

Robust Standard Errors in Small Samples

Michal Kolesár

October 03, 2023

Contents

Description	1
Methods	3
Variance estimate	4
Degrees of freedom correction	5
Proof of Lemma 1	6

Description

This package implements small-sample degrees of freedom adjustments to robust and cluster-robust standard errors in linear regression, as discussed in Imbens and Kolesár [2016]. The implementation can handle models with fixed effects, and cases with a large number of observations or clusters¹.

```
library(dfadjust)
```

To give some examples, let us construct an artificial dataset with 11 clusters

```
set.seed(7)
d1 <- data.frame(y = rnorm(1000), x1 = c(rep(1, 3), rep(0,
  997)), x2 = c(rep(1, 150), rep(0, 850)), x3 = rnorm(1000),
  cl = as.factor(c(rep(1:10, each = 50), rep(11, 500))))
```

Let us first run a regression of y on x_1 . This is a case in which, in spite of moderate data size, the effective number of observations is small since there are only three treated units:

```
r1 <- lm(y ~ x1, data = d1)
## No clustering
dfadjustSE(r1)
#>
#> Coefficients:
#>           Estimate HC1 se HC2 se Adj. se      df p-value
#> (Intercept)  0.00266 0.0311  0.031  0.0311 996.00  0.932
#> x1          0.12940 0.8892  1.088  2.3743   2.01  0.916
```

¹We thank Ulrich Müller for suggesting to us the lemma below.

We can see that the usual robust standard errors (HC1 se) are much smaller than the effective standard errors (Adj. se), which are computed by taking the HC2 standard errors and applying a degrees of freedom adjustment.

Now consider a cluster-robust regression of y on x_2 . There are only 3 treated clusters, so the effective number of observations is again small:

```
r1 <- lm(y ~ x2, data = d1)
# Default Imbens-Kolesár method
dfadjustSE(r1, clustervar = d1$c1)
#>
#> Coefficients:
#>               Estimate HC1 se HC2 se Adj. se   df p-value
#> (Intercept)  -0.0236 0.0135 0.0169  0.0222 4.94  0.2215
#> x2           0.1778 0.0530 0.0621  0.1157 2.43  0.0826
# Bell-McCaffrey method
dfadjustSE(r1, clustervar = d1$c1, IK = FALSE)
#>
#> Coefficients:
#>               Estimate HC1 se HC2 se Adj. se   df p-value
#> (Intercept)  -0.0236 0.0135 0.0169  0.0316 2.42  0.2766
#> x2           0.1778 0.0530 0.0621  0.1076 2.70  0.0731
```

Now, let us run a regression of y on x_3 , with fixed effects. Since we're only interested in x_3 , we specify that we only want inference on the second element:

```
r1 <- lm(y ~ x3 + c1, data = d1)
dfadjustSE(r1, clustervar = d1$c1, ell = c(0, 1, rep(0,
  r1$rank - 2)))
#>
#> Coefficients:
#>               Estimate HC1 se HC2 se Adj. se   df p-value
#> Estimate    0.0261 0.0463 0.0595  0.0928 3.23  0.688
dfadjustSE(r1, clustervar = d1$c1, ell = c(0, 1, rep(0,
  r1$rank - 2)), IK = FALSE)
#>
#> Coefficients:
#>               Estimate HC1 se HC2 se Adj. se   df p-value
#> Estimate    0.0261 0.0463 0.0595  0.0928 3.23  0.688
```

Finally, an example in which the clusters are large. We have 500,000 observations:

```
d2 <- do.call("rbind", replicate(500, d1, simplify = FALSE))
d2$y <- rnorm(length(d2$y))
r2 <- lm(y ~ x2, data = d2)
summary(r2)
#>
#> Call:
#> lm(formula = y ~ x2, data = d2)
#>
```

```

#> Residuals:
#>      Min       1Q   Median       3Q      Max
#> -5.073 -0.675  0.000  0.675  4.789
#>
#> Coefficients:
#>              Estimate Std. Error t value Pr(>|t|)
#> (Intercept) -0.000991   0.001535  -0.65    0.52
#> x2          -0.003590   0.003963  -0.91    0.37
#>
#> Residual standard error: 1 on 499998 degrees of freedom
#> Multiple R-squared:  1.64e-06,   Adjusted R-squared:  -3.59e-07
#> F-statistic: 0.821 on 1 and 5e+05 DF,  p-value: 0.365
# Default Imbens-Kolesár method
dfadjustSE(r2, clustervar = d2$c1)
#>
#> Coefficients:
#>              Estimate HC1 se  HC2 se Adj. se   df p-value
#> (Intercept) -0.000991 0.00133 0.00168 0.00294 2.66   0.603
#> x2          -0.003590 0.00483 0.00568 0.00997 2.65   0.578
# Bell-McCaffrey method
dfadjustSE(r2, clustervar = d2$c1, IK = FALSE)
#>
#> Coefficients:
#>              Estimate HC1 se  HC2 se Adj. se   df p-value
#> (Intercept) -0.000991 0.00133 0.00168 0.00315 2.42   0.607
#> x2          -0.003590 0.00483 0.00568 0.00984 2.70   0.577

