OVERIDENTIFICATION TESTS WITH GROUPED DATA

Caroline Hoxby
M. Daniele Paserman

Technical Working Paper **223**

TECHNICAL WORKING PAPER SERIES

OVERIDENTIFICATION TESTS WITH GROUPED DATA

Caroline Hoxby
M. Daniele Paserman

Technical Working Paper 223
http://www.nber.org/papers/T0223

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
February 1998

Overidentification Tests with Grouped Data
Caroline Hoxby and M. Daniele Paserman
NBER Technical Working Paper No. 223
February 1998

ABSTRACT

This paper examines the validity of overidentification tests and exogeneity tests in the presence of grouped data. We find that even a small intra-group correlation, when instruments do not vary within groups, may generate a substantial bias in the standard overidentification tests described in textbooks.

Caroline Hoxby
Department of Economics
Harvard University
Cambridge, MA 02138
and NBER
choxby@harvard.edu

M. Daniele Paserman
Department of Economics
Harvard University
Cambridge, MA 02138

Overidentification Tests With Grouped Data*

Caroline Hoxby and M. Daniele Paserman

This Version: December, 1997

Abstract

This paper examines the validity of overidentification tests and exogeneity tests in the presence of grouped data. We find that even a small intra-group correlation, when instruments do not vary within groups, may generate a substantial bias in the standard overidentification tests described in textbooks.

1 Introduction

This paper demonstrates that standard overidentification tests for instrumental variables estimation are invalid in the presence of grouped data, and proposes a simple remedy. When instrumental variables vary only between groups, the presence of a small amount of intra-group correlation generates substantial bias in both commonly used overidentification tests, the "omnibus" test (Anderson and Rubin, 1950; Basmann, 1960; Hausman, 1983) and the Hausman test for the validity of a specified subset of instrumental variables (Hausman, 1983; Spencer and Berk, 1981). Our result is a direct extension of work by Moulton (1986) and

*Preliminary and incomplete

Shore-Sheppard (1996). They show that a grouped error structure combined with no within-group variation leads to severe underestimation of standard errors and incorrect statistical inference in, respectively, ordinary least squares and instrumental variables estimation. We show that overidentification tests that do not allow for a grouped error structure can reject the overidentifying restrictions far too often, and we generate the correct test statistic using Generalized Method of Moments (GMM) techniques (Hansen, 1982, Newey, 1985). Using Monte-Carlo simulations and two recent empirical articles, we demonstrate the seriousness of this problem and the usefulness of the correct test statistic for common empirical applications. We view this contribution as a practical one.

The recent literature on instrumental variables emphasizes how important it is that the identifying restrictions are credible–that is, that the model is correctly specified. As a result, the two overidentification tests mentioned are increasingly likely to be used and reported. Each is a partial test of whether the identifying restrictions are satisfied. From Moulton and Shore-Sheppard, we know that OLS and IV standard errors can be severely biased in the presence of grouped data and even small intra-group correlation; furthermore, we know that the bias grows with the sample size and the number of groups. It is reasonable to suspect that similar biases affect the overidentification tests, which are based on the residuals from the instrumental variables estimation.

There is a long empirical tradition of using instrumental variables that do not vary within groups. State laws may be used to form instruments for an individual's marginal tax rate (Feldstein and Eissa,), an individual's job benefits (Gruber, 1992), a company's unionization status (Brown and Medoff, 1988), the starting wage a firm offers (Katz and Krueger, 1992), or teen fertility (Kane and Staiger, 1996; Levine, Trainor, and Zimmerman, 1996). Unemployment rates by industry-occupation cell have been used as instruments for an individual's own unemployment status (Murphy and Topel, 1987). Exchange rates by industry-country-of-origin cell have been used as instruments for the degree of import penetration an area

experiences (Revenga, 1992). All these examples share the feature that the identifying instrumental variables are measured at a higher level of aggregation than the unit of analysis. The potentially endogenous variable may or may not be aggregated at the same level as the instruments. Although, as described above, empirical use of instrumental variables has changed considerably over the past several years, grouped instrumental variables have continued to be useful. For instance, in a study of the effect of teen childbearing on schooling, Angrist and Evans (1996) use variation in the state abortion laws to which teens were exposed to instrument for their fertility.

The first application in this paper is based on Cutler and Glaeser (1997), who use historical metropolitan area characteristics to instrument for the effect of segregation on the economic performance of minorities. This application, which employs data from the Public Use Micro Samples of the U.S. Census of Population (PUMS), illustrates how severely the overidentification tests are biased when the number of observations (individuals) within each group (metropolitan area) is very large. The second application, Hoxby (1998), also uses a common type of data. Using data on students from the National Education Longitudinal Survey (NELS), Hoxby instruments for measures of the degree of competition among public school districts in a metropolitan areas using topographic data also measured at the metropolitan area level. This application forms a useful comparison with the other, because the instrumental variables strategies are similar but the number of individuals within each group is two orders of magnitude smaller than in Cutler and Glaeser's.

The rest of the paper is structured as follows. Section 2 describes the problem in a GMM framework, illustrates that the standard overidentification tests are inconsistent, and shows how to calculate the correct test statistic. In Section 3 we report the results of several Monte Carlo simulations, with which we numerically assess the magnitude of the bias under various assumptions about the number of observations, number of groups, and intra-group correlation. In Section 4 we present the results from the two applications, and in Section 5

we conclude.

