Clustered Regression, Variance Estimation, and the Jackknife

Bruce E. Hansen

Winter School 2022

Delhi School of Economics

Clustered Regression

- Clustered regression is the most common sampling structure in current applied econometrics at top economics journals.
- There are n observations, separated into G clusters.
  - The observations are assumed to be possibly correlated within each cluster, but uncorrelated across clusters.
  - Common choices for cluster level: family, firm, school, classroom, village, district, state.
- A common example is panel regressions (individual/firm observed over time), which are clustered at the level of the individual/firm.
- Standard regression (independent sampling) is a special case where each cluster has one observation.

What Level to Cluster?

- In general, unknown.
- Open research question.
- Current theory requires that the research knows the correct level of clustering.
- In practice, the cluster level is selected by the researcher.
- If the level is picked at too fine a level, between-cluster correlation will be missed.
- If the level is picked at too course a level, variance estimation will be imprecise, confidence intervals bloated, and inference inaccurate.
- Advice: Pick clusters so that
  - You can reasonably argue that observations beween clusters are uncorrelated
  - There are a sufficient number G of clusters to apply large sample approxmations.
  - The sample is not dominated by a small number of clusters.

What Level to Cluster - Advice?

Pick clusters so that

- Observations beween clusters are reasonably uncorrelated.
- There are a sufficient number G of clusters to apply large sample approximations.
- The sample is not dominated by a few large clusters.

Individual-Level Notation

- Observations on the ith individual in the gth cluster is (Y_{ig}, X_{ig}).
  - Y_{ig} is scalar
  - X_{ig} is a k-vector
- Clusters: $g = 1, ..., G$
- Individuals in each cluster: $i = 1, ..., n_g$
- n_g is the number of observations in cluster g
- $n = \sum_{g=1}^{G} n_g$ is total number of observations.
- In applications, n_g is heterogeneous (clusters are of unequal sizes).
- Classical regression: $n_g = 1$ (one observation per cluster).

Regression model and estimation

- Individual-level notation:

$$\begin{aligned} Y_{ig} &= X_{ig}'\beta + e_{ig} \\ \mathbb{E}\left[X_{ig}e_{ig}\right] &= 0 \end{aligned}$$

- Full-sample notation

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

- Least squares (OLS) estimator

$$\begin{aligned} \widehat{\beta} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{Y}\right) \\ &= \left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}X_{ig}'\right)^{-1}\left(\sum_{g=1}^{G}\sum_{i=1}^{n_g} X_{ig}Y_{ig}\right) \end{aligned}$$

- Sum across all observations = Sum over clusters + Sum over observations within each cluster
- OLS estimator is not affected by clustering.

Cluster-Level Notation

- Stack observations by cluster
- $\mathbf{Y}_g = (Y_{1g}, \ldots, Y_{n_g g})'$
  - $n_g \times 1$ vector
- $\mathbf{X}_g = (X_{1g}, \ldots, X_{n_g g})'$
  - $n_g \times k$ matrix
- $\mathbf{e}_g = (e_{1g}, \ldots, e_{n_g g})'$
  - $n_g \times 1$ vector
- Regression model using cluster notation

$$\mathbf{Y}_g = \mathbf{X}_g \beta + \mathbf{e}_g$$

- Identical to individual-level model, just different notation

Least Squares Estimator - Notation

- Using cluster-level notation

$$\widehat{\beta} = \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{Y}_g \right)$$

- Sum over clusters of cluster sums

Typical Assumptions

$$\mathbf{Y}_g = \mathbf{X}_g \beta + \mathbf{e}_g$$

- Error has zero mean
$$\mathbb{E}\left[\mathbf{e}_g\right] = \mathbf{0}$$
- Zero correlation across clusters
$$\mathbb{E}\left[\mathbf{e}_g \mathbf{e}_h'\right] = \mathbf{0}$$
for $g \neq h$
- Cluster-level covariance matrices
$$\mathbb{E}\left[\mathbf{e}_g \mathbf{e}_g'\right] = \mathbf{\Sigma}_g$$
- Often: exogenous regressors
    - Dynamic panels also allowed

