# Inference for Clustered Data

Chang Hyung Lee and Douglas G. Steigerwald
Department of Economics
University of California, Santa Barbara

November 6, 2017

**Abstract**

This article introduces `clusteff`, a new Stata command for checking the severity of cluster heterogeneity in cluster robust analyses. Cluster heterogeneity can cause a size distortion leading to underrejection of the null hypothesis. Carter, Schnepel, and Steigerwald (2015) develop the *effective* number of clusters to reflect a reduction in the degrees of freedom, thereby mirroring the distortion caused by assuming homogenous clusters. `clusteff` generates the effective number of clusters. We provide a decision tree for cluster robust analysis, demonstrate the use of `clusteff`, and recommend methods to minimize the size distortion.

## 1 Model

The basic setting is to consider a specification, for $n$ observations grouped into $G$ clusters, of the form

$$y_{ig} = x_{ig}^T \beta + u_{ig} \tag{1}$$

where observation $i$ belongs to cluster $g$ with $n_g$ observations in cluster $g$. We assume $\mathbb{E}[u_{ig}|x_{ig}] = 0$, so that (1) captures the conditional mean of $y_{ig}$. The error term $u_{ig}$ is allowed to have arbitrary correlation within a cluster, where $\Omega_g$ is the covariance matrix for cluster $g$ conditional on $x_g$, but is assumed to be independent across clusters. In this paper, we provide a Stata program that estimates the effective number of clusters, a diagnostic tool used to measure severity of cluster heterogeneity (including lack of balance in the covariate matrix) derived by Carter, Schnepel, and Steigerwald (CSS) (2015).

The question of interest is to test the null hypothesis $H_0 : a^T\beta = a^T\beta_0$, where $\beta_0$ is the value of $\beta$ under the null hypothesis and $a$ is a vector selecting the coefficients to be included in the test. We focus on the conventional test statistic constructed from $\widehat{\beta}$ - the OLS estimator of $\beta$ in (1):

$$t = \frac{a^T(\widehat{\beta} - \beta_0)}{\sqrt{a^T\widehat{V}a}} \tag{2}$$

where $\hat{V}$ is a cluster-robust estimator of $V$ - the variance of $\widehat{\beta}$ conditional on the covariate matrix $X$. The cluster-robust estimator of $V$ is

$$\widehat{V} = c(X^TX)^{-1}(\sum_{g=1}^{G} X_g^T\widehat{u}_g\widehat{u}_g^{\mathrm{T}}X_g)(X^TX)^{-1},$$

where $X_g$ and $u_g$ are the covariate matrix and error, respectively, for cluster $g$ and $c = \frac{G(n-1)}{(G-1)(n-k)}$ is designed to (partially) offset the downward bias in $\widehat{V}$.

The consistency of $\widehat{V}$ and the asymptotic normality of $t$ is established under general conditions in CSS (2015). As CSS describe, consistency of $\widehat{V}$ cannot be established simply by allowing the number of observations $n$ to grow but rather depends crucially on allowing the number of clusters $G$ to grow. To understand why this is so, consider a data set with a fixed number of clusters but an increasing number of observations in each cluster. As more observations are added to each cluster, the dimension of $\widehat{u}_g$ grows and more parameters are added to $\Omega_g$. In consequence $\widehat{u}_g\widehat{u}_g^{\mathrm{T}} := \widehat{\Omega}_g$ is not a consistent estimator of $\Omega_g$ and consistency of $\widehat{V}$ can only be obtained by averaging $\widehat{\Omega}_g$ over an increasingly large number of clusters. Thus the size of $G$ is often advocated as a guide to inference. According to this guide, if $G$ is large (say greater than 50), then the appropriate critical values to use when assessing $t$ are obtained from a normal distribution.

The standard practice of using $G$ as the sole criterion when selecting critical values relies on an assumption that clusters are homogenous in the sense that $\mathbb{E}\left(X_g^{\mathrm{T}}\Omega_g X_g\right)$ is identical over clusters. A sufficient condition for this assumption is that all clusters have the same: size, $n_g = \frac{n}{G}$; covariate matrices, $X_g$, that are identical over $g$; and covariance matrices, $\Omega_g$, that are identical over $g$. As these sufficient conditions rarely occur in practice, CSS investigate the behavior of $t$ when clusters are heterogeneous. They find that the test often falsely rejects (that is, the critical values from a normal distribution are too small) under cluster heterogeneity.

