Improved, Nearly Exact, Statistical Inference with Robust and Clustered Covariance Matrices using Effective Degrees of Freedom Corrections

Alwyn Young
London School of Economics

This Draft: January 2016

Abstract

I examine bias and effective degrees of freedom corrections, based upon mimicking the first two moment properties of a chi-squared variable, to statistical inference using robust and clustered covariance matrices. Simulation, using 1378 practical regression examples found in 44 experimental papers, shows that these corrections render the test statistics nearly exact in the face of ideal iid normal errors and provide large improvements in the accuracy of statistical inference in the presence of distinctly non-iid non-normal errors.

I. Introduction

Use of the Eicker (1963)-Hinkley (1977)-White (1980) robust covariance estimate and its clustered extension to correct for unknown and unspecified heteroskedasticity or within cluster correlation has become widespread in economics. For example, standard errors are calculated using one variation or another of these matrices in 1002 of 1378 OLS regressions found in a sample of 44 experimental papers examined further below. Not long after White (1980) established the asymptotic consistency of the robust variance estimate in greatest generality, it was quickly recognized that in finite samples statistical inference based upon these covariance matrices produces empirical rejection rates greater than nominal size. MacKinnon and White (1985) provided simulation evidence and, linking the problem, in the first instance, to the reduction in error variance brought about by least squares fitting, proposed various adjustments to correct for bias. These corrections, however, are inadequate, as shown in simulations by Angrist and Pischke (2009), who note that the covariance estimates are not merely biased but also much more variable than default OLS estimates, which contributes to their high rejection rates. This paper links the bias and variability of robust and clustered covariance matrix estimates to the interaction of regression design with hypothesis tests, developing easily calculable bias and "effective degrees of freedom" adjustments that, as supported by simulations using 1378 practical regressions, render test statistics using these covariance estimates nearly exact in the face of ideal iid normal errors and provide substantial improvements in the accuracy of inference in situations with non-iid non-normal errors.

The corrections suggested in this paper are motivated by two observations and two claims. The observations are trivial, noting that the chi-squared variable that underlies the t-statistic has a variance equal to twice its mean and that the robust and clustered covariance matrices can be re-expressed as quadratic forms whose first two moments, following suitable adjustment, mimic the chi-squared property. The claims are that in the face of iid normal errors the resulting test statistic is very nearly distributed t- with the degrees of freedom of its pseudo chi-squared denominator, and that the same distribution improves the evaluation of the test statistic in non-normal non-iid situations. The method works for the very simple reason that the degrees of freedom calculation for the conventional t-statistic reflects the number of orthogonal linear combinations of the disturbances used in its calculation, and the effective degrees of freedom correction similarly calculates how regression design interacts with hypothesis tests to make the robust and clustered covariance matrices dependent upon a reduced number of orthogonal combinations of the disturbances. This then identifies both the bias and variability of the covariance estimates in the case of ideal iid normal errors. The reduced dimensionality remains relevant, however, regardless of the characteristics of the error term, explaining the value of the method when extended to non-iid non-normal settings.

This paper builds on a number of earlier theoretical results. Chesher and Jewitt (1987) identified the link between leverage and bias bounds for robust covariance estimates. Chesher (1989) then went on to note that these matrices could be re-expressed as quadratic forms and, using Hajek's (1962) inequalities, established that the tails of the distribution of the test statistic are bounded by t-distributions with degrees of freedom determined by regression design. The bounds on potential bias and effective degrees of freedom corrections further below extend these results to clustered covariance estimates and a variant of the robust measure not considered in these earlier papers. Welch, first in a particular example (1936) and then in increasing generality (1938, 1947), developed the idea of approximating the distribution of a statistic based upon the sum of unequally weighted chi-squared variables using adjustments that mimic the first two moment properties of a chi-squared variable, in the process calculating "effective degrees of freedom" which, in essence, identify the variance of the variance estimate. It is but a small step to realize that the robust and clustered variance estimates are a version of Welch's problem, and that further analysis of the quadratic forms first examined by Chesher would yield the precise degrees of freedom correction required for each hypothesis test, transforming Chesher's broad bounds into a nearly exact distribution based upon the interaction of hypothesis tests with regression design.

The idea of using effective degrees of freedom (edf) corrections for the robust and clustered variance estimates has been explored in earlier papers. Kott (1994, 1996) proposed such corrections for his own bias corrected refinements of the clustered covariance estimate, as did Bell and McCaffrey (2002) using extensions of the MacKinnon and White robust bias correction methods to the clustered case. In an effort to popularize these improvements, Imbens and Kolesar (2015) promote the Bell and McCaffrey approach. This paper extends these earlier analyses, which typically only consider a small handful of artificial examples, by applying these techniques to 1378 practical regression designs used in published papers. I show that the MacKinnon and White and Bell and McCaffrey corrections frequently cannot be applied as, given regression design, in 1/3rd of published robust or clustered regressions they would require inverting a singular matrix. I find, however, that bias corrections based upon the baseline robust and clustered covariance estimates, which can be easily calculated for any regression, yield identical results to these other methods when both can be applied. Finally, I explicitly link the edf literature to regression design, both in establishing theoretical bounds and providing illustrative examples of how the interaction between hypothesis tests and regression design determine bias and effective degrees of freedom.

The paper proceeds as follows: Section II below shows how quadratic forms that are not based on idempotent matrices can be adjusted to mimic the moment properties of the chi-squared variable, in the process producing bias and effective degrees of freedom corrections. Section III presents convenient computational formulae and derives bounds on the bias and edf corrections for the robust and clustered matrices and their diverse variations. Maximal regression leverage plays an important role in determining these bounds and I provide examples showing how the interaction of the hypothesis test with regression design determines to what degree these bounds are attained. Section IV presents empirical simulations using the 44 paper sample mentioned earlier above, which provides a wide array of practical applications. I begin by showing that regression design in the typical published regression is quite poor, with about 1/3rd of regressions having a maximal leverage of 1, which renders the MacKinnon and White and Bell and McCaffrey corrections unusable, whatever their merits. I then show the extraordinary performance achieved by the bias and edf corrections in simulations with iid normal and decidedly non-iid non-normal disturbances. Section V concludes by considering the possibility and value of extensions to non-OLS settings. Stata ado files on my website allow users to ask for these corrections in their robust and clustered regressions.

II. Effective Degrees of Freedom Corrections

I use the familiar presentation of the t-statistic to establish notation and motivate the bias and edf corrections of the robust and clustered covariance matrices. In the n-observation k-regressor OLS regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,[1] with $\boldsymbol{\varepsilon}$ iid ~ $N(0, \sigma^2)$, let **b** denote the

[1] I follow standard notation, with bold capital and lowercase letters denoting matrices and column vectors, respectively.

estimated coefficients and **e** the estimated residuals. The symmetric and idempotent matrix **M** = **I**(n) **– X(X′X)**$^{-1}$**X′** is the "residual maker" as **e** = **My** = **Mε**. We (correctly) estimate the variance of **b** using V= **(X′X)**$^{-1}$**e′e**/(n-k) and test whether a linear combination **w** of **b** is significantly different from a null value w_0 using the statistic

$$(1)\quad \frac{\mathbf{w}'\mathbf{b}-\mathrm{w}_0}{\sqrt{\mathbf{w}'\mathbf{V}\mathbf{w}}} = \frac{\dfrac{\mathbf{w}'\mathbf{b}-\mathrm{w}_0}{\sqrt{\sigma^2\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}}}{\sqrt{\dfrac{1}{\mathrm{n}-\mathrm{k}}\dfrac{\mathbf{e}'\mathbf{e}}{\sigma^2}}}$$

Since, under the null, **w′b** ~ N(w_0,σ^2**w′(X′X)**$^{-1}$**w)**, the numerator is a standard normal variable while the denominator is the square root of an idempotent quadratic form in the standard normal vector **ε**/σ as

$$(2)\quad \frac{\mathbf{e}'\mathbf{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'\mathbf{M}'\mathbf{M}\boldsymbol{\varepsilon}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{M}\frac{\boldsymbol{\varepsilon}}{\sigma}$$

With trace(**M**) = n-k,[2] this is a chi-squared variable with n-k degrees of freedom and, consequently, has expectation n-k and variance 2(n-k). As the estimates **b** are statistically independent of **e**, the test statistic (1) is the ratio of a standard normal variable to the square root of an independently distributed chi-squared variable divided by its degrees of freedom and, consequently, follows the t-distribution.

Consider now the case where the variance of **b** is estimated using an alternative estimator $\mathbf{V}_i \neq \mathbf{V}$. The usual test statistic is given by

$$(3)\quad \frac{\mathbf{w}'\mathbf{b}-\mathrm{w}_0}{\sqrt{\mathbf{w}'\mathbf{V}_i\mathbf{w}}} = \frac{\dfrac{\mathbf{w}'\mathbf{b}-\mathrm{w}_0}{\sqrt{\sigma^2\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}}}{\sqrt{\dfrac{\mathbf{w}'\mathbf{V}_i\mathbf{w}}{\sigma^2\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}}} = \frac{\dfrac{\mathbf{w}'\mathbf{b}-\mathrm{w}_0}{\sqrt{\sigma^2\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}}}{\sqrt{\dfrac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{B}_i\dfrac{\boldsymbol{\varepsilon}}{\sigma}}}$$

where, in the last expression, I have assumed that $\mathbf{V}_i$ is such that the term in the square root in the denominator can be re-expressed as a standard normal quadratic form with matrix $\mathbf{B}_i$. Following the usual properties of quadratic forms, if **ε** is iid normal this has mean μ = trace($\mathbf{B}_i$) and variance ν = 2*trace($\mathbf{B}_i\mathbf{B}_i$).[3] If $\mathbf{B}_i$ is idempotent, this quadratic form is a chi-squared variable with ν = 2μ. If $\mathbf{B}_i$ is not idempotent, it can be modified to mimic this

[2] Trace(**M**) = n - trace(**X(X′X)**$^{-1}$**X′**) = n - trace(**X′X(X′X)**$^{-1}$) = n-k.

[3] Iid is sufficient for the mean property, while the addition of normality establishes the variance.

moment property by multiplying it by 2μ/ν, so that its mean is $2\mu^2/\nu$ and variance $4\mu^2/\nu$. Recognizing that we will want to divide the resulting "chi-squared" variable by its degrees of freedom, we see that we should divide $\mathbf{V}_\mathrm{i}$ by μ to form the test statistic:

$$(4)\quad \frac{\mathbf{w'b}-\mathrm{w}_0}{\sqrt{\mathbf{w}'\dfrac{\mathbf{V}_\mathrm{i}}{\mu}\mathbf{w}}} = \frac{\dfrac{\mathbf{w'b}-\mathrm{w}_0}{\sqrt{\sigma^2\mathbf{w}'(\mathbf{X'X})^{-1}\mathbf{w}}}}{\sqrt{\dfrac{1}{\mu}\dfrac{\mathbf{w}'\mathbf{V}_\mathrm{i}\mathbf{w}}{\sigma^2\mathbf{w}'(\mathbf{X'X})^{-1}\mathbf{w}}}} = \frac{\dfrac{\mathbf{w'b}-\mathrm{w}_0}{\sqrt{\sigma^2\mathbf{w}'(\mathbf{X'X})^{-1}\mathbf{w}}}}{\sqrt{\dfrac{1}{2\mu^2/\nu}\dfrac{\boldsymbol{\varepsilon}'}{\sigma}\left(\mathbf{B}_\mathrm{i}\dfrac{2\mu}{\nu}\right)\dfrac{\boldsymbol{\varepsilon}}{\sigma}}}$$

Since the denominator, with normal disturbances, remains independent of the numerator, we can say that this test statistic is distributed t with $2\mu^2/\nu$ "effective degrees of freedom".

