Inference for Clustered Data

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

January 30, 2017

Abstract

This article provides the applied econometrics community, and empirical researchers, with a guide to recent advances in cluster robust inference. Recent work emphasizing the importance of measuring cluster heterogeneity to determine the appropriate method of inference is featured. Cluster heterogeneity can cause a size distortion leading to under-rejection 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. We provide a decision tree for cluster robust analysis designed to minimize the size distortion, present a (Stata) program to implement inference, and demonstrate how to perform inference on microeconometric data sets. Interestingly, we find the presence of fixed effects can lead to substantial reductions in the effective number of clusters.

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 but is assumed to be independent across clusters. In this paper, we provide a decision tree for cluster robust inference that highlights the importance of 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 β_0 is the value of β 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 β 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

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

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 Ω_g. In consequence $\widehat{u}_g\widehat{u}_g^{\mathrm{T}} := \widehat{\Omega}_g$ is not a consistent estimator of Ω_g and consistency of $\widehat{V}$ can only be obtained by averaging $\widehat{\Omega}_g$ over an increasingly large number of clusters. Thus $\widehat{V}$ depends only on the variation between clusters and does not depend on the variation across observations within a cluster, so 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 is identical over g; and covariance matrices, Ω_g is 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 γ_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. Because a can be canceled out in the null hypothesis, it is possible to impose a restriction that $||a||^2 = 1$ without loss of generality. This additional restriction sets the Euclidean distance of a equal to unity prohibiting the magnitude of a from distorting G^*.

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 40.

In practice G^* must be estimated because it is a function of the unknown within-cluster error covariance matrix Ω. Unfortunately, we cannot use the residuals to estimate G^*, because use of the residuals to construct both the critical values and the test statistic renders the test invalid. 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. 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 effective number of treated clusters (and the effective number of control clusters).

In Section 2 we provide a decision tree for selecting the appropriate method of inference. We present examples on use of the decision tree in Section 3. The Appendix contains a Stata program to calculate the effective number of clusters.

2 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 50, then one should use the statistic (2) with critical values from a normal distribution. If G^{*A} is less than 50, 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 20, even if the overall effective number of clusters exceeds 50, then again the wild bootstrap could be used to obtain more accurate critical values.[1]

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 bootstrapped samples. Complete details are provided in Cameron, Gelbach, and Miller (2008), Cameron and Miller (2015), and MacKinnon and Webb (2016). Implementation of the wild bootstrap in STATA is aided by the *.do* file constructed by Miller (which is a modified version of the file used in Cameron, Gelbach and Miller).[2] The file contains codes generating the bootstrapped p-values, and

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

[2] The file can be found at http://faculty.econ.ucdavis.edu/faculty/dlmiller/statafiles/

Stata users can modify the codes to implement the wild bootstrap. There does not appear to be a simple to use *.ado* file.

For data sets that have a small effective number of clusters, either overall or within the treatment group (while rare, a simliar 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 α and δ 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 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 β and variance σ_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 σ_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

and it is called *bs_ example.do*

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 (2014) 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 β from

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

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

We provide a simple decision rule based on the effective number of clusters, G^*, and the number of observations within cluster, n_g.

1. Is G^{*A} large?

   (a) Yes $\rightarrow$ Use OLS with cluster robust standard error and normal critical values. Carter, Schnepel, and Steigerwald (2015) prove that $t_{OLS} \rightarrow_p \mathcal{N}(0,1)$ as $G^* \rightarrow \infty$.

   (b) No $\rightarrow$ Go to the second decision rule.

2. Is $min(n_g)$ large?

   (a) Yes $\rightarrow$ Use Ibragimov and Müller (2010) with $G-1$ degrees of freedom if the covariate of interest is varying within cluster or Donald and Lang (2007) with $G^{*A}-1$ degrees of freedom if the covariate of interest is fixed within cluster. Refer to section 2.2 and 2.3 respectively for guidance. Ibragimov and Müller performs better than the bootstrap with small G^* and cluster heterogeneity

and Donald and Lang performs on par with bootstrap with small G^* and $n_g = 40$.