2 Overidentification Tests with Grouped Data

2.1 Some basic GMM results

Let $r_{ij} = \left(y_{ij}, \mathbf{x}'_{ij}, \mathbf{z}'_{ij}\right)$ be a vector of observed variables for unit i in group j, with $i = 1, 2, ...N_j$, and $j = 1, 2, ..., M$. Let $T = \sum_{j=1}^{M} N_j$ be the total number of observations in the sample. Assume that the statistical model implies a set of orthogonality conditions:

$$E\left\{\boldsymbol{\psi}\left(\theta_0, r_{ij}\right)\right\} = 0$$

where θ_0 is the true value of an unknown $p \times 1$ parameter vector, and $\psi(\cdot)$ is a differentiable q-dimensional vector valued function, with $q \geq p$. The optimal Generalized Method of Moments (GMM) estimator $\hat{\theta}_{GMM}$ is the value of θ that minimizes

$$\mathbf{g}(\boldsymbol{\theta}; \mathbf{r})' \hat{\mathbf{W}}^{-1} \mathbf{g}(\boldsymbol{\theta}; \mathbf{r})$$

where

$$\mathbf{g}(\boldsymbol{\theta}; \mathbf{r}) = \frac{1}{T} \sum_{j=1}^{M} \sum_{i=1}^{N_j} \boldsymbol{\psi}\left(\theta, r_{ij}\right)$$

and $\hat{\mathbf{W}}$ is an estimate of the asymptotic covariance matrix of the sample mean of $\boldsymbol{\psi}\left(\theta_0, r_{ij}\right)$

$$\mathbf{W} = \lim_{T \to \infty} \frac{1}{T} \sum_{j=1}^{M} \sum_{i=1}^{N_j} \left\{ \sum_{n=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} E\left[\boldsymbol{\psi}\left(\theta_0, r_{ij}\right) \boldsymbol{\psi}\left(\theta_0, r_{i-n,j-m}\right)'\right] \right\}$$

Under standard regularity conditions, the asymptotic distribution of the GMM estimator is:

$$\sqrt{T}\left(\hat{\theta}_{GMM} - \theta_0\right) \to^d N\left(0, \left(\mathbf{D}'\mathbf{W}^{-1}\mathbf{D}\right)^{-1}\right)$$

where

$$\mathbf{D}' = E\left(\left.\frac{\partial \mathbf{g}(\theta, r_{ij})}{\partial \theta'}\right|_{\theta=\theta_0}\right)$$

Moreover, under the null hypothesis that the orthogonality conditions truly hold, $\mathbf{g}(\hat{\boldsymbol{\theta}}; \mathbf{r}) = \frac{1}{T}\sum_i \sum_j \boldsymbol{\psi}(\hat{\theta}, r_{ij})$ has the following asymptotic distribution:

$$\sqrt{T}\left(\mathbf{g}(\hat{\boldsymbol{\theta}}; \mathbf{r})\right) \rightarrow^d N(0, \mathbf{W})$$

It then follows that the score statistic $\bar{m}_T = T\mathbf{g}(\hat{\boldsymbol{\theta}}; \mathbf{r})'\mathbf{W}^{-1}\mathbf{g}(\hat{\boldsymbol{\theta}}; \mathbf{r})$ has an asymptotic chi-square distribution with $q - p$ degrees of freedom.

2.2 Overidentification Tests in Linear Models

Now consider a linear model

$$y_{ij} = \mathbf{x}_{ij}'\beta + u_{ij}$$

where $\mathbf{x}_{ij}$ is a $p \times 1$ vector of potentially endogenous explanatory variables, so that $E(\mathbf{x}_{ij}u_{ij}) \neq 0$ and OLS produces inconsistent estimators. Let $\mathbf{z}_{ij}$ be a $q \times 1$, $(q \geq p)$, vector of instruments, that are correlated with $\mathbf{x}_{ij}$ but uncorrelated with u_{ij}. There are q resulting orthogonality conditions:

$$E\left(\boldsymbol{\psi}(\beta_0, r_{ij})\right) = E\left(\mathbf{z}_{ij}u_{ij}\right) = E\left(\mathbf{z}_{ij}(y_{ij} - \mathbf{x}_{ij}'\beta_0)\right) = 0$$

If the disturbances are homoskedastic, uncorrelated, and have variance σ_u^2, then $\mathbf{W} = \sigma_u^2(\mathbf{Z}'\mathbf{Z})$ (in obvious notation); and the efficient GMM estimator reduces to the Two Stage Least Squares (2SLS) estimator, $\hat{\beta}_{2SLS} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y}$, where $\mathbf{P}_Z$ is the projection matrix $\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. Moreover, the chi-squared statistic $\overline{m}_T$ takes the form:

$$\overline{m}_T = \frac{\hat{u}'\mathbf{P}_Z\hat{u}}{\hat{\sigma}_u} = T\frac{\hat{u}'\mathbf{P}_Z\hat{u}}{\hat{u}'\hat{u}}$$

where $\hat{u}$ is the vector of 2SLS residuals. Therefore, $\bar{m}_T$ is equivalent to T times the uncentered R^2 in a regression of the residuals $\hat{u}$ on the set of instruments $\mathbf{Z}$. This is Hausman's (1983) form of the "omnibus" test of overidentifying restrictions. A large value of the test

statistic (a rejection of the null) is interpreted as evidence that some of the exclusion restrictions identifying the estimates are inappropriate–that is, there are variables that have been inappropriately excluded from the model. The test is known as the "omnibus" test precisely because the inappropriately excluded variables are not distinguished in any way.

