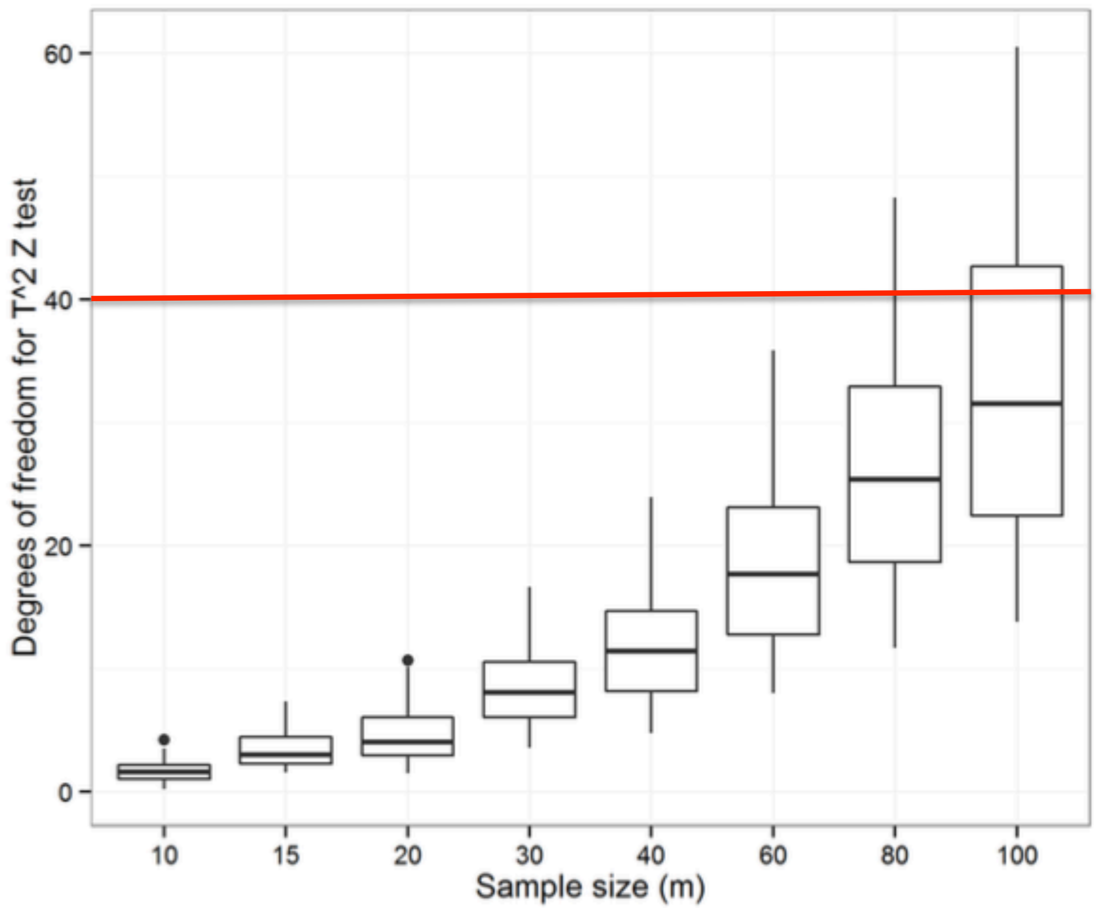
- If the policy is divided evenly across the m clusters, then $\nu \approx m - 1$.
- If the policy is rare, e.g, found in only 3 clusters, then ν can be quite small.

Importantly, in multiple regression, the degrees of freedom can vary considerably from covariate to covariate.