Intuition for why this approach improves statistical inference, even in the presence of non-iid non-normal disturbances, can be acquired by thinking about the quadratic forms that lie in the denominators of (1), (3) and (4). When the disturbances are known to be iid, the quadratic form in the denominator of (1) has an expectation of 1, i.e. the sum of squared residuals divided by degrees of freedom provides an unbiased estimate of the variance σ^2. To allow for the possibility that the disturbances are not iid, we use "robust" covariance estimates in (3). Unfortunately, as shown further below, depending upon the interaction of the hypothesis test with regression design, these place uneven weight on estimated residuals. Moreover, the estimated residuals themselves tend to be large or small in a manner that depends not only on the heteroskedasticity of the disturbance process, but also on regression design. By calculating the bias μ of the variance estimate in the presence of iid disturbances and using it to divide the variance estimate in (4), we remove the systematic bias in the estimate of variance associated with regression design. This generally provides gains, even when the error process is not iid and hence the true finite sample bias is unknown.

Turning to effective degrees of freedom, symmetric matrices allow the decomposition $\mathbf{U\Lambda U'}$, where $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues and $\mathbf{U}$ is the matrix whose columns are the corresponding mutually orthogonal eigenvectors ($\mathbf{U'U} = \mathbf{I}$). Consequently, a quadratic form $\boldsymbol{\varepsilon}'\mathbf{A}\boldsymbol{\varepsilon}$ can be reexpressed as $\Sigma\lambda_i\eta_i^2$, where, with $\mathbf{u}_\mathrm{i}$ denoting the i^th eigenvector and λ_i its corresponding eigenvalue, the $\eta_i = \mathbf{u}_\mathrm{i}'\boldsymbol{\varepsilon}$ are mutually orthogonal linear combinations of the disturbances. In the case of the denominator of (1), as re-expressed in (2), the quadratic form involves n-k eigenvalues equal to 1 and k equal to 0, i.e. the estimate of variance involves the square of n-k mutually independent variables. In the case of a general symmetric covariance estimate $\mathbf{V}_i$, the number and magnitude of non-zero eigenvalues will vary. However, the trace of a matrix $\mathbf{B}_i$ equals the sum of its eigenvalues, while the trace of $\mathbf{B}_i\mathbf{B}_i$ equals the sum of the squared eigenvalues, so the $2\mu^2/\nu$ effective degrees of freedom calculated above reduces to

$$(5)\quad \frac{\left(\sum_i \lambda_i\right)^2}{\sum_i \lambda_i^2}.$$

(5) is approximately equal to the number of equally "large" eigenvalues. By calculating this measure under the assumption of iid normal disturbances, edf corrections calculate the way in which the interaction between the hypothesis test and regression design places disproportionate weight on a reduced number of orthogonal combinations of the disturbances. Since the variance of the variance estimate is determined by the number of random variables that go into its computation, this provides information on the thickness of the tails of the test statistic distribution. Once again, by accounting for the systematic change in the distribution due to regression design, this calculation provides improvements even in cases where the disturbances are neither iid nor normal.

III. Formulae, Bounds and Intuitive Examples

In this section I present the formulae that underlie the calculation of the bias and effective degrees of freedom corrections associated with different versions of the robust and clustered covariance matrices, establish theoretical bounds on these measures, and provide specific examples that show how they are, in practice, determined by the interaction between the hypothesis test and regression design. Bias and effective degrees of freedom are calculated for each quadratic form using the assumption of ideal iid normal errors, but, as argued above and shown below, these provide substantial improvements to statistical inference in less than ideal circumstances.

The formulas for the robust and clustered covariance estimates are given by:

$$(6)\quad \mathbf{V}_{\mathrm{hc1}} = \mathrm{c}_{\mathrm{hc1}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{e_i^2\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \mathbf{V}_{\mathrm{cc1}} = \mathrm{c}_{\mathrm{cc1}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\mathbf{e_g}\mathbf{e_g}'\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{V}_{\mathrm{hc2}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left\{\frac{e_i^2}{m_{ii}}\right\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \mathbf{V}_{\mathrm{cc2}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\mathbf{M}_{\mathbf{gg}}^{-1/2}\mathbf{e_g}\mathbf{e_g}'\mathbf{M}_{\mathbf{gg}}^{-1/2}\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$$\mathbf{V}_{\mathrm{hc3}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\left\{\frac{e_i^2}{m_{ii}^2}\right\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \quad \mathbf{V}_{\mathrm{cc3}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\mathbf{M}_{\mathbf{gg}}^{-1}\mathbf{e_g}\mathbf{e_g}'\mathbf{M}_{\mathbf{gg}}^{-1}\}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

where I use the notation {} to denote a diagonal or block-diagonal matrix, while m_{ii} denotes the ith diagonal element of the residual maker **M**, $\mathbf{M_{gg}}$ the block diagonal element associated with cluster g, and e_i and $\mathbf{e}_{\mathrm{g}}$ the estimated observation and cluster residuals. $\mathbf{V}_{\mathrm{hc1}}$ is the baseline “heteroskedasticity-consistent” robust estimate of covariance with the finite sample correction $\mathrm{c_{hc1}}$=n/(n-k) originally suggested by Hinkley (1977) to counteract the mean reduction in squared residuals brought on by OLS fitting. $\mathbf{V}_{\mathrm{hc2}}$ and $\mathbf{V}_{\mathrm{hc3}}$ are alternative corrections proposed by MacKinnon and White (1985).[4] $\mathbf{V}_{\mathrm{hc2}}$ divides by the mean bias in the face of iid errors in the variance estimate of observation i itself,[5] while $\mathbf{V}_{\mathrm{hc3}}$ overcorrects in an attempt to improve the poor performance of $\mathbf{V}_{\mathrm{hc2}}$. $\mathbf{V}_{\mathrm{cc1}}$ - $\mathbf{V}_{\mathrm{cc3}}$ are corresponding “cluster-consistent” versions of these matrices, with the introduction of cc1 generally attributed to Liang and Zeger (1986)[6] and the cc2 and cc3 corrections to Bell and McCaffrey (2002). It goes without saying that the hc2/hc3 and cc2/cc3 bias corrections can only be applied when the minimum value of m_{ii} and the minimum eigenvalue of the $\mathbf{M_{gg}}$ matrices are greater than zero, respectively. I take this as given when discussing their characteristics. As shown in the next section, however, in practice these restrictions are often not met, which limits their usefulness.

For the hypothesis test $\mathbf{w'b}$ = $\mathrm{w_0}$, these matrices produce quadratic forms in standard normal variates of the kind described in the denominator of (3) above

$$(7)\quad \frac{\mathbf{w'V_{\mathrm{i}}w}}{\sigma^2\mathbf{w'(X'X)^{-1}w}} = \frac{\boldsymbol{\varepsilon}'}{\sigma}\mathbf{B}_{\mathrm{i}}\frac{\boldsymbol{\varepsilon}}{\sigma}$$

If one defines

[4] The hc3 correction, as proposed by MacKinnon and White (1985), was actually the jackknife, but subsequent simulations found it produced results very similar to $\mathbf{V}_{\mathrm{hc3}}$ above, which has come to be known as hc3 in texts (Davidson and MacKinnon 1993) and computer software (Stata).

[5] $\mathrm{E}(e_i^2) = \mathrm{E}(\mathbf{m}_{\mathrm{i}}'\boldsymbol{\varepsilon\varepsilon}'\mathbf{m}_{\mathrm{i}}) = \mathbf{m}_{\mathrm{i}}'\mathrm{E}(\boldsymbol{\varepsilon\varepsilon}')\mathbf{m}_{\mathrm{i}} = \mathbf{m}_{\mathrm{i}}'\{\sigma^2\}\mathbf{m}_{\mathrm{i}} = \sigma^2 m_{ii}$, where $\mathbf{m}_{\mathrm{i}}$ denotes the ith column of **M** and I have made use of the fact that the symmetry and idempotency of **M** imply that $\mathbf{m}_{\mathrm{i}}'\mathbf{m}_{\mathrm{i}} = \sigma^2 m_{ii.}$

[6] The appropriate clustered finite sample correction $\mathrm{c_{cc1}}$ is a matter of some debate and confusion. With $\mathrm{n_c}$ denoting the number of clusters, Stata, for example, uses $\mathrm{c_{cc1}} = (\mathrm{n_c}(\mathrm{n}-1))/((\mathrm{n_c}-1)(\mathrm{n}-\mathrm{k}))$ in its reg and areg commands, even when including fixed effects, and yet $\mathrm{c_{cc1}} = (\mathrm{n_c}(\mathrm{n}-1))/((\mathrm{n_c}-1)(\mathrm{n}+\mathrm{k_{fe}}-\mathrm{k}))$, where $\mathrm{k_{fe}}$ denotes the number of fixed effects, when executing identical regressions in its xtreg fixed effects command. This is a somewhat moot point, as bias varies in $\mathrm{V_{cc1}}$, so no fixed correction can eliminate it. In using $\mathrm{V_{cc1}}$ below, I apply Stata’s baseline reg/areg $\mathrm{c_{cc1}}$ correction in all regressions. In practice, as n and $\mathrm{n_c}$ are typically large, this amounts to an n/(n-k) correction and produces, on average, an approximately unbiased covariance estimate.

