2013

Asymptotic Behavior of a t Test Robust to Cluster Heterogeneity

Douglas G Steigerwald, University of California, Santa Barbara

Available at: https://works.bepress.com/douglas_steigerwald/26/
Asymptotic Behavior of a $t$ Test Robust to Cluster Heterogeneity

Andrew V. Carter
Department of Statistics
University of California, Santa Barbara

Kevin T. Schnepel and Douglas G. Steigerwald
Department of Economics
University of California, Santa Barbara

September 6, 2013

Abstract

We study the behavior of a cluster-robust $t$ statistic and make two principle contributions. First, we relax the restriction of previous asymptotic theory that clusters have identical size, and establish that the cluster-robust $t$ statistic continues to have a Gaussian asymptotic null distribution. Second, we determine how variation in cluster sizes, together with other sources of cluster heterogeneity, affect the behavior of the test statistic. To do so, we determine the sample specific measure of cluster heterogeneity that governs this behavior and show that the measure depends on how three quantities vary over clusters: cluster size, the cluster specific error covariance matrix and the actual value of the covariates. Because, in the absence of a fixed design, the third quantity will always vary over clusters, the vast majority of empirical analyses have test statistics whose finite sample behavior is impacted by cluster heterogeneity. To capture this impact, we develop the effective number of clusters, which scales down the actual number of clusters by the measure of cluster heterogeneity. Through simulation we demonstrate this effect and find rejection rates as high as 30 percent for a nominal size of 5 percent. We then apply our measure of cluster heterogeneity in several empirical settings to show how observable variation over clusters impacts the performance of a cluster-robust test.

Keywords: Cluster robust, heteroskedasticity, $t$ test

*We thank Dick Startz, together with members of the Econometrics Research Group at UC Santa Barbara, Colin Cameron, and Ulrich Müller, who served as discussant at the 2013 Econometric Society Winter Meetings, for helpful comments. Corresponding author: doug@econ.ucsb.edu
1 Introduction

In conducting inference with a cluster-robust \( t \) statistic, researchers often rely on the result that the statistic has a Gaussian asymptotic null distribution. The existing result is derived for the specific case in which clusters are equal in size. Because in many applications clusters are unequal in size, there is a gap between the existing result and empirical practice. We fill this gap by establishing that the conventional cluster-robust \( t \) statistic has a Gaussian asymptotic null distribution for the more general case in which clusters can vary in size. In so doing, we determine a sample specific measure of cluster heterogeneity that governs the behavior of this cluster-robust \( t \) statistic. From the sample specific measure we construct the effective number of clusters, which scales down the actual number of clusters by the measure of cluster heterogeneity. It is the effective number of clusters that governs inference: If the effective number of clusters is large, then Gaussian critical values are appropriate.

The conventional cluster-robust \( t \) statistic is based on the ordinary least squares coefficient estimator from the entire sample, together with a cluster-robust variance estimator based on the outer product of the residuals.\(^1\) The original asymptotic theory, due to White (1984, Theorem 6.3, p. 136), applies to clusters of equal size that satisfy a further assumption of cluster homogeneity. Under cluster homogeneity White establishes two principle results. First, that the cluster-robust \( t \) statistic has a Gaussian asymptotic null distribution. Second, that the variance component, which appears in the denominator of the test statistic, is consistently estimated through the use of the cluster-robust variance estimator. Consistent estimation of the variance component is also established in Hansen (2007), who maintains the assumption that clusters have equal size while relaxing White’s further assumption of cluster homogeneity. We allow both for unequal cluster size and for heterogeneity of clusters. Under these more general assumptions we establish that the cluster-robust variance estimator can be used to consistently estimate the variance component that appears in the denominator of the test statistic. We further establish that the cluster-robust \( t \) statistic has a Gaussian asymptotic null distribution.

To understand why variation in cluster sizes impacts the behavior of the cluster-robust \( t \) statistic, consider a sample of 20 observations divided into two clusters. Because observations are assumed to be independent across clusters, the potentially unique error covariance terms are all contained within the diagonal blocks that capture the correlation within clusters. If the clusters are equally sized, there are 110 potentially unique terms. As the size of one cluster grows, the number of potentially unique terms grows and reaches a maximum of 191 when one group contains 19 observations. Variation in cluster size, keeping fixed the total number of observations, alters the number of non-zero error covariance terms. Because each of these non-zero terms must be accounted for to avoid upward bias in the test statistic, as Kloek (1981) was among the first to show, the behavior of the cluster-robust \( t \) statistic is impacted by variation.

\(^1\)In what follows we refer to this test statistic simply as the cluster-robust \( t \) statistic.
in cluster size. Cameron, Gelbach and Miller (2008) find via simulation that for a small number of clusters, allowing clusters to have differing numbers of observations can substantially increase the size of a cluster-robust t test.

As we will show, use of the outer product of the residuals implies that the cluster-robust variance estimator is a function only of between cluster variation and, hence, that consistency of the variance estimator requires that the number of clusters grows without bound. One immediate consequence is that it is not possible to conduct valid inference on cluster fixed effects with the cluster-robust t statistic. A fixed effect limited to a single cluster has only one cluster from which to estimate the variance of the fixed effect, so the estimator of the variance is undetermined.

Because estimation and inference in practice are conditioned on the observed value of the covariates, the measure of cluster heterogeneity we derive is specific to each sample. The measure depends on how three quantities vary over clusters: cluster size, the cluster specific error covariance matrix and the observed value of the covariates. The measure of cluster heterogeneity scales down the number of clusters to produce the effective number of clusters. A low effective number of clusters leads to a higher mean-squared error for the cluster-robust variance estimator, which in turn affects the behavior of the cluster-robust t statistic.

The effective number of clusters can be thought of as a generalization of the correction formula reported in Moulton (1986). The correction formula, which requires that all observations be equally correlated within clusters, indicates how to increase standard error estimators to account for neglected cluster correlation. The effective number of clusters, which does not require equal correlation of all observations within clusters, does not alter the cluster-robust standard error estimator but rather alerts the researcher to the need for conservative critical values. Because the underlying measure of cluster heterogeneity depends on the value of the covariates under study, the effective number of clusters should be reported for virtually all studies that employ the cluster-robust t statistic, even if clusters do not vary in size.

Through simulation we demonstrate this point and find that in many settings, while cluster heterogeneity reduces the effective number of clusters, the reduction results in only a moderate increase in the rejection rate for the test. In these cases, a researcher can report the effective number of clusters and proceed with Gaussian critical values. For settings with severe heterogeneity and substantial cluster correlation, the effective number of clusters can fall below 20. When this is the case we find a downward bias in the cluster-robust standard errors, which in turn leads to rejection rates of up to 30 percent for a nominal size of 5 percent. In practice calculation of the effective number of clusters depends on the unknown error correlations. We show how to overcome this difficulty through use of an approximate measure that depends only on the observed covariates and cluster sizes. The simulations reveal that the approximate measure, while conservative, closely tracks the effective number of clusters in precisely the situations where the calculation is of most importance, namely where correlation within clusters is substantial.
The paper is organized as follows. In Section 2 we define the general class of models under study and define the measure of cluster heterogeneity. We relate the measure to the mean-squared error of the cluster-robust variance estimator, establish that the asymptotic null distribution of the cluster-robust \( t \) statistic is Gaussian and show that consistent testing of fixed effects is not possible. In Section 3, we define the effective number of clusters and emphasize, through simulation, that the effective number of clusters is a sample specific measure that varies with the coefficient under test. For several empirical settings we report an effective number of clusters for the key hypotheses under test and discuss appropriate inference, in Section 4. While not our principle focus, we discuss how to select conservative critical values in Section 5.

2 Asymptotic Behavior

We consider a set of \( n \) observations from the linear model

\[
y = X\beta + u,
\]

where the covariate matrix \( X \) consists of \( k \) linearly independent columns. The key feature of the model is that the observations can be sorted into \( G \) clusters, where the errors are independent between clusters. Given \( X \) the covariance matrix of \( u, \Omega \), is then a block diagonal matrix with blocks of the form \( \Omega_g \), which corresponds to the covariance matrix for cluster \( g \). Because \( \Omega \) is block diagonal, the variance of the ordinary least squares estimator \( \hat{\beta} \) can be written as the sum of the \( G \) cluster specific variance components. We have

\[
V := \text{Var} \left[ (X^T X)^{-1} X^T u \right] = \sum_{g=1}^G \text{Var} \left[ (X^T X)^{-1} X_g^T u_g \right],
\]

where \( X_g \) and \( u_g \) are the covariate matrix and error vector for cluster \( g \), respectively.