With grouped data, we are generally unwilling to assume that the disturbances are homoskedastic and uncorrelated. In this case, the above statistic is misspecified and does not have an asymptotic chi-square distribution.

With cross-sectional grouped data, one typically assumes that there may be some degree of intra-group correlation between disturbances, while maintaining the assumption of zero inter-group correlation[1].

In this case, one should form an estimate of $\mathbf{W}$ in order to obtain the efficient GMM estimator and to construct the appropriate overidentification test statistic. This is usually performed in two steps, although it is also possible to iterate the steps until convergence. The first step involves obtaining any consistent, but possibly inefficient, estimate of β. This can be achieved by performing GMM estimation using any weighting matrix: the identity matrix and $(\mathbf{Z}'\mathbf{Z})^{-1}$ are natural candidates. The residuals from this first stage, $\hat{u}_{GMM}$, are then calculated and used to construct the new weighting matrix $\hat{\mathbf{W}}$ which can be expressed as :

$$\hat{\mathbf{W}} = \frac{1}{T}\sum_{j}\left[\left(\sum_{i}^{N_j}\mathbf{z}_{ij}\hat{u}_{ij}\right)\left(\sum_{i}^{N_j}\hat{u}_{ij}\mathbf{z}'_{ij}\right)\right]$$

[1] A special case of this assumption is the error components model,

$$\begin{aligned} u_{ij} &= \eta_j + \varepsilon_{ij} \\ E(\eta_j) &= 0, \quad V(\eta_j) = \sigma_\eta^2 \\ E(\varepsilon_{ij}) &= 0, \quad V(\varepsilon_{ij}) = \sigma_\varepsilon^2 \end{aligned}$$

where disturbances are assumed to be equicorrelated within a group. However, most recent cross-sectional studies do not impose this restriction.

This form of $\hat{\mathbf{W}}$ takes the grouping of the data into account. It also takes into account our assumption that there is no correlation between disturbances across groups, but that there may be intra-group correlation. Note, however, that we are not restricting the covariance matrix of the disturbances to have the block-diagonal form implied by the error components model.

The GMM methods described above can also be applied to the Hausman test for the exogeneity of a subset of the explanatory variables, taking the exogeneity of the others as given. This test is often used when several instrumental variables of differing credibility are available. The most credible instrumental variables are assumed to be exogenous, and the exogeneity of the other potential instruments is tested. Under the null hypothesis that the questionable variables are exogenous, both the vector of questionable instruments $\mathbf{z}_{ij}$ and the explanatory variable vector $\mathbf{x}_{ij}$ are uncorrelated with u_{ij}. We would then have $p+q$ orthogonality conditions, so that we can derive a GMM estimator and a test of overidentifying restrictions. With homoskedastic and uncorrelated disturbances, the overidentification test is asymptotically equivalent to Spencer and Berk's (1981) single-equation version of the test[2]. However, if disturbances are correlated and/or heteroskedastic, the standard test will be misspecified; and the GMM framework provides a convenient way of reformulating this test.

[2]The test is a simple Wald test of the null $\gamma = 0$ in the regression:

$$y = X_1\beta_1 + X_2\beta_2 + \hat{X}_2\gamma + u$$

where X_1 is the matrix of explanatory variables assumed to be exogenous, X_2 is the matrix of questionable explanatory variables, and $\hat{X}_2$ is the matrix of predicted values from a regression of X_2 on all the exogenous variables.

See Greene (1993) p. 618, or Pindyck and Rubinfeld (1991) pp.303-304.

3 Monte Carlo Experiments

To numerically illustrate the problems described above, we ran a series of Monte Carlo experiments, varying the intra-group correlation parameter ρ, the sample sizes, and the grouping of the instrumental variables.

The data were generated as follows:

$$\begin{aligned} y_{ij} &= \alpha + \beta_1 x_{1_{ij}} + \beta_2 x_{2_{ij}} + u_{ij} \\ u_{ij} &= \eta_j + \varepsilon_{ij} \\ x_{2_{ij}} &= \lambda_j + \delta_{ij} + \eta_j + \varepsilon_{ij} \\ z_{1_{ij}} &= \lambda_j + k_1\delta_{ij} + \mu_{1_j} + \tau_{1_{ij}} \\ z_{2_{ij}} &= \lambda_j + k_2\delta_{ij} + \mu_{2_j} + \tau_{2_{ij}} \\ i &= 1, ..., N \qquad j = 1, ..., M \end{aligned}$$

The error components, $\eta, \varepsilon, \lambda, \delta, \mu_1, \mu_2, \tau_1$, and τ_2 all have mean zero and variance σ_θ^2, where θ denotes the appropriate Greek letter. k_1 and k_2 are either 1 or 0, depending on whether or not we assume that the instrument is grouped. We fix σ_ε^2 at 1 and let σ_η^2 vary between 0 and 0.2. Define $\rho = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2}$ to be the intra-group correlation parameter. For simplicity, we keep the number of observations in each group constant, and let N vary between runs.[3]

In order to investigate the performance of the Hausman test, we slightly modify the data generating process, letting $\tilde{x}_{2_{ij}} = \lambda_j + \delta_{ij}$ and $\tilde{y}_{ij} = \alpha + \beta_1 x_{1ij} + \beta\tilde{x}_{2_{ij}} + u_{ij}$. (Hence, $\tilde{x}_2$ is uncorrelated with the error term in this case).