Cluster Covariance

$$\mathbb{E}\left[\mathbf{e}_g \mathbf{e}_g'\right] = \mathbf{\Sigma}_g$$

- $\mathbf{\Sigma}_g$ is unstructured - arbitrary correlation allowed
- Within-cluster observations can be uncorrelated
- Within-cluster observations can be fully correlated
- $\mathbf{\Sigma}_g$ can be a function of the regressors (conditionally heteroskedastic)
- $\mathbf{\Sigma}_g$ can be a function of the cluster g (unconditionally heteroskedastic)
- Example: Observations are families, clusters are villages. Choices may be correlated due to kinship relationships, social relationships, etc., and these are specific connections between families i and j in the cluster. These relationships, and hence correlations, vary across villages.

Clustered Dependence vs Heteroskedasticity

- Classical regression (no clustering) is the special case where $n_g = 1$
- $Y_i = X_i'\beta + e_i$
- $\mathbb{E}\left[e_i^2\right] = \sigma_i^2$
- Heteroskedastic regression allows unstructured variances σ_i^2
- Clustered regression allows unstructured covariance matrices $\boldsymbol{\Sigma}_g$
  - Heteroskedastic regression is a special case.

Fixed Effects

- Components model

$$Y_{ig} = X_{ig}'\beta + u_i + e_{ig}$$

- Individual effect u_i may correlated with X_{ig} (fixed effect assumption)
- Within transformation:
  - Group means:

$$\overline{Y}_g = \overline{X}_g'\beta + u_i + \overline{e}_g$$

  - Differences

$$\begin{aligned} Y_{ig} - \overline{Y}_g &= \left(X_{ig} - \overline{X}_g\right)'\beta + (u_i - u_i) + \left(e_{ig} - \overline{e}_g\right) \\ \dot{Y}_{ig} &= \dot{X}_{ig}'\beta + \dot{e}_{ig} \end{aligned}$$

- Fixed effects (within) estimator is least squares on transformed model
- The transformed error $\dot{e}_{ig}$ is correlated within each cluster, even if the original error e_{ig} is uncorrleated
- Thus the within-transformed model has clustered dependence, at the level of the group g

Bias and Variance

- Assume strictly exogenous regressors
- Least squares is unbiased:

$$\begin{aligned} \mathbb{E}\left[\widehat{\beta}\right] &= \beta + \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{X}_g\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbb{E}\left[\mathbf{e}_g\right]\right) \\ &= \beta \end{aligned}$$

- Covariance matrix equals

$$\begin{aligned} \mathbf{V} &= \operatorname{var}\left[\widehat{\beta}\right] \\ &= \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \Sigma_g \mathbf{X}_g\right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}. \end{aligned}$$

- Standard errors require an estimate of **V**.

Variance Estimation Without Clustering

- Let $\widehat{e}_i = Y_i - X_i'\widehat{\beta}$ denote the least squares residual
- Eicker-Huber-White or HC0

$$\widehat{\mathbf{V}}_{\mathrm{HC0}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- HC1, Hinkley (1977)

$$\widehat{\mathbf{V}}_{\mathrm{HC1}} = \frac{n}{n-k}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{i=1}^{n} X_i X_i' \widehat{e}_i^2\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- The degree-of-freedom correction $\frac{n}{n-k}$ has limited justification.
- HC1 is the default robust estimator in Stata ("r" option), R, and elsewhere.