The hypotheses under test are formed from subsets of the coefficients in (1). The general form of null hypothesis is \( H_0: a^T \beta = a^T \beta_0 \), where \( a \) is a selection vector of dimension \( k \). Because any factor that multiplies the selection vector cancels out of the test statistic, we assume, without loss of generality, that \( ||a||^2 = 1 \), where \( ||a|| \) is the Euclidean norm of the vector \( a \). The cluster-robust \( t \) statistic is

\[
Z = \frac{a^T (\hat{\beta} - \beta_0)}{\sqrt{\text{Var} \left( a^T \hat{\beta} \right)}},
\]

where the variance component is \( \text{Var} \left( a^T \hat{\beta} \right) = a^T \hat{V} a \) and \( \hat{V} \) is the cluster-robust variance estimator. The cluster-robust variance estimator, which Shah, Holt and Folsom (1977) are among the first to use, is the sample analog for \( V \).
where the observed residuals $\hat{u}_g$ replace the errors $u_g$:

$$
\hat{V} = (X^T X)^{-1} \sum_{g=1}^{G} X_g^T \hat{u}_g \hat{u}_g^T X_g (X^T X)^{-1}.
$$

White establishes asymptotic results for the cluster-robust $t$ statistic and for the variance component $\hat{V}_a := a^T \hat{V} a$. White’s proof has two key assumptions: 1) that all clusters have an identical, fixed, number of observations and 2) that $\mathbb{E} (X_g^T \Omega_g X_g)$ not vary over $g$. He then proves that, if $G \to \infty$ as $n \to \infty$ then $Z$ has a Gaussian asymptotic null distribution and $\hat{V}_a$ is a consistent estimator of $V_a$. We relax both of White’s key, cluster homogeneity, assumptions. We allow the cluster size, $n_g$, to vary: over clusters, so that clusters need not be of identical size, and to vary with the sample size, so that cluster sizes need not be fixed. We also allow $\mathbb{E} (X_g^T \Omega_g X_g)$ to vary over $g$. We then prove that, if $G \to \infty$ as $n \to \infty$, then $Z$ has a Gaussian asymptotic null distribution and $\hat{V}_a$ is a consistent estimator of $V_a$.

Importantly, we establish consistency of $\hat{V}_a / V_a$ rather than $\hat{V}_a - V_a$. We do so because if $\hat{\beta}$ is a consistent estimator of $\beta$, then the elements of $V$ converge to zero and do so at a rate that depends on the behavior of the cluster sizes. The rate of convergence of $V$ to zero must be explicitly accounted for in $\hat{V}_a - V_a$, while it is implicitly controlled in $\hat{V}_a / V_a$. This point is clearly revealed in Hansen, who studies $\hat{V}_a - V_a$ and so must establish separate results depending on the rate at which $n_g$ grows with the sample size. Under the assumption that clusters have an identical number of observations, but where $\mathbb{E} (X_g^T \Omega_g X_g)$ is allowed to vary over $g$, Hansen establishes that $\hat{V}_a$ is a consistent estimator of $V_a$ for two rates of growth of $n_g$. The situation becomes more complex if $n_g$ varies over $g$, as the appropriate result depends on assumptions governing the growth of specific cluster sizes. Through study of $\hat{V}_a / V_a$ we avoid the need for rate-specific results and our theorem accommodates a wide range of behavior for $n_g$.

Because $\Omega_g$ is restricted only by the requirements of a positive definite matrix, the test statistic $Z$ is robust to a wide range of correlated processes. But this general robustness has an important implication: $\hat{V}$ is a function only of between cluster variation. It immediately follows that first, consistency of $\hat{V}$ requires that the number of clusters grow without bound, and second, that the behavior of $Z$, even for hypothesis tests of coefficients on covariates that vary within clusters, is governed by the number of clusters, not the total number of observations.\footnote{If there is no cluster correlation, then each observation effectively becomes a cluster, $G = n$, and the behavior of $Z$ is governed by the total number of observations.}

To establish these facts, we first show that the variance of $\hat{\beta}$ can be expressed as a weighted sum of the variances for the ordinary least squares estimators based only on the observations for cluster $g$, $\hat{\beta}_g$. In doing so, we note that the cluster specific estimator $\hat{\beta}_g$ is constructed with a generalized inverse (see
the Appendix for details) to allow both for cluster invariant covariates and for clusters with \( n_g < k \). We then show that it follows that \( \hat{V} \) is a function only of between cluster variation. We collect these findings in the following result (algebraic details that verify the result are contained in the Appendix).

**Result 1:**

a) The covariance matrix \( V \), together with the estimator \( \hat{V} \), can be expressed as functions of \( \hat{\beta}_g \):

\[
V = \sum_g A_g \text{Var} \left( \hat{\beta}_g \right) A_g^T,
\]

\[
\hat{V} = \sum_g A_g \left( \hat{\beta}_g - \hat{\beta} \right) \left( \hat{\beta}_g - \hat{\beta} \right)^T A_g^T,\tag{4}
\]

where \( A_g = (X^T X)^{-1} X_g^T X_g \).

b) The estimator \( \hat{V} \) is a function only of between cluster variation.

**Remarks:** The cost of the general robustness of \( Z \), even under cluster homogeneity, is reflected in Result 1b. Because \( \hat{V} \) is a function only of between cluster variation, consistency of \( \hat{V} \) requires that the number of clusters grow without bound. Thus, to ensure we have a consistent test, we require that the selection vector \( a \) include only covariates for which the number of clusters in which the covariate takes non-zero values grows without bound. Corollary 1, below, formalizes this remark.

To establish the behavior of \( \hat{V}_a/V_a \) observe that

\[
\frac{\hat{V}_a - V_a}{V_a} = \frac{\hat{V}_a - V_a}{V_a} + \frac{\hat{V}_a - V_a}{V_a},
\]

where \( \hat{V}_a \) is constructed from the unbiased function

\[
\hat{V} = \sum_g A_g \left( \hat{\beta}_g - \hat{\beta} \right) \left( \hat{\beta}_g - \hat{\beta} \right)^T A_g^T.
\]

When \( G \) is large the variance of \( \left( \hat{\beta}_g - \hat{\beta} \right) \) is much larger than the variance of \( \left( \hat{\beta} - \hat{\beta} \right) \), so the variation in \( \frac{\hat{V}_a - V_a}{V_a} \) is driven by the variation in \( \frac{\hat{V}_a - V_a}{V_a} \), while the bias in \( \frac{\hat{V}_a - V_a}{V_a} \) arises from the bias in \( \frac{\hat{V}_a - V_a}{V_a} \). To determine the sample specific features that govern the performance of the cluster robust variance estimator, we derive: i) the mean-squared error of \( \frac{\hat{V}_a - V_a}{V_a} \) conditionally on \( X \) (Lemma 1) and ii) the bias of \( \frac{\hat{V}_a - V_a}{V_a} \) conditionally on \( X \) (Lemma 2). In what follows we assume that the error has a conditionally normal distribution; this assumption is not necessary but simplifies several calculations.

**Assumption 1:** Conditional on the covariate matrix \( X \), the error vector \( u \):

(i) has mean zero,
(ii) is normally distributed,
(iii) has a block diagonal covariance matrix. Specifically, the error vector can be heteroskedastic and have cluster correlation that varies both within and across clusters.

**Lemma 1**: Under Assumption 1,

$$\mathbb{E}\left\{ \left[ \frac{\hat{V}_a - V_a}{V_a} \right]^2 \right\} = \frac{2}{G} \left( 1 + \Gamma(\Omega, X) \right),$$

where the quantity $\Gamma$ is defined by

$$\gamma_g(\Omega, X) = a^T A_g \text{Var}\left( \beta_g \right) A_g^T a,$$

$$\Gamma(\Omega, X) = \frac{1}{G} \sum_{g=1}^{G} \frac{(\gamma_g - \gamma)^2}{\gamma^2},$$

with $\gamma(\Omega, X) = \frac{1}{G} \sum \gamma_g(\Omega, X)$.

**Proof**: See Appendix.

**Remarks**: Because $\hat{V}_a$ is an unbiased estimator of $V_a$, the (relative) mean-squared error in Lemma 1 consists entirely of the variation in $\hat{V}_a$. The quantity $\Gamma(\Omega, X)$, which is the squared coefficient of variation for $\gamma(\Omega, X)$, is the measure of cluster heterogeneity that drives the variation in $\hat{V}_a$. In detail, if $\Gamma(\Omega, X) = 0$, then $\hat{V}_a \sim \chi^2(G)$ and $\mathbb{E}\left\{ \left[ \frac{\hat{V}_a - V_a}{\sqrt{\hat{V}_a}} \right]^2 \right\} = \frac{2}{G}$. If $\Gamma(\Omega, X) \neq 0$, then $\hat{V}_a \sim \chi^2(G)$ and the mean-squared error increases by the factor $(1 + \Gamma(\Omega, X))$.

We next turn to the bias of $\frac{\hat{V}_a - V_a}{\sqrt{\hat{V}_a}}$. To determine the bias, we need two technical conditions that ensure that both $\hat{V}_a$ and $(X^TX)$ exist and are well defined.