Importantly, CSS report a simple measure that can detect the extent to which cluster heterogeneity affects the test statistic. The measure adjusts the number of clusters downward to reflect the degree of cluster heterogeneity, such that the larger the amount of cluster heterogeneity, the greater the downward adjustment in the number of clusters. The resultant adjusted measure is the *effective number of clusters*. If the effective number of clusters is small, regardless of the magnitude of $G$, critical values that are larger than those from a normal distribution should be employed. These critical values may be obtained from a student's $t$ distribution or from bootstrapping, as explained below.

Observe that $V = \sum \gamma_g$ with $\gamma_g = a^T(X^TX)^{-1}(X_g^T\Omega_g X_g)(X^TX)^{-1}a$. Following CSS, we denote the effective number of clusters as $G^*$ and define it as

$$G^* = \frac{G}{1+\Gamma}, \qquad \Gamma = \frac{1}{G}\sum_{g=1}^{G}(\frac{\gamma_g - \overline{\gamma}}{\overline{\gamma}})^2, \tag{3}$$

with $\overline{\gamma} = G^{-1} \sum \gamma_g$. Simply put, cluster homogeneity requires $\gamma_g = \gamma$ for all clusters, so variation in $\gamma_g$ arises from cluster heterogeneity. If the clusters are homogenous, then $\Gamma = 0$ and $G^* = G$. If the clusters are heterogeneous, then $\Gamma > 0$ and $G^* < G$. A greater difference between $G^*$ and $G$ is indicative of more heterogeneous clusters.

Special attention to $a$, a selection vector of length $k$, is required here. The selection vector is derived from the hypothesis to be tested, $H_0 : a^T\theta = a^T\theta_0$. Consequently, a unique value of $G^*$ is generated based on each hypothesis to be tested. To be clear, the method is appropriate for tests of hypotheses on single coefficients, for example, $H_0 : \beta_1 = 0$, as well as linear combination of coefficients, $H_0 : \beta_1 + \beta_2 = 0$.

If $G^*$ is small, inference should be undertaken with care. CSS (2015) show that the test statistic using $\widehat{V}$ is normal as $G^* \to \infty$, which means the normal approximation should work well if $G^*$ is large. If $G^*$ is small, then the appropriate critical values are larger than those from a normal distribution, and mistakenly applying the normal critical values leads to incorrectly rejecting the null hypothesis far too often (the empirical size of the test exceeds the nominal size of the test). They find that the empirical size of a test to remains close to the nominal size using Gaussian critical values for $G^*$ greater than 25.

In practice $G^*$ must be estimated because it is a function of the unknown within-cluster error covariance matrix $\Omega$. Unfortunately, we cannot use the residuals to estimate $G^*$, because use of the residuals to construct both the critical values and the test statistic induces pre-test bias. Rather, $G^*$ is estimated by $G^{*A}$, which is constructed under the assumption of perfect within-cluster error correlation. (The estimation procedure for $G^{*A}$ employed by the Stata program is further discussed in the next section.) Because increasing the within-cluster correlation tends to increase cluster heterogeneity, the estimate $G^{*A}$ is designed to guard against this "worst-case scenario" in which the errors are perfectly correlated within clusters.

We recommend estimating $G^*$ as a first step in testing a model with a clustered error structure in order to credibly rule out size distortion from a small effective number of clusters. Application of the effective number of clusters need not be limited to small to moderate $G$ because a large $G$ does not guarantee $G^*$ to be large under cluster heterogeneity. CSS (2015) demonstrate the fallibility of assuming large $G^*$ based on large $G$ using the data set clustered at the industry level from Hersch (1998). The data set contains 5960 observations in 211 clusters. Conventional wisdom suggests that the number of clusters in this case is large enough to assume an approximately normal distribution for the test statistic. Calculating the effective number of clusters, however, reveals that the data set suffers from severe cluster heterogeneity with $G^{*A} = 19$, and the normal critical values are likely too small. In essence, variation in the covariate matrix across clusters yields substantial variability in the estimator of the standard error that appears in the denominator of the test statistic. Accounting for this variability enlarges the critical values. We also note that in applications where the key question of interest involves the response to treatment in specific clusters, the key criterion is not the overall value of $G^{*A}$, but rather the effective number of treated clusters (and the effective number of control clusters).

