The Loss in Efficiency from Using Grouped Data to Estimate Coefficients of Group Level Variables

Kathleen M. Lang*
Boston College

and

Peter Gottschalk
Boston College

Abstract

We derive the efficiency loss from using grouped data to estimate coefficients of variables that vary across groups but not individuals within a group (e.g., state unemployment rates) when micro data are unavailable on the dependent variable. We present an empirical example of our theoretical results, and show that the efficiency loss in this application is small. JEL Classification Number C10.

Key Words: Grouped Data, Relative Efficiency

* Corresponding author: Department of Economics, Boston College, Chestnut Hill, MA 02167.

We are grateful for helpful comments from David Belsley and Bruce Hansen on an earlier draft.

I. Introduction

Researchers often have access only to grouped data or imperfect micro data on variables used as dependent variables in regression. For example, non-wage compensation data are difficult to obtain at the individual level, but some state and industry averages are available. The researcher often must choose between using the grouped data or gathering micro data at considerable additional expense without prior knowledge of the gain in efficiency from gathering micro data. In this paper, we consider the special case where the gain in efficiency can be estimated directly with the data at hand. Specifically, we derive the efficiency loss from using grouped data to estimate the coefficient of a variable that varies across groups but not individuals within a group (e.g., the state unemployment rate). We show that the added precision in the estimation of parameters of micro level variates carries over to the precision of estimates of parameters of group specific variables only if there is correlation between the micro and group level explanatory variables. This might occur, for example, if state unemployment rates are correlated with mean education or other demographic variables in a state. In an empirical application, we show that the loss in efficiency from using state level data to estimate the impact of state-mandated workers' compensation insurance rates on employee compensation is small.

The key distinction between our work and previous work on grouped data estimation is that we focus on estimates of coefficients of variables that vary across groups but not within groups, such as state tax rates. Standard treatments of OLS on grouped data have addressed the issues of the loss in efficiency from estimating parameters of variables that vary across individuals within a group and the heteroskedasticity introduced by moving to a grouped model with different group sizes.[1] Focusing on coefficients of variables that vary only across groups, we show that the potential benefit from gathering, organizing, and utilizing micro data on the dependent variable can be computed ex ante, and that it is not likely to be large if one is interested in the coefficient of a variable that varies only across groups.

II. Relative Efficiency of Individual Versus Group Level Regression Estimates

Suppose a researcher is interested in estimating the following model

$$(1) \quad Y_{ig} = \boldsymbol{X}_{ig}\boldsymbol{\beta} + Z_g\Gamma + \varepsilon_{ig}$$

where

$E(\varepsilon_{ig}\varepsilon_{jk}) = \sigma_\varepsilon^2$ for i = j and g = k

= 0 otherwise.

[1] See Theil (1971). In estimating models with group specific unobservables, Moulton (1987) shows that standard errors can be seriously underestimated if the nonspherical nature of the errors is ignored.

In this model, *i* denotes an individual observation (e.g., person or firm) and *g* denotes a group (e.g., geographical location, cohort, or time period). Z_g is the group level variable of interest, and $\boldsymbol{X}_{ig}$ is a (K-1) vector of individual and, possibly, other group level covariates. There are G groups with N observations per group.[2] The problem faced by the researcher is that micro data are available on $\boldsymbol{X}$, but not *Y*. For example, individuals' wages are often available, but their non-wage compensation is not in the micro data set. To gauge the efficiency loss from using the available grouped data, we compare the variances of two estimators of Γ: $\hat{\Gamma}_m$, which would result if the researcher could estimate the model with micro data, and $\hat{\Gamma}_{gr}$, which can be estimated with available grouped data.