**Assumption 2**: As $n \to \infty$:

(i) The covariance matrix $V$, which is conditional on the design $X$, is well formed in that the largest eigenvalue of $V$, $\lambda_{V_{\max}}$, is bounded relative to $V_a$,

$$\sup_{\Omega} \left( \frac{\lambda_{V_{\max}}}{V_a} \right) \leq M_0.$$

(ii) The covariate matrix $X$ is well formed in that the ratio of the largest and smallest eigenvalues of $X^TX$, $\lambda_{X_{\max}}$ and $\lambda_{X_{\min}}$, respectively, is bounded,

$$\frac{\lambda_{X_{\max}}}{\lambda_{X_{\min}}} \leq M_1.$$

Note that in part (i) we do not require that $\lambda_{V_{\min}} > 0$, so that the model can include cluster fixed effects. (For cluster fixed effects the variance is undetermined, so $|V| = 0$ and $\lambda_{V_{\min}} = 0.$) Instead we require that $V_a > 0$, which
implies that the selection vector \( a \) excludes cluster fixed effects. Note further that the maximal eigenvalue for \( V \) is the maximum over \( a \) of the quadratic form \( V a \), hence \( V a \leq \|a\|^2 \lambda_{\text{max}} \).

**Lemma 2:** Under Assumptions 1-2,

\[
\mathbb{E} \left\{ \left| \frac{\hat{V}_a - \bar{V}_a}{V_a} \right| X \right\} \leq \frac{1}{G} \left( 1 + \Gamma^* (I,X) \cdot \zeta (\Omega, X) \right),
\]

where

\[
\zeta (\Omega, X) = M_0 M_1 \left( M_1 + k \left[ \frac{\Gamma^* (\Omega, X)}{\Gamma^* (I,X)} \right]^{1/2} \right),
\]

and the quantity \( \Gamma^* \) is defined for general \( \Omega \) by

\[
\Gamma^*_g (\Omega, X) = \frac{1}{G} \sum_{g=1}^{G} \frac{\| (\gamma_g^* - \hat{\gamma}_g^*) \|^2}{\| \hat{\gamma}_g^* \|^2}.
\]

**Proof:** See Appendix.

**Remarks:** The quantity \( \Gamma^* (\Omega, X) \) differs slightly from the quantity \( \Gamma (\Omega, X) \) in that covariances among the estimated coefficients are now included. Specifically, \( \gamma_g (\Omega, X) \) is a scalar that measures the term in \( A_g \text{Var} (\hat{\beta}_g) A_g^T \) corresponding to the variance of the coefficient under test while \( \gamma_g (\Omega, X) \) is a vector that contains \( \gamma_g (\Omega, X) \) along with the terms in \( A_g \text{Var} (\hat{\beta}_g) A_g^T \) corresponding to the covariances of the coefficient under test with the other coefficients. Also, \( \zeta (\Omega, X) \) includes \( \Gamma^* (I,X) \), which is a diagonalized quantity that captures solely the heterogeneity of the covariates.

We are now able to establish our principle asymptotic result that the cluster-robust test statistic has a Gaussian asymptotic null distribution. This result is not sample specific and so is not conditional on \( X \).

**Assumption 3:**

(i) As \( n \to \infty \) the number of clusters is increasing \( G \to \infty \).

(ii) As \( G \to \infty \), \( \sup_{\Omega} \mathbb{E} [\Gamma^* (\Omega, X)] \to 0 \).

(iii) There exist \( M_0 \) and \( M_1 \) such that Assumption 2 holds except on a set \( \mathcal{B} \) such that \( \mathbb{P} (\mathcal{B}) \to 0 \) as \( n \to \infty \).

**Theorem 1:** If Assumptions 1-3 hold, then \( \hat{V}_a \) is a consistent estimator of \( V_a \) and, under \( H_0 \),

\[
Z \sim N (0,1),
\]

where \( \sim \) denotes convergence in distribution.

**Proof:** See Appendix.
Remarks: Parts (i) and (ii) of Assumption 3 govern the growth rate of cluster sizes. Part (i) rules out the case in which all clusters remain a constant proportion of the sample as \( n \) grows, because the number of clusters must go to infinity. Part (ii) rules out the case in which any of the clusters remains a constant proportion of the sample as \( n \) grows, but does allow cluster sizes to grow with \( n \). Specifically, because \( \text{Var} \left( \tilde{\beta}_g \right) = O \left( \frac{1}{n^g} \right) \) while \( \text{Var} \left( \tilde{\beta}_g \right) = O \left( \frac{1}{n^g} \right) \), the quantity \( \| (\gamma_g^* - \gamma^*) \|^2 = O \left( \frac{n^g_{\max}}{n^g} \right) \) and so \( \frac{\text{Var}^* (G, X)}{c} = O \left( \frac{n^g_{\max}}{n} \right) \), where \( n^g_{\max} \) is the size of the largest cluster. If \( n^g_{\max} = o(n) \), then part (ii) of Assumption 3 is satisfied. Thus, Theorem 1 encompasses both the case in which cluster sizes are fixed as the number of clusters grows and the case in which the cluster sizes and the number of clusters go to infinity jointly.

Part (ii) of Assumption 3 controls heterogeneity across clusters. While it is difficult to relate this condition to the work of Hansen, who does not have an explicit condition controlling cluster heterogeneity, it is possible to relate the condition to the work of Rogers (1993). Although he does not derive an asymptotic null distribution, Rogers conjectured that a Gaussian approximation would be adequate for \( Z \) if \( \max n_g / n < 0.05 \). To link the conjecture to Assumption 3(ii), consider a model with only an intercept and common intracluster correlation, so that \( \gamma_g^* = \alpha^2 \left( \frac{n^g}{n^g_{\max}} \right)^2 \). We see that the adequacy of a Gaussian approximation does depend on \( n^g_{\max} \), albeit through the squared coefficient of variation, rather than the maximal value.

To obtain consistency for \( \tilde{V} \), the number of clusters must grow without bound. In consequence for any covariate that takes non-zero values over only a fixed subset of clusters, consistent testing is not possible.

**Corollary 1:**

For coefficient estimators that depend only on a fixed subset of clusters, the elements of \( \tilde{V} \) that correspond to these estimators are inconsistent.

**Proof:** Because \( \tilde{V} \) is a function only of between cluster variation, consistency of \( \tilde{V} \) requires information from a growing number of clusters. If a coefficient estimator depends only on a finite set of clusters, the requirement is not met. Consider a covariate that takes non-zero values for a fixed subset \( m \) of the clusters. (For a cluster specific control, \( m = 1 \).) The element of \( A_g \) that corresponds to this covariate is zero for all clusters other than the set of \( m \), so \( \gamma_g^* \) is nonzero on \( m \) elements. Hence \( \gamma^* = O \left( \frac{n^g}{n^g_{\max}} \right) \) and \( \Gamma^* = O \left( \frac{n^g}{n^g_{\max}} \right) \), so that \( \frac{1}{2} \Gamma^* = O \left( \frac{n^g}{n^g_{\max}} \right) \) which does not tend to zero as \( G \to \infty \). Q.E.D.

Leading examples of such covariates are cluster specific controls (most often termed cluster fixed effects) or controls that correspond to a group of clusters.

### 3 Effective Number of Clusters

We have established conditions under which the cluster-robust variance estimator \( \tilde{V} \) is consistent and the test statistic \( Z \) has a Gaussian asymptotic null distribution. We now turn to the question: How should a researcher use the
results to inform empirical analysis? An important component to the answer for this question is contained in Lemma 1, where we establish that the mean-squared error of $\hat{V}$ is inversely proportional to

$$G^* = \frac{G}{1 + \Gamma(\Omega, X)},$$

so $G^*$ is the key measure of the adequacy of the asymptotic results. A researcher should calculate $G^*$ to determine if the measure of $G^*$ obtained from their sample is large enough to use Gaussian critical values. We refer to $G^*$ as the effective number of clusters to reflect the fact that the results in Section 2 extend the conventional analysis (under cluster homogeneity) in which the number of clusters is the measure of the adequacy of the asymptotic results.

To calculate $G^*$, recall from Lemma 1 that

$$\Gamma(\Omega, X) = \frac{\sigma^2}{\gamma^2},$$

where $\sigma^2 = \frac{1}{G} \sum_{\gamma=1}^{G} (\gamma - \bar{\gamma})^2$. Because $\Gamma(\Omega, X) \geq 0$, $G^* \leq G$ so that the effective number of clusters is no larger than the actual number of clusters. Importantly the magnitude of the difference between $G^*$ and $G$ increases nonlinearly in the measure of cluster heterogeneity $\Gamma(\Omega, X)$. To construct $\Gamma(\Omega, X)$ note that the cluster-specific component $\gamma_g$, defined in Lemma 1, can also be written as

$$\gamma_g = a^T (X^T X)^{-1} X_g^T \Omega_g X_g (X^T X)^{-1} a. \quad (5)$$

From this expression we see that $X_g^T \Omega_g X_g$ is the quantity that drives cluster heterogeneity, so variation in cluster size is not required for cluster heterogeneity. Cluster heterogeneity can arise with clusters of equal size, but where the cluster error covariance matrix differs over clusters. Moreover, even if $\Omega_g$ is identical across clusters, the fact that the covariates differ over clusters induces heterogeneity. For this reason the vast majority of empirical analyses with cluster-robust inference are characterized by heterogeneous clusters.