Table 1a reports the results of 1,000 Monte Carlo simulations. The first two rows in each block display the coverage rate of the 95% confidence interval for the coefficient on

[3]We also ran some Monte Carlo simulations in which we allowed N to vary between groups, but the results did not change substantially.

the endogenous variable x_2[4], using standard errors constructed with the unadjusted and the adjusted covariance matrix. The second two rows in each block show the rejection rates of the unadjusted and adjusted "omnibus" overidentification test statistic. The final two rows show the rejection rates of the unadjusted and adjusted Spencer-Berk test for the exogeneity of $\tilde{x}_2$.

The first column of the table demonstrates that, if there is no intra-group correlation, then standard Two Stage Least Squares estimation leads to correct statistical inference, with the 95% confidence interval covering the "true" parameter 95% of the time and the numerical size of the tests approximating their nominal size. The adjusted confidence intervals and the adjusted overidentification tests achieve nearly identical results. However, as we increase the intra-group correlation parameter ρ, we see that the performance of the standard procedures worsens considerably, especially if the instruments are grouped. For a value of ρ as small as 0.01 the standard overidentification test still does reasonably well if the instruments vary at the individual level. However, with one instrument grouped, the standard overidentification test rejects the null hypothesis nearly 25% of the time even though the confidence interval coverage rate may indicate that the bias is not too severe. With ρ greater or equal to 0.1, standard 2SLS estimation can lead to grossly incorrect statistical inference.

The Monte Carlo results for the exogeneity test tell a similar story, with the standard test rejecting the null more than 70% of the time for large ρ. Simulations for both the omnibus test and exogeneity test demonstrate that performance of the standard procedures does not vary much depending on whether only one or both instruments are grouped.

Tables 1b and 1c are analogous to Table 1a, except that they replicate the number of

[4]The regression equation was

$$y_{ij} = -5 + 0.14x_1 + 0.9x_2 + u_{ij}$$

as in Shore-Sheppard (1996). The coverage rate was calculated as the percentage of times that the 95% confidence interval around $\hat{\beta}_2$ actually included 0.9.

units in each group. In Table 1b, there are fewer (100) units in group than Table 1a; in Table 1c, there are more (500). The tables show that the misspecification problem does indeed become more severe as the number of observations in each group increases. With $N = 500$, small intra-group correlation of 0.01 leads to significant overstatement of the test statistics and to incorrect rejection of the null hypothesis approximately 40% of the time when one instrument is grouped. Even with $\rho = 0.001$ the rejection rate is slightly too high when at least one instrument is grouped.

Tables 2a and 2b further investigate the effects of sample size and the number of units in each group on the rejection rates of the overidentification test statistics. In Table 2a, we keep the intra-group correlation ρ equal to 0.01, and we compare the performance of the unadjusted and adjusted test statistics with no grouped instruments and with only one grouped instrument. Table 2b examines the effect of reducing intra-group correlation by one order of magnitude to 0.001.

The first panel of each table holds the number of groups constant and lets the number of units in each group increase from 100 to 500. The second panel holds the total sample size constant but lets the ratio of groups to units per group vary. The third panel holds the number of units per group constant, and lets the number of groups vary. Comparing the first and second panel of each table to the third panel, we see that the unadjusted test statistics reject too often as the number of units per group increases, but are relatively unaffected by the number of groups. As before, the performance of the unadjusted test statistics is always much worse when one of the instruments is grouped. This result is most clear in Table 2a, in which $\rho = 0.01$, but one can also discern a very slight pattern of excessive rejection rates with ρ as small as 0.001 in Table 2b.

4 Two Applications

Two types of data in which the problem of grouped instruments and a clustered error structure commonly arise are longitudinal datasets with 5,000-50,000 observations (such as the NELS, Panel Survey of Income Dynamics, Survey of Income and Program Participation, and National Longitudinal Surveys) and cross-section surveys with 100,000 to 12 million observations (such as the Current Population Survey and the Micro Samples of the Census of Population). We have chosen two applications with the purpose of illustrating the scale of the problem in each type of data.

4.1 "Are Ghettos Good or Bad?"[5].