In Section 2 we detail the program. In Section 3 we follow with a decision tree for selecting the appropriate method of inference. We present an example on use of the decision tree in Section 4.

# 2 Program Specification

## 2.1 Syntax

```
clusteff varlist [if] [in] , cluster(varname) [test(varname) selection(string)
    noconstant covariance(real)]
```

## 2.2 Description

`clusteff` estimates the effective number of clusters ($G^*$) devised by Carter, Schnepel, and Steigerwald (2015) using a vector of independent variables, a clustering variable, and a selection vector. The program uses *varlist* as a list of variables to be included in the estimation procedure with the data clustered by the variable specified in the `cluster` option and the hypothesis test of interest defined by either the `selection` or `test` option.

## 2.3 Options

`cluster(`*varname*`)` states the clustering variable and must be specified.

`test(`*varname*`)` specifies a selection vector if the null hypothesis of interest involves a single covariate. Suppose a user aims to test the null hypothesis, $H_0 : \beta_2 = 0$, using a linear model of the following form: $y = \beta_0 + \beta_1 x + \beta_2 z + u$. Then,

```
clusteff x z, cluster(clustervar) test(z)
```

generates the relevant effective number of clusters.

`selection`(*string*) allows users to define their own selection vector. The input is a vector of values selecting the coefficients to be tested corresponding to the vector $a$ in the null hypothesis, $H_0 : a^T\beta = 0$. The order of covariates in *varlist* must match the order of elements in the selection vector. This option is especially useful if the null hypothesis of interest involves more than one covariate. For example, if a user is testing the null, $H_0 : \beta_1 + \beta_2 = 0$, stating

```
clusteff x z, cluster(clustervar) selection(1 1)
```

estimates the appropriate effective number of clusters.

The number of elements in a selection vector may not exceed the number of variables. The number of specified elements in a selection vector, however, is allowed to be smaller than the number of variables. The program fills empty elements with zeros such that `selection(1 0)` or `selection(1)` generate the effective number of coefficients under the null hypothesis, $H_0$ :

$\beta_1 = 0$.

`test` and `selection` options may not be specified simultaneously. If a user omits both the `test` and `selection` options, the program estimates an effective number of clusters under an assumption that the first variable in *varlist* is the covariate of interest. In the above example, omitting both of the options is equivalent to specifying `test(x)`, `selection(1)`, or `selection(1 0)`.

`noconstant` determines whether a linear model to be tested contains a vector of constants. If this option is specified, the program estimates an effective number of clusters without a vector of constants. Use this option when testing a linear model whose intercept is restricted at zero.

`covariance(`*real*`)` allows user to specify any real number between zero and one as the within-cluster covariance of the error used to estimate the effective number of clusters. If the option is left unspecified, the covariance between error terms within a cluster defaults to one.[1] The covariance of less than one estimates a less conservative effective number of clusters relative to the default in which perfect within-cluster error correlation is imposed.

## 2.4 Estimation Procedure

Generating a true value of an effective number of clusters ($G^*$) requires the underlying error structure, $u_g u_g^T$, to be known. Using residuals from a regression, $\hat{u}_g$, to construct critical values, however, renders a test invalid (Carter, Schnepel, and Steigerwald, 2015). Instead, CSS suggest using a 1-by-$n_g$ vector of ones, $\iota_g$, in place of $u_g$ to impose a perfect within-cluster error correlation as a conservative approach. `clusteff` uses the above estimation procedure to generate an estimate of $G^*$, $G^{*A}$, as outlined below.

[1] The program limits the maximum covariance at 0.9999 instead of 1 due to limits on floating value precision in MATA. This produces a more stable estimator compared to allowing perfect correlation.

$$G^{*A} = \frac{G}{1+\Gamma^A}$$

where $\Gamma^A = \frac{1}{G}\sum_{g=1}^{G}(\frac{\gamma_g^A - \bar{\gamma}^A}{\bar{\gamma}^A})^2$
and $\gamma_g^A = a^T(X^TX)^{-1}(X_g^T\iota_g\iota_g^TX_g)(X^TX)^{-1}a.$

Any valid input in `selection`(*string*) or `test`(*string*) is converted to a selection vector, $a$, used to generate $G^{*A}$. The program performs a matrix multiplication estimating a scalar value of $G^{*A}$.

