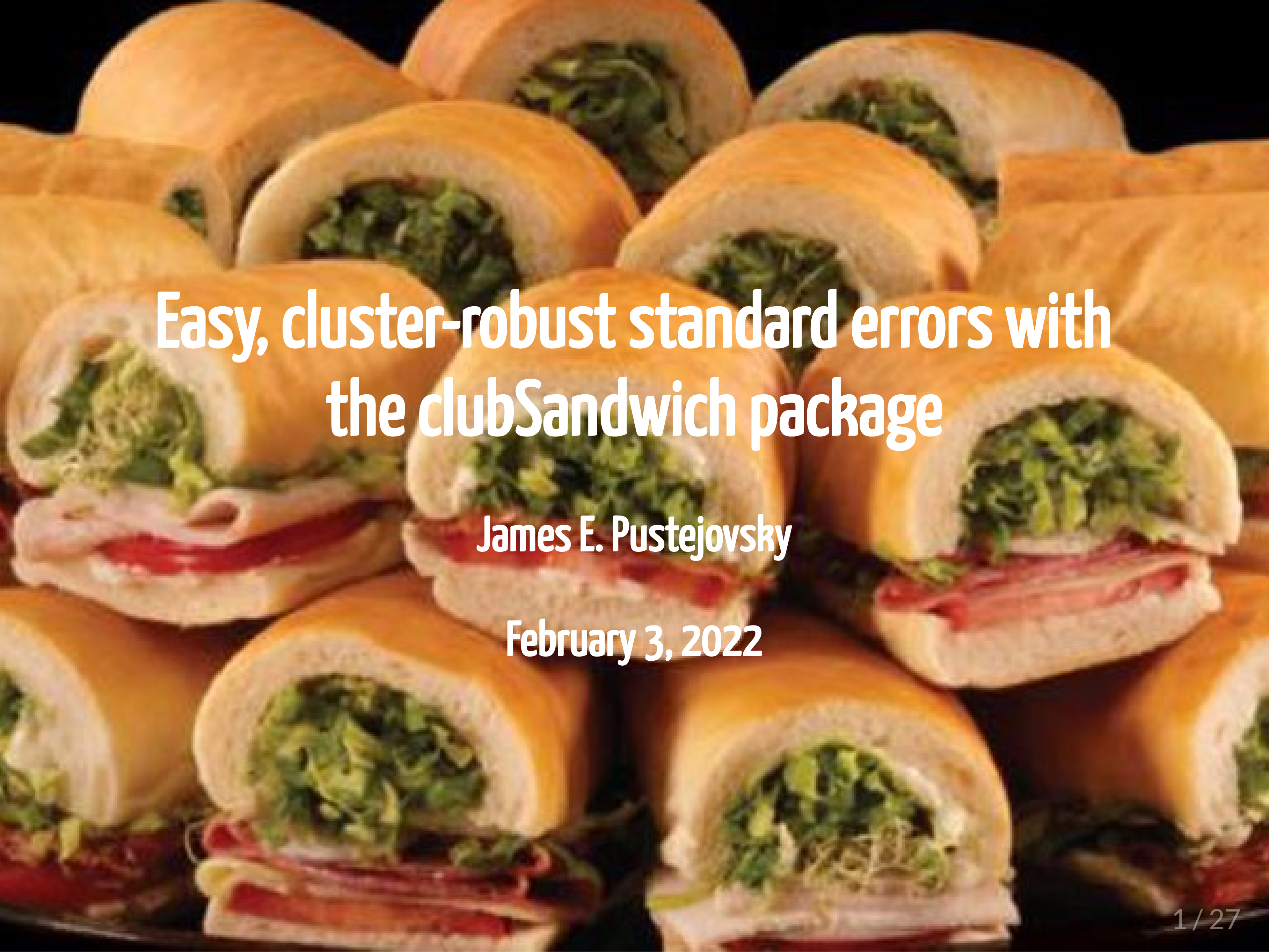


Easy, cluster-robust standard errors with the clubSandwich package

James E. Pustejovsky

February 3, 2022

Agenda

Basics of cluster-robust variance estimation (CRVE)