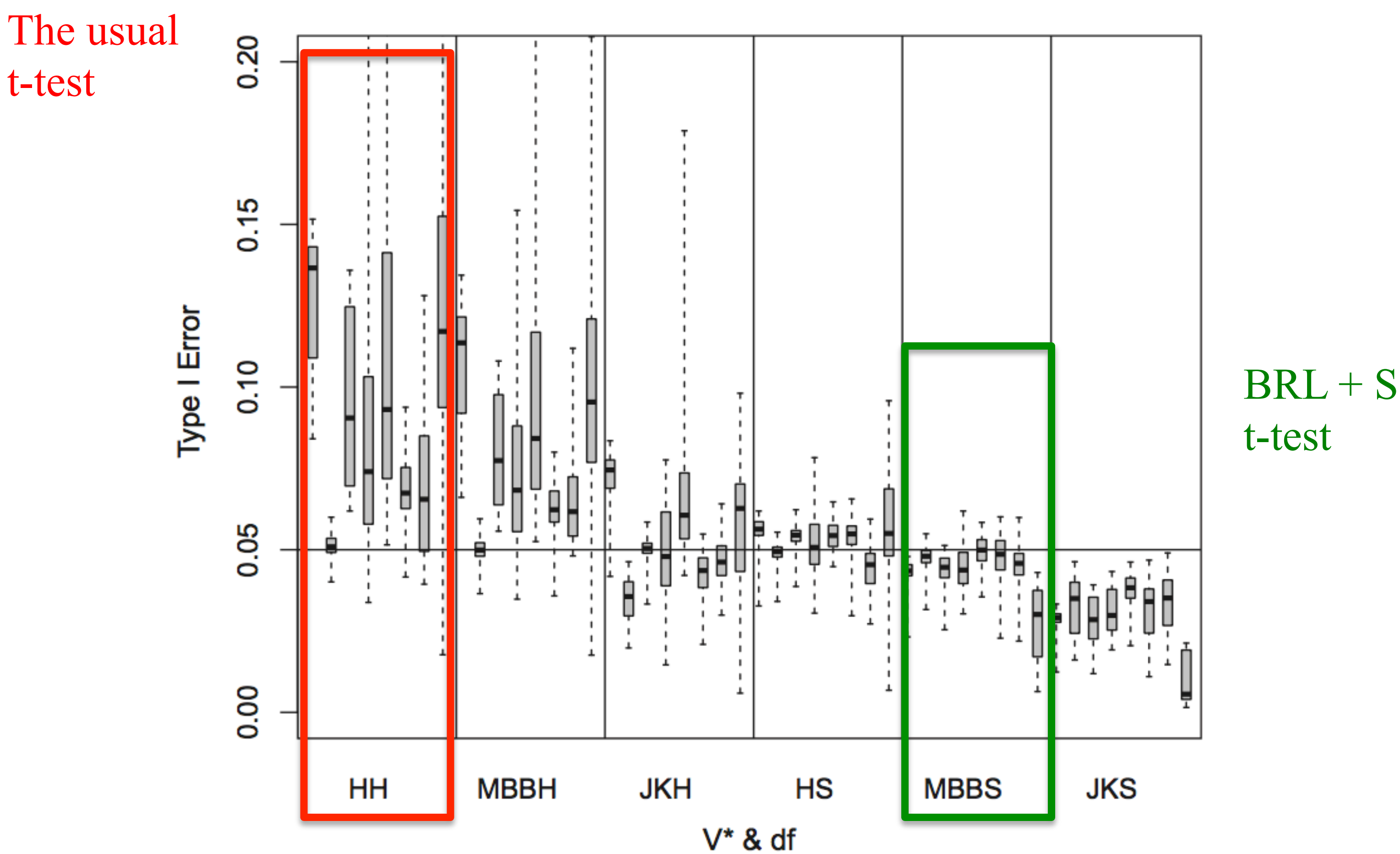
Degrees of freedom



Degrees of freedom



BRL + Satterthwaite



Other research

The BRL + S t-test has been shown to perform well under a wide variety of conditions:

- Simulations in survey-sampling conditions (Bell & McCaffrey, 2002; McCaffrey, Bell, and Botts, 2001).
- Simulations in meta-analytic conditions (Tipton, 2015);
- Simulations in econometric conditions (Imbens & Kolesar, 2015; Cameron & Miller, 2015).

This paper

What about economics?