$$(8)\quad \mathbf{z}' = \mathbf{z}'_{\mathrm{hc1}} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad \mathbf{z}'_{\mathrm{hc2}} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{m_{ii}^{-½}\}, \quad \mathbf{z}'_{\mathrm{hc3}} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{m_{ii}^{-1}\},$$

$$\mathbf{z}' = \mathbf{z}'_{\mathrm{cc1}} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}', \quad \mathbf{z}'_{\mathrm{cc2}} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\mathbf{M}_{\mathbf{gg}}^{-½}\}, \quad \mathbf{z}'_{\mathrm{cc3}} = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\{\mathbf{M}_{\mathbf{gg}}^{-1}\},$$

$$\text{and } \mathrm{c}_{\mathrm{hc2}} = \mathrm{c}_{\mathrm{hc3}} = \mathrm{c}_{\mathrm{cc2}} = \mathrm{c}_{\mathrm{cc3}} = 1,$$

with some algebraic manipulation the relevant $\mathbf{B}_{\mathrm{i}}$ are found to be given by

$$(9)\quad \mathbf{B}_x = \frac{\mathrm{c}_x}{\mathbf{z}'\mathbf{z}}\mathbf{M}\{z_{x,i}^2\}\mathbf{M} \quad (x = \mathrm{hc1, hc2, or\ hc3}),$$

$$\mathbf{B}_x = \frac{\mathrm{c}_x}{\mathbf{z}'\mathbf{z}}\mathbf{M}\{\mathbf{z}_{x,\mathbf{g}}\mathbf{z}'_{x,\mathbf{g}}\}\mathbf{M} \quad (x = \mathrm{cc1, cc2, or\ cc3}),$$

where $\mathbf{z}_{x,i}$ denotes the i$^{\mathrm{th}}$ term of $\mathbf{z}_x$ and $\mathbf{z}_{x,\mathbf{g}}$ the terms associated with cluster group g. Using μ_x = trace($\mathbf{B}_x$) and ν_x = 2*trace($\mathbf{B}_x\mathbf{B}_x$), the mean and variance of these quadratic forms are easily calculated using the formulas:

$$(10)\quad \mu_x = \frac{\mathrm{c}_x}{\mathbf{z}'\mathbf{z}}\sum_i z_{x,i}^2 m_{ii}, \quad \nu_x = \frac{2\mathrm{c}_x^2}{(\mathbf{z}'\mathbf{z})^2}\sum_i\sum_j (z_{x,i} m_{ij} z_{x,j})^2 \quad (x = \mathrm{hc1, hc2, or\ hc3}),$$

$$\mu_x = \frac{\mathrm{c}_x}{\mathbf{z}'\mathbf{z}}\sum_{\mathbf{g}} \mathbf{z}'_{x,\mathbf{g}}\mathbf{M}_{\mathbf{gg}}\mathbf{z}_{x,\mathbf{g}}, \quad \nu_x = \frac{2\mathrm{c}_x^2}{(\mathbf{z}'\mathbf{z})^2}\sum_{\mathbf{g}}\sum_{\mathbf{h}} (\mathbf{z}'_{x,\mathbf{g}}\mathbf{M}_{\mathbf{gh}}\mathbf{z}_{x,\mathbf{h}})^2 \quad (x = \mathrm{cc1, cc2, or\ cc3}).$$

with effective degrees of freedom given by $2\mu^2/\nu$, as described earlier above.

Putting aside finite sample corrections, for any given hypothesis test $\mathbf{w}'\mathbf{b}$ = w_0 the following inequalities hold (as proven in the appendix):

$$(11)\quad \text{(a)}\ \ 0 \leq m_{ii}^{\min} \leq \mu_{\mathrm{hc1}} \leq m_{ii}^{\max} \leq 1, \quad \mu_{\mathrm{hc2}} = 1, \quad 1 \leq (m_{ii}^{\max})^{-1} \leq \mu_{\mathrm{hc3}} \leq (m_{ii}^{\min})^{-1}$$

$$\text{(b)}\ \ 0 \leq \lambda^{\min}(\mathbf{M}_{\mathbf{gg}}) \leq \mu_{\mathrm{cc1}} \leq \lambda^{\max}(\mathbf{M}_{\mathbf{gg}}) \leq 1, \quad \mu_{\mathrm{cc2}} = 1, \quad 1 \leq \lambda^{\max}(\mathbf{M}_{\mathbf{gg}})^{-1} \leq \mu_{\mathrm{cc3}} \leq \lambda^{\min}(\mathbf{M}_{\mathbf{gg}})^{-1}$$

$$\text{(c)}\ \ \lambda^{\min}(\mathbf{M}_{\mathbf{gg}}) \leq m_{ii}^{\min}, \ m_{ii}^{\max} \leq \lambda^{\max}(\mathbf{M}_{\mathbf{gg}}),$$

$$\text{with} > \text{whenever } m_{ij} \neq 0 \ \forall\, j \neq i \ \text{ in the } \mathbf{M}_{\mathbf{gg}} \text{ associated with } m_{ii}^{\min} \text{ or } m_{ii}^{\max}$$

$$\text{(d)}\ \ n-k \geq \mathrm{edf}_x \geq 1 \quad (x = \mathrm{hc1, hc2\ or\ hc3})$$

$$\min(n_{\mathrm{c}}-1, n-k) \geq \mathrm{edf}_{\mathrm{cc1}} \geq 1, \ \min(n_{\mathrm{c}}, n-k) \geq \mathrm{edf}_{\mathrm{cc1}} \geq 1 \quad (x = \mathrm{cc2\ or\ cc3})$$

where $m_{ii}^{\min}$ and $m_{ii}^{\max}$ are the minimum and maximum diagonal elements of the residual maker $\mathbf{M}$, $\lambda^{\min}(\mathbf{M}_{\mathbf{gg}})$ and $\lambda^{\max}(\mathbf{M}_{\mathbf{gg}})$ the minimum and maximum eigenvalues of the cluster sub-matrices $\mathbf{M}_{\mathbf{gg}}$ of $\mathbf{M}$, and $\mathrm{n_c}$ equals the number of clusters.

Result (11a) indicates that, absent the degrees of freedom correction, $\mathbf{V}_{\mathrm{hc1}}$ is downward biased because of the reduction in error variance brought about by OLS fitting. As $1 \geq m_{ii} \geq 0$, with the typical n/(n-k) correction it may be upward or downward biased. $\mathbf{V}_{\mathrm{hc2}}$ is unbiased for all **w**, while $\mathbf{V}_{hc3}$ overcompensates and is upward biased. (11b) shows that similar inequalities hold for the cci cluster consistent versions, with bias bounds determined by the minimum and maximum eigenvalues of the cluster sub-matrices $\mathbf{M_{gg}}$ of the residual maker **M**. (11c) indicates that these eigenvalue limits lie strictly outside the limits of the diagonal elements of **M** whenever there exists a non-zero off-diagonal element m_{ij} in the cluster sub-matrix $\mathbf{M_{gg}}$ associated with $m_{ii}^{\max}$ or $m_{ii}^{\min}$. This indicates that, with iid errors, the bias of the clustered covariance estimate has greater potential dispersion than its robust counterpart, although there is no ordering for any given hypothesis test. (11d) establishes bounds on effective degrees of freedom. There is no theoretical or empirical ordering between the edf of the various matrices. In practice, however, the edf for the various hci (or cci) are virtually identical (see Section IV), i.e. increased bias is almost exactly offset by increased variance. This reflects the common way in which they depend upon a limited number of residuals, as seen in the examples which follow below.

The practical characteristics of the robust and clustered covariance matrices derive from the interaction between a given hypothesis test and regression design. To explore this, it might be helpful to first remind the reader of some terminology. The hat matrix $\mathbf{H} = \mathbf{X(X'X)^{-1}X'}$ puts the "hat" on **y,** as $\mathbf{\hat{y} = X\beta = Hy}$. The element h_{ij} is the derivative of the predicted value of y_i with respect to observation y_j. h_{ii}, the influence of observation y_i on its own predicted value, is known as the leverage of observation i. Leverage ranges between 0 and 1, as **H** is idempotent and symmetric and hence:

$$(12)\quad h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2$$

Leverage averages k/n[7] and when h_{ii} = k/n for all i, the regression is considered perfectly balanced. The residual maker **M** equals **I**(n) – **H**, so **H** appears implicitly in the results in (11) above. When, for example, regression design is perfectly balanced, m_{ii} = 1- h_{ii} = 1- k/n for all i and with the typical n/(n-k) finite sample correction the V_{hc1} estimate of variance is unbiased for all hypothesis tests **w**. Leverage features prominently in the analysis of the "robustness" of regressions, i.e. the sensitivity of coefficient estimates to particular observations, where it is easily shown that its influence depends upon an interaction with the error term (see Huber

[7] $\Sigma_i\, h_{ii}$ = trace(**H**) = trace($\mathbf{X(X'X)^{-1}X'}$) = trace($\mathbf{X'X(X'X)^{-1}}$) = k, so $\Sigma_i\, h_{ii}$/n = k/n.

1981, Fox 2008). In a similar fashion, the impact of leverage on bias and, most importantly, effective degrees of freedom depends upon its interaction with the hypothesis test $\mathbf{w}$.

Consider the case of the robust hc1 estimate of the variance of a linear combination $\mathbf{w}$ of the estimated coefficients $\mathbf{b}$, $\mathbf{w}'\mathbf{V}_{\text{hc1}}\mathbf{w}$. From (6) above, we see that, putting aside the finite sample correction, this is given by $\mathbf{z}'\{e_i^2\}\mathbf{z}$, where $\mathbf{z}' = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\{e_i^2\}$ is the diagonal matrix composed of estimated residuals. If $\mathbf{w}' = \mathbf{x}_\text{i}'$, where $\mathbf{x}_\text{i}'$ is the i^th^ observation row of $\mathbf{X}$, then $\mathbf{z} = \mathbf{h}_\text{i}'$, the i^th^ row of $\mathbf{H}$. Thus, in this case, the estimate of variance is a leverage weighted average of the estimated disturbances. As h_{ii} increases, the weight on the residual for observation i increases and the magnitude of the weights on the other observations decreases, as can be seen by examining (12) above. In the limit, h_{ii} equals 1, all h_{ij} $j \neq i$ equal 0, and the variance estimate depends upon only one residual and has one effective degree of freedom. The weighting implicitly created by leverage in this hypothesis test also determines bias, as the residuals have unequal expected variance.[8] The hc2 and hc3 corrections, dividing each e_i^2 by m_{ii} or m_{ii}^2, can eliminate or overcorrect bias, but they cannot correct for the weights placed on a limited number of residuals and hence retain roughly the same miniscule degrees of freedom.

As a counter-example, consider the hypothesis test that uses $\mathbf{w} = n\bar{\mathbf{x}}$, so $\mathbf{w}'\mathbf{b} = n\bar{\mathbf{x}}'\mathbf{b} = \text{w}_0$ is a test of whether the predicted mean of $\mathbf{y}$ equals w_0/n. If the regression contains a constant term, it can be shown, using standard results on the inverses of partitioned matrices, that $\mathbf{z}' = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ equals a row vector of 1s. Thus, $\mathbf{z}'\{e_i^2\}\mathbf{z}$ is an equally weighted sum of all residuals and the effective degrees of freedom is n-k, the theoretical maximum. Moreover, with the n/(n-k) finite sample correction, the expected bias is zero. This example shows that, with an appropriately chosen hypothesis test, regression design has no influence whatsoever on bias or the effective degrees of freedom of the regression.