To understand how features of the data impact the magnitude of $G^*$, for a given value of $G$, we turn to simulations. The simulations also reveal, at what magnitude of $G^*$ the empirical size of the test statistic exceeds the nominal size to such a degree that a researcher should consider the use of conservative critical values. The data generating process for the simulations is

$$y_{gi} = \beta_0 + \beta_1 x_{gi} + \beta_2 z_{gi} + u_{gi}, \quad (6)$$

together with the error-components model

$$u_{gi} = \varepsilon_{gi} + v_{gi}, \quad (7)$$

where the cluster component $\varepsilon_{gi} \mid X \sim i.i.d. N(0, 1)$ is independent of the individual component $v_{gi} \mid X \sim N(0, c_{gi}^2)$. To capture the richness of empirical settings in which $\hat{V}$ is typically employed, it is important that $Var(u_{gi} \mid X)$ be
heteroskedastic. If $\text{Var}(u_g|X)$ is homoskedastic, then the conditional cluster correlation is constant both across and within clusters. Under our specification if $c > 0$ the $\text{Var}(u_g|X)$ is heteroskedastic and the conditional cluster correlation, $\frac{1}{\sqrt{1+c^2 z_{gi}^2}}$, differs both across and within clusters. As $c$ increases, the idiosyncratic component grows in importance and the within cluster correlation declines.

As (5) reveals, the effective number of clusters varies with the hypothesis under test through the selection vector $a$. (Because the researcher selects $a$ through specification of the null hypothesis, we do not explicitly include $a$ as an argument in $\Gamma(\Omega, X)$.) For this reason, we investigate how the within-cluster correlation of the covariate under test impacts the effective number of clusters and so, the size of the test. In the simulations the cluster-invariant covariate $x_g$ is perfectly correlated within clusters and the cluster-varying covariant $z_g$, is uncorrelated within clusters. Importantly, because the number of clusters in which $x_g$ takes non-zero values grows with the sample size, the cluster-invariant covariate is distinct from a cluster-specific fixed effect and the statistic $Z$ is consistent for hypothesis testing on $\beta_1$.

Each simulated data set consists of 2500 observations divided into 100 clusters. For each simulated data set, the covariates are divided into clusters according to one of the following cluster-size designs. In the first cluster-size design, each cluster contains 25 observations. In each succeeding cluster-size design, one observation is subtracted from clusters 2 through 100 and these observations are added to the first cluster. We construct a sequence of 10 cluster-size designs, in which the proportion of the sample in the first cluster grows monotonically from 1 percent to 37 percent. The full description of design variation is contained in Table 1.

<table>
<thead>
<tr>
<th>Cluster Sizes</th>
<th>$n_1$</th>
<th>$n_2 = \cdots = n_{100}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design 1</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>Design 2</td>
<td>124</td>
<td>24</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design 10</td>
<td>916</td>
<td>16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Error Cluster Correlation</th>
<th>within cluster correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c = 500$</td>
<td>heteroskedastic 0.01</td>
</tr>
<tr>
<td>$c = 50$</td>
<td>0.06</td>
</tr>
<tr>
<td>$c = 5$</td>
<td>0.38</td>
</tr>
<tr>
<td>$c = 0$</td>
<td>homoskedastic 1.00</td>
</tr>
</tbody>
</table>

For each of the 40 pairs of cluster-size design and $c$, we generate 1000 simulated data sets and compute the effective number of clusters. In Figure 1, we plot the effective number of clusters as a function of the cluster-size design.

The cluster-invariant covariate is generated as a sequence of independent Bernoulli random variables with equal probability of 0 or 1; the cluster-varying covariate is generated as a sequence of independent Uniform(0,1) random variables.
The panels of the figure in the left column correspond to hypothesis tests on the coefficient for the cluster-invariant covariate, $x_g$, those in the right column correspond to hypothesis tests on the coefficient for the cluster-varying covariate, $z_{gi}$. The rows of the figure correspond to differing degrees of within cluster error correlation; in the top row $c = 500$, while in the bottom row, $c = 0$. Within each panel, for each set of cluster sizes, the bar spans the range from the minimum to the maximum effective number of clusters across the 1000 simulated data sets and the cross hatch marks the median value.

Figure 1

Several features are immediate. First, increasing the cluster size variation, by moving from design 1 to design 10, dramatically reduces the effective number of clusters. As the top row of Figure 1 reveals, this holds even for data sets that have little within cluster error correlation. Moreover, for the coefficient on the covariate that is perfectly correlated within clusters, there is a dramatic reduction in the effective number of clusters in moving only from design 1 to design 2. This suggests that the higher the within cluster correlation of the covariate under test, the more sensitive the test will be to cluster size variation.

A second feature that emerges in many designs, is that the realized values
of \( \{X_g\} \) can play an important role. As it is variation in the realized values of the covariates that is the source of the difference between the minimum and maximum values of \( G^* \) for a given design, we can see this impact through the length of the bar. The key component of the realized value of the covariates is the variation in the value of \( x_g \) across clusters. Data sets in which most of the clusters have the same value for \( x_g \), have a greatly reduced value for the effective number of clusters. The dependence of \( G^* \) on the observed values of the covariates underscores the need to calculate the effective number of clusters as a general rule.

We next investigate how the magnitude of \( G^* \) impacts the empirical size of the test statistic \( Z \). The goal is to determine the range of \( G^* \) over which the use of Gaussian critical values is appropriate. The hypothesis under test is \( H_0 : \beta_i = 0 \) for \( i = 1, 2 \), which is rejected if the calculated value of \( Z \) exceeds the Gaussian critical value of 1.96 in absolute value (which corresponds to a nominal size of 5 percent). To compute the empirical test size, we first generate \( X \) from one of the 40 cluster designs (10 cluster-size designs, each paired with 4 values of \( c \)). For the given value of \( X \) we generate 1000 error vectors, \( u \), each of length 2500. From each data set \( \{X, u\} \) we compute the empirical test size as the rejection probability over the 1000 data sets that share a common \( X \). We do this 5 times for each of the 40 cluster designs, resulting in 200 pairings of \( G^* \) with an empirical test size.

Figures 2 and 3 contain the empirical test size as a function of \( G^* \). For the coefficient on the covariate that is perfectly correlated within clusters, in Figure 2, the empirical size begins to rise above the nominal size of 5 percent when the
effective number of clusters falls below 40, although substantial size distortion
only occurs when the effective number of clusters falls below 20. In Figure 3 size
distortion is again apparent when $G^*$ falls below 20, although the pattern is less
pronounced due in large part to the fact that very small values of $G^*$ are rare for
test of the coefficient on the covariate that is uncorrelated within clusters.
(Interestingly there is one set of designs for which the test is conservative, those
in which $c = 0$ - we explore this in more detail below.) Cameron, Gelbach
and Miller, who focus on the impact of a small number of clusters on inference,
also find that size distortion increases with the within cluster correlation of the
covariate under test. In sum, if the effective number of clusters exceeds 20,
then the empirical size of the cluster-robust $t$ test is close to the nominal size
with the use of Gaussian critical values. Lower values of the effective number of
clusters alert the researcher to the need to use more conservative critical values
to ensure the empirical size remains close to the nominal size.

Recall that the effective number of clusters depends on two sources of heter-
ogeneity observable to the researcher, cluster sizes and the observed values of
the covariates, and from one source of heterogeneity that is latent, the error co-
variance matrix. In order to estimate the effective number of clusters, we must
make an assumption as to the structure of the error covariance matrix. It might
be tempting to use the estimated residuals to form the estimator $\hat{\Omega}_g = \hat{u}_g \hat{u}_g^T$,
from which a feasible estimator of $G^*$ could be formed. Yet this would imply
that the sample data are used both to construct the test statistic $Z$ and to
construct critical values for $Z$, which invalidates inference. In practice, absent
knowledge of $\Omega_g$, can the assumption that the errors are perfectly correlated
within clusters ($c = 0$) provide a useful approximate measure of $G^*$? Let $G^{*A}$ denote the measure of the effective number of clusters for which $\gamma_g$ is replaced with

$$
\gamma_g^A = a^T (X^T X)^{-1} X^T (g_i g^T) X_g (X^T X)^{-1} a,
$$

with $g_i$ a vector of length $n_g$ with each element equal to one. For each of the simulation designs underpinning Figures 2 and 3, we construct $G^{*A}$ and plot this measure of the effective number of clusters against the test size in Figure 4.

The upper panel combines the results of Figures 2 and 3, showing the relation of $G^*$ to the empirical test size. The bottom panel shows the relation between $G^{*A}$ and the empirical test size. For data sets in which the empirical test size is high, $G^{*A}$ is very similar to $G^*$, indicating that the approximate measure performs well in precisely the situations it is most needed, where the empirical test size exceeds 10 percent.

### 3.1 Bias and Variation

To determine how the behavior of the test statistic can be separated into the influence from the bias of $\widehat{V}$ and the influence of the variation in $\widehat{V}$, we explore the behavior of the cluster-robust estimator $\widehat{V}$ in more detail. We first examine the link between the effective number of clusters and the (relative) mean-squared error of $\widehat{V}$. From Lemma 1, we know that as the effective number of clusters
decreases, the mean-squared error of $\hat{V}$ increases. To measure the strength of this relation with a smaller number of clusters, we compare $G^*$ with the mean-squared error for $\hat{V}$, both constructed from the same realized value of $X$.