(b) No $\rightarrow$ Use wild cluster bootstrap. Refer to section 2.1 for further detail.

3 Empirical Implementation

We recommend using the program in the Appendix 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. We provide two empirical examples.

3.1 Voena (2015)

One example comes from Voena (2015), a paper studying changes in household savings and female employment as a result of introducing unilateral divorce laws under different property division regimes. With regards to the household savings, the paper posits that a switch to unilateral divorce laws is more likely to affect asset accumulation in states with more equal division of property at divorce. Couples separately accumulate assets if the division of assets is title-based (spouse who has title ownership of the asset is entitled to it at divorce) whereas couples make a joint decision in asset accumulation if the division of assets is more equitable (either court decides the division or equally divided between spouses). Therefore, asset accumulation for couples in more equitable property division regime are more responsive to increase in the threat of divorce. It also follows that the spouse with lower marital resources gains more from the switch to unilateral divorce laws under more equitable regime of property division. Female labor market participation, therefore, is likely to be more responsive to the divorce law reform in states with more equitable property division.

The paper tests this theory using the National Longitudinal Survey of Young and Mature Women (NLSW) and the Panel Study of Income Dynamics (PSID) in a simple linear model with fixed effects. The model is given in (13).

$$y_{ist} = \sum_{i=1}^{3} \beta_i (uni_{st} * d^i_{st}) + \beta_4 d^1_{st} + \beta_5 d^2_{st} + \gamma' Z_{it} + \delta_t + f_i + c_s + \epsilon_{ist}$$

The estimator is OLS with fixed effects for individual, state, and year. Unilateral divorce law is represented by uni_{st} and property regimes are denoted as d^1, d^2, and d^3 where d^1 is the most equitable and d^3 is the least equitable division of properties upon divorce. Two outcomes of interest are household asset accumulation and probability of employment of female member of household. Voena (2015) uses PSID to test the model with household asset accumulation as the outcome and NLSW to test the model with female employment as the outcome.

The null hypothesis to be tested is given by $H_0 : \beta_1 = \beta_3$ as the author is interested in studying the differences in the effect of introducing the unilateral divorce law under different division of assets. Because the individuals living in a same state are likely to be exposed to unobserved shocks of each other that may affect the outcome variables of interest, appropriate level of clustering is state in this case. Effective number of clusters can be estimated in STATA using [program_name].

```
[program_name] uni_d1 uni_d2 uni_d3 d1 d2 d_age i.year i.state i.person,
cluster(state) test(uni_d1-uni_d3)
```

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 lists two numbers in following format.

```
Number of cluster is 43
```

```
Effective number of cluster is 7.061544
```

We also estimate the effective number of clusters for PSID. The Stata code required follows the same format as PSID.

```
Number of cluster is 51
Effective number of cluster is 7.884968
```

Clusters are heterogeneous and a severe size distortion is possible if test statistic using OLS estimate of coefficients and cluster robust standard errors is assumed to be normally distributed with $G^{*A} \approx 7$ and $G^{*A} \approx 8$ for NLSW and PSID respectively. There are two possible alternatives to the test statistic using the cluster robust standard errors in this case. As the treatment variables of interest are varying within state, it is possible to improve the analysis using the Ibragimov and Müller estimator. Additionally, the conventional t statistic with the wild bootstrap is a valid method to reduce size distortion from cluster heterogeneity.

3.2 Autor and Dorn (2013)

Autor and Dorn (2013) hypothesizes "consumer preference" and "non-neutral technological progress" induces divergence in growth rate of service and non-service occupations with low skill requirement. Given technological advancements are concentrated in the goods market, demand for service outputs increases relatively as it becomes cheaper to consume same amount of goods. Additionally, outputs requiring services tasks such as "restaurant meals, house-cleaning, security services, and home health assistance" may not have a non-human substitute due to strong consumer preference for personalized care.