The two examples given above can also be interpreted in terms of the quadratic forms. As noted earlier, the robust variance estimate involves the quadratic form $(\boldsymbol{\varepsilon}/\sigma)'\mathbf{B}_x(\boldsymbol{\varepsilon}/\sigma)$, with $\mathbf{B}_x$ equal to a constant times $\mathbf{M}\{\text{z}_{x,i}^2\}\mathbf{M}$. When $\mathbf{w} = n\bar{\mathbf{x}}$, $\mathbf{z}$ is a row vector of 1s, $\mathbf{B}_x$ equals $\mathbf{M}$, and the robust variance estimate actually reduces to the default OLS covariance estimate. With $\mathbf{M}$'s n-k eigenvalues equal to 1 and k eigenvalues equal to 0, the dimensionality of the disturbances affecting the variance estimate is n-k. If $\mathbf{w}' = \mathbf{x}_\text{i}'$, $\mathbf{z} = \mathbf{h}_\text{i}'$ and as the number of

[8] As can be seen from (10), the bias in the estimated variance for this hypothesis test is given by $\Sigma_\text{j} h_{ij}^2 m_{jj}/h_{ii}$.

non-zero elements of $\mathbf{h}_{\mathrm{i}}$ falls below n-k the number of non-zero eigenvalues in $\mathbf{M}\{\mathrm{z}_{x,i}^{2}\}\mathbf{M}$ falls with them. In the limit, as h_{ii} equals 1, $\mathbf{z}$ has only one non-zero element, so $\mathbf{M}\{\mathrm{z}_{x,i}^{2}\}\mathbf{M}$ has only one non-zero eigenvalue and the dimensionality of disturbances affecting the variance estimate is 1. It is helpful, in thinking about this, to recall that a linear combination of iid normal variables behaves like a single normal variable, so it is not the number of disturbances ($\boldsymbol{\varepsilon}$) that goes into the calculation that matters but the number of effective orthogonal linear combinations. Moreover, to facilitate intuition I have couched the discussion in terms of the number of distinct non-zero eigenvalues, but it is fairly obvious that what matters is the number of "large" eigenvalues, as discussed in the previous section. The hc2/hc3 adjustments, by reweighting the residuals, can influence the relative size of eigenvalues and hence the effective degrees of freedom, but this effect is dominated by the fact that they depend upon the same $\mathbf{z}$ weighted combination of a limited number of residuals.

The extension of these ideas to clustered covariance estimates is fairly straightforward. The symmetry and idempotency of $\mathbf{H}$ implies that for the block element $\mathbf{H_{gg}}$ we have $\mathbf{H_{gg}} = \mathbf{H_{gg}H_{gg}} + \mathbf{H_{\sim g\sim g}}$, where $\mathbf{H_{\sim g\sim g}} = \Sigma_{\mathbf{h}}\mathbf{H_{gh}H_{gh}}'$ and $\mathbf{H_{gh}}$ equals the elements of $\mathbf{H}$ associated with the $\mathbf{g}$ x $\mathbf{h}$ cluster observations. Let λ_i denote an eigenvalue of $\mathbf{H_{gg}}$ and $\mathbf{u}_{\mathrm{i}}$ the corresponding eigenvector. Using $\mathbf{H_{gg}u}_{\mathrm{i}} = \lambda_i\mathbf{u}_{\mathrm{i}}$, we have:

$$(13)\quad \mathbf{H_{gg}u}_i = \lambda_i\mathbf{u}_i = \mathbf{H_{gg}H_{gg}u}_i + \mathbf{H_{\sim g\sim g}u}_i = \mathbf{H_{gg}}\lambda_i\mathbf{u}_i + \mathbf{H_{\sim g\sim g}u}_i$$

$$\rightarrow \mathbf{u}_i'\lambda_i\mathbf{u}_i = \mathbf{u}_i'\lambda_i^2\mathbf{u}_i + \mathbf{u}_i'\mathbf{H_{\sim g\sim g}u}_i$$

As λ_i goes to 1, $\mathbf{u}_{\mathrm{i}}'\mathbf{H_{\sim g\sim g}u}_{\mathrm{i}}$ is driven to 0. Next, consider the hypothesis test $\mathbf{w} = \mathbf{X_g u}_{\mathrm{i}}$, so that $\mathbf{z}' = \mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{u}_{\mathrm{i}}'\mathbf{H_g}$, where $\mathbf{H_g}$ equals the rows of $\mathbf{H}$ associated with cluster g. The quadratic form for the clustered cc1 case involves a constant times the matrix $\mathbf{M}\{\mathbf{z}_{x,\mathbf{h}}\mathbf{z}_{x,\mathbf{h}}'\}\mathbf{M}$. The non-zero eigenvalues of $\mathbf{z}_{x,\mathbf{h}}\mathbf{z}_{x,\mathbf{h}}'$ equal those of $\mathbf{z}_{x,\mathbf{h}}'\mathbf{z}_{x,\mathbf{h}}$, so we see that we are actually considering the non-zero values of the scalars $\mathbf{u}_{\mathrm{i}}'\mathbf{H_{gh}H_{gh}u}_{\mathrm{i}}'$. As λ_i goes to 1, these go to zero for all $\mathbf{h} \neq \mathbf{g}$ and the cluster covariance estimate has only one non-zero eigenvalue and depends upon only one of the many possible orthogonal combinations of the disturbances. Once again, the cc2 and cc3 corrections adjust for the bias in the variance of the disturbances, but do not eliminate the reduced dimensionality.

The preceding examples show how regression design interacts with hypothesis tests to produce reduced degrees of freedom. Poor regression design is a necessary but not sufficient condition for a reduction in effective degrees of freedom. This allows the following bounds, proven in the appendix:

$$(14)\quad \text{edf}_x \geq \max\left(1,(h_{ii}^{\max})^{-1}-1\right)\ (x=\text{hc1},\text{hc2}),\quad \text{edf}_{\text{hc3}} \geq \max\left(1,\frac{1-h_{ii}^{\max}}{1-h_{ii}^{\min}}[(h_{ii}^{\max})^{-1}-1]\right)$$

$$\text{edf}_x \geq \max\left(1,\lambda^{\max}(\mathbf{H_{gg}})^{-1}-1\right)\ (x=\text{cc1},\text{cc2}),\quad \text{edf}_{\text{cc3}} \geq \max\left(1,\frac{1-\lambda^{\max}(\mathbf{H_{gg}})}{1-\lambda^{\min}(\mathbf{H_{gg}})}[\lambda^{\max}(\mathbf{H_{gg}})^{-1}-1]\right)$$

An example where these bounds are attained is where there are two regressors, a constant and an independent variable which equals 1 for half of the observations and -1 for the other half. This is a perfectly balanced regression design, with $h_{ii}^{\max}=h_{ii}^{\min}=k/n=2/n$. The matrix $\mathbf{(X'X)^{-1}X'}$ is now made up of two column vectors, $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, one for each type of observation. There exist hypothesis tests **w** such that $\mathbf{w}'\boldsymbol{x}_1$ or $\mathbf{w}'\boldsymbol{x}_2 = 0$. In these cases, half of the elements of $\mathbf{z}' = \mathbf{w}'\mathbf{(X'X)^{-1}X'}$ equal zero, so the robust covariance estimate depends upon only half of the residuals and the effective degrees of freedom for hc1, hc2 and hc3 can be shown to equal $n/2-1$, attaining the bound predicted above. If each cluster contains two observations, one of each type, then $\lambda^{\max}(\mathbf{H_{gg}})=\lambda^{\min}(\mathbf{H_{gg}})=2/n$[9] and the effective degrees of freedom for cc1, cc2, and cc3 are all also equal to $n/2-1$, attaining the indicated bound for these estimates as well.

IV. Empirical Results

I demonstrate the improvements in statistical inference afforded by bias and effective degrees of freedom corrections using 1378 OLS regression specifications taken from 44 experimental papers published in the journals of the American Economic Association.[10] The robust and clustered covariance matrices are used in 167 and 835 of these regressions, respectively, so these papers provide practical examples of the conditions under which these covariance matrices are used. The number of robust regressions is, however, somewhat small. Since the robust covariance matrix could, in principle, have been used in any of these regressions, I enlarge the robust sample by considering its application in all 1378

[9] As the paired cluster observations are orthogonal, so $\mathbf{H_{gg}}$ is diagonal.

[10] These papers are part of a comprehensive 53 experimental paper sample used in a separate study of randomization inference (Young 2015, which provides details regarding the selection criteria). 9 of the 53 papers do not contain OLS regressions and hence to not appear in the simulations above.

regressions.[11] In the case of 591 regressions the dependent variable is binary, which allows me to evaluate performance in the face of non-normal heteroskedastic errors.

The papers in my sample typically present a regression specification of the form $y_i = \mathbf{t}_i'\boldsymbol{\beta}_\mathbf{t} + \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$, where $\mathbf{t}_i$ is a vector of treatment characteristics, $\mathbf{x}_i$ a vector of other determinants of the outcome y_i, and $\boldsymbol{\beta}_\mathbf{t}$ and $\boldsymbol{\beta}_\mathbf{x}$ are the coefficients associated with each type of regressor. The coefficients of primary interest in these papers are those associated with treatment variables and the t-statistics test the null that each of these coefficients is zero. To this end, I use the non-treatment variables to generate complex simulated data that satisfies the null of no treatment effect. I initially run the baseline equation $y_i = \mathbf{x}_i'\boldsymbol{\beta}_\mathbf{x} + \varepsilon_i$ and then use the predicted values and random draws of normal iid errors, with standard deviation equal to that of the estimated standard error of the baseline equation, to produce 10,000 sets of simulated data. To allow for non-iid disturbances, I take the estimated standard error, mechanically divide its variance into cluster and iid components (with the cluster share of variance ranging from .2 to .8), and produce another 10,000 sets of data with normal but cluster-correlated disturbances. To allow for non-normal heteroskedastic disturbances, I take the equations in which the dependent variable is binary, estimate a baseline probit equation using $\mathbf{x}_i$, and then use this to produce simulated binary outcomes. The unconditional variance of the simulated dependent variable is given by P(1-P), where P is the predicted probability of the baseline probit equation. OLS fitting eliminates some of this inherent heteroskedasticity, but the departure from iid disturbances remains large, as illustrated by the poor performance of the baseline OLS estimate of covariance in the simulations below. To add cluster level correlation to these binary outcomes, I divide the standard normal disturbance that underlies the probit determination of 0/1 outcomes into cluster and iid components, with the cluster component share again ranging from .2 to .8, producing cluster correlated non-normal heteroskedastic disturbances. On all of this simulated data based upon baseline equations without treatment regressors, I then run a full OLS regression, including the original author specified treatment variables, and use t-tests to evaluate the null that each treatment coefficient is zero. Since the nulls are by construction true, the test statistics should reject each null 1 percent of the time at the .01 level.

[11] The clustered sample cannot be enlarged quite as simply, as it would require specifying a variable to cluster on in the non-clustered regressions.

Table I summarizes key characteristics of the sample. As shown in panel (a), the number of observations ranges from 40 to 450,000, and the number of clusters, in clustered regressions, from 11 to 5648, with mean values of 5314 and 215, respectively. Maximum leverage and the maximum eigenvalue of the cluster components of the hat matrix both range from nearly 0 to 1. 24 percent of the regressions have a maximum leverage of 1 and 38 percent of the clustered regressions have a maximum cluster eigenvalue of 1. Application of the hc2/hc3/cc2/cc3 corrections of the robust and clustered covariance matrices in these cases involves division by zero, [12] and hence is not practical, whatever the merits of these methods. As noted earlier in the Introduction, Bell and McCaffrey (2002) implement clustered edf corrections by first applying the cc2 bias correction, which ensures that the covariance estimate in the face of iid errors is unbiased for all hypothesis tests, and then calculating the edf correction for the particular hypothesis test. In contrast, when working with cc1, I calculate both bias and edf corrections separately for each hypothesis test. The Bell and McCaffrey approach, endorsed by Imbens and Kolesar (2015), depends upon computationally costly inversions of potentially large cluster sub-matrices ($\mathbf{M_{gg}}$), inversions which are, moreover, impossible in more than a third of the practical cases in my experimental sample. As seen further below, the cc1 (and similarly hc1) based approach yields virtually identical results to cc2 and cc3 when all can be applied, but working directly with cc1 is computationally simpler and much more widely implementable.