To compute $E \left\{ \left[ \frac{\hat{V} - V_a}{V_a} \right]^2 \right\} X$, we first generate $X$ from one of the 40 cluster designs (10 cluster-size designs, each paired with 4 values of $c$). For the given value of $X$ we generate 1000 error vectors, $u$, each of length 2500. From each data set $\{X, u\}$ we compute $\hat{V}$ as in (3) and then compute the average of $\left[ \frac{\hat{V} - V_a}{V_a} \right]^2$ over the 1000 data sets that share a common $X$. We do this 5 times for each of the 40 cluster designs, resulting in 200 pairs of $G^*$ and the mean-squared error for $\hat{V}$.

We report the results for the coefficient for the covariate that is perfectly correlated within clusters in Figure 5 and for the covariate that is uncorrelated within clusters in Figure 6. Each figure contains a plot of the 200 calculated pairs together with a fitted curve. Figure 5 reveals a striking pattern. As the effective number of clusters drops below 20, the mean-squared error of $\hat{V}$ rises sharply.

![MSE as a function of ENC for Covariate X](image)

Figure 5

In Figure 6 the pattern is less pronounced, due in large part to the fact that $G^*$ rarely falls below 20 for the coefficient on the covariate that is uncorrelated within clusters. To confirm that the relation between $G^*$ and the mean-squared
error of $\hat{V}$ is similar for the two coefficients, we note that the point estimates for the two curves are similar:

$$\ln (mse) = -3.91 - 0.88G^* \text{ for the coefficient on } x_g,$$

$$\ln (mse) = -3.91 - 0.79G^* \text{ for the coefficient on } z_{gi}.$$

In sum, we find that low values of $G^*$ indicate a large mean-squared error for $\hat{V}$. Larger values of $G^*$ generally indicate a small mean-squared error although, if the within cluster error correlation is high, then the mean-squared error for $\hat{V}$ can be large even if $G^*$ is as large as 40.

Interestingly, for the designs in which the test is conservative in Figure 3, the mean-squared error is substantially larger than would be predicted on the basis of $G^*$ in Figure 6. To understand the link between the mean-squared error and the size of the test, we investigate the proportion of the mean-squared error that is attributable to the bias of $\hat{V}$. The bias, conditional on $X$, is computed as the average of $\left(\frac{\hat{V}}{V} - 1\right)$ over the 1000 simulated data sets $\{X, u\}$ that share a common $X$. We then calculate the proportion of the mean-squared error accounted for by the bias. To accord with the figures for the mean-squared error, we group the calculated proportions by $c$ (so that we have 50 estimates of the proportion for each value of $c$).
For each value of \( c \), in Figure 7 we represent the distribution of proportion measurements with a box plot (the shaded area corresponds to the interquartile range, the dark line is the median and the bars represent the remainder of the distribution excluding outliers, which are denoted with circles). For the covariate that is perfectly correlated within clusters, bias accounts for slightly less than half the mean-squared error of the cluster-robust standard error estimator. For the covariate that is uncorrelated within clusters, for which the mean-squared error is smaller, the bias generally accounts for a smaller share, with the exception of the designs in which \( c = 0 \).

In Figure 8 we present the bias. For the coefficient on the covariate that is perfectly correlated within clusters, a small value for the effective number of clusters is associated with substantial downward bias in the estimated standard error, which is the source of the size inflation noted earlier. The magnitude of the bias (the median bias when \( c = 0 \) is \(-0.4\)) suggests that the bias correction proposed by Hansen for the case of equal cluster sizes, which is to multiply \( \hat{V} \) by \( \frac{G}{G^*} = 1.01 \), will not reliably reduce the bias arising from unequal cluster sizes. For the coefficient on the covariate that is uncorrelated within clusters, the mean-squared error and the bias are far smaller. If, however, \( c = 0 \) there is a substantive upward bias in the estimated standard errors, which is the source of the conservative findings reported in Figure 3. We conclude that for the coefficient on a covariate that is highly correlated within clusters, if variation in cluster sizes reduces the effective number of clusters to a small value, then the
estimated standard error is characterized both by a downward bias and by large variation over samples. If the cluster correlation is confined to the error, so that the covariate under test is uncorrelated within clusters, we find that cluster heterogeneity leads to larger variation in cluster-robust standard errors without a downward bias, and so without a systematic inflation of test size. This last result accords with the formula in Moulton, under which the downward bias in a standard error estimator that ignores within cluster correlation vanishes for a covariate that is uncorrelated within clusters.

Figure 8

In Figure 9, we report the mean-squared error of $\hat{V}$ as a function of the estimated effective number of clusters. Because the estimated effective number of clusters assume $c = 0$, the estimated effective number of clusters generally forms a lower bound for the actual value of the effective number of clusters. In comparison with Figure 2, we see that the only points that lie far from the curve are associated with $c = 500$. If $c = 500$, then there is virtually no within cluster correlation, so it is perhaps not surprising that the assumption of perfect within cluster correlation often provides a conservative bound.

Remarks: While the simulations correspond to a single model, they reveal that the effective number of clusters is a useful guide to the finite-sample behavior of the cluster-robust $t$ statistic. Importantly, as the approximate measure of the effective number of clusters falls below 20, the empirical test size rises substantially above the nominal level.
4 Empirical Settings

To illustrate how the research design impacts the effective number of clusters, we calculate the effective number of clusters for two empirical settings in which unobserved shocks that are common within a cluster naturally arise: data on children grouped by classroom and workers grouped by industry. Importantly, growth of the sample size can occur through the addition of classrooms or industries, so that each of these settings accommodates the assumption that the number of clusters grows with the sample size.

The first setting corresponds to measurement of the impact of class size on student achievement. Krueger (1999) analyzes data from the STAR experiment in which students were randomly assigned to classrooms of different sizes, identifying the class size effect using the following regression model

\[ a_{gi} = \beta_0 + \beta_1 s_g + z_{gi}' + u_{gi}, \]

where \( a_{gi} \) is the test score of student \( i \) in classroom \( g \), \( s_g \) is the number of students in classroom \( g \) and \( z_{gi} \) captures other observed determinants of student performance, including the race, gender and socioeconomic status of student \( i \). For kindergarten students, the public use version of the data employed by Krueger contains 5,743 students grouped into 318 classrooms. In describing
regression results Krueger reports a sample size corresponding to the number of children (Table V, p. 513). Yet for the purpose of inference, even regarding a coefficient on a cluster-varying covariate, the appropriate sample size is based on the number of classrooms.

As classrooms form the clusters, the data set has $G = 318$, which appears to be well in excess of the number needed to use Gaussian critical values. Yet the number of students varies across classrooms, from a low of 9 to a high of 27. The mean number of students per classroom is 18 with a variance of 15.7. To determine how the variation in cluster sizes, together with other sources of variation in the design, impacts inference, we compute the effective number of clusters for test of hypotheses on $\beta_1$ and find $G^* = 192$. While the variation in the design across clusters has reduced the effective number of clusters to 60 percent of the actual number of clusters, the initial large number of clusters leaves the effective number of clusters sufficiently large that Gaussian inference is reliable.

The second setting corresponds to measurement of the impact of injury risk on wages. Hersch (1998) analyzes data on individual wages from the Current Population Survey, together with injury rates for workers by industry:

$$w_{gi} = \beta_0 + \beta_1 r_g + z_{gi} + u_{gi},$$

where $w_{gi}$ is the (logarithm of the) wage for individual $i$ working in industry $g$, $r_g$ is the industry-specific injury rate and $z_{gi}$ captures other observed determinants of individual wages. For male workers, the Hersch data set (Table 3, Panel B, column 1) contains 5,960 workers grouped into 211 industries.\(^4\)

As industries form clusters, the data set has $G = 211$, which again appears to be well in excess of the number needed to use Gaussian critical values. The number of workers varies dramatically across industries, ranging from a low of 1 to a high of 517. The mean number of workers per industry is 28 with a variance of 2,474. For test of hypotheses on $\beta_1$, we compute $G^* = 19$, which indicates caution in using Gaussian critical values. In this setting the degree of variation in cluster sizes, together with other sources of variation in the design, is large enough to drive the effective number of clusters into a warning area, even though the actual number of clusters is quite large.\(^5\)

The effective number of clusters calculated in these empirical examples is in line with our simulation results presented in Section 3. The Krueger setting (the coefficient of variation for cluster sizes is $cv = 22$) contains less cluster heterogeneity than the first unbalanced simulation design of one large group with 124 observations and 99 groups with 24 observations ($cv = 40$). The cluster size heterogeneity in Hersch ($cv = 176$) is similar to the variation in the designs including one large group of more than 420 observations and 99 groups with 21 observations.

\(^4\)We thank Colin Cameron for providing the data needed to replicate the Hersch results.

\(^5\)In some specifications Hersch (1998) also includes occupation-specific injury rates and clusters by either occupation or industry. Cameron, Gelbach and Miller (2011) replicate Hersch’s results for men and compare clustering on industry and occupation with clustering by industry or occupation. The impact of cluster heterogeneity in multi-way clustering scenarios is left for future research.
or less observations. For these designs, the effective number of clusters is very small compared to the actual number of clusters. Hersch provides an empirical setting in which the degree of cluster heterogeneity can lead to large increases in the mean squared error of the conventional cluster-robust variance estimator and a downward bias of test statistics. Along with the simulation results, these examples help emphasize the importance of calculating the effective number of clusters—even when the number of clusters is large—to gauge whether inference using the cluster-robust $t$ statistic is appropriate.