While the BRL+S approach is promising, there are three problems that limit it's application:

1. To date, there is no multi-parameter F-test.
2. The adjustment ($\mathbf{A}_j$) matrices are not defined when fixed effects are included in a model (the Angrist-Pischke problem).
3. The degrees of freedom (ν) can differ depending on the estimation strategy (the Cameron-Miller problem).

This paper

We solve these problems.

The result is a unified framework for hypothesis testing with CRVE in finite samples.

F-test

Previous papers by Bell & McCaffrey, Imbens & Kolesar, and Cameron & Miller all focus on small-sample corrections to t-tests.

But analysts often also conduct F-tests:

- In experiments with multiple arms;
- In approximate Hausman tests;
- When comparing the joint influence of covariates on a model;
- When testing hypotheses about categorical variables;
- When testing baseline equivalence.

Standard F-test

Consider a hypothesis test of general form,

$$H_0\text{: } \mathbf{C}\boldsymbol{\beta} = \mathbf{c}$$

C is a q x p contrast matrix and **c** is a q x 1 vector.

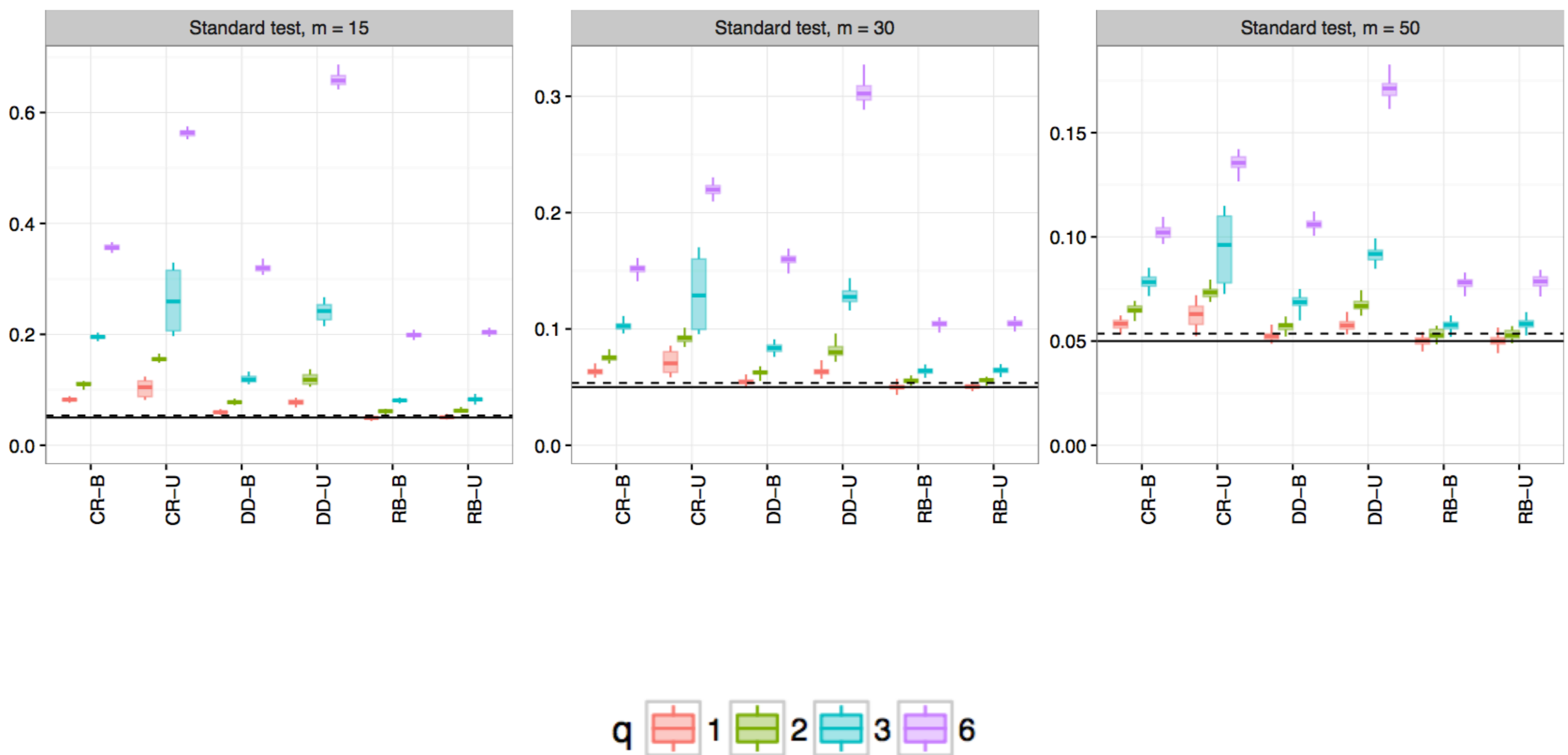
This results in the "standard" F-test (based on the Wald test),

