The results with little heteroskedasticity, reported in the second panel, show that conventional standard errors are still too low; this bias is now in the order of 15%. HC_0 and HC_1 are also too small, about like before in absolute terms, though they now look worse relative to the conventional standard errors. The HC_2 and HC_3 standard errors are still larger than the conventional standard errors, on average, but empirical rejection rates are higher for these two than for conventional standard errors. This means the robust standard errors are sometimes too small "by accident," an event that happens often enough to inflate rejection rates so that they exceed the conventional rejection rates.

The lesson we can take a away from this is that robust standard errors are no panacea. They can be smaller than conventional standard errors for two reasons: the small sample bias we have discussed and the higher sampling variance of these standard errors. We therefore take empirical results where the robust standard errors fall below the conventional standard errors as a red flag. This is very likely due to bias or a chance occurrence that is better discounted. In this spirit, we like the idea of taking the maximum of the conventional standard error and a robust standard error as your best measure of precision. This rule of thumb helps on two counts: it truncates low values of the robust estimators, reducing bias, and it reduces variability. Table 8.1.1 shows the empirical rejection rates obtained using $Max(HC_j, Conventional)$. The empirical rejection rates using this rule of thumb look pretty good in the first two panels and greatly improve on the robust estimators alone.[8]

Since there is no gain without pain, there must be some cost to using $Max(HC_j, Conventional)$. The cost is that the best standard error when there is no heteroskedasticity is the conventional OLS estimate. This is documented in the bottom panel of the table. Using the maximum inflates standard errors unnecessarily under homoskedasticity, depressing rejection rates. Nevertheless, the table shows that even in this case rejection rates don't go down all that much. We also view an underestimate of precision as being less costly than an over-estimate. Underestimating precision, we come away thinking the data are not very informative and that we should try to collect more data, while in the latter case, we may mistakenly draw important substantive conclusions.

A final comment on this Monte Carlo investigation concerns the sample size. Labor economists like us are used to working with tens of thousands of observations or more. But sometimes we don't. In a study of the effects of busing on public school students, Angrist and Lang (2004) work with samples of about 3000 students grouped in 56 schools. The regressor of interest in this study varies within grade only at the school level, so some of the analysis in this paper uses 56 school means. Not surprisingly, therefore, Angrist and Lang (2004) obtained HC_1 standard errors below conventional OLS standard errors when working with school-level data. As a rule, even if you start with the micro data on individuals, when the regressor of interest varies at a higher level of aggregation - a school, state, or some other group or cluster - effective sample sizes are much closer to the number of clusters than to the number of individuals. Inference procedures for clustered data are discussed in detail in the next section.

8.2 Clustering and Serial Correlation in Panels

8.2.1 Clustering and the Moulton Factor

Bias problems aside, heteroskedasticity rarely leads to dramatic changes in inference. In large samples where bias is not likely to be a problem, we might see standard errors increase by about 25 percent when moving from the conventional to the HC_1 estimator. In contrast, clustering can make all the difference.

The clustering problem can be illustrated using a simple bivariate regression estimated in data with a group structure. Suppose we're interested in the bivariate regression,

$$\text{Y}_{ig} = \beta_0 + \beta_1 x_g + e_{ig}, \tag{8.2.1}$$

where Y_{ig} is the dependent variable for individual i in cluster or group g, with G groups. Importantly, the regressor of interest, x_g, varies only at the group level. For example, data from the STAR experiment analyzed by Krueger (1999) come in the form of Y_{ig}, the test score of student i in class g, and class size, x_g.

Although students were randomly assigned to classes in the STAR experiment, the data are unlikely to be independent across observations. The test scores of students in the same class tend to be correlated because students in the same class share background characteristics and are exposed to the same teacher

[8] Yang, Hsu, and Zhao (2005) formalize the notion of test procedures based on the maximum of a a set of test statistics with differing efficiency and robustness properties.

and classroom environment. It's therefore prudent to assume that, for students i and j in the same class, g,

$$E[e_{ig}e_{jg}] = \rho\sigma_e^2 > 0, \tag{8.2.2}$$

where ρ is the intra-class correlation coefficient and σ_e^2 is the residual variance.[9]

Correlation within groups is often modeled using an additive random effects model. In particular, we assume that the residual, e_{ig}, has a group structure:

$$e_{ig} = v_g + \eta_{ig}. \tag{8.2.3}$$

where v_g is a random component specific to class g and η_{ig} is a mean-zero student-level component that's left over. We focus here on the correlation problem, so both of these error components are assumed to be homoskedastic.