HC2

- MacKinnon and White (1985)
- Define the leverage values

$$h_i = X_i' \left(\mathbf{X}'\mathbf{X}\right)^{-1} X_i$$

- Define the HC2 estimator

$$\widehat{\mathbf{V}}_{\mathrm{HC2}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{i=1}^{n} X_i X_i' \frac{\widehat{e}_i^2}{(1-h_i)} \right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- It is unbiased for $\mathbf{V}$ when the errors e_i are homoskedastic

HC3

- MacKinnon and White (1985), Andrews (1991)

$$\widehat{\mathbf{V}}_{\mathrm{HC3}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \frac{\widehat{e}_i^2}{(1-h_i)^2} \right) (\mathbf{X}'\mathbf{X})^{-1} \qquad (1)$$

- An alternative expression is

$$\widehat{\mathbf{V}}_{\mathrm{HC3}} = (\mathbf{X}'\mathbf{X})^{-1} \left(\sum_{i=1}^{n} X_i X_i' \widetilde{e}_i^2 \right) (\mathbf{X}'\mathbf{X})^{-1} \qquad (2)$$

where $\widetilde{e}_i = Y_i - X_i'\widehat{\beta}_{-i}$ is the leave-one-out prediction error and $\widehat{\beta}_{-i}$ is the leave-one-out least squares estimator

$$\widehat{\beta}_{-i} = \left(\sum_{j \neq i} X_j X_j' \right)^{-1} \left(\sum_{j \neq i} X_j Y_j \right)$$

Theorem: (1)=(2)

Proof: Holds if

$$\widetilde{e}_i = \frac{\widehat{e}_i}{1 - h_i} \tag{3}$$

To see this, take

$$\widehat{\beta}_{-i} = \left(\mathbf{X}'\mathbf{X} - X_i X_i'\right)^{-1} \left(\mathbf{X}'\mathbf{Y} - X_i Y_i\right).$$

Multiply each side by $\left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}'\mathbf{X} - X_i X_i'\right)$ to find

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\mathbf{X}'\mathbf{X} - X_i X_i'\right) \widehat{\beta}_{-i} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{Y} - \left(\mathbf{X}'\mathbf{X}\right)^{-1} X_i Y_i$$

Reorganizing, $\widehat{\beta}_{-i} - \widehat{\beta} = -\left(\mathbf{X}'\mathbf{X}\right)^{-1} X_i \widetilde{e}_i$. Premultiply by X_i' to find

$$X_i'\widehat{\beta}_{-i} = X_i'\widehat{\beta} - h_i \widetilde{e}_i$$

Reorganizing, we find $\widehat{e}_i = (1 - h_i)\widetilde{e}_i$, which is (3).

Jackknife Variance Estimator

$$\widehat{\mathbf{V}}_{\text{jack}} = \sum_{i=1}^{n} \left(\widehat{\beta}_{-i} - \widehat{\beta}\right)\left(\widehat{\beta}_{-i} - \widehat{\beta}\right)'$$

- Conceptually simple
- Easily generalizable to a variety of estimators
- Straightforward to obtain in Stata using `jackknife` command
- In the linear regression context, the jackknife equals VC3:

Theorem: $\widehat{\mathbf{V}}_{\text{HC3}} = \widehat{\mathbf{V}}_{\text{jack}}$

Proof

We showed in the previous proof that

$$\widehat{\beta}_{-i} - \widehat{\beta} = -\left(\mathbf{X}'\mathbf{X}\right)^{-1} X_i \widetilde{e}_i$$

Substituting into the jackknife formula

$$\begin{aligned}
\widehat{\mathbf{V}}_{\text{jack}} &= \sum_{i=1}^{n} \left(\widehat{\beta}_{-i} - \widehat{\beta}\right)\left(\widehat{\beta}_{-i} - \widehat{\beta}\right)' \\
&= \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{i=1}^{n} X_i X_i' \widetilde{e}_i^2\right) \left(\mathbf{X}'\mathbf{X}\right)^{-1} \\
&= \widehat{\mathbf{V}}_{\text{HC3}}
\end{aligned}$$