CRVE for multi-level models

Small-sample refinements

clubSandwich features

Agenda

Basics of cluster-robust variance estimation (CRVE)

CRVE for multi-level models

Small-sample refinements

clubSandwich features

Conventional regression analysis

A generic regression model:

$$Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

Statistics 101 regression analysis makes two strong assumptions:

1. Errors are **independent**, so that $\text{corr}(e_i, e_j) = 0$ when $i \neq j$
2. Errors are **homoskedastic**, so $\text{Var}(e_i) = \sigma^2$ for all i

Many situations where these assumptions are untenable:

- Multi-stage survey data
- Longitudinal/repeated measures/panel data
- Cluster-randomized experiments or quasi-experiments

Escape homoskedasticity and independence assumptions with sandwich estimators

- Calculate regression coefficient estimates $\hat{\boldsymbol{\beta}}$ per usual (ordinary least squares)

- Use sandwich estimators for standard errors of $\hat{\boldsymbol{\beta}}$.

- Sandwich estimators are based on the ***weaker assumption*** that observations can be grouped into J clusters of independent observations:

$$Y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \dots + \beta_p x_{pij} + e_{ij}$$

  - $\mathrm{cor}(e_{hj}, e_{ik}) = 0$ if observations are in different clusters $(j \neq k)$
  - $\mathrm{cor}(e_{hj}, e_{ij}) = \rho_{hij}$ for observations in the same cluster
  - $\mathrm{Var}(e_{ij}) = \phi_{ij}$, allowing for heteroskedasticity

- Cameron and Miller (2015) give an in-depth survey of cluster-robust variance estimation.

Plain sandwich estimators

- Actual variance of coefficient estimate $\hat{\boldsymbol{\beta}}$:

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \frac{1}{J}\mathbf{B}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\boldsymbol{\Phi}_j\mathbf{X}_j\right)\mathbf{B}$$

  where $\boldsymbol{\Phi}_j = \text{Var}(\mathbf{e}_j)$ and $\mathbf{B} = \left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{X}_j\right)^{-1}$.

- The plain sandwich estimator:

$$\mathbf{V}^{plain} = \frac{1}{J}\mathbf{B}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{e}_j\mathbf{e}_j'\mathbf{X}_j\right)\mathbf{B}$$

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

- Relies on a **large-sample approximation** (weak law of large numbers).

U.S. Sustaining Effects Study

- Repeated measures of student mathematics performance in Kindergarten through 5th grade.
- 1721 students from 60 schools
- Indicators for grade retention, sex, race.
- School size, percentage of low income students, mobility index.

```
library(dplyr)
data("egsingle", package = "mlmRev")
egsingle_clean <- egsingle %>%
  mutate(
    retained = as.numeric(retained),
    female = if_else(female == "Female", 1L, 0L),
    grade = factor(grade),
    size = size / 100,
    lowinc = lowinc / 100,
    mobility = scale(mobility)
  )

USSE_lm <- lm(math ~ 0 + grade + size + lowinc + mobility
              + female + black + hispanic + retained,
              data = egsingle_clean)
```

A plain sandwich

```
library(clubSandwich)

# type = "CR0" is the plain sandwich variance estimator
V_plain <- vcovCR(USSE_lm, cluster = egsingle_clean$schoolid,
                  type = "CR0")
```

```
coef_test(USSE_lm, vcov = V_plain, test = "z", coefs = 7:9)
```

```
##     Coef. Estimate     SE t-stat d.f. (z) p-val (z) Sig.
##      size -0.00793 0.0179 -0.443      Inf     0.658
##    lowinc -0.65578 0.1794 -3.655      Inf    <0.001  ***
##  mobility -0.11710 0.0739 -1.584      Inf     0.113
```

- Similar methods implemented in the `sandwich` package (Zeileis, Koll, & Graham, 2020).