Panel (b) of Table I reports the bias and edf hc1 and cc1 measures for the treatment coefficients in my sample regressions. The bias of the robust and clustered covariance matrices, with Stata's finite sample corrections, averages close to 1, but ranges quite widely. Effective degrees of freedom in robust regressions average only 40 percent of each regression's putative n-k degrees of freedom, while in clustered regressions they are typically just under 50 percent of the theoretical n_c-1 limit. Panel (c) focuses on regressions where the hc2/hc3/cc2/cc3 corrections can be implemented, and shows the patterns described by the theory of the last section. The hc2/cc2 corrections eliminate bias, but at the cost of increasing the variance of the variance estimate, while the hc3/cc3 overcorrection has a positive, and occasionally quite large, bias and also a substantially larger variance. All three techniques have quite similar effective degrees of freedom, as the movements in bias and variance offset

[12] As m_{ii} = 1-h_{ii} = 0 and $\lambda^{min}(\mathbf{M_{gg}}) = 1 - \lambda^{max}(\mathbf{H_{gg}}) = 0$.

Table I: Characteristics of the Sample

	mean	s.d.	min	max		mean	s.d.	min	max
(a) 1378 regressions					(a) 835 clustered regressions				
n	5314	$2.5e^4$	40	$4.5e^5$	n_c	215	417	11	5648
h_{ii}^{max}	.383	.395	$3.8e^{-5}$	1	$\lambda^{max}(\mathbf{H_{gg}})$	.616	.376	$9.3e^{-4}$	1
(b) 3665 coefficients in 1378 regressions					(b) 1897 coefficients in 835 clustered regressions				
μ_{hc1}	.990	.064	.138	1.24	μ_{cc1}	.976	.130	.060	1.93
n-k	5209	$2.9e^4$	22	$4.5e^5$	n_c-1	266	511	10	5647
edf_{hc1}	2102	$1.5e^4$	1.11	$3.9e^5$	edf_{cc1}	126	378	1.17	5000
(c) 2819 coef. in 1053 regressions with $h_{ii}^{max} < 1$					(c) 929 coef. in 514 regressions with $\lambda^{max}(\mathbf{H_{gg}}) < 1$				
μ_{hc1}	.990	.043	.520	1.06	μ_{cc1}	.939	.083	.479	1.01
μ_{hc2}	1	0	1	1	μ_{cc2}	1	0	1	1
μ_{hc3}	1.07	.100	1.00	2.02	μ_{cc3}	1.12	.193	1.00	2.47
ν_{hc1}	.018	.040	$5.1e^{-6}$	.493	ν_{cc1}	.034	.043	$4.0e^{-4}$	.429
ν_{hc2}	.026	.127	$5.1e^{-6}$	1.90	ν_{cc2}	.048	.095	$4.0e^{-4}$	1.55
ν_{hc3}	.060	.499	$5.1e^{-6}$	7.78	ν_{cc3}	.110	.386	$4.0e^{-4}$	6.21
edf_{hc1}	2591	$1.7e^4$	1.11	$3.9e^5$	edf_{cc1}	195	529	1.56	5000
edf_{hc2}	2590	$1.7e^4$	1.05	$3.9e^5$	edf_{cc2}	193	529	1.29	4999
edf_{hc3}	2589	$1.7e^4$	1.03	$3.9e^5$	edf_{cc3}	191	529	1.15	4999

Notes: s.d. = standard deviation; n = number of observations; n_c = number of clusters; k = number of regressors; h^{max} = maximum leverage of the regression; λ^{max}= maximum cluster eigenvalue of the hat matrix; μ, ν and edf = bias, variance and effective degrees of freedom of the variance estimate for the coefficient; ae^b = $a*10^b$; coef. = coefficients. hc1 and cc1 corrections are calculated with Stata's finite sample correction, so the relative bias bounds listed earlier in (11) do not apply.

each other. As noted earlier, fundamentally the three estimates place weight on the same disturbances, albeit somewhat transformed, and hence depend roughly upon the same linear combinations of the error process.[13]

Figures I and II below graph the actual ln reduction in degrees of freedom against the theoretical bound for the hc1 and cc1 estimators. As noted in the preceding section, leverage creates the possibility of reduced degrees of freedom, but the actual reduction depends upon the specific hypothesis test. As the theoretical bound falls, however, hypothesis tests, even if randomly selected across the universe of possible tests, produce results that range within the increased bounds. Consequently, edf fall with increased maximal leverage. Figures III and IV graph the bias of the variance estimate against its theoretical lower bound. Here we see

[13] The pairwise correlation of the different hci (cci) edf with each other is 1.0000. Transformed into lns to moderate the influence of large values, edf_{hc1} has a correlation of .9999 with edf_{hc2} and .9980 with edf_{hc3}, while edf_{cc1} has a correlation of .9996 with edf_{cc2} and .9976 with edf_{cc3}.

Figure I: Robust Effective Degrees of Freedom
and Bounds Created by Maximum Leverage

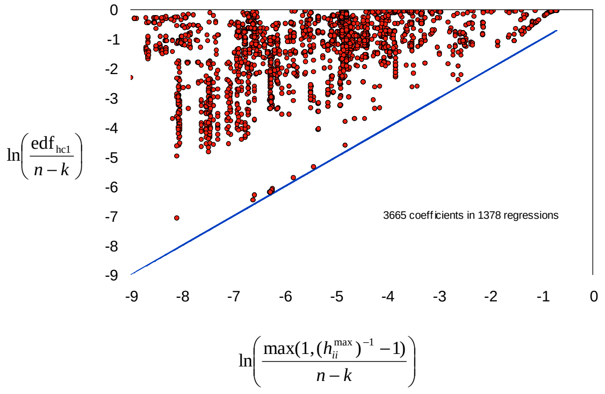


Figure II: Clustered Effective Degrees of Freedon
and Bounds Created by Maximum Leverage

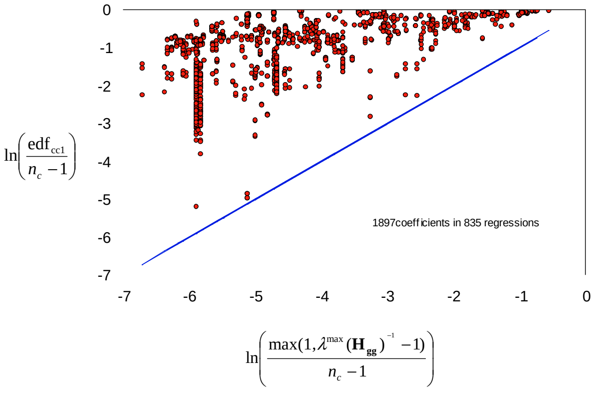


Figure III: Robust Bias and
Bounds Created by Maximum Leverage

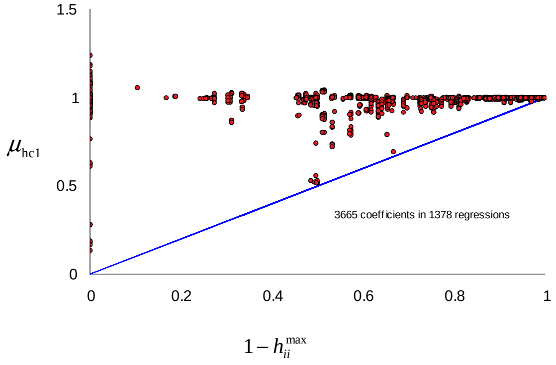
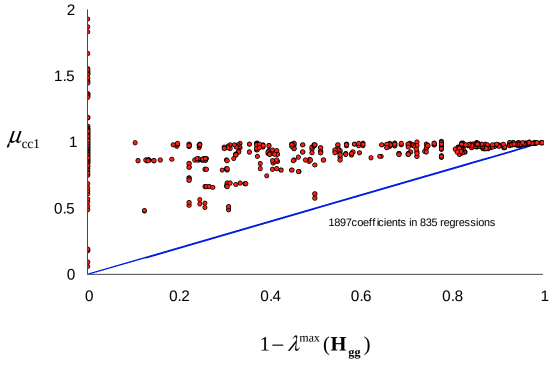


Figure IV: Clustered Bias and
Bounds Created by Maximum Leverage



that greater potential bias translates into increased dispersion, but with little change in average bias. This stems from the finite sample corrections used in the hc1 and cc1 estimates. For example, in the case of hc1 the range of bias is $m_{ii}^{\min}=1-h_{ii}^{\max}$ to $m_{ii}^{\max}=1-h_{ii}^{\min}$, but the average of m_{ii} is always (n-k)/n. Different hypotheses tests use different linear combinations of residuals, each of which underestimates its own variance (with iid disturbances) by m_{ii}. Not surprisingly, bias on average is close to (n-k)/n, so the hc1 finite sample correction of n/n-k approximately eliminates average bias, which is after all the whole point of the finite sample correction.[14]

Figures I-IV explain why, in Young (2015), I find that coverage bias is increasing in maximal leverage. As maximal leverage increases, effective degrees of freedom typically fall. Consequently, test statistics evaluated with the default n-k or n_c-1 degrees of freedom

[14] In examining Figures III and IV the reader might note the points along the y-axis. These represent the large mass of regressions where maximal leverage h_{ii} or the maximum eigenvalue of the sub-matrices $\mathbf{H_{gg}}$ equals 1. The average bias across all of these extreme cases is still close to one, but there is a great deal of dispersion.

have thicker tails than expected, producing rejection probabilities for each hypothesis test that are greater than nominal size. As maximal leverage increases, the dispersion of the bias of the covariance estimate increases. As the absolute value of the t-statistic is a convex function of the covariance estimate, this increased dispersion raises its average value across all hypothesis tests, raising average rejection probabilities above nominal size.[15] This effect can be seen in the tables below by comparing hc1 and cc1 with hc2 and cc2, which have virtually the same effective degrees of freedom and average bias, but eliminate the variation in bias and, consequently, have somewhat lower rejection rates (without bias and edf corrections).