5 Remarks

Consistency of the cluster-robust variance estimator, together with a null distribution for the resultant $t$ test statistic as the number of clusters grows large, have previously been established under the assumption of equally sized clusters. We allow the size of clusters to vary and establish conditions under which parallel asymptotic results hold. Our theory yields a sample specific adjustment to the number of clusters, which we term the effective number of clusters. The key innovation is that it is the effective number of clusters that must grow without bound. The effective number of clusters replaces the number of clusters; if the effective number of clusters is large, then the asymptotic theory provides a reliable guide to inference.

The effective number of clusters depends on two sample specific measures in addition to variation in cluster sizes. First, the measure depends on the cluster-specific error covariance matrices. As these matrices are latent, direct calculation of the effective number of clusters is infeasible. The assumption of perfect within-cluster error correlation provides a useful lower bound on the effective number of clusters. When this feasible measure of the effective number of clusters is large, Gaussian critical values can be used with the cluster-robust $t$ test statistic.

Second, the effective number of clusters depends on how the realized values of the covariates are distributed across clusters. This is the essence of the sample specific nature of the effective number of clusters. Because in virtually all data sets the realized values of the covariates are not identical across clusters, the effective number of clusters will be less than the number of clusters. In consequence, the effective number of clusters should be measured in virtually all studies that use cluster-robust inference.

A natural question arises: If the effective number of clusters is not large, then how should critical values be obtained? Use of critical values from a $t$ distribution is argued for by Kott (1994). Although his analysis does not contain formal asymptotic results, he suggests that the degrees of freedom should be selected to mirror the variation of the cluster-robust variance estimator. As we establish that the variation of the cluster-robust variance estimator depends on the effective number of clusters, the logical implication would be to set the degrees of freedom for the $t$ distribution equal to the effective number of clusters.

The appeal of this approach would be enhanced by the ability to bound the
error introduced by use of the $t$ distribution to approximate the finite sample distribution of $Z$. To understand the difficulty in constructing such a bound, consider the behavior of $	ilde{Z} = \frac{\alpha^T(\beta - \beta_0)}{\sqrt{V_a}}$, which uses the (infeasible) unbiased estimator $\tilde{V}_a$. Even under homogeneous clusters, for which $G \tilde{V}_a \sim \chi^2(G)$, $\tilde{Z} \sim t(G)$ because the numerator and denominator of $\tilde{Z}$ are correlated. The error from approximating $\tilde{Z}$ by a $t$ distribution is magnified under cluster heterogeneity because $G^* \tilde{V}_a$ is not a $\chi^2(G^*)$ random variable. A further source of approximation error is introduced by use of $\tilde{V}_a$, rather than $\tilde{V}_a$, to construct the test statistic $Z$. Because it is difficult to bound the approximation error that these three sources induce, use of critical values from a $t(G^*)$ distribution could lead to difficulty in controlling the size of the test.

An alternative approach is to use critical values from a re-sampling method, as Cameron, Gelbach and Miller recommend when clusters are equal in size and $G$ is small. Formal treatment of these approaches under a full range of cluster heterogeneity remains a topic for further study.
6 Appendix

Verification of Result 1: Let $X^*_g$ be the $n \times k$ covariate matrix with all rows that do not correspond to cluster $g$ set to zero. 

**Part a:** Observe that because $X^T y = \sum_g X^*_g T y$,

$$\hat{\beta} = \sum_g (X^T X)\gamma X^*_g (X^*_g X_g)\gamma X^*_g T y = \sum_g A_g \hat{\beta}_g,$$

(8)

where $(X^*_g X_g)\gamma$ is a generalized inverse.\(^6\) As $V \equiv Var\left(\hat{\beta} \mid X\right)$, the cluster representation of $V$ in (4) follows directly from (8). To derive the cluster representation of $\tilde{V}$ in (4), note that $X^*_g X_g = X^*_g T X = X^*_g T X$. Hence

$$X^*_g T \left( y - X \hat{\beta} \right) = \left[ X^*_g T - X^*_g T \left( X^T X \right)^{-1} X^T \right] y$$

$$= X^*_g T \left( (X^*_g X_g)^{-1} X^*_g T - (X^T X)^{-1} X^T \right) y$$

$$= X^*_g T \left( \hat{\beta} - \hat{\beta} \right).$$

Thus

$$A_g \left( \hat{\beta}_g - \hat{\beta} \right) = (X^T X)^{-1} X^*_g T \left( y - X \hat{\beta} \right) = (X^T X)^{-1} X^*_g T \tilde{u}_g,$$

(9)

because $(y - X \hat{\beta}) = \tilde{u}$ and $X^*_g T \tilde{u} = X^*_g T \tilde{u}_g$. Hence the cluster representation of $\tilde{V}$ in (4) follows directly from (9).

**Part b:** The estimator $\tilde{V}$ is a function of the residuals

$$y - X (X^T X)^{-1} X^T y = (I_n - \Pi_X) y = \tilde{u},$$

where $\Pi_X = X (X^T X)^{-1} X^T$. These residuals can be decomposed into two components

$$\tilde{u} = (I_n - \Pi_G + \Pi_G - \Pi_X) y = (I_n - \Pi_G) y + (\Pi_G - \Pi_X) y = \tilde{u}_W + \tilde{u}_B,$$

where $\Pi_G = \sum_g X^*_g \left[ X^*_g T X^*_g \right]^{-1} X^*_g T$ is the projection operator onto the cluster specific models. The residual component $\tilde{u}_W$ captures the within cluster variation while the residual component $\tilde{u}_B$ captures the between cluster variation.

The quantity $\tilde{V}$ depends on the residuals through the linear function $X^*_g T \tilde{u}_g = X^*_g T \tilde{u}$. Hence

$$X^*_g T \tilde{u}_g = X^*_g T \tilde{u}_W + X^*_g T \tilde{u}_B.$$

\(^6\)Because the covariate matrix may not be of full column rank within cluster $g$, we use the generalized inverse $(X^*_g X_g)\gamma$ defined such that $(X^*_g X_g)\gamma X^*_g = X^*_g \gamma$ (Harville 1997, Theorem 12.3.4 part (5), p. 167). The generalized inverse, for which $(X^*_g X_g)\gamma X^*_g = X^*_g \gamma$ also holds, presents the issue that $\hat{\beta}_g$ is not uniquely defined, but any convenient choice of generalized inverse results in an identical variance estimator.
Because the least squares residuals are orthogonal to the corresponding model space,

\[ X_g^T \hat{u}_W = X_g^T (I_n - \Pi_G) y = (X_g^T - X_g^T y = 0, \]

and

\[ X_g^T \hat{u}_B = X_g^T (\Pi_G - \Pi_X) y = (X_g^T - A_g^T X^T) y \neq 0. \]

Thus, \( \hat{V} \) is only a function of between cluster variation.

**Proof of Lemma 1:** Let \( Q_g := a^T A_g \left( \hat{\beta}_g - \beta \right) \). Under part (i) of Assumption 1, \( Q_g X \sim N\left( 0, a^T A_g \text{Var} \left( \hat{\beta}_g \right) A_g^T a \right) \). Because the components \( \left( \hat{\beta}_g - \beta \right) \) are uncorrelated across clusters, \( \hat{V}_a \) is the sum of \( G \) squared, independent, normal random variables. Under normality \( \text{Var} \left( Q_g^2 \right) = 2 \left[ \text{Var} \left( Q_g \right) \right]^2 \), so

\[ \text{Var} \left( \hat{V}_a \right) = 2 \sum_g \left[ a^T A_g \text{Var} \left( \hat{\beta}_g \right) A_g^T a \right]^2. \]

The magnitude of this quantity is relative to \( V_a^2 \), which is \( \left( \sum_g a^T A_g \text{Var} \left( \hat{\beta}_g \right) A_g^T a \right)^2 \). Thus

\[ \mathbb{E} \left\{ \left[ \frac{\hat{V}_a - V_a}{V_a} \right]^2 \left| X \right. \right\} = \frac{2 \sum_g \left[ a^T A_g \text{Var} \left( \hat{\beta}_g \right) A_g^T a \right]^2}{\left( \sum_g a^T A_g \text{Var} \left( \hat{\beta}_g \right) A_g^T a \right)^2}. \]

From the definition of \( \gamma_g (\Omega, X) \),

\[ \mathbb{E} \left\{ \left[ \frac{\hat{V}_a - V_a}{V_a} \right]^2 \left| X \right. \right\} = \frac{2}{G} \left( 1 + \Gamma (\Omega, X) \right). \]

Q.E.D.

We next introduce a technical lemma that will be useful in the proof of Lemma 2. While similar results have been established by direct calculation (e.g. Ruben 1960) we wish to establish a result that specifically governs the behavior of the covariance matrix of \( \left( \hat{\beta}_g - \beta \right) \).