## 3 Decision Tree

What is the correct approach for a practitioner with clustered data? As noted above, a key quantity in determining the best method of inference is the effective number of clusters. Thus, the decision begins with an estimate of this quantity for a given sample. If the estimated effective number of clusters, $G^{*A}$ is at least 25, then one should use the statistic (2) with critical values from a normal distribution. If $G^{*A}$ is less than 25, then a leading approach would be to use (2) but with critical values obtained in a different way. Cameron, Gelbach, and Miller (2008) and MacKinnon and Webb (2016) find that the wild bootstrap, which delivers critical values that are larger than those from a normal distribution, brings the empirical size of the test much closer to the nominal size.

Note, that for models where the coefficient of interest is a cluster-level treatment, $G^{*A}$ should be calculated separately for both the treated clusters and the control clusters. If either of these measures of $G^{*A}$ is less than 25, even if the overall effective number of clusters exceeds 50, then again the wild bootstrap could be used to obtain more accurate critical values.[2]

[2] With clusters identical to the size of U.S. states, MacKinnon and Webb (2016) show

The wild bootstrap begins by drawing, with replacement, from the collection of cluster residual vectors $\{\widehat{u}_g\}_{g=1}^{G}$. Each residual vector is multiplied by either 1 or $-1$ with equal probability. Then, the resultant residual vectors are combined with the observed regressors to produce bootstrap samples. Complete details are provided in Cameron, Gelbach, and Miller (2008), Cameron and Miller (2015), and MacKinnon and Webb (2016). A couple of user-written programs, `cgmwildboot` by Judson Caskey and `boottest` by David Roodman, can be used to generate $p$-values via wild bootstrap.[3]

For data sets that have a small effective number of clusters, either overall or within the treatment group (while rare, a similar issue arises if the control group has a small effective number of clusters) there are alternatives to the wild bootstrap. If interest centers on the coefficient of a covariate that varies within clusters, and there are a large number of observations in each cluster, then Ibragimov and Müller (2010) propose an alternative test statistic. To illustrate their method we first rewrite (1) to distinguish an observation-level covariate, $x_{ig}$ from a cluster-level covariate, $z_g$,

$$y_{ig} = \alpha + \beta x_{ig} + \delta z_g + u_{ig}. \tag{4}$$

The test statistic is derived by first estimating $\widehat{\beta}_g$ separately for each cluster. Note that $\alpha$ and $\delta$ are both absorbed in the cluster level intercept and so are not separately identified. The test statistic is

$$t_{IM} = \frac{\sqrt{G}\left(\overline{\widehat{\beta}} - \beta\right)}{s_{\widehat{\beta}}},$$

where $\overline{\widehat{\beta}} = \frac{1}{G}\sum_{g=1}^{G}\widehat{\beta}_g$ and $s^2 = \frac{1}{G-1}\sum_{g=1}^{G}\left(\widehat{\beta}_g - \overline{\widehat{\beta}}\right)^2$. Under the cluster

that severe under-rejection can occur if there are fewer than 7 treated or untreated clusters. Ferman and Pinto (2015) study the case of a small number of treated clusters in a difference-in-differences setting.

[3] `cgmwildboot` can be found at https://sites.google.com/site/judsoncaskey/data and `boottest` can be found at https://ideas.repec.org/c/boc/bocode/s458121.html.

assumption, $\widehat{\beta}_g$ is independent of $\widehat{\beta}_h$ and, if $n_g$ is sufficiently large, then $\widehat{\beta}_g$ has a normal asymptotic null distribution with mean $\beta$ and variance $\sigma_g^2$. Of course, if $\widehat{\beta}_g$ is a normal random variable and $\sigma_g^2 = \sigma^2$ then $t_{IM} \sim t(G-1)$. One would think that allowing $\sigma_g^2$ to vary would result in a test statistic with larger critical values than those from the student-$t\,(G-1)$. What is surprising is that for a test with nominal size of 5 percent, the critical values for $t_{IM}$ are *smaller* than the critical values from a student-$t\,(G-1)$. Thus combining $t_{IM}$ with the critical values from a $t\,(G-1)$ yields a test whose size will not exceed the nominal size of 5 percent. Note, such a result does not hold for a test with a nominal size of 10 percent, so selection of a nominal size of 5 percent is important. In comparing this method to the wild bootstrap, Ibragimov and Müller (2016) find that $t_{IM}$ is better at eliminating the size distortion for a very small number of heterogeneous clusters with large $n_g$.