Using data on individuals from the 1% Public Use Micro Sample of the Census of Population, Cutler and Glaeser attempt to determine whether there is any differential effect of housing segregation on the educational, social and labor market outcomes of blacks. Cutler and Glaeser's basic strategy is to estimate equations of the form:

$$Y_{ij} = \beta_1 \times Segregation_j + \beta_2 \times (Segregation_j * Black_{ij}) + X'_{ij}\gamma + \varepsilon_{ij}$$

where i indexes individuals, j indexes metropolitan areas, $Segregation_j$ is a measure of housing segregation that varies only between metropolitan areas, and X_{ij} is a set of variables[6]. Segregation may be the result of poor economic outcomes, or may reflect omitted city characteristics. The authors therefore use a set of fiscal variables (the number of governments in 1962, the percentage of revenue from inter-governmental transfers in 1962, and the interactions of these two variables with an indicator variable for black) as instrumental variables to

[5]We are very grateful to the authors for making their data available.

[6]The control variables are age, sex and race indicators, log of MSA population, log MSA median income, percentage blacks and percentage employed in manufacturing in an MSA, and the interaction of these last 4 variables with the black dummy variable.

identify potentially endogenous segregation. Their data are clustered and their instruments are grouped so that any test of overidentifying restrictions that they performed would be likely to be misspecified, as described above.[7].

In Table 3, we replicate some of Cutler and Glaeser's results: the first two panels show the estimated coefficients and standard errors for the two potentially endogenous variables under OLS and 2SLS estimation. the next part of Table 3 displays the omnibus overidentification test statistic, which is distributed asymptotically as a χ^2 random variable with 2 degrees of freedom. We can see that the standard test statistic strongly rejects the null hypothesis in every case but one, whereas the test statistic that adjusts for the clustered error structure, while close to the critical value, fails to reject in every case but one.

We also report the results of tests for the null hypothesis of exogeneity of the two segregation variables. The unadjusted test statistic assumes a homoskedastic error structure, while the adjusted statistic takes the clustered nature of the data into account. The test statistic is distributed asymptotically as a χ^2 random variable with 4 degrees of freedom. The final rows of Table 3 show that the unadjusted test would reject the null hypothesis of exogeneity in all cases but one, whereas the results of the adjusted test are ambiguous.

4.2 Competition Among Public Schools

Hoxby (1998) uses the data on individual students from the National Education Longitudinal Survey in an attempt to determine whether competition among public schools affects per-pupil expenditure or student achievement. Hoxby measures competition among public school districts with a Herfindahl index of school districts' shares of total metropolitan area

[7]Cutler and Glaeser are aware of this problem, and choose not to report the overidentification test statistic. They explain in a footnote that the "...instruments generally fail the standard test of overidentifying restrictions,...,[but the test] is based on an assumption of uncorrelated error terms, which is plainly violated in our data."

enrollment. She is concerned that observed school district concentration is not wholly exogenous: a district's enrollment may reflect its success, and district consolidation may also be a function of a school's success or of the heterogeneity of a metropolitan area's population. She therefore uses variation in the number of natural bodies of water in a metropolitan area to instrument for school concentration, based on the fact that areas with many natural boundaries (such as rivers and streams) tended to be divided into more, smaller school districts when boundaries were initially created. The two instrumental variables are the number of large rivers and the number of smaller streams, so that there is one degree of freedom in the test of overidentifying restrictions[8]. The endogenous variable and the instruments are grouped at the metropolitan area level, inducing the potential for bias discussed above. The OLS and 2SLS coefficients and standard errors are presented in the first two panels of Table 4, and the unadjusted and adjusted overidentification test statistic are presented just below. In this case, the adjusted test statistics are substantially smaller than the unadjusted test statistics, but even the unadjusted test statistics fail to reject the null hypothesis. These results illustrate one of the key conclusions of the Monte Carlo exercise: the larger the number of units in each group, the greater is the extent of the bias. Comparing the Hoxby application to the Cutler and Glaeser application, we conclude that incorrect calculation of the overidentification tests is much more likely to be a serious practical problem in analysis of large datasets, like the Census or Current Population Survey, in which the common levels of grouping for instrumental variables are the metropolitan area, county, or state.

We also report the results the test for the exogeneity of the topographic instrumental variables. These results suggest the same conclusion: the adjusted test statistics are substantially smaller than the unadjusted ones, but difference is not as likely to be decisive when there are a smaller number of units in each group.

[8]The reasoning is that large rivers divide counties, across whose boundaries school districts are almost never consolidated. Small streams often were of sufficient importance to create natural boundaries when districting was performed, but are irrelevant to transportation costs and higher-level jurisdictional boundaries.

5 Conclusion

In this paper we show how to calculate the correct overidentification and exogeneity test statistics in the presence of a grouped error structure. The Monte Carlo experiments illustrate that the standard test statistics that ignore the grouped structure of the data may be severely biased, leading one to reject the null hypothesis far too often. We show that this problem is exacerbated when the instruments themselves are grouped and do not vary at the individual level. Finally, using data from two recent applications that are representative of current empirical methods and common datasets, we illustrate how the incorrect test statistic may lead to incorrect statistical inference and the usefulness of the corrected test statistic.

References

[1] Anderson, T, and H. Rubin, (1950), "The Asymptotic Properties of Estimators of the Parameters of a Single Equation in a Complete System of Stochastic Equations," *Annals of Mathematical Statistics*, 21:570-582.

[2] Angrist, Joshua D., and William N. Evans, (1996), "Schooling and Labor Market Consequences of the 1970 State Abortion Reforms", NBER WP #5406.