$$F = Q/q = (\mathbf{Cb} - \mathbf{c})'[\mathbf{C}\mathrm{v}(\mathbf{b})\mathbf{C}']^{-1}(\mathbf{Cb} - \mathbf{c})/q$$

And under H_0, in large samples it is assumed that

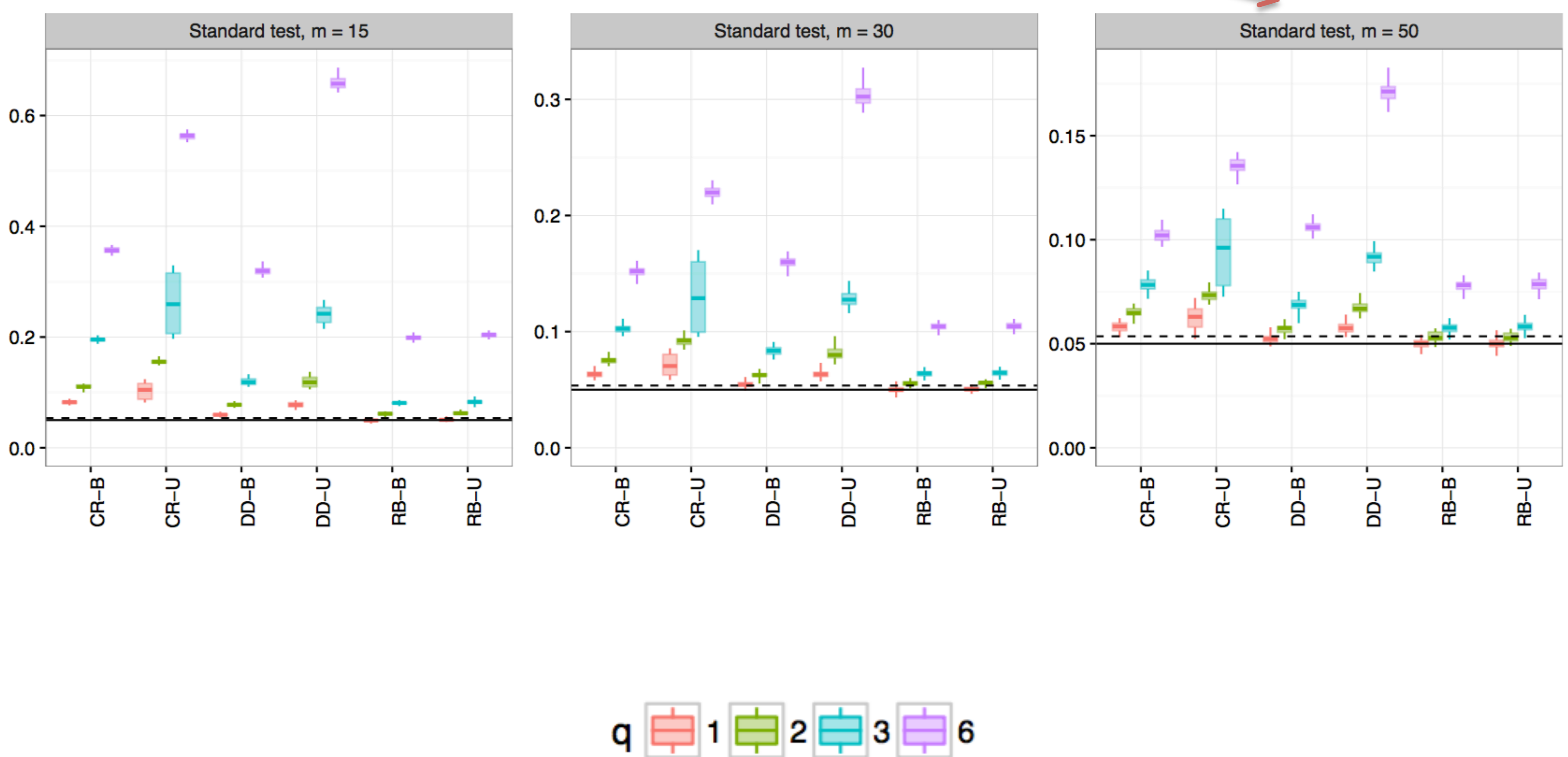
$$F \sim F(q, m - 1).$$

But this test is no good



But this test is no good

Even with 50 clusters!



The AHT Test

We propose instead the Approximate Hotelling's T^2 test,

$$F = [(\eta + q - 1)/\eta]\ Q/q$$