If interest centers on the coefficient of a covariate that does not vary within clusters, and $n_g$ is large, then Donald and Lang (2007) propose an alternative test statistic. To illustrate their method begin with the regression (4) where the error has an error-components structure

$$u_{ig} = \rho_g + \epsilon_{ig}.$$

The first step is to construct the OLS fixed effects estimator from

$$y_{ig} = \beta x_{ig} + c_g + \epsilon_{ig},$$

yielding $\{\widehat{c}_g\}_{g=1}^{G}$. The second step is to construct the OLS estimator of $\beta$ from

$$\widehat{c}_g = a + \delta z_g + v_g,$$

yielding $\widehat{\delta}$. For the $H_0 : \delta = \delta_0$ the test statistic is

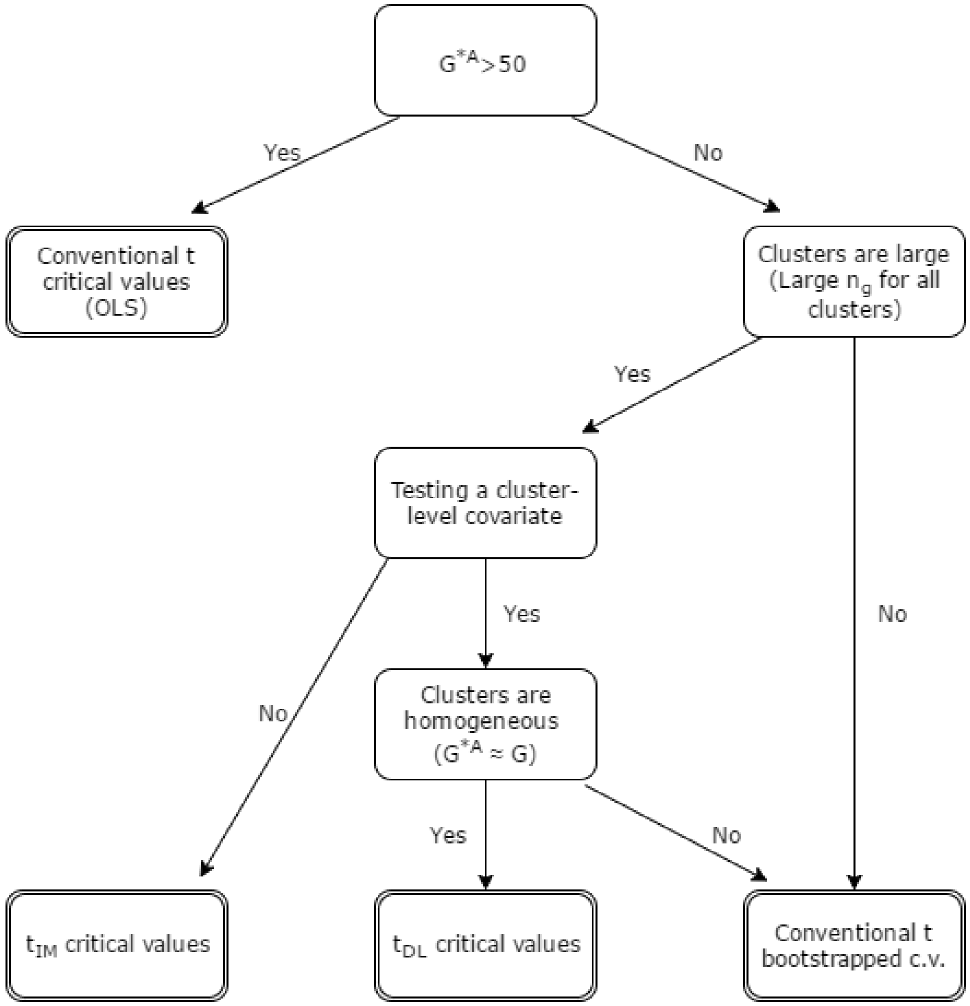
$$t_{DL} = \frac{(\hat{\delta} - \delta_0)}{s_{\widehat{\delta}}},$$

where $s^2_{\widehat{\delta}} = \frac{s^2}{\sum_{i=1}^{G}(z_g-\overline{z})^2}$ and $s^2 = \frac{1}{G-2}\sum_{g=1}^{G}(\widehat{v}_g\widehat{v}_g^T)$. The distribution of $t_{DL}$ is approximately student-$t\,(G-2)$, so again the critical values are larger than those from a normal distribution.