Agenda

Basics of cluster-robust variance estimation (CRVE)

CRVE for multi-level models

Small-sample refinements

clubSandwich features

Multi-level models

- A generic multi-level model for clustered data:

$$Y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \mathbf{z}_{ij}\mathbf{u}_j + e_{ij}$$

  where

  - Random effects vector for cluster j: $\mathbf{u}_j \sim N\left(\mathbf{0}, \mathbf{T}\right)$
  - Individual-level error for unit i in cluster j: $e_{ij} \sim N\left(0, \sigma^2\right)$

- Typically, hypothesis tests and confidence intervals for $\boldsymbol{\beta}$ are **model-based**.
  - Usual approximations require a "large enough" number of clusters.
  - ***Contingent on having correctly specified the variance structure***.

Estimation in multi-level models

- Estimate the parameters of the variance structure (using ML or REML).
  - This gives us estimates $\hat{\mathbf{T}}$ and $\hat{\sigma}^2$.
  - Can now estimate the variance of the errors:

$$\mathbf{V}_j = \widehat{\text{Var}}(\mathbf{Y}_j|\mathbf{X}_j) = \mathbf{Z}_j\hat{\mathbf{T}}\mathbf{Z}_j' + \hat{\sigma}^2\mathbf{I}_j$$

- Estimate $\boldsymbol{\beta}$ by weighted least squares:

$$\hat{\boldsymbol{\beta}} = \frac{1}{J}\mathbf{B}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{V}_j^{-1}\mathbf{Y}_j, \qquad \text{where} \qquad \mathbf{B} = \left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{V}_j^{-1}\mathbf{X}_j\right)^{-1}$$

- Estimate uncertainty of $\hat{\boldsymbol{\beta}}$ based on estimated variance structure:

$$\widehat{\text{Var}}(\hat{\boldsymbol{\beta}}) \approx \mathbf{B}$$

  - Contingent on correct specification of the random effects and level-1 error structure.

Potential pitfalls in multi-level model specification

- Inadvertently omitting a random slope
    - Your model: $Y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + e_{ij}$
    - True process: $Y_{ij} = \beta_0 + \beta_1 x_{1ij} + u_{0j} + u_{1j} x_{1ij} + e_{ij}$
- Heterogeneous random effects
    - Your model: $\mathbf{u}_j \sim N(\mathbf{0}, \mathbf{T})$
    - True process: $\mathbf{u}_j \sim N\left(\mathbf{0}, \mathbf{T} \times f(\mathbf{X}_j)\right)$
- Not modeling an intermediate level
    - Your model: $Y_{ik} = \mathbf{x}_{ik}\boldsymbol{\beta} + \mathbf{z}_{ik}\mathbf{u}_k + e_{ik}$
    - True process: $Y_{ijk} = \mathbf{x}_{ijk}\boldsymbol{\beta} + \mathbf{z}_{1ijk}\mathbf{u}_k + \mathbf{z}_{2ijk}\mathbf{u}_{jk} + e_{ijk}$
- Mis-specifying the individual-level error structure
    - Your model: $\mathrm{Var}(e_{ij}) = \sigma^2, \quad \mathrm{cor}(e_{hj}, e_{ij}) = 0.$

Avoid model-contingency with sandwich estimators

- Suppose that, under the true process, $\mathrm{Var}(\mathbf{Y}_j|\mathbf{X}_j) = \boldsymbol{\Omega}_j$.
  - Not necessarily compatible with your assumed structure, so $\boldsymbol{\Omega}_j \neq \mathbf{Z}_j\hat{\mathbf{T}}\mathbf{Z}_j' + \sigma^2\mathbf{I}_j$.
- True variance of coefficient estimate $\hat{\boldsymbol{\beta}}$:

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) \approx \frac{1}{J}\mathbf{B}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{V}_j^{-1}\boldsymbol{\Omega}_j\mathbf{V}_j^{-1}\mathbf{X}_j\right)\mathbf{B}$$

- The plain sandwich estimator:

$$\mathbf{V}^{plain} = \frac{1}{J}\mathbf{B}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{V}_j^{-1}\mathbf{e}_j\mathbf{e}_j'\mathbf{V}_j^{-1}\mathbf{X}_j\right)\mathbf{B}$$