where η is empirically estimated using a Satterthwaite approach.

Under H_0, we show that $F \sim F(q, \eta + q - 1)$.

Degrees of freedom

The degrees of freedom are a function of η, which is estimated.

Like with the t-test, $\eta \ll m - 1$, especially when the covariates tested are unbalanced.

Unbalance or skewness are harder to detect in multivariate form.

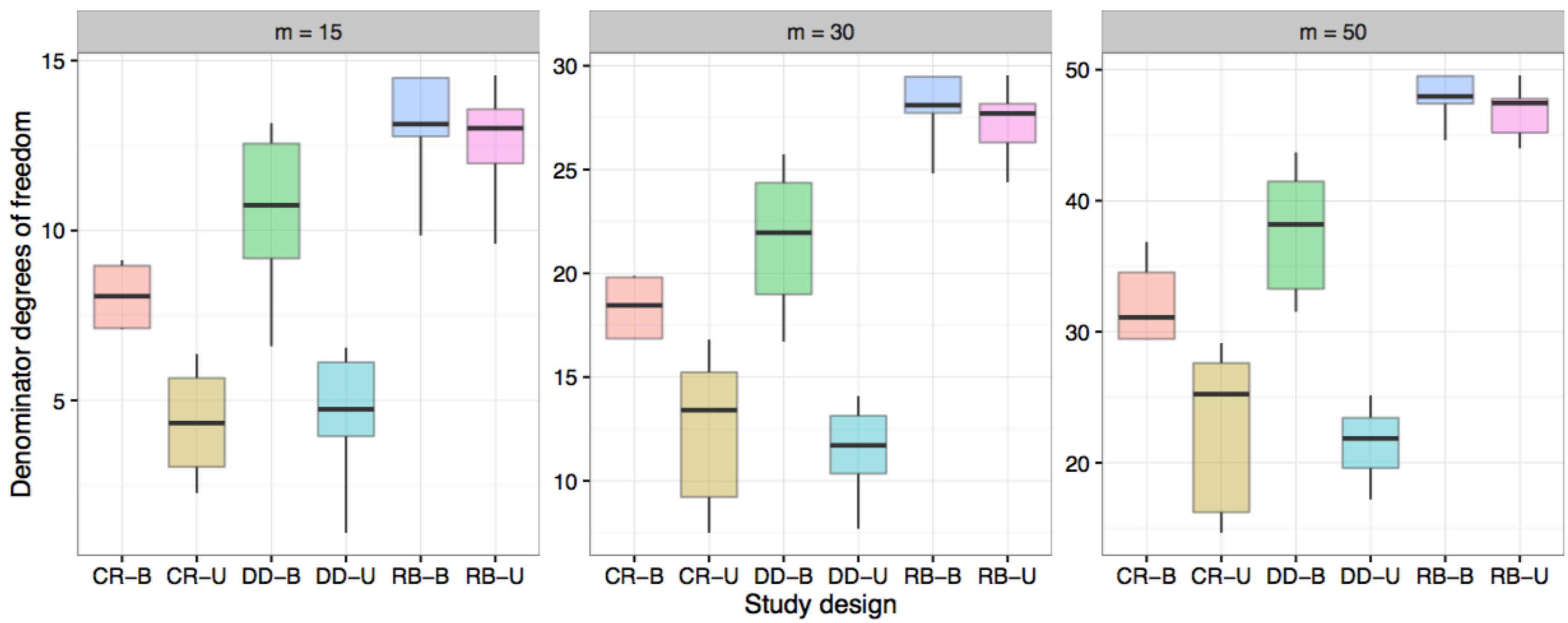
The simplest case is a generalization of the t-test: three policies being compared.

