Corrigendum: Small sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models

James E. Pustejovsky *
University of Wisconsin - Madison
and
Elizabeth Tipton †
Northwestern University

January 12, 2023

Abstract

Pustejovsky and Tipton (2018) considered how to implement cluster-robust variance estimators for fixed effects models estimated by weighted (or unweighted) least squares. Theorem 2 of the paper concerns a computational short cut for a certain cluster-robust variance estimator in models with cluster-specific fixed effects. It claimed that this short cut works for models estimated by generalized least squares, as long as the weights are taken to be inverse of the working model. However, the theorem is incorrect. In this corrigendum, we review the CR2 variance estimator, describe the assertion of the theorem as originally stated, and explain the error. We then provide a revised version of the theorem, which holds only under more limited conditions, for models estimated by ordinary least squares.

Keywords:

*Department of Educational Psychology, University of Wisconsin - Madison, 1025 West Johnson Street, Madison, WI 53706. Email: pustejovsky@wisc.edu

†Department of Statistics, Northwestern University. Email: tipton@northwestern.edu

1 A fixed effects model

For data that can be grouped into m clusters of observations, Pustejovsky and Tipton (2018) considered the model

$$\mathbf{y}_i = \mathbf{R}_i\boldsymbol{\beta} + \mathbf{S}_i\boldsymbol{\gamma} + \mathbf{T}_i\boldsymbol{\mu} + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\mathbf{y}_i$ is an $n_i \times 1$ vector of responses for cluster i, $\mathbf{R}_i$ is an $n_i \times r$ matrix of focal predictors, $\mathbf{S}_i$ is an $n_i \times s$ matrix of additional covariates that vary across multiple clusters, and $\mathbf{T}_i$ is an $n_i \times t$ matrix encoding cluster-specific fixed effects, all for $i = 1, ..., m$. The cluster-specific fixed effects satisfy $\mathbf{T}_h\mathbf{T}_i' = \mathbf{0}$ for $h \neq i$. Interest centers on inference for the coefficients on the focal predictors $\boldsymbol{\beta}$.

Pustejovsky and Tipton (2018) considered estimation of Model 1 by weighted least squares (WLS). Let $\mathbf{W}_1, ..., \mathbf{W}_m$ be a set of symmetric weight matrices used for WLS estimation. The CR2 variance estimator involves specifying a working model for the structure of the errors. Consider a working model $\text{Var}\left(\boldsymbol{\epsilon}_i|\mathbf{R}_i, \mathbf{S}_i, \mathbf{T}_i\right) = \sigma^2\boldsymbol{\Phi}_i$ where $\boldsymbol{\Phi}_i$ is a symmetric $n_i \times n_i$ matrix that may be a function of a low-dimensional, estimable parameter. In some applications, the weight matrices might be taken as $\mathbf{W}_i = \hat{\boldsymbol{\Phi}}_i^{-1}$, where $\hat{\boldsymbol{\Phi}}_i$ is an estimate of $\boldsymbol{\Phi}_i$. In other applications, the weight matrices may be something else, such as diagonal matrices consisting of sampling weights or identity matrices (i.e., ordinary least squares).

2 The CR2 variance estimator

Pustejovsky and Tipton (2018) provided a generalization of the bias-reduced linearization estimator introduced by McCaffrey, Bell, and Botts (2001) and Bell and McCaffrey (2002) that can be applied to Model 1, referred to as the CR2 variance estimator. We follow the same notation as Pustejovsky and Tipton (2018) to define CR2. Let $N = \sum_{i=1}^m n_i$ be the total sample size. Let $\mathbf{U}_i = [\mathbf{R}_i\ \mathbf{S}_i]$ be the set of predictors that vary across clusters and $\mathbf{X}_i = [\mathbf{R}_i\ \mathbf{S}_i\ \mathbf{T}_i]$ be the full set of predictors. Let $\mathbf{R}$, $\mathbf{S}$, $\mathbf{T}$, $\mathbf{U}$, $\mathbf{X}$, and $\mathbf{y}$ denote the stacked versions of the cluster-specific matrices (i.e., $\mathbf{R} = [\mathbf{R}_1'\ \mathbf{R}_2'\ \cdots\ \mathbf{R}_m']'$, etc.). Let $\mathbf{W} = \bigoplus_{i=1}^m \mathbf{W}_i$ and $\boldsymbol{\Phi} = \bigoplus_{i=1}^m \boldsymbol{\Phi}_i$. For a generic matrix $\mathbf{Z}$, let $\mathbf{M}_Z = (\mathbf{Z}'\mathbf{W}\mathbf{Z})^{-1}$ and $\mathbf{H}_\mathbf{Z} = \mathbf{Z}\mathbf{M}_\mathbf{Z}\mathbf{Z}'\mathbf{W}$. Let $\mathbf{C}_i$ be the $n_i \times N$ matrix that selects the rows of cluster i from the

full set of observations, such that $\mathbf{X}_i = \mathbf{C}_i\mathbf{X}$. Finally, let $\mathbf{D}_i$ be the upper-right Cholesky factorization of $\mathbf{\Phi}_i$.