Variance Estimation Under Clustering

- Let $\widehat{\mathbf{e}}_g = \mathbf{Y}_g - \mathbf{X}_g \widehat{\beta}$ denote the least squares residual vector for the gth cluster.
  - These are the residuals for each observation, grouped by cluster
- Cluster-Robust Variance estimator (CRVE)

$$\widehat{\mathbf{V}}_{\mathrm{CR}} = \frac{G(n-1)}{(G-1)(n-k)} \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \widehat{\mathbf{e}}_g \widehat{\mathbf{e}}_g' \mathbf{X}_g \right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}.$$

- Liang and Zeger (1986) and Arellano (1987).
- The $\frac{G(n-1)}{(G-1)(n-k)}$ constant is an *ad hoc* degree-of-freedom correction, introduced by Stata "`cluster`".
  - Specializes to $\frac{n}{n-k}$ when $G = n$
- These standard errors are ubiquitous in current econometric practice

CRVE

2

- Bell and McCaffrey (2002)
- Partial projection matrices

$$\mathbf{M}_g = \mathbf{I}_{n_g} - \mathbf{X}_g \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}_g'$$

  diagonal blocks of $\mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$

- Generalization of HC2

$$\widehat{\mathbf{V}}_2 = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{M}_g^{-1/2} \widehat{\mathbf{e}}_g \widehat{\mathbf{e}}_g' \mathbf{M}_g^{-1/2} \mathbf{X}_g \right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- It is unbiased for $\mathbf{V}$ when $\mathbf{\Sigma}_g = \sigma^2 \mathbf{I}_{n_g}$ (i.i.d. errors)
- Relatively computationally expensive
  - Especially when cluster sizes n_g are large, as $\mathbf{M}_g$ is $n_g \times n_g$

CRVE

3

- Bell and McCaffrey (2002)
- Generalization of HC3

$$\widehat{\mathbf{V}}_3 = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{M}_g^{-1} \widehat{\mathbf{e}}_g \widehat{\mathbf{e}}_g' \mathbf{M}_g^{-1} \mathbf{X}_g \right) \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Cluster Jackknife

- Delete-one-cluster estimators

$$\widehat{\beta}_{-g} = \left(\sum_{j \neq g} \mathbf{X}_j' \mathbf{X}_j \right)^{-1} \left(\sum_{j \neq g} \mathbf{X}_j' \mathbf{Y}_j \right)$$

- Jackknife variance estimator

$$\widehat{\mathbf{V}}_{\text{jack}} = \sum_{g=1}^{G} \left(\widehat{\beta}_{-g} - \widehat{\beta} \right) \left(\widehat{\beta}_{-g} - \widehat{\beta} \right)'$$

- **Theorem**: In linear regression model $\widehat{\mathbf{V}}_3 = \widehat{\mathbf{V}}_{\text{jack}}$
  - MacKinnon, Nielsen, Webb: "Fast and reliable jackknife and bootstrap methods for cluster-robust inference," working paper

Cluster Jackknife

- Conceptually simple
- Easily generalizable
- MacKinnon, Nielsen, & Webb demonstrate that the jackknife formula is computationally much faster than the CRVE_3 formula when $n_g \geq 4$.
  - Essentially, the jackknife involves G re-computations of $\widehat{\beta}$.
  - This can be done computationally efficiently.
  - The CRVE_3 formula requires computation of $n_g \times n_g$ matrices $\mathbf{M}_g$.

Challenges with Cluster Inference

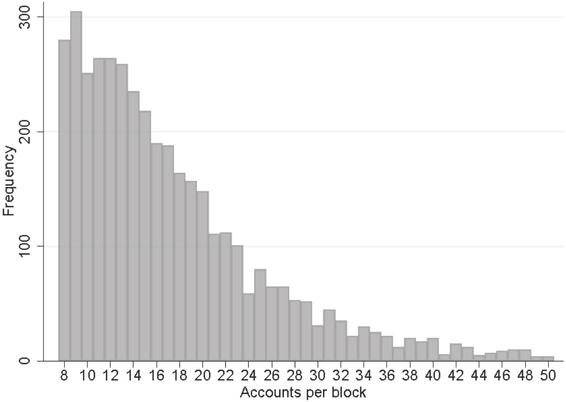
- The CR variance estimators work well when cluster sizes n_g are small.
  - In this case, we have a small number of error correlations to estimate.
  - And the number of independent clusters G is large.
  - This is the context of many traditional panel applications.
- The methods work less well when cluster sizes n_g are large.
  - In this case, there are a large number of error correlations to estimate.
  - And the number of clusters G is small.
  - This is common in many contemporary clustered regressions.