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

U.S. Sustaining Effects Study

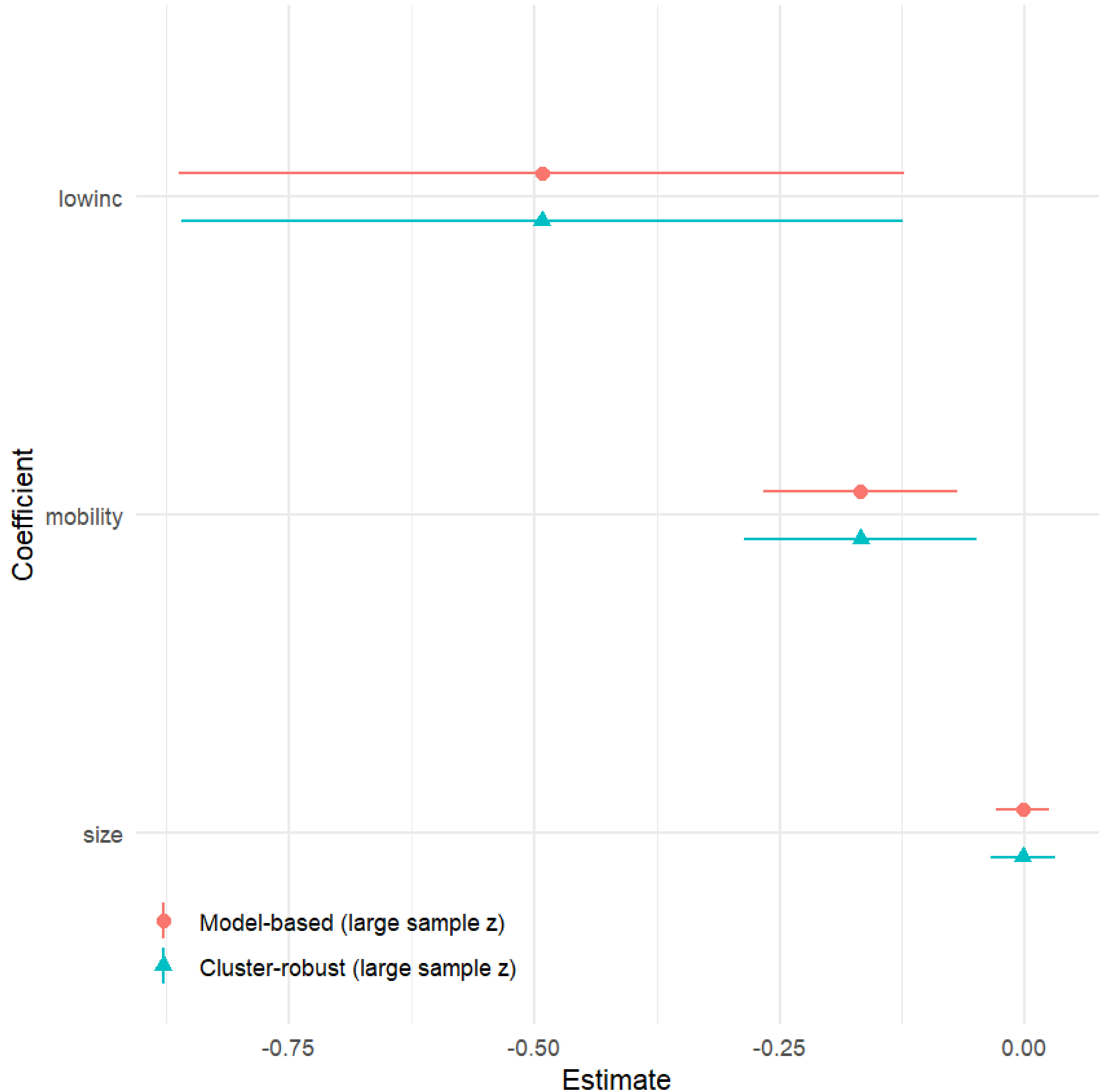
A random intercepts model:

```
library(lme4)
USSE_ri <- lmer(math ~ 0 + grade + size + lowinc + mobility
                + female + black + hispanic + retained
                + (1 | schoolid) + (1 | childid),
                data = egsingle_clean)
```