When the regressor of interest varies only at the group level, an error structure like (8.2.3) can increase standard errors sharply. This unfortunate fact is not news - Kloek (1981) and Moulton (1986) both made the point - but it seems fair to say that clustering didn't really become part of the applied econometrics zeitgeist until about 15 years ago.

Given the error structure, (8.2.3), the intra-class correlation coefficient becomes

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2}.$$

where σ_v^2 is the variance of v_g and σ_η^2 is the variance of η_{ig}. A word on terminology: ρ is called the *intra-class correlation coefficient* even when the groups of interest are not classrooms.

Let $V_c(\widehat{\beta}_1)$ be the conventional OLS variance formula for the regression slope (generated using Ω_c in the previous section), while $V(\widehat{\beta}_1)$ denotes the correct sampling variance given the error structure, (8.2.3). With regressors fixed at the group level and groups of equal size, n, we have

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + (n-1)\rho, \tag{8.2.4}$$

a formula derived in the appendix to this chapter. We call the square root of this ratio the Moulton factor, after Moulton's (1986) influential study. Equation (8.2.4) tells us how much we over-estimate precision by ignoring intra-class correlation. Conventional standard errors become increasingly misleading as n and ρ increase. Suppose, for example, that $\rho = 1$. In this case, all the errors within a group are the same, so the Y_{ig}'s are the same as well. Making a data set larger by copying a smaller one n times generates no new information. The variance $V_c(\widehat{\beta}_1)$ should therefore be scaled up by a factor of n. The Moulton factor increases with group size because with a fixed overall sample size, larger groups means fewer clusters, in which case there is less independent information in the sample (because the data are independent across clusters but not within).[10]

Even small intra-class correlation coefficients can generate a big Moulton factor. In Angrist and Lavy (2007), for example, 4000 students are grouped in 40 schools, so the average n is 100. The regressor of interest is school-level treatment status - all students in treated schools were eligible to receive cash rewards for passing their matriculation exams. The intra-class correlation in this study fluctuates around .1. Applying formula (8.2.4), the Moulton factor is over 3: the standard errors reported by default are only one-third of what they should be.

Equation (8.2.4) covers an important special case where the regressors are fixed within groups and group size is constant. The general formula allows the regressor, x_{ig}, to vary at the individual level and for different group sizes, n_g. In this case, the Moulton factor is the square root of

$$\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)} = 1 + \left[\frac{V(n_g)}{\overline{n}} + \overline{n} - 1\right]\rho_x\rho, \tag{8.2.5}$$

where $\overline{n}$ is the average group size, and ρ_x is the intra-class correlation of x_{ig}:

$$\rho_x = \frac{\sum_g \sum_{i\neq k} (x_{ig} - \overline{x})(x_{kg} - \overline{x})}{V(x_{ig}) \sum_g n_g(n_g - 1)}.$$

[9] This sort of residual correlation structure is also a consequence of stratified sampling (see, e.g., Wooldridge, 2003). Most of the samples that we work with are close enough to random that we typically worry more about the dependence due to a group structure than clustering due to stratification.

[10] With non-stochastic regressors and homoscedastic residuals, the Moulton factor is a finite-sample result. Survey statisticians call the Moulton factor the *design effect* because it tells us how much to adjust standard errors in stratified samples for deviations from simple random sampling (Kish, 1965).

Note that ρ_x does not impose a variance-components structure like (8.2.3) - here, ρ_x is a generic measure of the correlation of regressors within groups. The general Moulton formula tells us that clustering has a bigger impact on standard errors with variable group sizes and when ρ_x is large. The impact vanishes when $\rho_x = 0$. In other words, if the x_{ig}'s are uncorrelated within groups, the grouped error structure does not matter for the estimation of standard errors. That's why we worry most about clustering when the regressor of interest is fixed within groups.