**Lemma A1:** If \( A \) is a \( k \times k \) matrix and \( D \) is a \( k \)-dimensional Gaussian random variable with mean 0 and variance \( \Sigma \) such that \( \Sigma A \) is diagonalizable, then the quadratic form \( D^T A D \) has the same distribution as

\[ \sum_{i=1}^k \lambda_i Z_i^2, \]

25
where \( \{Z_i\} \) is a sequence of independent standard Gaussian random variables and \( \{\lambda_i\} \) are the eigenvalues of the matrix \( \Sigma A \). Further, if \( A \) is a nonnegative definite matrix, then for all \( i \), \( \lambda_i \geq 0 \) and \( \mathbb{E} (D^T AD) \to 0 \) is sufficient to imply that \( D^T AD \overset{P}{\to} 0 \).

**Proof:** The moment-generating function of the quadratic form is

\[
\mathbb{E} \left[ \exp \left( tX^T AX \right) \right] = \int \int (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} x^T \Sigma^{-1} x + tx^T Ax \right] dx
\]

\[
= \int \int (2\pi)^{-k/2} |\Sigma|^{-1/2} \exp \left[ -\frac{1}{2} x^T (\Sigma^{-1} - 2tA) x \right] dx
\]

\[
= |\Sigma|^{-1/2} |\Sigma^{-1} - 2tA|^{-1/2}
\]

\[
= |I - 2t\lambda\Sigma|^{-1/2}.
\]

Because \( A\Sigma \) is diagonalizable, \( A\Sigma = Q\Lambda Q^{-1} \) for a diagonal matrix \( \Lambda \), so

\[
|I - 2t\lambda\Sigma|^{-1/2} = |QQ^{-1} - Q (2t\Lambda) Q^{-1}|^{-1/2}
\]

\[
= |Q (I - 2t\Lambda) Q^{-1}|^{-1/2}
\]

\[
= |Q|^{-1/2} |Q^{-1}|^{-1/2} |I - 2t\Lambda|^{-1/2}
\]

\[
= \Pi_{i=1}^k (1 - 2t\lambda_i)^{-1/2},
\]

which is the moment-generating function of the sum of \( \text{Gamma} \left( \frac{1}{2}, 2\lambda_i \right) \) distributions.

To prove the second part of the lemma, note that if \( \lambda_i \geq 0 \), then

\[
\sum_i |\lambda_{i,n}| = \text{tr} (A\Sigma) = \mathbb{E} (D^T AD).
\]

Hence \( \mathbb{E} (D^T AD) \to 0 \) implies \( \sum_i |\lambda_{i,n}| \to 0 \). Let \( W_n := \sum_{i=1}^k \lambda_{i,n} Z_{i,n}^2 \). Then \( \sum_i |\lambda_{i,n}| \to 0 \) implies

\[
\mathbb{E} W_n^2 = \left[ \sum_i \lambda_{i,n} \right]^2 + 2 \sum_i \lambda_{i,n}^2 \leq \left[ \sum_i |\lambda_{i,n}| \right]^2 + 2 \left[ \sum_i |\lambda_{i,n}| \right]^2 \to 0,
\]

which, in turn, implies that \( W_n \overset{P}{\to} 0 \).

**Remark:** If \( \Sigma A \) cannot be diagonalized, then note that \( x^T Ax = x^T A^T x \) so that

\[
x^T Ax = x^T \left( \frac{1}{2} A + \frac{1}{2} A^T \right) x
\]

can be used to derive an identical quadratic form with a matrix that can be diagonalized.
Proof of Lemma 2: We wish to establish that
\[
\mathbb{E}\left\{ \left[ \frac{\hat{V}_a - \tilde{V}_a}{V_a} \right] X \right\} \leq \frac{1}{G} \left( 1 + \Gamma^* (\Omega, X) \cdot \zeta (\Omega, X) \right). \tag{10}
\]
The setting of the problem follows from the expansion
\[
a^T \left( \hat{V} - \tilde{V} \right) a = \sum_g a^T A_g \left[ \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T - 2 \left( \hat{\beta}_g - \beta \right) \left( \hat{\beta} - \beta \right)^T \right] A_g^T a.
\]
We use the fact that \( \sum_g A_g = I \), to introduce the matrix \( (A_g - \frac{1}{G} I) \), together with the fact \( \sum_g A_g \hat{\beta}_g = \hat{\beta} \) to obtain
\[
a^T \left( \hat{V} - \tilde{V} \right) a = \frac{1}{G} a^T \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T a + \sum_g a^T A_g \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T \left[ A_g - \frac{1}{G} I \right]^T a + \frac{2}{G} a^T \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T a - 2 \sum_g a^T A_g \left( \hat{\beta}_g - \beta \right) \left( \hat{\beta} - \beta \right)^T \left[ A_g - \frac{1}{G} I \right]^T a.
\]
Combining terms on the right side yields
\[
a^T \left( \hat{V} - \tilde{V} \right) a = -\frac{1}{G} a^T \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T a + \sum_g a^T A_g \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T \left[ A_g - \frac{1}{G} I \right]^T a + \frac{2}{G} a^T \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T a - 2 \sum_g a^T A_g \left( \hat{\beta}_g - \beta \right) \left( \hat{\beta} - \beta \right)^T \left[ A_g - \frac{1}{G} I \right]^T a. \tag{11}
\]
The bias that arises from the first term in (11) is
\[
\mathbb{E}\left\{ -\frac{1}{G} a^T \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T a \right| X \right\} = -\frac{V_a}{G}, \tag{12}
\]
which is the downward bias present even when clusters are homogeneous.

The bias of the second term in (11) is
\[
\mathbb{E}\left\{ \sum_g a^T A_g \left( \hat{\beta} - \beta \right) \left( \hat{\beta} - \beta \right)^T \left[ A_g - \frac{1}{G} I \right]^T a \right| X \right\} = \sum_g a^T A_g V \left[ A_g - \frac{1}{G} I \right]^T a \tag{13}
\]
\[
= \sum_g a^T \left[ A_g - \frac{1}{G} I \right] V \left[ A_g - \frac{1}{G} I \right]^T a
\]
\[
\leq \lambda_{\text{max}}^V \sum_g \left\| \left[ A_g - \frac{1}{G} I \right]^T a \right\|^2,
\]
which is an upward bias due to the heterogeneity of the covariate matrices across clusters.
For the third term in (11), for which we make use of Lemma A1, rewrite the term as

\[ \sum_g a^T A_g (\hat{\beta}_g - \beta) (\hat{\beta} - \beta)^T \left[ A_g - \frac{1}{G} I \right] a = \sum_g (\hat{\beta}_g - \beta)^T A_g^T a a^T \left[ A_g - \frac{1}{G} I \right] (\hat{\beta} - \beta) \]  

(14)

\[ = \sum_g (\hat{\beta}_g - \beta)^T A_g^T a a^T \left[ A_g - \frac{1}{G} I \right] \sum_j A_j (\hat{\beta}_j - \beta) \]

\[ = \sum_j \sum_g (\hat{\beta}_g - \beta)^T A_g^T a a^T \left[ A_g - \frac{1}{G} I \right] A_j (\hat{\beta}_j - \beta). \]

To see that this is a quadratic form that corresponds to Lemma A1, let \( D \) from the lemma be the \( kG \times 1 \) vector

\[ \left( \begin{array}{c} A_1 (\hat{\beta}_1 - \beta) \\ A_2 (\hat{\beta}_2 - \beta) \\ \vdots \\ A_G (\hat{\beta}_G - \beta) \end{array} \right), \]

with \( A \) the \( kG \times kG \) matrix that consists of \( G \) columns of the matrices \( a a^T \left[ A_g - \frac{1}{G} I \right] \). Because the covariance matrix of \( D \) is block diagonal, with covariance matrix blocks \( B_g := A_g V a r (\hat{\beta}_g) A_g^T \), by Lemma A1 the distribution of the quadratic form can be expressed with the eigenvalues from

\[ \begin{pmatrix} B_1 a a^T [A_1 - \frac{1}{G} I] & B_1 a a^T [A_1 - \frac{1}{G} I] & \cdots & B_1 a a^T [A_1 - \frac{1}{G} I] \\ B_2 a a^T [A_2 - \frac{1}{G} I] & B_2 a a^T [A_2 - \frac{1}{G} I] & \cdots & B_2 a a^T [A_2 - \frac{1}{G} I] \\ \vdots & \vdots & \ddots & \vdots \\ B_G a a^T [A_G - \frac{1}{G} I] & B_G a a^T [A_G - \frac{1}{G} I] & \cdots & B_G a a^T [A_G - \frac{1}{G} I] \end{pmatrix}. \]

(15)

The spectrum of this matrix has \( kG - k \) zeros (which can be most easily seen by noting that (15) can be expressed as the product of a block diagonal matrix - with \( B_g a a^T \left[ A_g - \frac{1}{G} I \right] \) on the diagonal - and a projection matrix \( \Pi \) - which consists of a \( G \times G \) tiling of \( I_k \) - where \( \text{rank} (\Pi) = k \)). The eigenvectors that correspond to the non-zero eigenvalues for (15) each consist of \( G \) copies of \( v \in \mathbb{R}^k \), where \( v \) is an eigenvector of \( \sum_g B_g a a^T \left[ A_g - \frac{1}{G} I \right] \). Thus, under Lemma A1, the quadratic form (14) has the same distribution as

\[ Z^T \left[ \sum_g B_g a a^T \left[ A_g - \frac{1}{G} I \right] \right] Z \]

\[ = Z^T \left[ \sum_g \left[ B_g - \frac{1}{G} V \right] a a^T \left[ A_g - \frac{1}{G} I \right] \right] Z, \]