Confidence intervals with a plain sandwich estimator:

```
conf_int(USSE_ri, vcov = "CR0", test = "z", coefs = 7:9)
```

```
##     Coef. Estimate     SE d.f. Lower 95% CI Upper 95% CI
##      size -0.00214 0.0167  Inf      -0.0349       0.0307
##    lowinc -0.49241 0.1875  Inf      -0.8599      -0.1250
##  mobility -0.16748 0.0606  Inf      -0.2863      -0.0486
```



lowinc
mobility
size
Coefficient
Model-based (large sample z)
Cluster-robust (large sample z)
-0.75
-0.50
-0.25
0.00
Estimate

Agenda

Basics of cluster-robust variance estimation (CRVE)

CRVE for multi-level models

Small-sample refinements

clubSandwich features

Problems with plain sandwich estimators

- **Require a large number of clusters** to work well.
    - Downward bias if the number of clusters is not big enough.
    - Hypothesis tests have inflated type-I error.
    - Confidence intervals have less-than-advertised coverage.
- What counts as "large enough" depends on:
    - **number of clusters**, not number of observations
    - distribution of predictors $\mathbf{X}$ within and across clusters

How can you tell whether your plain sandwich estimators are edible?

Fancy sandwiches

- Adjust the residuals so that they are unbiased under a working model (Bell & McCaffrey, 2002, 2006; Pustejovsky & Tipton, 2018):

$$\mathbf{V}^{club} = \frac{1}{J}\mathbf{B}\left(\frac{1}{J}\sum_{j=1}^{J}\mathbf{X}_j'\mathbf{V}_j^{-1}\mathbf{A}_j\mathbf{e}_j\mathbf{e}_j'\mathbf{A}_j\mathbf{V}_j^{-1}\mathbf{X}_j\right)\mathbf{B}$$

- Adjustment matrices calculated so that

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

if the model is correctly specified.

- Even when the model is mis-specified, $\mathbf{V}^{club}$ has drastically reduced bias (Pustejovsky & Tipton, 2018).

Degrees of freedom adjustment

- Typical methods involve large-sample normal approximations.
- `clubSandwich` implements Satterthwaite-type degrees-of-freedom adjustments for hypothesis tests and confidence intervals.
    - Generalization of the Welch-Satterthwaite t-test (allowing for unequal variances).
    - Approximate Hotelling's T^2 test for multiple-contrast hypothesis tests (Tipton & Pustejovsky, 2015).
- These approximations work well ***even when*** J ***is small*** and even when the working model isn't correct.
- Degrees-of-freedom are *diagnostic*, so low d.f. implies:
    - little information available for variance estimation
    - asymptotic approximations haven't "kicked in"