[3] Basmann, R, (1960), "On Finite Sample Distributions of Generalized Classical Linear Identifiability Test Statistics," *Journal of the American Statistical Association*, 55: 650-59.

[4] Brown, Charles, and James Medoff, (1988), "Employer Size, Pay, and the Ability to Pay in the Public Sector," in R. Freeman and C. Ichniowski, eds., *When Public Sector Workers Unionize*, Chicago: University of Chicago Press for the NBER.

[5] Cutler, David M., and Edward L.Glaeser (1997), "Are Ghettos Good or Bad?", *Quarterly Journal of Economics*, 112(3):827-872.

[6] Eissa, Nada, (1995), "Taxation and the Labor Supply of Married Women: The Tax Reform Act of 1986 as a Natural Experiment," NBER Working Paper No. 5023.

[7] Greene, William H. (1993), *Econometric Analysis*, Second Edition, New York: Macmillan Publishing Company.

[8] Gruber, Jonathan, (1992), "The Efficiency of a Group-Specific Mandated Benefit: Evidence from Health Insurance Benefits for Maternity," NBER Working Paper No. 4157.

[9] Hansen, Lars Peter, (1982), "Large Sample Properties of Generalized Method of Moments Estimators", *Econometrica*, 50(4):1029-1054.

[10] Hausman, J, (1983), "Specification and Estimation of Simultaneous Equations Models," in Z. Griliches and M. Intriligator, eds., *Handbook of Econometrics*, Amsterdam: North Holland.

[11] Hoxby, Caroline M., (1998), "Does Competition Among Public Schools Benefit Students and Taxpayers? Evidence from Natural Variation in School Districting", NBER Working Paper No. 4979.

[12] Kane, Thomas, and Douglas Staiger, (1996), "Teen Motherhood and Abortion Access," *Quarterly Journal of Economics*, 111(2): 467-506.

[13] Katz, Lawrence, and Alan Krueger, (1992), "The Effect of the Minimum Wage on the Fast Food Industry," *Industrial and Labor Relations Review*, 46(1): 6-21.

[14] Levine, Phillip, Amy Trainor, and David Zimmerman, (1996), "The Effects of Medicaid Abortion Funding Restrictions on Abortions, Pregnancies, and Births," *Journal of Health Economics*, 15(5): 555-78.

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

[16] Murphy, Kevin, and Robert Topel, (1987), "Unemployment, Risk, and Earnings: Testing for Equalizing Differences in the Labor Market," in K. Lang and J. Leonard, eds., *Unemployment and the Structure of Labor Markets.*

[17] Newey, Whitney K., "Generalized Method of Moments Specification Testing", *Journal of Econometrics*, 29:229-256.

[18] Pindyck, R., and D. Rubinfeld., *Econometric Models and Economic Forecasts*, Third Edition, New York: McGraw-Hill.

[19] Revenga, Ana, (1992), "Exporting Jobs? The Impact of Import Competition on Employment and Wages in U.S. Manufacturing," *Quarterly Journal of Economics*, 107(6), 255-84.

[20] Shore-Sheppard, Lara, (1996), "The Precision of Instrumental Variables Estimates with Grouped Data", Princeton Industrial Relations Section, Working Paper No. 374.

[21] Spencer, David E.., and Kenneth N. Berk, (1981), "A Limited Information Specification Test", *Econometrica,* 49(4):1079-1085.

Table 1a [9]

$M = 50$, $N = 200$	ρ					
	0	0.0001	0.001	0.01	0.1	0.2
No grouped instruments						
95% CI (unadj.)	.953	.945	.941	.881	.605	.467
95% CI (adj.)	.939	.939	.932	.925	.952	.925
Overidentification Rejection rate (unadj.)	.051	.052	.049	.109	.333	.497
Overidentification Rejection rate (adj.)	.050	.052	.037	.062	.060	.055
Exogeneity Rejection Rate (unadj.)	.049	.063	.063	.129	.530	.706
Exogeneity Rejection Rate (adj.)	.052	.054	.051	.049	.046	.047
z_2 grouped						
95% CI (unadj.)	.956	.940	.934	.850	.522	.385
95% CI (adj.)	.930	.930	.932	.928	.945	.927
Overidentification Rejection rate (unadj.)	.060	.064	.078	.242	.591	.736
Overidentification Rejection rate (adj.)	.060	.059	.054	.060	.037	.059
Exogeneity Rejection Rate (unadj.)	.046	.062	.082	..262	.729	.856
Exogeneity Rejection Rate (adj.)	.050	.059	.058	.052	.045	.049
z_1 and z_2 grouped						
95% CI (unadj.)	.949	.943	.925	.741	.369	.256
95% CI (adj.)	.925	.935	.916	.920	.932	.900
Overidentification Rejection rate (unadj.)	.057	.050	.072	.334	.649	.726
Overidentification Rejection rate (adj.)	.061	.050	.044	.051	.052	.051
Exogeneity Rejection Rate	.042	.061	.085	.348	.833	.910
Exogeneity Rejection Rate	.042	.048	.051	.060	.054	.052

[9] Results from 1000 Monte Carlo simulations described in text. The entries in the table represent the coverage rate of the 95% confidence interval for $\hat{\beta}_2$, the rejection rate of the overidentification test statistic with nominal size 5%, and the rejection rate of the exogeneity test statistic with nominal size 5%.