28
where $\sum_g B_g = V$ and the last equality follows from the fact that $\sum_g \left[ A_g - \frac{1}{G} I \right] = 0$. Standard inequalities imply

$$Z^T \left[ \sum_g \left[ B_g - \frac{1}{G} V \right] a a^T \left[ A_g - \frac{1}{G} I \right] \right] Z \leq \sum_g \left| Z^T \left[ B_g - \frac{1}{G} V \right] a a^T \left[ A_g - \frac{1}{G} I \right] Z \right| \leq \sum_g \| Z \left[ B_g - \frac{1}{G} V \right] a \| \left\| A_g - \frac{1}{G} I \right\| a^T \| Z \| \leq \| Z \|^2 \sum_g \left[ B_g - \frac{1}{G} V \right] a \| \left\| A_g - \frac{1}{G} I \right\| a^T \| Z \|.$$  \hfill (16)

Because $\| Z \|^2$ has a $\chi^2_{(k)}$ distribution

$$\mathbb{E} \left\{ \| Z \|^2 \sum_g \left[ B_g - \frac{1}{G} V \right] a \| \left\| A_g - \frac{1}{G} I \right\| a^T X \right\} = k \sum_g \left[ B_g - \frac{1}{G} V \right] a \| \left\| A_g - \frac{1}{G} I \right\| a^T \| X \right\} \leq k \left[ \sum_g \left[ B_g - \frac{1}{G} V \right] a \right] \left[ \sum_g \left[ A_g - \frac{1}{G} I \right] a \right]^{1/2} \left[ \sum_g \left[ A_g - \frac{1}{G} I \right] a \right]^{1/2}.$$  \hfill (16)

From (12), (13) and (16) we have

$$\mathbb{E} \left\{ \left| \frac{\bar{V}_a - \bar{V}_a}{V_a} \right| X \right\} \leq \frac{1}{G} + \frac{X_{\max}}{V_a} \sum_g \left\| A_g - \frac{1}{G} I \right\| a \| \left\| \left[ X^T X \right]^{-1} a \right\| ^2 + \frac{k}{V_a} \left[ \sum_g \left[ B_g - \frac{1}{G} V \right] a \right] \left[ \sum_g \left[ A_g - \frac{1}{G} I \right] a \right]^{1/2} \left[ \sum_g \left[ A_g - \frac{1}{G} I \right] a \right]^{1/2}.$$  \hfill (16)

Observe that

$$\Gamma^* (I, X) = G \left\| \left[ X^T X \right]^{-1} a \right\| ^2 \sum_g \left\| \left[ X^T X \right]^{-1} \left[ A_g - \frac{1}{G} I \right] a \right\| ^2,$$

hence

$$\sum_g \left\| \left[ A_g - \frac{1}{G} I \right] a \right\| ^2 = \sum_g \left\| \left[ X^T X \right] \left[ X^T X \right]^{-1} \left[ A_g - \frac{1}{G} I \right] a \right\| ^2 \leq \frac{\left( \lambda_{\max} \right)^2}{G} \left\| \left[ X^T X \right]^{-1} a \right\| ^2 \Gamma^* (I, X),$$

and that

$$\Gamma^* (\Omega, X) = G \left\| V_a \right\| ^2 \sum_g \left\| \left[ B_g - \frac{1}{G} V \right] a \right\| ^2,$$
thus
\[
E \left\{ \left| \frac{\tilde{V}_a - \bar{V}_a}{V_a} \right| \right\} \leq \frac{1}{G} + \frac{\lambda_{\text{max}} V}{\lambda_{\text{max}} X} \left( \frac{X^T X}{} \right)^{1/2} \left( I + \Gamma^* (I, X) \right) + \frac{k \lambda_{\text{max}}}{GV_a} \left( [X^T X]^{-1} a \right)\|V a\|^{1/2} \left( \Gamma^* (I, X) \cdot \Gamma^* (\Omega, X) \right)^{1/2}.
\]

To tie the bound more closely to Assumption 2:
\[
\lambda_{\text{max}}^V \left( \lambda_{\text{max}} X \right)^2 \left( [X^T X]^{-1} a \right)^2 \leq \left( \lambda_{\text{max}} V a \|a\|^2 \right) \left( \lambda_{\text{max}} X \right)^2 \leq M_0 M_1^2,
\]
because \( \|a\|^2 = 1 \). Further,
\[
\frac{\lambda_{\text{max}}^V \left( [X^T X]^{-1} a \right)}{V a} \leq \lambda_{\text{max}} V a \frac{\lambda_{\text{max}} X}{\lambda_{\text{min}}^X} \leq M_0 M_1.
\]

We thus have
\[
E \left\{ \left| \frac{\tilde{V}_a - \bar{V}_a}{V_a} \right| \right\} \leq \frac{1}{G} \left( 1 + \Gamma^* (I, X) \cdot \zeta (\Omega, X) \right),
\]
where \( \zeta (\Omega, X) = M_0 \left( M_1^2 + k M_1 \left[ \frac{\Gamma^* (\Omega, X)}{1+\Gamma^* (I, X)} \right]^{1/2} \right) \).

**Q.E.D.**

**Proof of Theorem 1:** By our Assumption 1(i) the random variable
\[
W = a^T \left( \beta - \beta_0 \right) / \sqrt{a^T V a} \sim N (0, 1),
\]
under the null hypothesis. Thus the test statistic
\[
Z = W \left[ \frac{a^T V a}{a^T V a} \right]^{1/2},
\]
will converge in distribution to \( W \) if \( \frac{a^T V a}{a^T V a} \rightarrow 1 \), by Slutsky's lemma.

To imply convergence in probability,
\[
P \left\{ \left| \frac{\tilde{V}_a - \bar{V}_a}{V_a} - 1 \right| > \epsilon \right\} \leq P \left\{ \left| \frac{\tilde{V}_a - \bar{V}_a}{V_a} - 1 \right| > \epsilon \right\}^{B^c} + P (B)
\]
\[
= E \left[ P \left\{ \left| \frac{\tilde{V}_a - \bar{V}_a}{V_a} - 1 \right| > \epsilon \right\}^{B^c} \ | \ X \in B^c \right] + P (B)
\]
\[
\leq E \left[ E \left\{ \left| \frac{\tilde{V}_a - \bar{V}_a}{V_a} - 1 \right| \ | \ X \in B^c \right\}^{B^c} \ | \ X \in B^c \right] + P (B),
\]
where the last inequality follows from Markov’s inequality.

Because

\[
E \left\{ \left| \frac{\tilde{V}_a}{V_a} - 1 \right| X \right\} \leq E \left\{ \left| \frac{\tilde{V}_a}{V_a} - 1 \right| \right\} + E \left\{ \left| \frac{\tilde{V}_a - \tilde{V}_a}{V_a} \right| X \right\}
\]

\[
\leq \left[ E \left\{ \left( \frac{\tilde{V}_a}{V_a} - 1 \right)^2 \right| X \right\} \right]^{1/2} + E \left\{ \left( \frac{\tilde{V}_a - \tilde{V}_a}{V_a} \right) X \right\},
\]

the definition of \( \mathcal{B} \), together with application of Lemma 1 and Lemma 2 yields

\[
P \left\{ \left| \frac{\tilde{V}_a}{V_a} - 1 \right| > \varepsilon \right\} \leq \frac{\sqrt{2}}{\sqrt{G\varepsilon}} [E \{1 + \Gamma (\Omega, X)\}]^{1/2} + \frac{1}{G\varepsilon} \left( 1 + M_0M^2_1 \Gamma^* (I, X) + kM^2_0 \Gamma^* (I, X)^{1/2} \Gamma^* (\Omega, X)^{1/2} \right) + P (\mathcal{B}).
\]

Under Assumption 3, the second two terms on the right tend to zero. For the first term on the right side (recall \( \gamma_g = a^T B_g a \) and \( \Gamma = G^{-1} V_a \)):

\[
\Gamma (\Omega, X) = \frac{G}{V_a^2} \sum_g \left( a^T \left( B_g - \frac{1}{G} V \right) a \right)^2
\]

\[
\leq \frac{G}{V_a^2} \sum_g \|a\|^2 \left\| \left( B_g - \frac{1}{G} V \right) a \right\|^2
\]

\[
= \frac{\|V a\|^2}{V_a^2} \Gamma^* (\Omega, X)
\]

\[
\leq \left( \frac{\lambda_{\max}}{V_a} \right)^2 \Gamma^* (\Omega, X) \leq M^2_0 \Gamma^* (\Omega, X).
\]

Hence Assumption 3(ii) implies that the first term also tends to zero, hence

\[
\sup_{\Omega} P \left\{ \left| \frac{\tilde{V}_a}{V_a} - 1 \right| > \varepsilon \right\} \to 0,
\]

which implies convergence in distribution. \( Q.E.D. \)
References