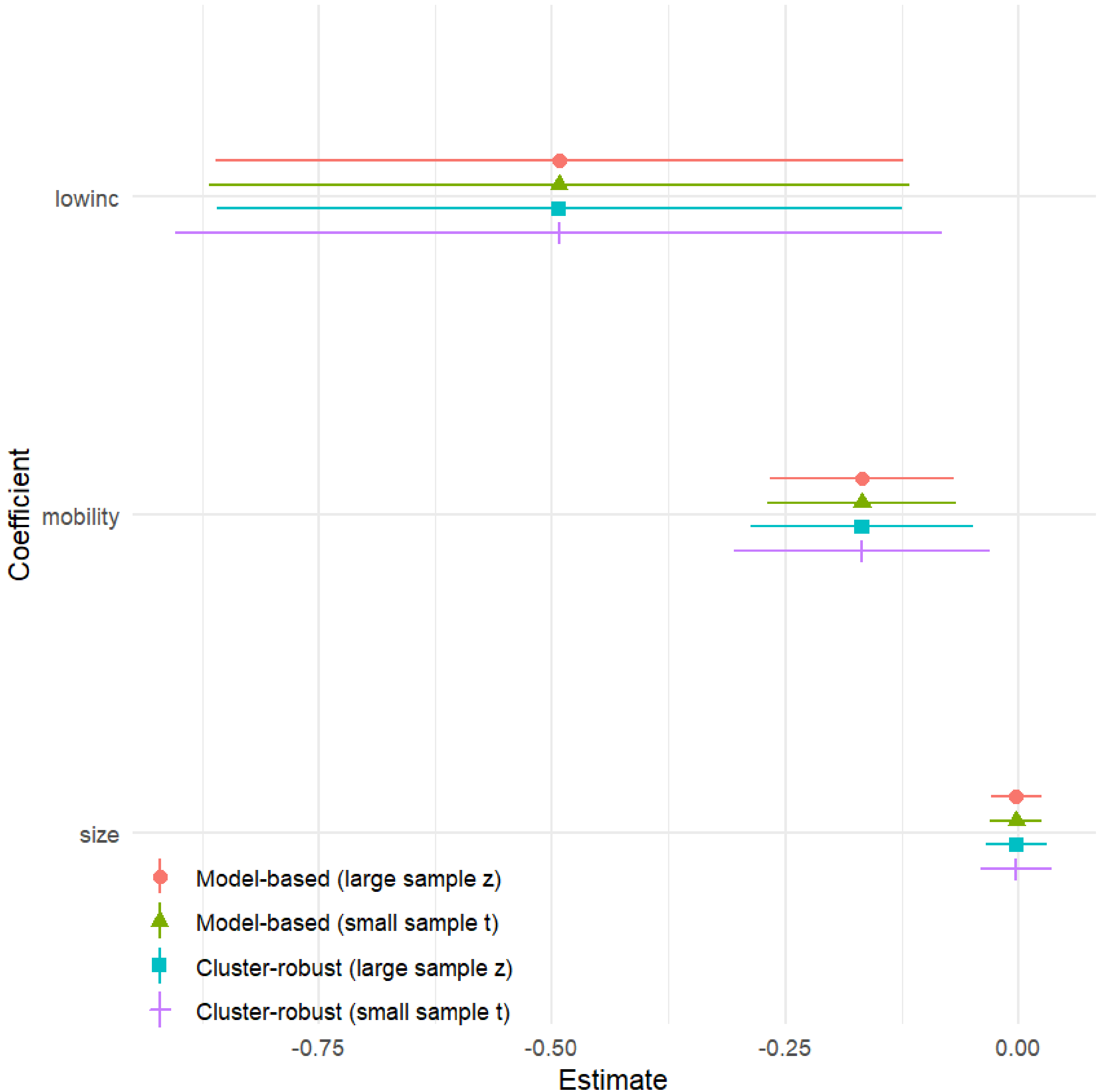
Plain vs. club sandwich estimators

```r
# cluster = egsingle$schoolid is automatically detected
coef_test(USSE_ri, vcov = "CR0", test = "z", coefs = 7:9)
```

```
##     Coef. Estimate     SE t-stat d.f. (z) p-val (z) Sig.
##      size -0.00214 0.0167 -0.128      Inf   0.89844
##    lowinc -0.49241 0.1875 -2.627      Inf   0.00862   **
##  mobility -0.16748 0.0606 -2.762      Inf   0.00575   **
```

```r
# "CR2" for small-sample adjustments
coef_test(USSE_ri, vcov = "CR2",
          test = "Satterthwaite", coefs = 7:9)
```

```
##     Coef. Estimate     SE t-stat d.f. (Satt) p-val (Satt) Sig.
##      size -0.00214 0.0181 -0.118        18.7       0.9072
##    lowinc -0.49241 0.1994 -2.469        25.3       0.0207    *
##  mobility -0.16748 0.0650 -2.575        17.8       0.0192    *
```


lowinc
mobility
size
Coefficient
Model-based (large sample z)
Model-based (small sample t)
Cluster-robust (large sample z)
Cluster-robust (small sample t)
-0.75
-0.50
-0.25
0.00
Estimate

Agenda

Basics of cluster-robust variance estimation (CRVE)

CRVE for multi-level models

Small-sample refinements

clubSandwich features

R package `clubSandwich`

Methods work with many sorts of regression models:

- linear regression with `stats::lm()`
- hierarchical linear models with `nlme::lme()` and `lme4::lmer()`
- generalized least squares with `nlme::gls()`
- logistic/generalized linear models with `glm()`
- multivariate regression with `mlm` objects
- instrumental variables with `AER::ivreg()`
- panel data models with `plm::plm()`
- meta-analysis with `metafor::rma()`, `metafor::rma.mv()`, `robumeta::robu()`

Object-oriented design for extensibility.