Challenges with Cluster Inference

- CR methods are also severely impacted when cluster sizes n_g are heterogenous.
- Some small clusters, a few large clusters.
- This has an effect similar to heteroskedasticity in regression models
- When the variation in n_g is very large, this is similar to severe heteroskedasticity, which is empirically uncommon in standard regression applications.

"Design of Two-Stage Experiments with an Application to Spillovers in Tax Compliance", Vazquez-Bare, Cruses, & Tortarolo

Figure 2: Distribution of accounts per block



Notes: This figure shows the distribution of accounts per block using data from the year 2019. We use these data design the experiment. Our sample size consists of 68,808 accounts distributed in 3,982 blocks.

Cluster Sums

- The reason is because the least squares estimator **sums** across observations in a cluster

$$\widehat{\beta} - \beta = \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{X}_g \right)^{-1} \left(\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{e}_g \right)$$

- The asymptotic normal approximation is obtained by the component

$$\sum_{g=1}^{G} \mathbf{X}_g' \mathbf{e}_g$$

- The summands $\mathbf{X}_g' \mathbf{e}_g$ are within-cluster **sums**, so their magnitudes are highly influenced by the cluster size n_g

Bias of Robust Covariance Matrix Estimators

- For simplicity, assume β is scalar (dimension 1)
- The **relative expectation** of an estimator $\widehat{\mathbf{V}}$ for $\mathbf{V}$ is

$$\frac{\mathbb{E}\left[\widehat{\mathbf{V}}\right]}{\mathbf{V}}$$

- The estimator is **downward biased** if $\frac{\mathbb{E}\left[\widehat{\mathbf{V}}\right]}{\mathbf{V}} < 1$
- We care about downward bias, as this means our standard errors are "too small"
- The **maximum downward bias** is $b = \inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}\right]}{\mathbf{V}}$, where the inf is taken over a set of distributions
- An estimator is **fully downwardly biased** if $b = 0$.
- An estimator is **conservative** if $b \geq 1$.

Theorem

In the clustered regression model,

$$\inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}_{\mathrm{CR}}\right]}{\mathbf{V}} = 0$$

$$\inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}_{\mathrm{VC2}}\right]}{\mathbf{V}} = 0$$

$$\inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}_{\mathrm{jack}}\right]}{\mathbf{V}} = 1$$

- The CRVE and CRVE_2 variance estimators are fully downwardly biased.
- The jackknife estimator is conservative.

Interpretation: CRVE

- There are settings where the CRVE estimator is highly biased downward.
- This also applies to the standard HC1 robust variance estimators.
- It means that in some contexts standard errors are much too small (misleadingly).
- This extreme bias arises when the regression is highly leveraged.
  - In the non-clustered setting, when the sample is dominated by a few large regressors.
  - This can occur when a regressor has a highly skewed distribution.
  - It also arises when a dummy variable is highly sparse (equals one for only a few observations).
  - Sparse dummies can easily occur in saturated designs with several dummy variables.