Tables II - IV show the remarkable improvement in the accuracy of statistical inference allowed by bias and effective degrees of freedom corrections. Table II begins with the application of the robust covariance estimate to all 3665 treatment coefficients in 1378 regressions. Panel (a) of the table uses simulations with ideal normal iid disturbances. Under such circumstances statistical inference based upon the default OLS covariance estimate is exact, as confirmed by the simulation results, which reject an average of .01 of the time, with a standard deviation of .001, i.e. precisely the predicted values with 10,000 simulations per equation. Statistical inference based upon the hc1 robust covariance estimate, however, is biased and wildly inaccurate, rejecting the (true) null of a zero coefficient an average of .0132 of the time, with rejection probabilities reaching as high as .4253, producing a standard deviation of the rejection rate more than 20 times that of the OLS estimates. With bias and effective degrees of freedom corrections, however, inference based upon the hc1 estimate is nearly exact, rejecting the null .0099 of the time with a standard deviation of .0013. In panel (b) of the table I restrict attention to those regressions where the hc2 and hc3 corrections can be applied. As shown, these provide some improvement over hc1, producing more accurate mean rejection rates, but extreme outcomes are still present and the standard deviation of the rejection rate remains 9 to 12 times greater than that of the default OLS method. With bias and edf corrections, however, the three robust methods are indistinguishable and nearly exact, as shown in the right-hand panel.

Panels (c) and (d) of Table II examine results when the error generating process is non-normal and heteroskedastic, a consequence of the binary character of the dependent

[15] The rejection probability for an individual hypothesis test might be lower than nominal size (due to an upward biased covariance estimate), but averaged across all hypothesis tests the absolute value of the t-statistic increases with greater bias dispersion.

Table II: Empirical Size at the .01 Level with Robust Covariance Estimates
(10000 simulations per regression)

	uncorrected, df = n-k				bias and edf corrected			
	mean	s.d.	min	max	Mean	s.d.	min	max
(a) ideal normal iid disturbances - 3665 coefficients in 1378 regressions								
default OLS	.0100	.0010	.0066	.0133				
hc1	.0132	.0224	.0056	.4253	.0099	.0013	.0000	.0135
(b) ideal normal iid disturbances - 2819 coefficients in 1053 regressions with $h^{max} < 1$								
default OLS	.0100	.0010	.0066	.0133				
hc1	.0124	.0160	.0065	.2697	.0099	.0013	.0000	.0135
hc2	.0117	.0118	.0065	.2013	.0099	.0013	.0000	.0134
hc3	.0094	.0085	.0012	.1477	.0099	.0013	.0000	.0138
(c) non-normal heteroskedastic disturbances - 1522 coef. in 591 regressions with binary y								
default OLS	.0136	.0373	.0000	.5853				
hc1	.0169	.0459	.0000	.6408	.0111	.0082	.0000	.1655
(d) non-normal heteroskedastic disturbances - 1181 coef. in 453 reg. with $h^{max} < 1$ & binary y								
default OLS	.0144	.0421	.0000	.5853				
hc1	.0178	.0515	.0000	.6408	.0106	.0052	.0000	.1315
hc2	.0172	.0505	.0000	.6409	.0106	.0051	.0000	.1313
hc3	.0150	.0485	.0000	.6405	.0106	.0051	.0000	.1313

Notes: df = degrees of freedom; edf = effective degrees of freedom; reg. = regressions; y = dependent variable; otherwise as in Table I.

variable. The baseline probit equations that are used as the data generating process produce substantial heteroskedasticity. In 190 of the 591 estimating equations, at least one observation has a predicted probability (P) less than .0001, and in 138 cases at least one observation has a predicted probability greater than .9999. The unconditional P(1-P) variance of these observations is close to zero. More generally, the standard deviation of the unconditional observation level variance averages .062 and ranges between 0 and .222. The deleterious impact this heteroskedasticity has on statistical inference is immediately apparent in panel (c). The default OLS estimate of covariance is biased, rejecting .0136 of the time, and extraordinarily variable, with a standard deviation of .0373, i.e. 37 times greater than the exact rate, and with rejection rates rising as high as .5853. Unfortunately, the hc1 robust "correction" for heteroskedasticity performs even worse, with a mean rejection rate of .0169 and a standard deviation of .0459. With bias and edf corrections, however, the mean and standard deviation of the robust rejection rate are reduced to .0111 and .0082, respectively. Panel (d) shows that the hc2 and hc3 corrections of the robust covariance matrix, when they

Table III: Empirical Size at the .01 Level with Clustered Covariance Estimates
(10000 simulations per regression)

	uncorrected, df = n_c-1				bias and edf corrected			
	mean	sd	min	max	mean	sd	min	max
(a) ideal normal iid disturbances – 1897 coefficients in 835 regressions								
default OLS	.0100	.0010	.0071	.0131				
cc1	.0157	.0255	.0008	.5277	.0097	.0013	.0000	.0132
(b) ideal normal iid disturbances – 929 coefficients in 514 regressions with $\lambda^{max} < 1$								
default OLS	.0100	.0010	.0072	.0127				
cc1	.0150	.0108	.0075	.1473	.0098	.0012	.0000	.0132
cc2	.0116	.0050	.0074	.1116	.0097	.0013	.0000	.0133
cc3	.0085	.0035	.0001	.0770	.0096	.0016	.0000	.0133
(c) non-normal heteroskedastic disturbances – 737 coef. in 324 regressions with binary y								
default OLS	.0174	.0053	.0000	.5853				
cc1	.0153	.0101	.0000	.1268	.0097	.0036	.0000	.0622
(d) non-normal heteroskedastic disturbances – 489 coef. in 235 reg. with $\lambda^{max} < 1$ & binary y								
default OLS	.0210	.0649	.0000	.5853				
cc1	.0145	.0097	.0000	.1268	.0098	.0039	.0000	.0622
cc2	.0115	.0052	.0000	.0621	.0097	.0038	.0000	.0621
cc3	.0084	.0039	.0000	.0621	.0096	.0039	.0000	.0618

Notes: as in Tables I and II.

can be applied, are not very effective, as these estimators remain substantially biased on average and remarkably unpredictable, with a standard deviation of rejection rates 50 times that of the exact test statistic. Once again, however, with bias and edf corrections all three methods are virtually identical, producing, in this sample of regressions, rejection rates that are only slightly biased on average (.0106) and have a standard deviation (.0052) that is one-tenth that of the unadjusted measures. These and later results show that bias and edf corrections, motivated with normal iid errors, substantially improve the accuracy of statistical inference in situations with less than ideal disturbances.

Table III above examines statistical inference using the clustered estimate of covariance in regressions that clustered in the original papers. In panel (a) we see, once again, that with ideal normal iid errors the baseline clustered covariance method, cc1, is biased and very variable, producing rejection rates as high as .5277 at the .01 level. With bias and edf corrections, however, it is virtually exact. The cc2/cc3 corrections, as shown in panel (b), reduce the average bias of the rejection rate, with cc3 overcorrecting and producing an average rejection rate of .0085 at the .01 level. With bias and edf corrections, however, all three methods are very similar, although corrected cc1 produces the lowest standard deviation of results, .0012, which approaches the .0010 level of exact statistics. Panels (c) and (d) examine performance with the non-normal heteroskedastic disturbances produced by binary dependent variables. Once again, without corrections coverage is upward biased or (in the case of cc3) downward biased, and quite variable. With bias and edf corrections, as shown in the right-hand side of panel (d), the three measures produce virtually identical unbiased rejection rates and standard deviations of the rejection rate (.0039) that, despite the non-normality and heteroskedasticity, are only four times that of an exact statistic.

Table IV below examines rejection rates when the error process has a cluster level random effect that accounts for either .2 or .8 of the total error variance. Treatment in these regressions is often administered at the cluster level and, hence, highly correlated with the cluster random effects. Under these circumstances, the default OLS covariance matrix grossly understates coefficient standard errors (Kloek 1981, Moulton 1986), producing rejection rates much greater than nominal size. As shown in the table, clustered standard errors improve, dramatically, on the performance of the default OLS estimator, but have somewhat biased coverage on average and produce highly volatile results. In the broadest samples, in panel (a) for example, the cc1 correction has a mean rejection rate of .0167 and a standard deviation 25 times that of the exact benchmark (.001), with rejection rates ranging as high as .5285. With bias and edf corrections, however, the mean rejection probability is .0103, with a standard deviation of only .0020. The problems of the different cci estimators are somewhat less in the smaller samples that allow calculation of cc2/cc3 or have binary dependent variables, as seen in the other panels of the table. Nevertheless, it is notable that average coverage bias is almost completely eliminated and the standard deviation of results substantially reduced with the application of the corrections.

Table IV: Empirical Size at the .01 Level with Cluster Random Effects and Clustered Covariance Estimates (10000 simulations per regression)

	uncorrected				bias and edf corrected			
	mean	sd	min	max	Mean	sd	min	max
ρ = .2								
(a) ideal normal iid disturbances – 1897 coefficients in 835 regressions								
default OLS	.1029	.1566	.0036	.8079				
cc1	.0167	.0250	.0010	.5285	.0103	.0020	.0000	.0333
(b) ideal normal iid disturbances – 929 coefficients in 514 regressions with $\lambda^{max} < 1$								
default OLS	.1831	.1879	.0036	.8079				
cc1	.0166	.0123	.0078	.1455	.0109	.0021	.0000	.0333
cc2	.0122	.0055	.0073	.1119	.0102	.0015	.0000	.0208
cc3	.0084	.0036	.0003	.0792	.0096	.0017	.0000	.0131
(c) non-normal heteroskedastic disturbances – 737 coef. in 324 regressions with binary y								
default OLS	.0848	.1359	.0000	.7557				
cc1	.0156	.0100	.0000	.1025	.0099	.0031	.0000	.0339
(d) non-normal heteroskedastic disturbances – 489 coef. in 235 reg. with $\lambda^{max} < 1$ & binary y								
default OLS	.1181	.1561	.0000	.7557				
cc1	.0150	.0093	.0000	.1025	.0102	.0029	.0000	.0339
cc2	.0115	.0042	.0000	.0345	.0097	.0026	.0000	.0230
cc3	.0082	.0028	.0000	.0226	.0093	.0026	.0000	.0229
ρ = .8								
(e) ideal normal iid disturbances – 1897 coefficients in 835 regressions								
default OLS	.2059	.2564	.0000	.9044				
cc1	.0170	.0234	.0001	.5326	.0106	.0027	.0000	.0365
(f) ideal normal iid disturbances – 929 coefficients in 514 regressions with $\lambda^{max} < 1$								
default OLS	.3491	.2813	.0000	.9044				
cc1	.0169	.0125	.0042	.1519	.0112	.0028	.0000	.0365
cc2	.0118	.0056	.0030	.1141	.0100	.0022	.0000	.0231
cc3	.0078	.0038	.0002	.0792	.0090	.0025	.0000	.0135
(g) non-normal heteroskedastic disturbances – 737 coef. in 324 regressions with binary y								
default OLS	.2080	.2356	.0000	.8921				
cc1	.0152	.0106	.0000	.0985	.0096	.0045	.0000	.0378
(h) non-normal heteroskedastic disturbances – 489 coef. in 235 reg. with $\lambda^{max} < 1$ & binary y								
default OLS	.2804	.2539	.0000	.8921				
cc1	.0150	.0102	.0000	.0985	.0102	.0041	.0000	.0378
cc2	.0110	.0052	.0000	.0335	.0094	.0036	.0000	.0276
cc3	.0074	.0035	.0000	.0270	.0086	.0036	.0000	.0278

Notes: ρ = share of error variance accounted for by cluster random effect; otherwise, as in Tables I and II.