Table 1b [10]

$M = 50, N = 100$	ρ					
	0	0.0001	0.001	0.01	0.1	0.2
No grouped instruments						
95% CI (unadj.)	.948	.954	.953	.922	.722	.608
95% CI (adj.)	.940	.943	.943	.945	.922	.939
Overidentification Rejection Rate (unadj.)	.059	.056	.051	.073	.238	.343
Overidentification Rejection Rate (adj.)	.057	.052	.046	.041	.048	.053
Exogeneity Rejection Rate (unadj.)	.056	.047	.063	.084	.349	.503
Exogeneity Rejection Rate (adj.)	.057	.050	.050	.041	.053	.046
z_2 grouped						
95% CI (unadj.)	.946	.950	.941	.902	.638	.511
95% CI (adj.)	.945	.940	.925	.935	.929	.922
Overidentification Rejection rate (unadj.)	.060	.043	.059	.110	.494	.619
Overidentification Rejection Rate (adj.)	.060	.049	.053	.037	.051	.072
Exogeneity Rejection Rate	.060	.045	.066	.123	.583	.730
Exogeneity Rejection Rate	.056	.037	.061	.042	.058	.060
z_1 and z_2 grouped						
95% CI (unadj.)	.948	.961	.925	.832	.443	.365
95% CI (adj.)	.937	.947	.912	.930	.904	.886
Overidentification Rejection rate (unadj.)	.059	.047	.069	.161	.544	.619
Overidentification Rejection Rate (adj.)	.057	.054	.049	.05	.045	.047
Exogeneity Rejection Rate (unadj.)	.053	.048	.075	.193	.728	.828
Exogeneity Rejection Rate (adj.)	.056	.044	.050	.045	.055	.058

[10] Results from 1000 Monte Carlo simulations described in text. The entries in the table represent the coverage rate of the 95% confidence interval for $\hat{\beta}_2$, the rejection rate of the overidentification test statistic with nominal size 5%, and the rejection rate of the exogeneity test statistic with nominal size 5%.

Table 1c [11]

$M = 50, N = 500$	ρ					
	0	0.0001	0.001	0.01	0.1	0.2
No grouped instruments						
95% CI (unadj.)	.958	.947	.929	.802	.407	.321
95% CI (adj.)	.945	.939	.931	.940	.937	.922
Overidentification Rejection rate (unadj.)	.045	.062	.050	.160	.507	.631
Overidentification Rejection rate (adj.)	.044	.056	.044	.056	.042	.062
Exogeneity Rejection Rate (unadj.)	.041	.060	.046	.256	.731	.834
Exogeneity Rejection Rate (adj.)	.040	.060	.034	.042	.050	.063
z_2 grouped						
95% CI (unadj.)	.959	.950	.927	.721	.319	.248
95% CI (adj.)	.946	.938	.934	.944	.941	.922
Overidentification Rejection rate (unadj.)	.052	.063	.110	.370	.758	.828
Overidentification Rejection rate (adj.)	.054	.053	.061	.054	.038	.073
Exogeneity Rejection Rate (unadj.)	.058	.068	.101	.441	.893	.924
Exogeneity Rejection Rate (adj.)	.044	.059	.044	.053	.045	.064
z_1 and z_2 grouped						
95% CI (unadj.)	.960	.940	.884	.569	.204	.169
95% CI (adj.)	.921	.925	.923	.932	.915	.888
Overidentification Rejection rate (unadj.)	.052	.054	.100	.417	.737	.838
Overidentification Rejection rate (adj.)	.062	.052	.048	.055	.046	.056
Exogeneity Rejection Rate (uandj.)	.053	.067	.136	.572	.931	.967
Exogeneity Rejection Rate (adj.)	.055	.049	.045	.048	.049	.060

[11] Results from 1000 Monte Carlo simulations described in text. The entries in the table represent the coverage rate of the 95% confidence interval for $\hat{\beta}_2$, the rejection rate of the overidentification test statistic with nominal size 5%, and the rejection rate of the exogeneity test statistic with nominal size 5%.

Table 2a [12]

$\rho = 0.01$	**No Grouped Instruments**				**One grouped instrument**			
	Overid.		Exog.		Overid.		Exog.	
M, N	Unadj.	Adj.	Unadj.	Adj.	Unadj.	Adj.	Unadj	Adj.
50,100	.085	.056	.104	.066	.136	.052	.157	.055
50,200	.089	.046	.123	.042	.213	.052	.237	.044
50,300	.129	.053	.164	.047	.285	.059	.304	.051
50,400	.152	.056	.193	.037	.326	.063	.372	.047
50,500	.167	.043	.246	.049	.355	.047	.446	.054
50,600	.192	.052	.266	.046	.411	.068	.474	.044
100,100	.089	.059	.091	.055	.121	.038	.150	.043
50,200	.093	.040	.145	.070	.229	.048	.289	.062
40,250	.104	.046	.155	.065	.251	.044	.281	.054
20,500	.152	.044	.230	.047	.369	.051	.427	.058
10,1000	.237	.049	.311	.027	.510	.058	.554	.033
100,50	.047	.038	.064	.046	.087	.046	.101	.044
200,50	.058	.051	.065	.051	.114	.056	.112	.058
300,50	.072	.053	.058	.047	.087	.051	.090	.051
400,50	.065	.050	.059	.047	.104	.050	.095	.060
500,50	.053	.042	.067	.056	.095	.050	.090	.043
600,50	.060	.047	.070	.045	.092	.051	.102	.046

[12] Results from 1000 Monte Carlo simulations described in text. The entries in the table represent the rejection rate of the overidentification and the exogeneity test statistics with nominal size 5%.