Extreme Leverage in Clustered Regressions

- Extreme leverage easily arises in clustered settings.
- When regressor cluster sums $\mathbf{X}_g'\mathbf{X}_g$ are highly heterogenous
- Occurs when cluster sizes n_g are highly heterogeneous

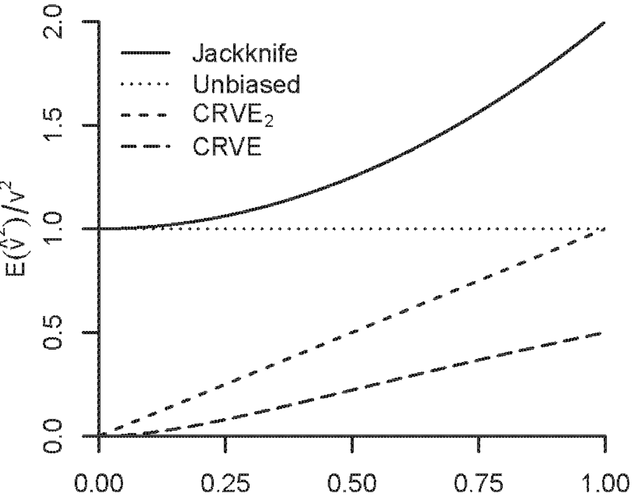
Example: Behrens-Fisher Problem with Two Treated Clusters

- $\mathbf{Y}_g = \mu + \mathbf{X}_g \beta + \mathbf{e}_g$
- $\mathbf{X}_g$ is binary at the cluster level ($\mathbf{X}_g = \mathbf{1}_g$ or $\mathbf{X}_g = \mathbf{0}_g$).
- Two treated clusters, $G - 2$ untreated clusters.
- Variance of cluster sums: $\sigma_g^2 = \text{var}\left[\mathbf{1}_g' \mathbf{e}_g\right] = \mathbf{1}_g' \boldsymbol{\Sigma}_g \mathbf{1}_g$
- Assume $\sigma_g^2 = 0$ for $g \geq 3$ (treated clusters have relative large variance, and/or are relatively large).
- Focus: Inference on treatment slope β (difference in treatment and nontreatment mean).

Behrens-Fisher Problem

- It turns out that the relevant distributions only depend on two parameters:
  - $\rho = \frac{n_2}{n_1}$, ratio of two treated cluster sizes
  - $\psi^2 = \frac{\sigma_2^2}{\sigma_1^2}$, ratio of two treated variances
- Variance Estimation Bias:
  - Plot $\frac{\mathbb{E}[\widehat{\mathbf{V}}]}{\mathbf{V}}$ as a function of ρ
  - (worst-case variance ratio ψ used)

Variance Bias in Behrens-Fisher Problem



Theorem, Revisited

In the clustered regression model,

$$\inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}_{\mathrm{CR}}\right]}{\mathbf{V}} = 0$$

$$\inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}_{\mathrm{VC2}}\right]}{\mathbf{V}} = 0$$

$$\inf \frac{\mathbb{E}\left[\widehat{\mathbf{V}}_{\mathrm{jack}}\right]}{\mathbf{V}} = 1$$

- Thus, the CRVE and CRVE_2 variance estimators are fully downwardly biased.
- The jackknife estimator is conservative.

Interpretation: Jackknife

- The jackknife estimator is never downward biased.
- It is globally conservative.
- This means jackknife standard errors are never expected to be too small.
- Conservative property means that they can be too large

Inference

- For finite sample theory, assume errors are normally distributed
- $\mathbf{e}_g \sim N(0, \mathbf{\Sigma}_g)$
- Normality should not be taken literally, but used to gain insight into exact distribution
- Define t-ratio statistics

$$T = \frac{\widehat{\beta} - \beta}{s\left(\widehat{\beta}\right)}$$

  where $s\left(\widehat{\beta}\right) = \sqrt{\widehat{\mathbf{V}}}$ is a standard error
- t-ratios: T_{CR}, T_2, and T_{jack}, using CRVE, CRVE_2, and jackknife standard errors
- Tests: Reject for $|T| > c$ for some c
  - normal approximation: $c = 1.96$
  - student t approximation: $c \simeq 2$

Theorem

In the clustered regression model, under null hypothesis, and for any c

$$\sup \mathbb{P}\left[\left|T_{\mathrm{CR}}\right| > c\right] = 1$$

$$\sup \mathbb{P}\left[\left|T_{2}\right| > c\right] = 1$$

$$\sup \mathbb{P}\left[\left|T_{\mathrm{jack}}\right| > c\right] \leq 2\left(1 - F(c)\right)$$

where $F(c)$ is the Cauchy distribution function.