The authors provide a model in which individuals consume goods and services, goods are produced using abstract labor and either capital or routine labor, services are produced using service labor, and price of capital falls as

technology advances. The model makes three alternative predictions. First, if elasticity of substitution is higher between alternative production inputs, capital and routine labor, compared to the elasticity of substitution between goods and services in consumption, capital replaces routine labor faster than goods replace services as capital becomes cheaper to produce. Price of routine labor, therefore, falls in comparison to the price of service labor, and low-skill workers move from goods industry to service industry. Second, if consumption elasticity is less than 1 in addition to being smaller than the goods market production elasticity, price paid to service labor remains constant or rises compared to the price paid to abstract labor. Third, if goo ds market production elasticity is lower than consumption elasticity, wage paid to service labor is not raised as consumption of services falls at least as quickly as the demand for routine labor.

Empirically, Autor and Dorn use the variation in adoption of computer among locations with differing share of routine labor across different commuting zones to test whether the share of service labor grows faster between 1980 and 2005 in commuting zones with greater routine labor share in 1980. We specifically focus on the linear regression of the following form:

$$\Delta SVC_{jst} = \delta_t + \beta_1 RSH_{jt_0} + X'_{jt_0}\beta_2 + \gamma_s + e_{jst} \tag{5}$$

where key covariate is the share of routine labor within commuting zone, j, at time t_0. Dependent variable is the change in the non-college service employment share within commuting zone in state s between time t_0 and t_1 where t denotes an interval of time within t_0 and t_1. The specification includes time and state indicators. Authors limit the time frame to between 1980 (t_0) and 2005 (t_1) in the analysis, and the variables are measured in 1980, 1990, 2000, and 2005. Therefore, there are intervals in time (1980-1990, 1990-2000, 2000-2005) and $t = \{1, 2, 3\}$, where t represents the interval. Additionally, X represents any additional characteristics about the commuting zone that

may affect the change in share of service occupations. The error is clustered at the state level.

The null hypothesis is $H_0 : \beta_1 = 0$. If H_0 is rejected based on the test statistic obtained, statistical evidence supports the relationship between the share of routine occupations and growth rate of service occupations. Autor and Dorn find the result to be statistically significant using the test statistics formed with cluster robust standard errors.[3] Given a limited number of states in the United States and varying numbers of commuting zone within each state based on population and size of the states, it is unlikely that the clusters are homogeneous. We, therefore, use [program_name] to test the severity of cluster heterogeneity.

Because Autor and Dorn include fixed effects for both time and state, we first expand the variables for state and time. The time is already in three indicator variables ($t1, t2, t3$) in the data provided by Autor and Dorn. For expansion of the state variable ($statefip$), we use the STATA command, 'xi'.

```
xi i.statefip
```

This creates $m - 1$ variables named $\_Istatefip_i$, where i indicates the value of $statefip$ and m is the number of unique values in $statefip$. One of the indicators are omitted.

Next, we use [program_name] to estimate the effective number of clusters. First, we do not include any additional controls in X to start. The code lists all covariates and restricts the sample to observations in or after 1980 Census. We are required to include the cluster option, and we list $statefip$ as the clustering variable. Finally, we include the covariate to be tested in the test option.[4]

```
>[program_name] l_sh_routine33a t2 t3 _Istatefip* if yr>=1980,
```

[3] Autor and Dorn (2015) presents results from the OLS specifications in Panel A of Table 5.