Under active development

- Available on CRAN: https://cran.r-project.org/package=clubSandwich
- Package website: https://jepusto.github.io/clubSandwich/
- Development repo: https://github.com/jepusto/clubSandwich
- ***Pull requests welcome!***

Functions

- `vcovCR()` to calculate robust variance-covariance matrix.
- Hypothesis tests for single regression coefficients: `coef_test()`
- Confidence intervals
  - for single regression coefficients: `conf_int()`
  - for linear combinations of coefficients: `linear_contrast()`
- Wald-tests for multi-parameter constraints (i.e., robust ANOVA/F-tests): `Wald_test()`

- Confidence intervals for linear combinations of coefficients:

```
linear_contrast(USSE_ri, vcov = "CR2",
                contrasts = constrain_pairwise(1:4))
```

```
##              Coef. Estimate     SE d.f. Lower 95% CI Upper 95% CI
##  grade1 - grade0    1.138 0.0474 40.7        1.042        1.234
##  grade2 - grade0    1.801 0.0479 40.6        1.704        1.898
##  grade3 - grade0    2.447 0.0584 40.5        2.329        2.565
##  grade2 - grade1    0.663 0.0388 40.9        0.584        0.741
##  grade3 - grade1    1.309 0.0526 40.5        1.202        1.415
##  grade3 - grade2    0.646 0.0362 40.7        0.573        0.719
```

- Wald-tests for multi-parameter constraints (i.e., robust ANOVA/F-tests):

```
Wald_test(USSE_ri,
          constraints = constrain_equal(1:6),
          vcov = "CR2", test = "HTZ")
```

```
##  test Fstat df_num df_denom  p_val sig
##   HTZ   468      5     24.7 <0.001 ***
```

Final thoughts

- Using robust methods simplifies analysis plans.
  - Great strategy for pre-registration/registered report!
- Other inferential methods may have advantages for very small samples, especially for multi-parameter hypothesis tests.
  - Cluster wild bootstrap (Cameron, Gelbach, & Miller, 2011) and randomization inference (Wu & Ding, 2020; Su & Ding, 2021).
  - Joshi, Pustejovsky, and Beretvas (2022) study cluster wild bootstrapping for meta-analysis.

Thanks!

pustejovsky@wisc.edu

http://jepusto.com

References

- Bell, R. M., & McCaffrey, D. F. (2002). *Survey Methodology.* http://www.statcan.gc.ca/pub/12-001-x/2002002/article/9058-eng.pdf
- Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). *Review of Economics and Statistics.* https://doi.org/10.1162/rest.90.3.414
- Cameron, A. C., & Miller, D. L. (2015). *Journal of Human Resources.* https://doi.org/10.3368/jhr.50.2.317
- Carpenter, C., & Dobkin, C. (2011). *Journal of Economic Perspectives.* http://doi.org/10.1257/jep.25.2.133
- Joshi, M., Pustejovsky, J. E., & Beretvas, S. N. (2022). https://osf.io/preprints/metaarxiv/x6uhk
- McCaffrey, D. F., & Bell, R. M. (2006). *Statistics in Medicine.* http://doi.org/10.1002/sim.2502
- Pustejovsky, J. E. & Tipton, E. (2018). *Journal of Business and Economic Statistics.* https://doi.org/10.1080/07350015.2016.1247004
- Su, F., & Ding, P. (2022). https://arxiv.org/abs/2104.04647
- Tipton, E., & Pustejovsky, J. E., (2015). *Journal of Educational and Behavioral Statistics.* http://doi.org/10.3102/1076998615606099
- Wu, J., & Ding., P. (2021). https://arxiv.org/abs/1809.07419
- Zeileis, A., Köll, S., & Graham, N. (2020). *Journal of Statistical Software.* https://doi.org/10.18637/jss.v095.i01