- The CRVE and CRVE_2 t-tests have worst-case size equalling one.
- This means that they have arbitrarily high size distortion.
- This is connected to the arbitrary bias of the CRVE and CRVE_2 variance estimators.
- This distortion arises despite normal errors, and for **any** sample size G
- The distortion in CRVE can occur entirely due to extreme regressor leverage
- The distortion in CRVE requires both extreme leverage and extreme heteroskedasticity

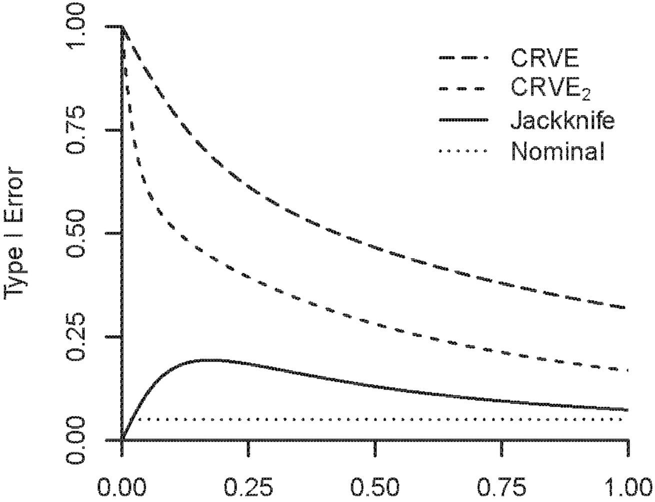
Jackknife t-test

- The distribution of the jackknife t-test is bounded by the Cauchy distribution.
- This means that while the test can still have size distortion, the magnitude is bounded.
- In principle, the Cauchy distribution permits uniformly valid inference.
  - In practice, this is not recommended.

Behrens-Fisher Problem

- We calculate the Type I error of nominal 5% tests
- In our example, it is a function only of ρ and ψ
- We fix $\rho = 0.2$ (one treated cluster is five times larger than the other)
- Plot as a function of variance ratio ψ

Type I Error in Behrens-Fisher Problem



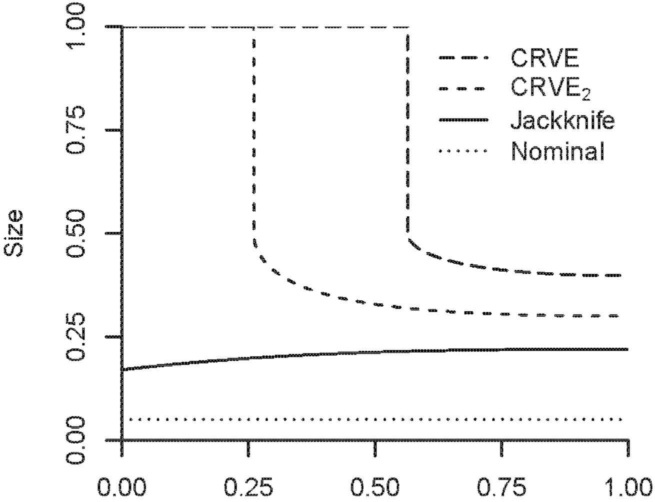
Discussion

- All three tests are oversized
- The CRVE and CRVE_2 tests are exceedingly oversized
- Their Type I error is non-similar: depends on the unknown ψ
- The jackknife test is relatively better performing.