[4] Share of routine occupation within the commuting zone is l_sh_routine33a in Autor and Dorn's data.

```
 cluster(statefip) test(l_sh_routine33a)
Number of cluster is 48
Effective number of cluster is 1.1297756
```

The program demonstrates cluster heterogeneity may be extremely severe in this case. This is likely driven by uneven cluster sizes as some states contain very small number of commuting zones. (note: does including cluster level covariates – state indicators – intensify the difference in cluster sizes? Excluding the cluster level indicators increases the estimated effective number of clusters to around 15.) Because the coefficient of interest that Autor and Dorn are testing varies both within and across clusters, it is not possible to construct a valid test statistic using Donald and Lang (2007). It may be possible to use the wild bootstrap to simulate the distribution of the test statistic or use Ibragimov and Müller (2010) to construct a $t(G-1)$ distributed test statistic. One concern is that n_g is not large enough for some groups and Ibragimov and Müller test statistic is rendered invalid.

```
>tab statefip if yr>=1980

   statefip |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |         42        1.94        1.94
          4 |         15        0.69        2.63
          5 |         54        2.49        5.12
          6 |         54        2.49        7.62
          8 |         51        2.35        9.97
          9 |          3        0.14       10.11
...
```

Just from looking at the first six states, number of observations within some of the clusters is clearly not large enough for Ibragimov and Müller test statistic to be valid. In particular, one of the states appear to contain single

commuting zone with three observations from each time periods within the time frame, implying within-cluster invariant RSH_{jt_0} for smallest states.

3.2.1 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(statefip);
sort clustvar;
local maxclustvar = clustvar[_N];
```

As discussed in section 2.2, it 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}.

```
. gen bhat = .;
(3610 missing values generated)

. forval i = 1(1)‘maxclustvar’ {;
  2.         qui regress d_shocc1_service_nc l_sh_routine33a t2 t3 [a
> w=timepwt48] if yr>=1980 & clustvar==‘i’;
  3.         qui replace bhat = _b[l_sh_routine33a] if clustvar==‘i’;
  4. };
```

```
. preserve;

. collapse bhat if yr>=1980, by(clustvar);

. qui sum bhat;

. local t_im = r(mean)/(r(sd)/sqrt(r(N)));

. di "Mean of betahat is " r(mean);
Mean of betahat is .10445227

. di "Standard error of betahat is " r(sd)/sqrt(r(N));
Standard error of betahat is .03070484

. di "Test statistic is " `t_im' " distributed t with " r(N)-1 " degr
> ees of freedom.";
Test statistic is 3.4018177 distributed t with 47 degrees of freedom.
```

The resulting test suggests that β_1 remain statistically different from 0. The two-sided 5-percent critical value of the t-distribution with 47 degrees of freedom (≈ 2.02) is smaller than $\hat{t}_{IM}$. Compared to $\hat{\beta}_1$ (0.105) and standard error (0.032), Ibragimov and Müller method produces an estimate very similar to the OLS estimate of the coefficient. It also estimates standard error to be close, but surprisingly it appears to be slightly less conservative compared to the clustered standard error. However, t_{IM} is unlikely to be consistently estimated given small n_g especially for smaller states where the number of observations is close to unity.

3.2.2 Wild Bootstrap

In this section we estimate the p-value for $\hat{\beta}_1$ using the wild bootstrap. A program generating the wild bootstrapped p-values does not exist, but Doug

Miller provides a .do file implementing three types of bootstrap procedure used in Cameron, Gelbach, and Miller (2008).[5] We modify Miller's code to generate bootstrapped p-value which we can compare with the p-value generated using cluster robust standard errors.

The code provided by Miller is fairly straightforward. Outside of replacing variable names, change we make to the code is limited to keeping only the wild bootstrap method part of the program and Miller's code we modify and use for this section is documented in the Appendix. The result we obtain from the modified code are as follows:

```
Number BS reps = 999, Null hypothesis = 0
Main beta   main T   Main %le   wild %le
 0.105       3.287    0.00101    0.00400
```

While bootstrapped p-value is greater than the p-value from cluster robust standard error, the estimated coefficient remains significant.

References

[1] Autor, David H. and David Dorn. 2013. "The Growth of Low-Skill Service Jobs and the Polarization of the US Labor Market." *American Economic Review* 103(5): 1553-1597.

[2] 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.

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

[5] As mentioned in 2.1, Miller posts the code on his website at http://faculty.econ.ucdavis.edu/faculty/dlmiller/statafiles/

[4] 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.*

[5] 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.

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

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

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

[9] MacKinnon, James G. and Matthew D. Webb. 2016. "Wild Bootstrap Inference for Wildly Different Cluster Sizes." *Queen's Economics Department Working Paper No. 1314.*

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

4 Appendix

4.1 Effective Number of Clusters Program Specification

4.2 Syntax

[program_name] *varlist* [*if*] [*in*] [*weights*] , cluster(*varname*) [noconstant test(*string*)]

4.3 Description

[program_name] uses a user-selected clustering variable, a vector of independent variables included in the linear model of interest, and a selection vector to approximate the effective number of clusters. *varlist* contains variables to be included in the analysis and at least one variable must be specified. If no options are specified, [program_name] generates a selection vector assuming a hypothesis of the following form $H_0 : \theta_0 + \theta_1 + ... + \theta_k = \theta_0^{H_0} + \theta_1^{H_0} + ... + \theta_k^{H_0}$, where θ_0 is the intercept.

4.4 Options

cluster(*varname*) specifies a clustering variable used to estimate the effective number of clusters. This is a required option and must be specified.
noconstant specifies whether the linear model to be tested contains a constant. If noconstant is specified in the options, the program estimates the effective number of clusters excluding the constant. The default is to include the constant in the estimation.
test(*value*) specifies the left hand side of the null hypothesis to be tested. Suppose $H_0 : \beta_2 + 2\beta_3 = 0$, then the correct specification of the test is test($`indepvar'_2 + 2 * `indepvar'_3$). If left unspecified, the program assumes the null hypothesis in which all coefficients are included with equal weight.

4.5 Estimation Procedure

[program_name] generates an estimate of the effective number of clusters in (3). As noted above, it is not possible to derive G^* unless the underlying error structure, $u_g u_g^T$, is known for each cluster. An intuitive solution would be to substitute a vector of residuals, $\widehat{u}_g$, in for u_g. Using $\widehat{u}_g$ to construct critical values in addition to constructing the test statistic, however, renders the test invalid (Carter, Schnepel, and Steigerwald, 2015). Instead, CSS suggest assuming the errors are perfectly correlated within each cluster. Following CSS, we use a 1-by-n_g vector of ones, ι_g, in place of u_g in the estimation procedure. An estimator of G^*, G^{*A}, is constructed by replacing γ_g with $\gamma_g^A = a^T(X^TX)^{-1}(X_g^T\iota_g\iota_g^{\mathrm{T}}X_g)(X^TX)^{-1}a$ in (3):

$$G^{*A} = \frac{G}{1+\Gamma^A}, \qquad \Gamma^A = \frac{1}{G}\sum_{g=1}^{G}(\frac{\gamma_g^A - \bar{\gamma}^A}{\bar{\gamma}^A})^2.$$

As discussed in the previous section, a is restricted by $||a||^2 = 1$. Therefore, any valid input in test(*string*) is converted to a selection vector meeting this restriction while preserving the relative weights. The program performs matrix multiplication estimating a scalar value of G^{*A}.

4.6 Modified Wild Bootstrap Program

The credit for the program goes to Douglas Miller. We make minor changes to the program and annotate the changes.

```
. #delimit ;
delimiter now ;
. version 9.0 ;

. clear ;

. set mem 5m ;
```

```
. set more off ;

. set seed 365476247 ;

. cap prog drop runme ;

. prog def runme ;
  1. local hypothesis = `0' ;
  2. tempfile main bootsave ;
  3. use "C:\Users\Chang\Dropbox\Ideas
\Stata Journal - Cluster Robust Inference\Example\Autor Dorn
\Autor-Dorn-LowSkillServices-FileArchive\dta\workfile2012.dta";
  4. xi i.statefip;
  5. /*      d_shocc1_service_nc - LHS var, find and replace them
>                 with appropriate LHS variable.
>     l_sh_routine33a - RHS var of interest, we test whether
>                 the coefficient on this variable is significantly
>                 different from zero. Find and replace with appropriate
>                 RHS variable of interest.
>         t2 t3 _Istatefip* - Additional RHS variables included
>                 in the regression. Find and replace with appropriate
>                 RHS variables.
>         statefip - Clustering variable. Find and replace with
>                 appropriate variable.
> */
>
> di ;
  6. drop if yr<1980;
  7. regress d_shocc1_service_nc l_sh_routine33a t2 t3 _Istatefip* if yr
```