V. Conclusion

Extensions of the OLS methods described above to non-OLS settings easily suggest themselves. In a GLS setting, the GLS transform in principle renders the error term iid. In practice, however, in my sample papers authors generally still use a clustered covariance estimate. The GLS transformed independent variables could be used to calculate leverage for the regression and make appropriate corrections. In maximum likelihood estimation, such as that of probit or logit models, authors, again, generally use the robust or clustered covariance estimate.[16] The scores of these generalized linear models can be reinterpreted as including error terms which, with a suitable GLS transform, are iid. Thus, leverage can be calculated, using the GLS transformed independent variables, and bias and edf corrections made. Unlike OLS, however, in both of these examples "leverage" is a function of the estimated GLS covariance matrix, and hence a function not merely of the regressors, but also of the realized disturbances. The sampling variation in the correction itself could easily undo any benefits gained by the attempt to understand how regression design favours certain residuals.

Such extensions also miss the implications of the results presented above. The argument presented in this paper is twofold: first, that an adjustment designed to mimic the moment properties of the chi-squared distribution closely approximates the actual distribution of the test statistic with ideal iid normal errors, and second, that the estimated distribution improves the accuracy of inference in the presence of non-ideal errors. The tables, with both iid normal and non-iid non-normal disturbances, establish that there is some validity to both statements. The first-step, however, does not have to be calculated using an approximation formula. It can be calculated by simulation, i.e. drawing enough ideal iid disturbances to get an estimate of the empirical distribution of the test statistic. In the spirit of the second-step, this distribution can then be used to evaluate the regression test statistic, based as it is upon whatever error process is actually at work in the data. In the case of OLS regressions, the bias and edf corrections are easily calculated and much less costly than simulating the distribution of the test statistic. In other applications, if an approximation formula is unknown or potentially less reliable, simulation based upon ideal disturbances as modelled by the baseline specification can provide insight into its distribution in less ideal circumstances.

[16] Implicitly suggesting that the likelihood may be mis-specified and their estimation procedures quasi-maximum likelihood.

Appendix: Proofs

I provide proofs of the results stated in (11) and (14). I make repeated use of two well-known matrix results, which for convenient reference I enumerate. First, regarding the Rayleigh quotient:

$$\text{(p1)}\;\; \max_{\mathbf{z}} \frac{\mathbf{z}'\mathbf{A}\mathbf{z}}{\mathbf{z}'\mathbf{B}\mathbf{z}} = \lambda^{\max}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}), \quad \min_{\mathbf{z}} \frac{\mathbf{z}'\mathbf{A}\mathbf{z}}{\mathbf{z}'\mathbf{B}\mathbf{z}} = \lambda^{\min}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})$$

where **z** is a non-zero vector, **A** a symmetric matrix, **B** a symmetric positive definite matrix, and $\mathbf{B}^{1/2}$ the symmetric root of **B**. Second, for matrices **A** and **B**:

(p2) the non-zero eigenvalues of **AB** = the non-zero eigenvalues of **BA**,

when **AB** and **BA** are both defined. I remind the reader that leverage is bounded by $1 \geq h_{ii} \geq 0$, so $m_{ii} = 1 - h_{ii}$ is bounded by $0 \leq m_{ii} \leq 1$. From (13) earlier, it can be seen that similar bounds apply to the eigenvalues of $\mathbf{H_{gg}}$ and $\mathbf{M_{gg}}$. In discussing hc2/hc3 and cc2/cc3, I take it as given that $m_{ii}^{\min} > 0$ and $\lambda^{\min}(\mathbf{M_{gg}}) > 0$, respectively, so the measures can actually be implemented.

Beginning with (11a) and (11b), substituting in for $\mathbf{z}_x$ in (10) using (8), and dropping the finite sample adjustments c_x, we see that the means are given by the Rayleigh quotients:

$$\text{(a1)}\;\; \mu_{\text{hc1}} = \frac{\mathbf{z}'\{m_{ii}\}\mathbf{z}}{\mathbf{z}'\mathbf{z}},\; \mu_{\text{hc2}} = \frac{\mathbf{z}'\mathbf{z}}{\mathbf{z}'\mathbf{z}},\; \mu_{\text{hc3}} = \frac{\mathbf{z}'\{m_{ii}^{-1}\}\mathbf{z}}{\mathbf{z}'\mathbf{z}},\; \mu_{\text{cc1}} = \frac{\mathbf{z}'\{\mathbf{M_{gg}}\}\mathbf{z}}{\mathbf{z}'\mathbf{z}},\; \mu_{\text{cc2}} = \frac{\mathbf{z}'\mathbf{z}}{\mathbf{z}'\mathbf{z}},\; \mu_{\text{cc3}} = \frac{\mathbf{z}'\{\mathbf{M_{gg}^{-1}}\}\mathbf{z}}{\mathbf{z}'\mathbf{z}}$$

From (p1) we see that the maxima and minima of these equal the maximum and minimum eigenvalues of the diagonal and block diagonal matrices in the numerators, proving (11a) and (11b). The inequalities on these means follow from the bounds on m_{ii} and $\lambda(\mathbf{M_{gg}})$ described above.

To prove (11c) I once again appeal to the Rayleigh quotient, noting that it implies that the maximum and minimum eigenvalues of **A** equal the maximum and minimum of $\mathbf{e}'\mathbf{A}\mathbf{e}$ across all vectors **e** such that $\mathbf{e}'\mathbf{e} = 1$. Let $\mathbf{e_i}$ denote the vector with a 1 in the ith position and 0s everywhere else. Since the diagonal of $\{\mathbf{M_{gg}}\}$ equals $\{m_{ii}\}$, one can see that $\mathbf{e_i}'\{\mathbf{M_{gg}}\}\mathbf{e_i} = \mathbf{e_i}'\{m_{ii}\}\mathbf{e_i}$, which shows that the maximum and minimum of $\mathbf{e}'\{\mathbf{M_{gg}}\}\mathbf{e}$ must lie weakly outside the bounds given by the maximum and minimum of $\mathbf{e}'\{m_{ii}\}\mathbf{e}$. Strict inequality comes when $\{\mathbf{M_{gg}}\}$ contains non-zero within-cluster off-diagonal elements in the rows associated with the maximum and minimum values of m_{ii}. To see this, let **e** have τ in the ith position, η in the jth position (both within the same cluster), and 0s everywhere else. With $\tau^2 + \eta^2 = 1$ (as $\mathbf{e}'\mathbf{e} = 1$), we have $d\tau/d\eta = -\eta/\tau$. We then have $f(\eta) = \mathbf{e}'\{\mathbf{M_{gg}}\}\mathbf{e} = \tau^2 m_{ii} + 2\tau\eta m_{ij} + \eta^2 m_{jj}$, with $f'(\eta) = 2(-\eta m_{ii} + m_{ij}\tau - m_{ij}\eta^2/\tau + \eta m_{jj})$, $f(0) = m_{ii}$ and $f'(0) = 2m_{ij}$. From this it follows that if m_{ij} is not zero there is a small deviation away from $\tau = 1$ (with η moving above or below zero) such that $f(\eta)$ is greater or less than m_{ii}. This indicates that the bounds for $\mathbf{e}'\{\mathbf{M_{gg}}\}\mathbf{e}$ lie strictly outside those for $\mathbf{e}'\{m_{ii}\}\mathbf{e}$ if there are any non-zero within cluster elements in the rows associated with the maximum and minimum values of m_{ii},[17] establishing (11c).

[17] In the proof the elements must be within cluster because $\{\mathbf{M_{gg}}\}$ is block diagonal, so that an **e** with τ in the ith position and η in the jth position (belonging to another cluster), produces $f(\eta) = \mathbf{e}'\{\mathbf{M_{gg}}\}\mathbf{e} = \tau^2 m_{ii} + \eta^2 m_{jj}$, with

For (11d), I begin by determining the rank of the matrix in each quadratic form:

$$\text{(a2)}\quad rank(\mathbf{B}_{\text{hci}}) = rank\left(\frac{c_x}{\mathbf{z'z}}\mathbf{M}\{z^2_{x,i}\}\mathbf{M}\right) \le \min(rank(\mathbf{M}), rank(\{z^2_{x,i}\})) \le \text{n}-\text{k}$$

$$rank(\mathbf{B}_{\text{cci}}) = rank\left(\frac{c_x}{\mathbf{z'z}}\mathbf{M}\{\mathbf{z}_{x,\mathbf{g}}\mathbf{z}'_{x,\mathbf{g}}\}\mathbf{M}\right) = rank(\mathbf{MR}_x) \le \min(\text{n}_\text{c}, \text{n}-\text{k})$$

where I have used the fact that rank(**M**) = n-k and re-expressed $\{\mathbf{z}_{x,\mathbf{g}}\mathbf{z}'_{x,\mathbf{g}}\}$ as $\mathbf{R}_x\mathbf{R}'_x$, where $\mathbf{R}_x$ is an n x n_c matrix composed of column vectors with $\mathbf{z}_{x,\mathbf{g}}$ in each column in the position associated with the observations for the cluster associated with the number of the column, and 0s everywhere else. $\mathbf{R}_x$ is at most of rank n_c and **M** of rank n-k. However, when *x* = cc1 the row sum of $\mathbf{MR}_{\text{cc1}}$ always equals a column of 0s, as (allowing $\mathbf{i}(\text{n}_\text{c})$ to denote an n_cx1 column vector of 1s):

$$\text{(a3)}\quad \mathbf{MR}_{\text{cc1}}\mathbf{i}(\text{n}_\text{c}) = \mathbf{Mz}_{\text{cc1}} = (\mathbf{I}(\text{n}) - \mathbf{X(X'X)}^{-1}\mathbf{X}')(\mathbf{X(X'X)}^{-1}\mathbf{w}) = \mathbf{0}$$

As the columns are co-linear, the rank of the product is no more than n_c-1. Hence, rank($\mathbf{B}_{\text{cc1}}$) ≤ min(n_c-1,n-k). The columns of $\mathbf{MR}_{\text{cc2}}$ and $\mathbf{MR}_{\text{cc3}}$, however, need not be collinear,[18] and so the rank upper bound for $\mathbf{B}_{\text{hc2}}$ and $\mathbf{B}_{\text{hc3}}$ is min(n_c,n-k).

Effective degrees of freedom equal $2\mu_x^2/\nu_x$, where μ_x = trace($\mathbf{B}_x$) and ν_x = 2*trace($\mathbf{B}_x\mathbf{B}_x$). Let λ_i denote the i^th non-zero eigenvalue of $\mathbf{B}_x$. There are r such eigenvalues, where r is the rank of $\mathbf{B}_x$. Writing the edf in terms of these eigenvalues we have:

$$\text{(a4)}\quad \text{edf}_x = \frac{\text{tr}(\mathbf{B}_x)^2}{\text{tr}(\mathbf{B}_x\mathbf{B}_x)} = \left(\sum_{i=1}^{r}\lambda_i\right)^2 \Big/ \sum_{i=1}^{r}\lambda_i^2$$

It is readily apparent that this attains a minimum value of 1 when there is only one non-zero eigenvalue. Maximizing with respect to λ_i we derive the first order condition:

$$\text{(a5)}\quad \lambda_i = \sum_{i=1}^{r}\lambda_i^2 \Big/ \sum_{i=1}^{r}\lambda_i \quad \forall\ i$$

which is satisfied when $\lambda_i = \lambda$ for all i, at which point edf_x = r. We conclude that edf_x ranges between 1 and r, which, using the rank bounds noted above, establishes the results reported in (11d). The examples later in the text show how these bounds can be attained.