Size

- The size is the worst-case Type I error
- We maximize the Type I error over ψ
- Plot as a function of cluster-size ratio ρ

Size in Behrens-Fisher Problem



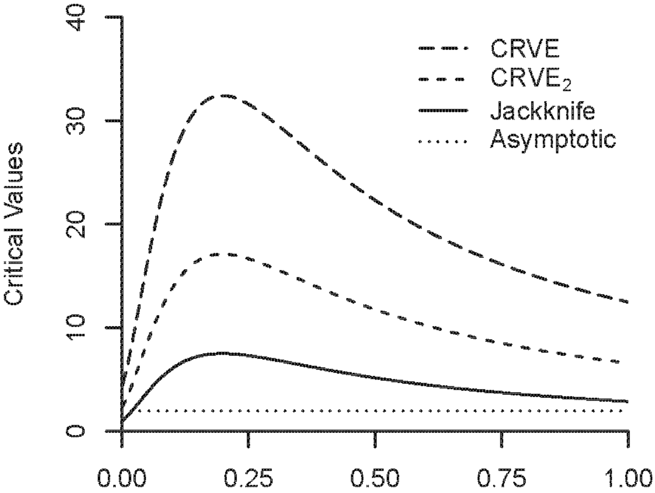
Discussion

- The CRVE and CRVE_2 have size equal to one for many values of ρ
- The size of the jackknife test is relatively invariant to ρ

Finite Sample Critical Values

- Another way of see this is to calculate the finite-sample critical value c which produces an exact 5% test
- First, we fix $\rho = 0.2$, and plot these critical values as a function of ψ

Exact Critical Values in Behrens-Fisher Problem



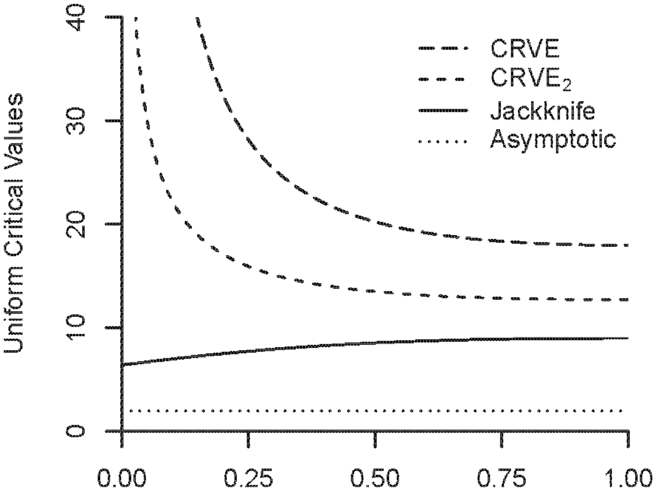
Discussion

- The CRVE and CRVE_2 tests require the largest critical values to achieve 5% size
- These two tests are highly non-similar
- The jackknife test is least sensitive to (unknown) ψ

Uinform Critical Values

- For a critical value to control **size** of the test, it needs to be the **largest** pointwise critical value
- This is the largest point on the previous graphs
- We calculate these uniform critical values, and plot against ρ

Uniform Critical Values in Behrens-Fisher Problem



Discussion

- The CRVE and CRVE_2 tests require extremely large critical values for small ρ
- The jackknife test is least sensitive

Summary

- In heteroskedastic regression and clustered regression (especially), we need accurate standard errors.
- The most common estimators are HC1 and the CRVE.
- A simple alternative is the jackknife (HC3).
  - Computationally straightforward
- Problems with HC1/CRVE
  - Can be arbitrarily downwardly biased
  - Can have arbitrarily large Type I error
  - Correct finite sample critical values can be arbitrarily large
- Properties of the jackknife
  - Conservative estimator (never downward biased)
  - Bounded size distortion and critical values

Computation in Stata

- Non-Clustered Regression
  - `reg y x, vce(hc3)`
- Clustered Regression
  - `reg y x, vce(jackknife,cluster(village) mse)`
- In preparation
  - Jackknife regression command `jregress`
  - Efficiently calculates regression coefficients & jackknife standard errors
  - Clustered and non-clustered regression
- Computation in R
  - In preparation: Similar jackknife regression command