```
>=1980 [aw=timepwt48], cluster(statefip) ;
  8. global mainbeta = _b[l_sh_routine33a] ;
  9. global maint = (_b[l_sh_routine33a] - ‘hypothesis’) / _se[l_sh_routine33a] ;
 10. predict epshat , resid;
 11. predict yhat , xb ;
 12. /* also generate
 "impose the null hypothesis" yhat and residual */
> gen temp_y = d_shocc1_service_nc - l_sh_routine33a * ‘hypothesis’ ;
 13. reg temp_y t2 t3 _Istatefip* ;
 14. predict epshat_imposed , resid ;
 15. predict yhat_imposed , xb ;
 16. qui replace yhat_imposed = yhat_imposed + l_sh_routine33a * ‘hypothesis’ ;
 17. sort statefip ;
 18. qui save ‘main’ , replace ;
 19. qui by statefip: keep if _n == 1 ;
 20. qui summ ;
 21. global numyears = r(N) ;
 22. cap erase ‘bootsave’ ;
 23. qui postfile bskeep t_wild
>         using ‘bootsave’ , replace ;
 24. forvalues b = 1/$bootreps { ;
 25. /* first do wild bootstrap */
> use ‘main’, replace ;
 26. qui by statefip: gen temp = uniform() ;
 27. qui by statefip: gen pos = (temp[1] < .5) ;
 28. /*  these two lines (and commendted t-stat below
> are if you don’t want to impose the null hypothesis*/
> /*
> gen wildresid = epshat * (2*pos - 1) ;
> gen wildy = yhat + wildresid ;
```

```
> */
> gen wildresid = epshat_imposed * (2*pos - 1) ;
 29. gen wildy = yhat_imposed + wildresid ;
 30. qui reg wildy l_sh_routine33a t2 t3 _Istatefip* [aw=timepwt48],
cluster(statefip) ;
 31. local bst_wild = (_b[l_sh_routine33a] - `hypothesis') /
_se[l_sh_routine33a] ;
 32. *local bst_wild = (_b[post_self] - $mainbeta) / _se[post_self] ;
. post bskeep (`bst_wild') ;
 33. } ;
 34.  /* end of bootstrap reps */
>
> qui postclose bskeep ;
 35. qui drop _all ;
 36. qui set obs 1 ;
 37. gen t_wild = $maint ;
 38. qui append using `bootsave' ;
 39. qui gen n = . ;
 40. foreach stat in t_wild { ;
 41. qui summ `stat' ;
 42. local bign = r(N) ;
 43. sort `stat' ;
 44. qui replace n = _n ;
 45. qui summ n if abs(`stat' - $maint) < .000001 ;
 46. local myp = r(mean) / `bign' ;
 47. global pctile_`stat' = 2 * min(`myp',(1-`myp')) ;
 48. } ;
 49. global mainp = norm($maint) ;
 50. global pctile_main = 2 * min($mainp,(1-$mainp)) ;
 51. local myfmt = "%7.5f" ;
```

```
 52. di ;
 53.  di "Number BS reps = $bootreps, Null hypothesis = `hypothesis'" ;
 54. display "Main beta" _column(13) "main T"        _column(22) "Main %le"
>         _column(33) "wild %le" ;
 55. di      %6.3f $mainbeta _column(13) %6.3f $maint  _column(23)
>         `myfmt' $pctile_main _column(34) `myfmt' $pctile_t_wild ;
 56. end ;

. global bootreps = 999 ;

. runme 0 ;
```