There are two caveats to the use of this test statistic. The first is that, as in the case of $t_{IM}$ the number of observations in each cluster must be large. The second is that, the distribution of the test statistic depends crucially on homogeneity across clusters (in essence, $n_g$ and $\overline{x}_g$ both identical across clusters). Thus, if $G^{*A}$ differs substantially from $G$, indicating that these homogeneity conditions do not hold, then it may not be appropriate to use $t_{DL}$.

MacKinnon and Webb (2016) investigate the relative performance of the wild bootstrap and $t_{DL}$. For data in which each cluster has 40 observations, but varying covariates across clusters, the wild bootstrap and $t_{DL}$ can have comparable empirical size. Importantly, the comparable size requires the use of $G^{*A}$ rather than $G$ when constructing the critical values from a student-$t$ distribution. In other words, if $t_{DL}$ is used with critical values from the $t(G-2)$ distribution, then the wild bootstrap outperforms it in the sense of more accurate size. A second set of simulations allow the cluster sizes to vary, together with varying covariates across clusters. In these models with more pronounced cluster heterogeneity, the wild bootstrap outperforms $t_{DL}$ and delivers the most accurate size.

In Figure 1 we provide a decision tree that encapsulates this discussion.



$G^{*A} > 50$
Yes
No
Conventional t critical values (OLS)
Clusters are large (Large $n_g$ for all clusters)
Yes
No
Testing a cluster-level covariate
Yes
No
Clusters are homogeneous ($G^{*A} \approx G$)
Yes
No
$t_{IM}$ critical values
$t_{DL}$ critical values
Conventional t bootstrapped c.v.

# 4 Example

We recommend using `clusteff` as a simple check to verify validity of analyses and to find an optimal method to use in order to minimize both the amount of computational power required and the size distortion. This section utilizes an example from the economics literature to demonstrate the use of `clusteff` in analysis of clustered samples.

## 4.1 Clustering at the State Level

Voena (2015) studies changes in the employment decisions of married women that result from the introduction of unilateral divorce laws. The introduction of unilateral divorce, under which divorce can be initiated without mutual consent of both partners, increases the probability of divorce. If women have fewer resources in divorce than in marriage, they may need to insure themselves against this potential loss of resources by working while married (thereby building their human capital). As states have different rules governing the distribution of property upon divorce, the strength of this effect is likely to vary across states. In states with "equitable distribution", under which women often have fewer resources after divorce, this effect is likely to be most pronounced. In states with community property, under which each partner gets an equal share of the resources, this effect is likely to be weaker. Female labor market participation, therefore, is likely to be more responsive to the divorce law reform in states with "equitable property" division.

To test the theory, a linear probability model is estimated for the labor force participation by women in household $i$, in state $s$, and year $t$. Key coefficients of interest are on the interaction covariates, which are indicators for whether state $s$ has unilateral divorce and (say) community property in year $t$. The corresponding component of the regression model is

$$\beta_1 \left(\{uni_{st}\} \cdot \{com_{st}\}\right) + \beta_2 \left(\{uni_{st}\} \cdot \{eq_{st}\}\right),$$

where $\{uni_{st}\}$ takes the value 1 if unilateral divorce is legal in state $s$ in year $t$, $\{com_{st}\}$ takes the value 1 if community property rules are used to govern divorce, and $\{eq_{st}\}$ takes the value 1 if equitable distribution rules are used to govern divorce. The individual hypotheses under test are $H_0 : \beta_i = 0$ $i = 1, 2$.