With regards to (14), I begin by noting that (a4) implies:

$$\text{(a6)}\quad \text{edf}_x = \frac{\left(\sum_{i=1}^{r}\lambda_i\right)^2}{\sum_{i=1}^{r}\lambda_i^2} \ge \frac{\left(\sum_{i=1}^{r}\lambda_i\right)^2}{\lambda^{\max}(\mathbf{B}_x)\left(\sum_{i=1}^{r}\lambda_i\right)} = \frac{\text{trace}(\mathbf{B}_x)}{\lambda^{\max}(\mathbf{B}_x)} = \frac{\mu_x}{\lambda^{\max}(\mathbf{B}_x)} \ge \frac{\mu_x^{\min}}{\lambda^{\max}(\mathbf{B}_x)}$$

where $\mu_x^{\min}$ is the minimum possible bias of $\mathbf{w}'\text{V}_x\mathbf{w}$, as established earlier in (11a) and (11b). All that remains, then, is to establish bounds on $\lambda^{\max}(\mathbf{B}_x)$. As seen in (9) earlier, each $\mathbf{B}_\text{x}$ is of

$f'(\eta) = 2(-\eta m_{ii} + \eta m_{jj})$, $f'(0) = 0$ and $f'' = 2(-m_{ii} + m_{jj})$, i.e. a local maximum or minimum as m_{ii} is the maximal or minimal diagonal element.

[18] And in fact are not in some of the practical cases examined in Section IV above.

the form **MAM**, where **A** and **M** are both positive semi-definite. The non-zero eigenvalues of this matrix satisfy $\mathbf{MAMu}_i = \lambda_i\mathbf{u}_i$, where $\mathbf{u}_i$ is the eigenvector associated with λ_i. Using the fact that **M** is idempotent, we pre-multiply by **M** and see that $\mathbf{M}\lambda_i\mathbf{u}_i = \mathbf{MMAMu}_i = \mathbf{MAMu}_i = \lambda_i\mathbf{u}_i$. With $\lambda_i > 0$, this implies $\mathbf{Mu}_i = \mathbf{u}_i$, so we pre-multiply by $\mathbf{u}_i'$ and see that $\mathbf{u}_i'\lambda_i\mathbf{u}_i = \mathbf{u}_i'\mathbf{MAMu}_i = \mathbf{u}_i'\mathbf{Au}_i$, yielding the result $\lambda_i = \mathbf{u}_i'\mathbf{Au}_i/\ \mathbf{u}_i'\mathbf{u}_i$. λ_i can be expressed as a Rayleigh quotient of **A**, and hence is bounded by the maximum eigenvalue of **A**.

In the case of hci, **A** is the diagonal matrix $\{z_{x,i}^2\}/(\mathbf{z}'\mathbf{z})$, with maximum eigenvalues given by the maximum values of the diagonal elements. Applying the formulas for **z** and $z_{x,i}$ as defined earlier in (8) we have:

$$\text{(a7)}\quad \frac{z_{x,i}^2}{(\mathbf{z}'\mathbf{z})} = \frac{\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i m_{x,i}^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}{\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}$$

where $m_{x,i} = 1$, $m_{ii}^{-½}$, or m_{ii}^{-1} as x = hc1, hc2 or hc3, respectively, and $\mathbf{x}_i$ is the ith row of **X**. Following (p1), (a7) attains its maximum value at the maximum eigenvalue of $(\mathbf{X}'\mathbf{X})^{½}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i m_{x,i}^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})^{½} = (\mathbf{X}'\mathbf{X})^{-½}\mathbf{x}_i m_{x,i}^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-½}$. Applying (p2), we see that this matrix has only one non-zero eigenvalue, equal to $m_{x,i}^2\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i = m_{x,i}^2h_{ii}$. As $m_{ii} = 1 - h_{ii}$, the maximum of this, considering in each case the expression for $m_{x,i}$, is attained when $h_{ii} = h_{ii}^{\max}$. Applying our knowledge of $\mu_x^{\min}$ from (11a), then gives the hc results reported in (14).

In the case of cci, we seek the maximum eigenvalues of the block diagonal matrix **A** $=\{\mathbf{z}_{x,\mathbf{g}}\mathbf{z}'_{x,\mathbf{g}}\}/(\mathbf{z}'\mathbf{z})$. Again, applying earlier formulas we have each block diagonal element given by:

$$\text{(a8)}\quad \frac{\mathbf{z}_{x,\mathbf{g}}\mathbf{z}'_{x,\mathbf{g}}}{(\mathbf{z}'\mathbf{z})} = \frac{\mathbf{M}_{x,\mathbf{g}}\mathbf{x}_\mathbf{g}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{ww}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_\mathbf{g}\mathbf{M}_{x,\mathbf{g}}}{\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}$$

where $\mathbf{M}_{x,\mathbf{g}} = 1$, $\mathbf{M}_\mathbf{gg}^{-½}$, or $\mathbf{M}_\mathbf{gg}^{-1}$ as x = cc1, cc2 or cc3, respectively, and $\mathbf{x}_\mathbf{g}$ is the matrix composed of the rows of **X** associated with cluster g. . Applying (p2), we realize there is only one non-zero eigenvalue, given by

$$\text{(a9)}\quad \frac{\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}'_\mathbf{g}\mathbf{M}_{x,\mathbf{g}}\mathbf{M}_{x,\mathbf{g}}\mathbf{x}_\mathbf{g}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}{\mathbf{w}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{w}}$$

Using (p1), we see that the maximum value of this is given by the maximum eigenvalue of $(\mathbf{X}'\mathbf{X})^{-½}\mathbf{x}_\mathbf{g}'\mathbf{M}_{x,\mathbf{g}}\mathbf{M}_{x,\mathbf{g}}\mathbf{x}_\mathbf{g}(\mathbf{X}'\mathbf{X})^{-½}$. Applying (p2) once again, we see that this equals the maximum eigenvalue of $\mathbf{M}_{x,\mathbf{g}}\mathbf{x}_\mathbf{g}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_\mathbf{g}'\mathbf{M}_{x,\mathbf{g}} = \mathbf{M}_{x,\mathbf{g}}\mathbf{H}_\mathbf{gg}\mathbf{M}_{x,\mathbf{g}}$. As $\lambda^{\max}(\mathbf{AB}) \leq \lambda^{\max}(\mathbf{A})\lambda^{\max}(\mathbf{B})$ if **A** is symmetric positive semi-definite and **B** symmetric positive definite (Roy 1954), we see that $\lambda^{\max}(\mathbf{M}_{x,\mathbf{g}}\mathbf{H}_\mathbf{gg}\mathbf{M}_{x,\mathbf{g}}) \leq \lambda^{\max}(\mathbf{M}_{x,\mathbf{g}})^2\lambda^{\max}(\mathbf{H}_\mathbf{gg})$. As $\lambda^{\max}(\mathbf{M}_\mathbf{gg}^{-½}) = (1- \lambda^{\max}(\mathbf{H}_\mathbf{gg}))^{-½}$, applying our knowledge of $\mu_x^{\min}$ from (11b) then gives the cc results reported in (14).

Bibliography

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton, NJ: Princeton University Press, 2009.

Bell, Robert M. and Daniel F. McCaffrey. 2002. "Bias Reduction in Standard Errors for Linear Regression with Multi-Stage Samples." *Survey Methodology* 28 (2): 169-181.

Chesher, Andrew and Ian Jewitt. 1987. "The Bias of a Heteroskedasticity Consistent Covariance Matrix Estimator." *Econometrica* 55(5): 1217-1222.

Chesher, Andrew. 1989. "Hajek Inequalities, Measures of Leverage and the Size of Heteroskedasticity Robust Wald Tests." *Econometrica* 57 (4): 971-977.

Davidson, Russell and James G. MacKinnon. 1993. Estimation and Inference in Econometrics. New York: Oxford University Press, 1993.

Eicker, F. 1963. "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions." *The Annals of Mathematical Statistics* 34 (2): 447-456.

Fox, John. 2008. Applied Regression Analysis and Generalized Linear Models. Second edition. Los Angeles: Sage Publications, 2008.

Hájek, Jaroslav. 1962. "Inequalities for the Generalized Student's Distribution and their Applications." *Selected Translations in Mathematical Statistics and Probability* (2): 63-74. Providence: American Mathematical Society, 1962.

Hinkley, David V. 1977. "Jackknifing in Unbalanced Situations." *Technometrics* 19 (3): 285-292.

Huber, Peter J. 1981. Robust Statistics. New York: John Wiley & Sons, 1981.

Imbens, Guido W. and Michal Kolesar. 2015. "Robust Standard Standard Errors in Small Samples: Some Practical Advice." *Review of Economics and Statistics*, forthcoming.

Kloek, Teun. 1981. "OLS Estimation in a Model Where a Microvariable is Explained by Aggregates and Contemporaneous Disturbances are Equicorrelated." *Econometrica* 49 (1): 205-207.

Kott, Phillip S. 1994. "A Hypothesis Test of Linear Regression Coefficients with Survey Data." *Survey Methodology*: 159-164.

Kott, Phillip S. 1996. "Linear Regression in the Face of Specification Error: A Model-Based Exploration of Randomization-Based Techniques." *Proceedings of the Survey Methods Section, Statistical Society of Canada*: 39-47.

Liang, Kung-Yee and Scott L. Zeger. 1986. "Longitudinal Data Analysis using Generalized Linear Models." *Biometrika* 73 (1): 13-22.

MacKinnon, James G. and Halbert White. 1985. "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties." *Journal of Econometrics* 29 (3): 305-325.

Moulton, Brent R. 1986. "Random Group Effects and the Precision of Regression Estimates." *Journal of Econometrics* 32 (3): 385-397.

Roy, S.N. 1954. “A Useful Theorem in Matrix Theory.” *Proceedings of the American Mathematical Society* (5): 635-638.

Welch, B.L. 1936. “The Specification of Rules for Rejecting Too Variable a Product, with Particular Reference to an Electric Lamp Problem.” *Supplement to the Journal of the Royal Statistical Society* (3): 29-48.

Welch, B.L. 1938. “The Significance of the Difference Between Two Means when the Population Variances are Unequal.” *Biometrika* (29): 350-362.

Welch, B.L. 1947. “The Generalization of `Student’s’ Problem when Several Different Population Variances are Involved.” Biometrika (34): 28-35.

White, Halbert. 1980. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica* 48 (4): 817-838.

Young, Alwyn. 2015. “Channelling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results.” Working paper, October 2015.