- Balanced means $m/3$ are allocated to each;
- If there is one policy that is rarer, unbalance results in smaller df.

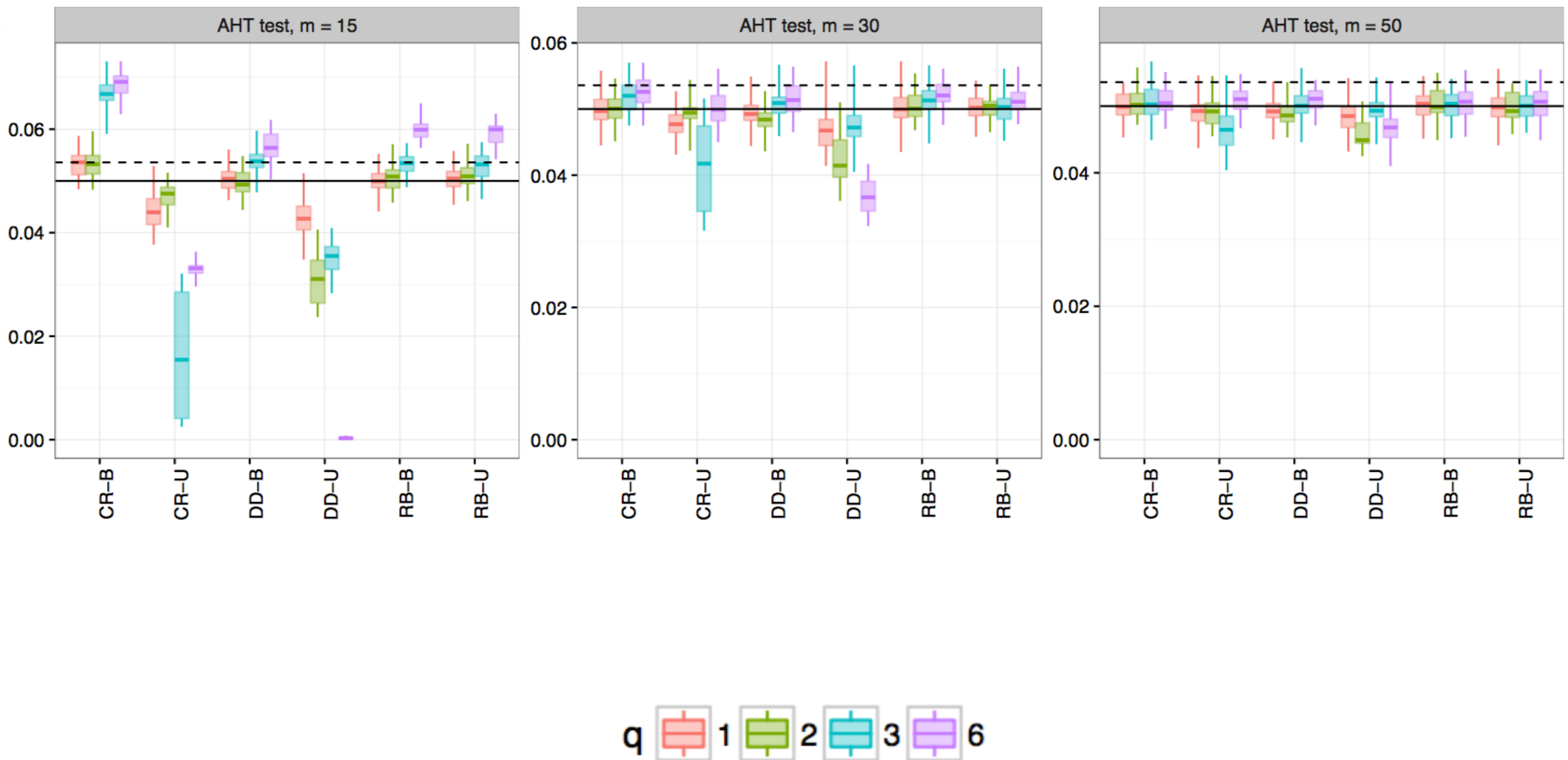
Degrees of freedom are typically:

- Largest for covariates varying *within* clusters; and
- Smaller for covariates at the cluster level.

Degrees of freedom smaller than $m - 1$



The AHT test is nearly level-α



Two other results

The BRL approach as developed was originally focused on problems in survey-sampling.

In econometric applications, following Bertrand, Duflo, & Mullainathan (2004), it is typical to account for clustering with *both*:

- The inclusion of fixed effects;
- AND the use of CRVE.

Angrist-Pischke problem

Problem: It is possible that there is a covariate that is constant within a cluster (e.g. the whole cluster receives a policy).

If dummy fixed effects are included in the model, there is an identification problem.

The result is that the $\mathbf{A}_j$ matrices cannot be defined (because the $(\mathbf{I} - \mathbf{H})_j$ matrix is not full rank, thus making inversion impossible).

In the paper, we provide a method for calculating $\mathbf{A}_j$ using the *generalized inverse*, and a theorem indicating the conditions under which this inverse is estimable.

Cameron-Miller problem

Problem: In practice, instead of including dummy fixed effects, for computational purposes the fixed effects are first "absorbed" (i.e., demeaned, the within estimator).

But the set of variables in **X** then changes depending upon the approach.

This means that you can get *different* degrees of freedom depending on the approach you use.

In the paper, we provide a theorem indicating the conditions under which results from absorption and dummy fixed effects are equivalent.

Does this matter in practice?

Angrist & Lavy example

Hypothesis	Test	F	df	p
ATE - upper half (q = 1)	Standard	5.746	34.00	0.02217
	AHT	5.169	15.86	0.03726
ATE - joint (q = 2)	Standard	3.848	34.00	0.03116
	AHT	3.371	15.46	0.06096
Moderation - upper half (q = 2)	Standard	3.186	34.00	0.05393
	AHT	0.091	3.19	0.91520
Moderation - joint (q = 4)	Standard	8.213	34.00	0.00010
	AHT	2.895	3.21	0.19446

Panel data example

Hypothesis	Test	F	df	p
Random effects	Standard	8.261	49.00	0.00598
	AHT	7.785	24.74	0.00999
Fixed effects	Standard	9.660	49.00	0.00313
	AHT	9.116	22.72	0.00616
Hausman test	Standard	2.930	49.00	0.06283
	AHT	2.489	8.69	0.13980

Conclusions

The standard t- and F-tests used in CRVE do not perform well in small ***or even moderate*** samples.

It is hard to detect a priori when they will fail, since “small” depends not only on the number of clusters, but also covariate features.

The AHT F-test performs well in a broad range of applications and is nearly always level-α. In large-samples it converges to the standard estimator.

Therefore we recommend analysts use the AHT F-test (and t-test) in *all* analyses, not just when the number of clusters seems “small”.

Future work:

- Will focus on comparing this approach to the cluster Wild bootstrap.
- Includes development of a Stata macro.

Contact information

Elizabeth Tipton

tipton@tc.columbia.edu

James Pustejovsky

pusto@austin.utexas.edu