These operators provide a means to define absorbed versions of the predictors and the outcome. Let $\ddot{\mathbf{S}} = (\mathbf{I} - \mathbf{H_T})\,\mathbf{S}$ be the covariates after absorbing the cluster-specific effects, let $\ddot{\mathbf{U}} = (\mathbf{I} - \mathbf{H_T})\,\mathbf{U}$ be an absorbed version of the focal predictors and the covariates, let $\ddot{\mathbf{R}} = (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}})\,(\mathbf{I} - \mathbf{H_T})\,\mathbf{R}$ be the focal predictors after absorbing the covariates and the cluster-specific fixed effects, and let $\mathbf{e} = (\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{R}}})\,(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{S}}})\,(\mathbf{I} - \mathbf{H_T})\,\mathbf{y}$ denote the vector of residuals, with $\mathbf{e}_i = \mathbf{C}_i\mathbf{e}$.

With this notation established, the CR2 variance estimator has the form

$$\mathbf{V}^{CR2} = \mathbf{M}_{\ddot{\mathbf{R}}} \left(\sum_{i=1}^{m} \ddot{\mathbf{R}}_i' \mathbf{W}_i \mathbf{A}_i \mathbf{e}_i \mathbf{e}_i' \mathbf{A}_i \mathbf{W}_i \ddot{\mathbf{R}}_i \right) \mathbf{M}_{\ddot{\mathbf{R}}}, \tag{2}$$

where $\ddot{\mathbf{R}}_i = \mathbf{C}_i\ddot{\mathbf{R}}$ is the cluster-specific matrix of absorbed focal predictors, $\mathbf{e}_i$ is the vector of weighted least squares residuals from cluster i, and $\mathbf{A}_1, ..., \mathbf{A}_m$ are a set of adjustment matrices that correct the bias of the residual cross-products.

The adjustment matrices are calculated as follows. Define the matrices

$$\mathbf{B}_i = \mathbf{D}_i\mathbf{C}_i\,(\mathbf{I} - \mathbf{H_X})\,\mathbf{\Phi}\,(\mathbf{I} - \mathbf{H_X})'\,\mathbf{C}_i'\mathbf{D}_i' \tag{3}$$

for $i = 1, ..., m$. The adjustment matrices are then calculated as

$$\mathbf{A}_i = \mathbf{D}_i'\mathbf{B}_i^{+1/2}\mathbf{D}_i, \tag{4}$$

where $\mathbf{B}_i^{+1/2}$ is the symmetric square root of the Moore-Penrose inverse of $\mathbf{B}_i$. Theorem 1 of Pustejovsky and Tipton (2018) shows that, if the working model $\mathbf{\Phi}$ is correctly specified and some conditions on the rank of $\mathbf{U}$ are satisfied, then the CR2 estimator is exactly unbiased for the sampling variance of the weighted least squares estimator of $\boldsymbol{\beta}$. Moreover, although the CR2 estimator is defined based on a working model, it remains close to unbiased and outperforms alternative sandwich estimators even when the working model is not correctly specified.

3 The original statement of Theorem 2

The adjustment matrices given in (4) can be expensive to compute directly because the $\mathbf{B}_i$ matrices involve computing a "residualized" version of the $N \times N$ matrix $\mathbf{\Phi}$ involving the

full set of predictors $\mathbf{X}$—including the cluster-specific fixed effects $\mathbf{T}_1, ..., \mathbf{T}_m$. Theorem 2 considered whether one can take a computational short cut by omitting the cluster-specific fixed effects from the calculation of the $\mathbf{B}_i$ matrices. Specifically, define the modified matrices

$$\tilde{\mathbf{B}}_i = \mathbf{D}_i \mathbf{C}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right) \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)' \mathbf{C}_i' \mathbf{D}_i' \tag{5}$$

and

$$\tilde{\mathbf{A}}_i = \mathbf{D}_i' \tilde{\mathbf{B}}_i^{+1/2} \mathbf{D}_i. \tag{6}$$