The conventional cluster robust $t$-statistic (2) is estimated, where clustering is at the state level. The number of clusters is 51, corresponding to the 50 states and the District of Columbia. The number of observations from each state varies widely, from 3 to 3,552. This large variation in cluster size indicates substantial cluster heterogeneity. As an initial indicator, we compute the effective number of clusters accounting only for variation in cluster sizes (that is, ignoring how the covariates change over clusters).[4] Such a calculation provides a quick indicator of the degree of cluster heterogeneity. For this data set, $G^{*A} = 13$, well below the cutoff for Gaussian inference. As noted above, this approximation of $G^*$ is likely to be conservative, as it is based on an intracluster correlation of 1. An alternative approximation, which assumes no intracluster correlation and so is much less conservative, can be constructed by replacing the unit matrix in $\gamma_g^A$ with the identity matrix. For this data set, this less conservative approximation yields $G^{*A} = 20$, again below the cutoff for Gaussian inference. All initial evidence points to the need to move away from the use of critical values from the normal distribution.

Because the form of the conditional expectation function is not known, Voena provides four regression approximations that differ in the number of controls (Table 2 columns 5-8, p. 2314). In the following table we present the OLS estimate and the cluster-robust standard error reported by Voena, followed by the bootstrapped confidence interval in brackets and the effective number of clusters.

[4] This computation corresponds to a test on the intercept.

Table 1: Replication Results

| | (1) | (2) | (3) | (4) |
|---|---|---|---|---|
| Variables | Employed | Employed | Employed | Employed |
| | | | | |
| Uni × ComProp | -0.0377* | -0.0389* | -0.0575** | -0.0488** |
| | (0.0164) | (0.0175) | (0.0175) | (0.0177) |
| $G^{*A}$ | 1.9191 | 1.9227 | 4.9457 | 5.0394 |
| Bootstrapped 95% C.I. | [-0.0868, 0.0056] | [-0.1096, 0.0073] | [-0.1205, -0.0204] | [-0.1181, -0.0092] |
| | | | | |
| Uni × EqDistr | -0.0279 | -0.0263 | -0.0265 | -0.0298 |
| | (0.0306) | (0.0314) | (0.0387) | (0.0414) |
| $G^{*A}$ | 4.9574 | 4.9630 | 13.3717 | 12.8005 |
| Bootstrapped 95% C.I. | [-0.1089, 0.0372] | [-0.1018, 0.0360] | [-0.1235, 0.0553] | [-0.1228, 0.0541] |
| | | | | |
| Year fixed effects | Yes | Yes | Yes | Yes |
| Age dummies | Yes | Yes | Yes | Yes |
| Children dummies | No | Yes | Yes | Yes |
| State fixed effects | No | No | Yes | Yes |
| Polyn yrs. married | No | No | No | Yes |
| Observations | 44,808 | 44,808 | 44,808 | 39,824 |
| Individual fixed effects | 3,437 | 3,437 | 3,437 | 2,607 |

Note: Replication of columns 5-8 from Table 2 of Voena (2015). Standard errors are clustered at the state level and critical values are generated by the wild bootstrap procedure with 1,000 replications. The third row estimates the effective number of clusters while the fourth row presents the wild bootstrap confidence interval between 2.5 and 97.5 percentiles.
*** p<0.01, ** p<0.05, * p<0.1

For each of the null hypotheses under test, the effective number of clusters is obtained within Stata using `clusteff`. For example, consider test of $\beta_1$ in column 1, for which the command is:

```
clusteff uni_comprop uni_title uni_eqdistr comprop eqdistr d_age
>> yrd* i.person, cluster(state) test(uni_comprop)
```

We list all covariates included in the model in *varlist*, specify *state* as clustering variable and include the null hypothesis to be tested. The program output is:

```
Number of clusters: 51
Estimated effective number of clusters: 1.919089
Warning: G* estimated to be below 50.
```

where the effective number of clusters corresponds to the coefficient under test.

With such a small value for $G^{*A}$, and such substantial cluster heterogeneity, from the decision tree there are two potential methods of inference. The first is to combine the standard test statistic $t$ with critical values obtained from the wild bootstrap. A second possibility, appropriate for regressors that vary within states, is to use $t_{IM}$ with critical values from the student-$t(50)$ distribution. To construct $t_{IM}$ we must be able to estimate $\beta_1$ and $\beta_2$ for each state separately. Yet for some states $\{uni_{st}\} \cdot \{com_{st}\}$ is always 0, rendering $\beta_1$ unidentified for these states.[5] Hence we report wild bootstrap critical values for $t$ below the approximations of $G^*$ in Table 1.