```

Methods

This section describes the implementation of the Imbens and Kolesár [2016] and Bell and McCaffrey [2002] degrees of freedom adjustments.

There are S clusters, and we observe n_s observations in cluster s , for a total of $n = \sum_{s=1}^S n_s$ observations. We handle the case with independent observations by letting each observation be in its own cluster, with $S = n$. Consider the linear regression of a scalar outcome Y_i onto a p -vector of regressors X_i ,

$$Y_i = X_i' \beta + u_i, \quad E[u_i | X_i] = 0.$$

We're interested in inference on $\ell' \beta$ for some fixed vector $\ell \in \mathbb{R}^p$. Let X , u , and Y denote the design matrix, and error and outcome vectors, respectively. For any $n \times k$ matrix M , let M_s denote the $n_s \times k$ block corresponding to cluster s , so that, for instance, Y_s corresponds to the outcome vector in cluster s . For a positive semi-definite matrix M , let $M^{1/2}$ be a matrix satisfying $M^{1/2'} M^{1/2} = M$, such as its symmetric square root or its Cholesky decomposition.

Assume that

$$E[u_s u_s' | X] = \Omega_s, \quad \text{and} \quad E[u_s u_t' | X] = 0 \quad \text{if } s \neq t.$$

Denote the conditional variance matrix of u by Ω , so that Ω_s is the block of Ω corresponding to cluster s . We estimate $\ell' \beta$ using OLS. In R, the OLS estimator is computed via a QR decomposition,

$X = QR$, where $Q'Q = I$ and R is upper-triangular, so we can write the estimator as

$$\ell' \hat{\beta} = \ell' \left(\sum_s X_s' X_s \right)^{-1} \sum_s X_s' Y_s = \tilde{\ell}' \sum_s Q_s' Y_s, \quad \tilde{\ell} = R^{-1} \ell.$$

It has variance

$$V := \text{var}(\ell' \hat{\beta} \mid X) = \ell' (X' X)^{-1} \sum_s X_s' \Omega_s X_s (X' X)^{-1} \ell = \tilde{\ell}' \sum_s Q_s' \Omega_s Q_s \tilde{\ell}.$$

Variance estimate

We estimate V using a variance estimator that generalizes the HC2 variance estimator to clustering. Relative to the LZ2 estimator described in Imbens and Kolesár [2016], we use a slight modification that allows for fixed effects:

$$\hat{V} = \ell' (X' X)^{-1} \sum_s X_s' A_s \hat{u}_s \hat{u}_s' A_s' X_s (X' X)^{-1} \ell = \ell' R^{-1} \sum_s Q_s' A_s \hat{u}_s \hat{u}_s' A_s' Q_s R^{-1} \ell = \sum_{s=1}^S (\hat{u}_s' a_s)^2,$$

where

$$\hat{u}_s := Y_s - X_s \hat{\beta} = u_s - Q_s Q' u, \quad a_s = A_s' Q_s \tilde{\ell},$$

and A_s is a generalized inverse of the symmetric square root of $I - Q_s Q_s'$, the block of the hat matrix corresponding to cluster s . In presence of cluster-specific fixed effects, $I - Q_s Q_s'$ is not generally invertible, which necessitates taking a generalized inverse. So long as the vector ℓ doesn't load on these fixed effects, \hat{V} will be unbiased under homoskedasticity, as the next result, which slightly generalizes the Theorem 1 in Pustejovsky and Tipton [2018], shows.²

Lemma 1. *Suppose that $X = (W, L)$ is full rank, and suppose that the vector ℓ loads only on elements of W . Let \tilde{W} denote the residual from projecting W onto L , and suppose that for each cluster s , (i) $L_s' \tilde{W}_s = 0$ and that (ii) $\sum_{k=1}^S I(k \neq s) \tilde{W}_k' \tilde{W}_k$ is full rank. Then \hat{V} is unbiased under homoskedasticity.*