Theorem 2 claimed that if the weight matrices are inverse of the working model, such that $\mathbf{W}_i = \mathbf{\Phi}_i^{-1}$ for $i = 1, ..., m$, then $\tilde{\mathbf{B}}_i^{+1/2} = \mathbf{B}_i^{+1/2}$ and hence $\tilde{\mathbf{A}}_i = \mathbf{A}_i$. The implication is that the cluster-specific fixed effects can be ignored when calculating the adjustment matrices. However, the claimed equivalence does not actually hold. The proof of Theorem 2 as given in the supplementary materials of Pustejovsky and Tipton (2018) relied on a Woodbury identity for generalized inverses that does not hold for $\mathbf{B}_i$ because necessary rank conditions are not satisfied.

We describe a simple numerical example that contradicts the original statement of Theorem 2. Consider a block-randomized experiment with $m = 3$ blocks, of sizes $n_1 = 2$, $n_2 = 4$, and $n_3 = 6$. In each block, a single observation receives treatment and the remaining observations receive the control condition, so that $\mathbf{R}_i = [1\ 0 \cdots 0]'$, $\mathbf{T}$ consists of indicators for each block, and there are no additional covariates. The model is then

$$\mathbf{y}_i = \mathbf{R}_i \beta + \mu_i + \boldsymbol{\epsilon}_i$$

Assume that treatment impacts are heterogeneous such that $\text{Var}(\boldsymbol{\epsilon}_{1i}) = \sigma^2 + \tau^2$ and $\text{Var}(\boldsymbol{\epsilon}_{ji}) = \sigma^2$ for $j > 1$, where τ^2 is known. Consider estimating the model using weighted least squares with inverse-variance weights.

```
library(clubSandwich)
```

```
## Registered S3 method overwritten by 'clubSandwich':
##   method    from
##   bread.mlm sandwich
```

```r
ni <- c(2,4,6)
m <- length(ni)
ri <- lapply(ni, \(x) c(1L, rep(0L, x - 1L)))
wi <- lapply(ri, \(r) length(r) * (r + (1 - r) / (length(r) - 1)))
rddi_ols <- lapply(ri, \(r) r - mean(r))
rddi_wt <- lapply(ri, \(r) r - 1 / 2)
clust <- rep(LETTERS[1:m], ni)
yi <- rnorm(sum(ni))
yddi <- tapply(yi, clust, \(x) x - mean(x))
dat <- data.frame(
  y = yi,
  y_dd = unlist(yddi),
  R = unlist(ri),
  R_dd_ols = unlist(rddi_ols),
  R_dd_wt = unlist(rddi_wt),
  w = unlist(wi),
  clust = clust
)

ols_fit <- lm(y ~ clust + R, data = dat)
A_ols_full <- attr(vcovCR(ols_fit, type = "CR2", cluster = dat$clust),"adjustments")
emat_ols_full <- mapply(\(r, a) a %*% r, r = rddi_ols, a = A_ols_full)

ols_absorbed <- lm(y_dd ~ 0 + R_dd_ols, data = dat)
A_ols_absorb <- attr(vcovCR(ols_absorbed, type = "CR2", cluster = dat$clust), "adjustmen
emat_ols_absorb <- mapply(\(r, a) a %*% r, r = rddi_ols, a = A_ols_absorb)
# si_ols_absorb <- mapply(\(x,e) sum(x * e), x = emat_ols_absorb, e = split(residuals(
# MR_ols <- 1 / sum(sapply(rddi_ols, \(r) sum(r^2)))
# MR_ols * si_ols_absorb
# vcovCR(ols_absorbed, type = "CR2", cluster = dat£clust, form = "estfun")
```

```r
# sum(si_ols_absorb^2) * MR_ols^2
# vcovCR(ols_absorbed, type = "CR2", cluster = dat$clust)
# vcovCR(ols_fit, type = "CR2", cluster = dat$clust)["R","R"]

wls_fit <- lm(y ~ clust + R, weights = w, data = dat)
A_wt_full <- attr(vcovCR(wls_fit, type = "CR2", cluster = dat$clust),"adjustments")
emat_wt_full <- mapply(\(r, w, a) a %*% (w * r), r = rddi_wt, w = wi, a = A_wt_full)
# si_wt_full <- mapply(\(x,e) sum(x * e), x = emat_wt_full, e = split(residuals(wls_fit
# MR_wt <- 1 / sum(mapply(\(r,w) sum(w * r^2), r = rddi_wt, w = wi))
# MR_wt * sum(si_wt_full^2) * MR_wt
# vcovCR(wls_fit, type = "CR2", cluster = dat$clust)["R","R"]

wls_absorbed <- lm(y_dd ~ 0 + R_dd_wt, weights = w, data = dat)
A_wt_absorb <- attr(vcovCR(wls_absorbed, type = "CR2", cluster = dat$clust), "adjustment
emat_wt_absorb <- mapply(\(r, w, a) a %*% (w * r), r = rddi_wt, w = wi, a = A_wt_absorb)
# si_wt_absorb <- mapply(\(x,e) sum(x * e), x = emat_wt_absorb, e = split(residuals(wls
# MR_wt * sum(si_wt_absorb^2) * MR_wt
# vcovCR(wls_absorbed, type = "CR2", cluster = dat$clust)

data.frame(
  ols_full = unlist(emat_ols_full),
  ols_absorb = unlist(emat_ols_absorb),
  wt_full = unlist(emat_wt_full),
  wt_absorb = unlist(emat_wt_absorb)
)
```

```
##      ols_full ols_absorb   wt_full  wt_absorb
## 1   0.5735393  0.5735393  1.043413  1.0434125
## 2  -0.5735393 -0.5735393 -1.043413 -1.0434125
## 3   0.9375000  0.9375000  1.026980  2.3898897
## 4  -0.3125000 -0.3125000 -1.026980 -0.7734339
```

```
## 5  -0.3125000 -0.3125000 -1.026980 -0.7734339
## 6  -0.3125000 -0.3125000 -1.026980 -0.7734339
## 7   1.0758287  1.0758287  1.150447  4.5774116
## 8  -0.2151657 -0.2151657 -1.150447 -1.0015292
## 9  -0.2151657 -0.2151657 -1.150447 -1.0015292
## 10 -0.2151657 -0.2151657 -1.150447 -1.0015292
## 11 -0.2151657 -0.2151657 -1.150447 -1.0015292
## 12 -0.2151657 -0.2151657 -1.150447 -1.0015292
```