We illustrate formula (8.2.1) using the Tennessee STAR example. A regression of Kindergartners' percentile score on class size yields an estimate of -0.62 with a robust (HC_1) standard error of 0.09. In this case, $\rho_x = 1$ because class size is fixed within classes while $V(n_g)$ is positive because classes vary in size (in this case, $V(n_g) = 17.1$). The intra-class correlation coefficient for residuals is .31 and the average class size is 19.4. Plugging these numbers into (8.2.1) gives a value of about 7 for $\frac{V(\widehat{\beta}_1)}{V_c(\widehat{\beta}_1)}$, so that conventional standard errors should be multiplied by a factor of $2.65 = \sqrt{7}$. The corrected standard error is therefore about 0.24.

The Moulton factor works similarly with 2SLS except that ρ_x should be computed for the instrumental variable and not the regressor. In particular, use (8.2.5) replacing ρ_x with ρ_z, where ρ_z is the intra-class correlation coefficient of the instrumental variable (Shore-Sheppard, 1996) and ρ is the intra-class correlation of the second-stage residuals. To understand why this works, recall that conventional standard errors for 2SLS are derived from the residual variance of the second-stage equation divided by the variance of the first-stage fitted values. This is the same asymptotic variance formula as for OLS, with first-stage fitted values playing the role of regressor.[11]

Here are some solutions to the Moulton problem:

1. Parametric: Fix conventional standard errors using (8.2.5). The intra-class correlations ρ and ρ_x are easy to compute and supplied as descriptive statistics in some software packages.[12]

2. Cluster standard errors: Liang and Zeger (1986) generalize the White (1980a) robust covariance matrix to allow for clustering as well as heteroskedasticity:

$$\begin{aligned} \hat{V}(\widehat{\beta}) &= (X'X)^{-1}\left(\sum_g X_g \widehat{\Psi}_g X_g\right)(X'X)^{-1} \qquad (8.2.6)\\ \widehat{\Psi}_g &= a\widehat{e}_g\widehat{e}_g' = a\begin{bmatrix} \widehat{e}_{1g}^2 & \widehat{e}_{1g}\widehat{e}_{2g} & \cdots & \widehat{e}_{1g}\widehat{e}_{n_g g} \\ \widehat{e}_{1g}\widehat{e}_{2g} & \widehat{e}_{2g}^2 & & \vdots \\ \vdots & & \ddots & \widehat{e}_{(n_g-1)g}\widehat{e}_{n_g g} \\ \widehat{e}_{1g}\widehat{e}_{n_g g} & \cdots & \widehat{e}_{(n_g-1)g}\widehat{e}_{n_g g} & \widehat{e}_{n_g g}^2 \end{bmatrix}. \end{aligned}$$

   where X_g is the matrix of regressors for group g, and a is a degrees of freedom adjustment factor similar to that in HC_1. The clustered variance estimator $\hat{V}(\widehat{\beta})$ is consistent as the number of groups gets large under any within-group correlation structure and not just the parametric model in (8.2.3). $\hat{V}(\widehat{\beta})$ is not consistent with a fixed number of groups, however, even when the group size tends to infinity. Clustered standard errors are therefore unlikely to be reliable with few clusters, a point we return to below.

3. Use group averages instead of micro data: let $\overline{\mathrm{Y}}_g$ be the mean of Y_{ig} in group g. Estimate

$$\overline{\mathrm{Y}}_g = \beta_0 + \beta_1 x_g + \overline{e}_g$$

   by weighted least squares using the group size as weights. This is equivalent to OLS using micro data but the standard errors are asymptotically correct given the group structure, (8.2.3). Again, the asymptotics here are based on the number of groups and not the group size. Importantly, however, because the group means are close to Normally distributed with modest group sizes, we can expect the good finite-sample properties of regression with Normal errors to kick in. The standard errors that come out of grouped estimation are therefore likely to be more reliable than clustered standard errors in samples with few clusters.

[11] Clustering can also be a problem in regression-discontinuity designs if the variable that determines treatment assignment varies only at a group level (see Card and Lee, 2008, for details).

[12] Use Stata's `loneway` command, for example.