Table 2b [13]

$\rho = 0.001$	**No grouped instruments**				**One grouped instrument**			
	Overid.		Exog.		Overid.		Exog.	
M, N	Unadj.	Adj.	Unadj.	Adj.	Unadj.	Adj.	Unadj.	Adj.
50,100	.066	.059	.062	.058	.064	.060	.074	.056
50,200	.052	.050	.050	.048	.070	.055	.070	.044
50,300	.073	.064	.063	.050	.082	.057	.084	.059
50,400	.057	.050	.057	.043	.091	.053	.082	.055
50,500	.064	.051	.068	.054	.081	.042	.087	.046
50,600	.058	.038	.052	.042	.104	.055	.118	.059
100,100	.055	.051	.056	.053	.062	.049	.051	.048
50,200	.064	.057	.073	.062	.092	.054	.071	.045
40,250	.067	.053	.066	.052	.071	.052	.075	.049
20,500	.062	.043	.070	.047	.081	.048	.089	.042
10,1000	.072	.054	.081	.042	.119	.042	.146	.026
100,50	.049	.054	.048	.052	.050	.050	.055	.047
200,50	.055	.052	.050	.048	.052	.048	.058	.057
300,50	.041	.040	.047	.049	.050	.048	.050	.049
400,50	.059	.058	.050	.049	.039	.037	.053	.048
500,50	.067	.071	.055	.058	.052	.047	.055	.049
600,50	.045	.043	.057	.053	.056	.056	.056	.044

[13] Results from 1000 Monte Carlo simulations described in text. The entries in the table represent the rejection rate of the overidentification and the exogeneity test statistics with nominal size 5%.

Table 3 [14]

	Age 20-24				Age 25-30			
Dep. Var:	High School Graduate	College Graduate	Idle	Unmarried Mother	High School Graduate	College Graduate	Idle	Unmarried Mother
OLS								
segregation	.0158	.067	-.006	.008	.021	-.014	.000	-.023
	(.033)	(.040)	(.019)	(.030)	(.025)	(.067)	(.025)	(.024)
segregation*black	-.323	-.081	.324	.355	-.257	-.048	.277	.047
	(.043)	(.035)	(.044)	(.063)	(.046)	(.052)	(.039)	(.059)
2SLS								
segregation	.129	.212	-.046	-.051	.077	.095	.005	.108
	(.044)	(.053)	(.025)	(.038)	(.032)	(.077)	(.028)	(.035)
segregation*black	-.405	-.202	.317	.327	-.231	-.121	.295	.582
	(.082)	(.056)	(.087)	(.101)	(.076)	(.068)	(.062)	(.116)
N	97,976	97,976	97,976	49,038	139,715	139,715	139,715	71,531
Overid. Test								
Unadjusted	**16.67**	**33.49**	**3.84**	**9.18**	**28.10**	**133.02**	**8.65**	**6.88**
Adjusted	**5.53**	**5.57**	**0.91**	**5.91**	**5.12**	**15.84**	**3.03**	**3.36**
Exog. Test								
Unadjusted	84.38	171.54	19.25	21.56	62.62	196.76	9.50	35.66
Adjusted	17.59	16.18	6.54	8.97	11.35	20.76	3.16	10.95

[14] All regression include also race, age and sex dummies, a set of metropolitan area controls (log population, percentage black, log median income and percentage employed in manufacturing), and trhe metropolitan area controls interacted with the black dummy variable.

Table 4 [15]

	8th Grade Standardized Score in:				12th Grade Standardized Score in:			
	math	reading	history	science	math	reading	history	science
OLS								
Index of public school	2.146	1.238	3.597	1.349	-2.481	2.541	-2.104	-2.965
competition	(1.523)	(1.391)	(1.727)	(1.490)	(3.116)	(3.088)	(3.039)	(3.078)
2SLS								
Index of public school	6.814	7.541	10.278	6.329	9.327	10.086	11.752	9.869
competition	(3.478)	(3.633)	(4.209)	(3.513)	(5.387)	(5.363)	(5.658)	(5.553)
Overid. Test:								
unadjusted	0.7914	0.4685	0.2981	0.3941	0.5269	0.4592	0.3819	0.4320
adjusted	0.1834	0.1819	0.1987	0.1479	0.1772	0.1728	0.2611	0.1622
Exog. Test								
unadjusted	2.1449	1.6766	1.2199	2.0745	2.1376	1.7904	1.3407	2.1902
adjusted	0.7183	0.6194	0.5987	0.5479	0.6747	0.6062	0.4887	0.6264

[15] For the full list of regressors, cfr. Hoxby (1998), Appendix Table 5. The index of public school competition for a metropolitan area is a Herfindahl index based on school districts' shares of metropolitan area enrollment.