4 A revised Theorem 2

The implication of the original Theorem 2 was that using the modified adjustment matrices $\tilde{\mathbf{A}}_i$ to calculate the CR2 estimator yields the same result as using the full adjustment matrices $\mathbf{A}_i$. Although this does not hold under the general conditions given above, a modified version of the theorem does hold for the more limited case of ordinary (unweighted) least squares regression with an "independence" working model. The precise conditions are given in the following theorem.

Theorem. *Let* $\mathbf{L}_i = \left(\ddot{\mathbf{U}}'\ddot{\mathbf{U}} - \ddot{\mathbf{U}}_i'\ddot{\mathbf{U}}_i\right)$ *and assume that* $\mathbf{L}_1, ..., \mathbf{L}_m$ *have full rank* $r + s$*. If* $\mathbf{W}_i = \mathbf{I}_i$ *and* $\mathbf{\Phi}_i = \mathbf{I}_i$ *for* $i = 1, ..., m$ *and* $\mathbf{T}_i\mathbf{T}_k' = \mathbf{0}$ *for* $i \neq k$*, then* $\mathbf{A}_i\ddot{\mathbf{R}}_i = \tilde{\mathbf{A}}_i\ddot{\mathbf{R}}_i$*, where* $\mathbf{A}_i$ *and* $\tilde{\mathbf{A}}_i$ *are as defined in (4) and (6), respectively.*

The implication of the revised theorem is that, for ordinary least squares regression with an "independence" working model, calculating the CR2 with the modified adjustment matrices $\tilde{\mathbf{A}}_i$ leads to the same result as using the full adjustment matrices $\mathbf{A}_i$. The equality does not hold for weighted or generalized least squares, nor for ordinary least squares with working models other than $\mathbf{\Phi}_i = \mathbf{I}_i$.

4.1 Proof

Setting $\mathbf{\Phi}_i = \mathbf{I}_i$ and observing that $\ddot{\mathbf{U}}_i'\mathbf{T}_i = \mathbf{0}$ for $i = 1, ..., m$, it follows that

$$\begin{aligned}
\mathbf{B}_i &= \mathbf{D}_i\mathbf{C}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right) \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{\Phi} \left(\mathbf{I} - \mathbf{H}_{\mathbf{T}}\right)' \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}}\right)' \mathbf{C}_i'\mathbf{D}_i' \\
&= \mathbf{C}_i \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}} - \mathbf{H}_{\mathbf{T}}\right) \left(\mathbf{I} - \mathbf{H}_{\ddot{\mathbf{U}}} - \mathbf{H}_{\mathbf{T}}\right) \mathbf{C}_i' \\
&= \left(\mathbf{I}_i - \ddot{\mathbf{U}}_i\mathbf{M}_{\ddot{\mathbf{U}}}\ddot{\mathbf{U}}_i' - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right) \qquad (7)
\end{aligned}$$

and similarly,

$$\tilde{\mathbf{B}}_i = \left(\mathbf{I}_i - \ddot{\mathbf{U}}_i\mathbf{M}_{\ddot{\mathbf{U}}}\ddot{\mathbf{U}}_i'\right). \qquad (8)$$