The proof is given in the last section. By definition of projection, L and \tilde{W} are orthogonal. Condition (i) of the lemma strengthens this requirement to orthogonality within each cluster. It holds if L corresponds to a vector of cluster fixed effects, or more generally if L contains cluster-specific variables. Condition (ii) ensures that after partialling out L , it is feasible to run leave-one-cluster-out regressions. Without clustering, the condition is equivalent to the requirement that the partial leverages associated with \tilde{W} are smaller than one.³

If the observations are independent, the vector of leverages $(Q_1' Q_1, \dots, Q_n' Q_n)$ can be computed directly using the `stats::hatvalues` function. In this case, we use this function to compute $A_i = 1/\sqrt{1 - Q_i' Q_i}$ directly, and we then compute $a_i = A_i Q_i' \tilde{\ell}$ using vector operations. For the case with clustering, computing an inverse of $I - Q_s Q_s'$ can be expensive or even infeasible if the cluster

²In particular, Pustejovsky and Tipton [2018] require A_s to be the pseudo-inverse. We show that it's sufficient for A_s to be a generalized inverse.

³To see this, let $H = \tilde{W}(\tilde{W}' \tilde{W})^{-1} \tilde{W}'$ denote the partial projection matrix. Since $H = H^2$,

$$H_{ii} - H_{ii}^2 = \sum_{j \neq i} H_{ij} H_{ji} = \tilde{W}_i' (\tilde{W}' \tilde{W})^{-1} [\sum_{j \neq i} \tilde{W}_j \tilde{W}_j'] (\tilde{W}' \tilde{W})^{-1} \tilde{W}_i.$$

so that $H_{ii} = 1$ iff $\sum_{j \neq i} \tilde{W}_j \tilde{W}_j'$ is reduced rank.

size n_s is large. We therefore use the following result, a version of which was suggested to us by Ulrich Müller, which allows us to compute a_s by computing a spectral decomposition of a $p \times p$ matrix.

Lemma 2. *Let $Q'_s Q_s = \sum_{i=1}^p \lambda_{is} r_{is} r'_{is}$ be the spectral decomposition of $Q'_s Q_s$. Then $a_s = Q_s D_s \tilde{\ell}$, where $D_s = \sum_{i: \lambda_i \neq 1} (1 - \lambda_i)^{-1/2} r_{is} r'_{is}$.*

The lemma follows from the fact that $I - Q_s Q'_s$ has eigenvalues $1 - \lambda_{is}$ and eigenvectors $Q_s r_{is}$. More precisely, let $Q_s = \sum_i \lambda_{is}^{1/2} u_{is} r'_{is}$ denote the singular value decomposition of Q_s , so that $I - Q_s Q'_s = \sum_i (1 - \lambda_{is}) u_{is} u'_{is}$, and we can take $A_s = \sum_{i: \lambda_{is} \neq 1} (1 - \lambda_{is})^{-1/2} u_{is} u'_{is}$. Then,

$$A'_s Q_s = \sum_{i: \lambda_i \neq 1} (1 - \lambda_{is})^{-1/2} \lambda_{is}^{1/2} u_{is} r'_{is} = Q_s \sum_{i: \lambda_i \neq 1} (1 - \lambda_{is})^{-1/2} r_{is} r'_{is},$$

where the second equality uses $Q_s r_{is} = \lambda_{is}^{1/2} u_{is}$.

Degrees of freedom correction

Let G be an $n \times S$ matrix with columns $(I - QQ')'_s a_s$. Then the Bell and McCaffrey [2002] adjustment sets the degrees of freedom to

$$f_{\text{BM}} = \frac{\text{tr}(G'G)^2}{\text{tr}((G'G)^2)}.$$

Since $(G'G)_{st} = a'_s (I - QQ')_s (I - QQ')'_t a_t = a_s (\mathbb{1}\{s = t\} - Q_s Q'_t) a_t$, the matrix $G'G$ can be efficiently computed as

$$G'G = \text{diag}(a'_s a_s) - BB' \quad B_{sk} = a'_s Q_{sk}.$$

Note that B is an $S \times p$ matrix, so that computing the degrees of freedom adjustment only involves $p \times p$ matrices:

$$f_{\text{BM}} = \frac{(\sum_s a'_s a_s - \sum_{s,k} B_{sk}^2)^2}{\sum_s (a'_s a_s)^2 - 2 \sum_{s,k} (a'_s a_s) B_{sk}^2 + \sum_{s,t} (B'_s B_t)^2}.$$