We use `boottest`, the aforementioned user-written program for STATA, to obtain the wild bootstrap critical values. The first line of the code runs a regression and the second line of the code performs wild bootstrap to generate critical values for the specified null.

```
regress participation uni_comprop uni_title uni_eqdistr comprop
>> eqdistr d_age* yrd* i.person chd*, cluster(state)
boottest {uni_comprop=0} {uni_eqdistr=0}
```

While the bootstrap procedure yields a wider confidence interval than the conventional $t$ critical values, the estimated coefficient remains significant.

# References

[1] Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *Review of Economics and Statistics* 90(3): 414-427.

[5] We provide the code to construct $t_{IM}$, the Ibragimov-Müller t-statistic in the Appendix.

[2] Cameron, A. Colin and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *Journal of Human Resources* 50(2): 317-373.

[3] Carter, Andrew V., Kevin T. Schnepel, and Douglas G. Steigerwald. 2015. "Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity." Forthcoming, *Review of Economics and Statistics.*

[4] Donald, Stephen G and Kevin Lang. 2007. "Inference with Difference-In-Differences and Other Panel Data." *Review of Economics and Statistics* 89(2): 221-233.

[5] Ferman, Bruno and Christine Pinto. 2015. "Inference in Differences-in-Differences with Few Treated Groups and Heteroskedasticity." *Sao Paulo School of Economics Working Paper.*

[6] Ibragimov, Rustam and Ulrich Müller. 2010. "t-Statistic Based Correlation and Heterogeneity Robust Inference." *Journal of Business and Economic Statistics* 28(4): 453-468.

[7] Ibragimov, Rustam and Ulrich K. Müller. 2016. "Inference with Few Heterogeneous Clusters." *Review of Economics and Statistics* 98(1): 83-96.

[8] MacKinnon, James and Matthew Webb. 2016. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Economics Department Working Paper No. 1314, Queens University.*

[9] Voena, Alessandra. 2015. "Yours, Mine, and Ours: Do Divorce Laws Affect the Intertemporal Behavior of Married Couples?" *American Economic Review* 105(8): 2295-2332.

# 5 Appendix

## 5.1 Ibragimov and Müller

Although the test using Ibragimov and Müller test statistic is unlikely to be valid, we show how to derive the Ibragimov and Müller test statistic, $t_{IM}$, to demonstrate implementation of $t_{IM}$ using Stata. First, we define cluster variable, *clustvar*, and find the number of clusters (denoted *maxclustvar* here):

```
egen clustvar = group(state);
sort clustvar;
local maxclustvar = clustvar[_N];
```

As discussed in section 3, $t_{IM}$ is derived by calculating the coefficient of interest individually and then assuming the derived coefficients to be approximately t-distributed with $G-1$ degrees of freedom. As far as the authors are aware, there is no Stata code for Ibragimov and Müller type analysis. It is, however, fairly simple to implement in Stata without a dedicated program. We use a loop to calculate the coefficients individually for each group, store the results, and calculate $t_{IM}$ using the dataset from Voena (2015). It must be noted that this exercise does not have an analytical power as the covariates of interest vary in some, but not all, clusters.[6]

```
gen bhat = .
forval i = 1(1)‘maxclustvar’ {
qui regress participation uni_comprop comprop d_age* yrd*
>> i.person if clustvar==‘i’
qui replace bhat = _b[uni_comprop] if clustvar==‘i’
}
```

[6] Only eight states had both unilateral divorce law and community property regime in the data. As such, all states without any variation in the interaction term must be eliminated to estimate $t_{IM}$ for $\beta_1$.

```
collapse bhat, by(clustvar)
qui sum bhat
local t_im = r(mean)/(r(sd)/sqrt(r(N)))
di "Mean of betahat is " r(mean)
di "Standard error of betahat is " r(sd)/sqrt(r(N))
"Test statistic is " ‘t_im’ " distributed t with " r(N)-1 "
>> degrees of freedom."
```

The above code produces the following output:

```
Mean of betahat is -.19478994
Standard error of betahat is .2571104
Test statistic is -.75761206 distributed t with 7 degrees of
>> freedom.
```