We now show that $\tilde{\mathbf{A}}_i\mathbf{T}_i = \mathbf{T}_i$. Denote the rank of $\ddot{\mathbf{U}}_i$ as $u_i \leq \min\{n_i, r+s\}$ and take the thin QR decomposition of $\ddot{\mathbf{U}}_i$ as $\ddot{\mathbf{U}}_i = \mathbf{Q}_i\mathbf{R}_i$, where $\mathbf{Q}_i$ is an $n_i \times u_i$ semi-orthonormal matrix and $\mathbf{R}_i$ is a $u_i \times r+s$ matrix of rank u_i, with $\mathbf{Q}_i'\mathbf{Q}_i = \mathbf{I}$. Note that $\mathbf{Q}_i'\mathbf{T}_i = \mathbf{0}$. From the observation that $\tilde{\mathbf{B}}_i$ can be written as

$$\tilde{\mathbf{B}}_i = \mathbf{I}_i - \mathbf{Q}_i\mathbf{Q}_i' + \mathbf{Q}_i \left(\mathbf{I} - \mathbf{R}_i\mathbf{M}_{\ddot{\mathbf{U}}}\mathbf{R}_i'\right)\mathbf{Q}_i',$$

it can be seen that

$$\tilde{\mathbf{A}}_i = \tilde{\mathbf{B}}_i^{+1/2} = \mathbf{I}_i - \mathbf{Q}_i\mathbf{Q}_i' + \mathbf{Q}_i \left(\mathbf{I} - \mathbf{R}_i\mathbf{M}_{\ddot{\mathbf{U}}}\mathbf{R}_i'\right)^{+1/2}\mathbf{Q}_i'. \qquad (9)$$

It follows that $\tilde{\mathbf{A}}_i\mathbf{T}_i = \mathbf{T}_i$.

Setting

$$\mathbf{A}_i = \tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i', \qquad (10)$$

observe that

$$\begin{aligned}
\mathbf{B}_i\mathbf{A}_i\mathbf{B}_i\mathbf{A}_i &= \left(\tilde{\mathbf{B}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right)\left(\tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right)\left(\tilde{\mathbf{B}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right)\left(\tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right) \\
&= \left(\tilde{\mathbf{B}}_i\tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right)\left(\tilde{\mathbf{B}}_i\tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right) \\
&= \left(\tilde{\mathbf{B}}_i\tilde{\mathbf{A}}_i\tilde{\mathbf{B}}_i\tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right) \\
&= \left(\tilde{\mathbf{B}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right) \\
&= \mathbf{B}_i.
\end{aligned}$$

It follows that $\mathbf{A}_i$ is the symmetric square root of the Moore-Penrose inverse of $\mathbf{B}_i$, i.e., $\mathbf{A}_i = \mathbf{B}_i^{+1/2}$. Finally, because $\mathbf{T}_i'\ddot{\mathbf{R}}_i = \mathbf{0}$, it can be seen that $\mathbf{A}_i\ddot{\mathbf{R}}_i = \left(\tilde{\mathbf{A}}_i - \mathbf{T}_i\mathbf{M}_{\mathbf{T}}\mathbf{T}_i'\right)\ddot{\mathbf{R}}_i = \tilde{\mathbf{A}}_i\ddot{\mathbf{R}}_i$.

References

Bell, Robert M, and Daniel F McCaffrey. 2002. "Bias reduction in standard errors for linear regression with multi-stage samples." *Survey Methodology* 28 (2): 169–81.

McCaffrey, Daniel F, Robert M Bell, and Carsten H Botts. 2001. "Generalizations of biased reduced linearization." In *Proceedings of the Annual Meeting of the American Statistical Association*. 1994.

Pustejovsky, James E, and Elizabeth Tipton. 2018. "Small-Sample Methods for Cluster-Robust Variance Estimation and Hypothesis Testing in Fixed Effects Models." *Journal of Business & Economic Statistics* 36 (4): 672–83. https://doi.org/10.1080/07350015.2016.1247004.