If the observations are independent, we compute B directly as $B \leftarrow a * Q$, and since a_i is a scalar, we have

$$f_{\text{BM}} = \frac{(\sum_i a_i^2 - \sum_{sk} B_{sk}^2)^2}{\sum_i a_i^4 - 2 \sum_i a_i^2 B'_i B_i + \sum_{i,j} (B'_i B_j)^2}.$$

The Imbens and Kolesár [2016] degrees of freedom adjustment instead sets

$$f_{\text{IK}} = \frac{\text{tr}(G' \hat{\Omega} G)^2}{\text{tr}((G' \hat{\Omega} G)^2)},$$

where $\hat{\Omega}$ is an estimate of the Moulton [1986] model of the covariance matrix, under which $\Omega_s = \sigma_\epsilon^2 I_{n_s} + \rho \iota_{n_s} \iota'_{n_s}$. Using simple algebra, one can show that in this case,

$$G' \Omega G = \sigma_\epsilon^2 \text{diag}(a'_s a_s) - \sigma_\epsilon^2 BB' + \rho (D - BF')(D - BF')',$$

where

$$F_{sk} = \iota'_{n_s} Q_{sk}, \quad D = \text{diag}(a'_s \iota_{n_s})$$

which can again be computed even if the clusters are large. The estimate $\hat{\Omega}$ replaces σ_ϵ^2 and ρ with analog estimates.

Proof of Lemma 1

The estimator of the block of V associated with W implied by \hat{V} is given by

$$(\ddot{W}'\ddot{W})^{-1} \sum_s \ddot{W}'_s A_s (I - QQ')_s u u' (I - QQ')'_s A'_s \ddot{W}_s (\ddot{W}'\ddot{W})^{-1},$$

which is unbiased under homoskedasticity if for each s ,

$$\ddot{W}'_s A_s (I - Q_s Q'_s) A'_s \ddot{W}_s = \ddot{W}'_s \ddot{W}_s. \quad (1)$$

We will show that (1) holds. To this end, we first claim that under conditions (i) and (ii), \ddot{W}_s is in the column space of $I - Q_s Q'_s$ (a claim that's trivial if this matrix is full rank). Decompose $I - QQ' = I - H_{\ddot{W}} - H_L$, where $H_{\ddot{W}}$ and H_L are hat matrices associated with \ddot{W} and L . The block associated with cluster s can thus be written as $I - Q_s Q'_s = I - L_s (L' L)^{-1} L'_s - \ddot{W}_s (\ddot{W}' \ddot{W})^{-1} \ddot{W}'_s$. Let $B_s = \ddot{W}_s (\ddot{W}' \ddot{W} - \ddot{W}'_s \ddot{W}_s)^{-1} \ddot{W}'_s \ddot{W}$, which is well-defined under condition (ii). Then, using condition (i), we get

$$\begin{aligned} (I - Q_s Q'_s) B_s &= (I - \ddot{W}_s (\ddot{W}' \ddot{W})^{-1} \ddot{W}'_s) B_s \\ &= \ddot{W}_s (I - (\ddot{W}' \ddot{W})^{-1} \ddot{W}'_s \ddot{W}_s) (\ddot{W}' \ddot{W} - \ddot{W}'_s \ddot{W}_s)^{-1} \ddot{W}'_s \ddot{W} = \ddot{W}_s, \end{aligned}$$

proving the claim. Letting C denote the symmetric square root of $I - Q_s Q'_s$, the left-hand side of (1) can therefore be written as

$$\ddot{W}'_s A_s (I - Q_s Q'_s) A'_s \ddot{W}_s = B'_s C C A_s C C A'_s C C B_s = \ddot{W}'_s \ddot{W}_s,$$

where the second equality follows by the definition of a generalized inverse, since $A_s = C^-$.

References

- Robert M. Bell and Daniel F. McCaffrey. Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, 28(2):169–181, December 2002.
- Guido W. Imbens and Michal Kolesár. Robust standard errors in small samples: Some practical advice. *Review of Economics and Statistics*, 98(4):701–712, October 2016. doi: 10.1162/REST_a_00552.
- Brent R. Moulton. Random group effects and the precision of regression estimates. *Journal of Econometrics*, 32(3):385–397, August 1986. doi: 10.1016/0304-4076(86)90021-7.
- James E. Pustejovsky and Elizabeth Tipton. Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36(4):672–683, October 2018. doi: 10.1080/07350015.2016.1247004.