Micro Data Estimator

Stacking the data by group gives

$$(2) \quad \boldsymbol{Y} = \boldsymbol{X}\beta + \boldsymbol{DZ}\Gamma + \varepsilon$$

where $\boldsymbol{X}$ is an NG x (K-1) matrix containing K-1 covariates that vary across individuals and groups; $\boldsymbol{D}$ is an NG x G matrix of group dummy variables, and $\boldsymbol{Z}$ is the G vector comprising the variable of interest that varies only across groups and not across individuals within a group. Let $[\hat{\beta}_m \hat{\Gamma}_m]'$ be the OLS estimates of β and Γ that would result from estimating (2). The corresponding variance-covariance matrix of $[\hat{\beta}_m \hat{\Gamma}_m]'$ is given

$$(3) \quad V(\hat{\beta}_m \hat{\Gamma}_m)' = \sigma_\varepsilon^2 \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{DZ} \\ \boldsymbol{Z}'\boldsymbol{D}'\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{D}'\boldsymbol{DZ} \end{bmatrix}^{-1} = \sigma_\varepsilon^2 \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{DZ} \\ \boldsymbol{Z}'\boldsymbol{D}'\boldsymbol{X} & N\boldsymbol{Z}'\boldsymbol{Z} \end{bmatrix}^{-1},$$

noting that $\boldsymbol{D}'\boldsymbol{D} = N\boldsymbol{I}$. Using standard results for partitioned matrices, the variance of $\hat{\Gamma}_m$ is

$$(4) \quad V(\hat{\Gamma}_m) = \frac{\sigma_\varepsilon^2}{[\boldsymbol{Z}'\boldsymbol{D}'\boldsymbol{M}\boldsymbol{D}\boldsymbol{Z}]}$$

$$= \frac{\sigma_\varepsilon^2}{(1-R^2{}_m)[\boldsymbol{Z}'\boldsymbol{D}'\boldsymbol{DZ}]} = \frac{\sigma_\varepsilon^2}{(1-R^2{}_m)N[\boldsymbol{Z}'\boldsymbol{Z}]}$$

where $\boldsymbol{M} = [\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']'$. $\boldsymbol{M}$ is an NGxNG symmetric, idempotent matrix that transforms the NG vector $\boldsymbol{DZ}$ into a vector of residuals from an auxiliary regression of $\boldsymbol{DZ}$ on the micro data matrix, $\boldsymbol{X}$.[3] The R^2_m in the final expression is the uncentered R-square from the auxiliary micro

[2] Expanding to an unbalanced panel would add complexity without adding further insight.

[3] There is no loss of generality in assuming one group level variate. With multiple $\boldsymbol{Z}$s, relative efficiency is determined separately for the coefficient of each $\boldsymbol{Z}$; the remaining $\boldsymbol{Z}$s are included in the $\boldsymbol{X}$ matrix, and the R^2_m represents the uncentered R-square from the auxiliary regression of the group specific variable under consideration on the $\boldsymbol{X}$s and the remaining $\boldsymbol{Z}$s. Note that the uncentered R-square is defined as the ratio of the sum of the squares of the predicted *Z*s to the sum of the squares of the actual *Z*s. This is distinct from the usual R-square, which is a ratio of predicted to actual sums of squared deviations from mean *Z*.

regression of ***DZ***, which consists of the group specific variable ***Z*** with repeated values for individuals from a given group, on ***X***.

Grouped Data Estimator

Let $\hat{\Gamma}_{gr}$ be the estimator from the grouped version of the model that the researcher is able to estimate with available data

$$(5) \quad \overline{\boldsymbol{Y}} = \overline{\boldsymbol{X}}\beta + \boldsymbol{Z}\Gamma + \overline{\varepsilon}$$

where

$\bar{\varepsilon}_g = \frac{1}{N}\sum_{i=1}^{N}\varepsilon_{ig}$

and $E(\bar{\varepsilon}_g\bar{\varepsilon}_k) = \sigma_{\bar{\varepsilon}}^2$ for g = k

= 0 otherwise.

$\overline{\boldsymbol{X}}$ is the G x (K-1) matrix of group means where means are taken over individuals within a group.

Using a procedure analogous to the preceding, the variance of $\hat{\Gamma}_{gr}$ is seen to be

$$(6) \quad V(\hat{\Gamma}_{gr}) = \frac{\left[\frac{\sigma_\varepsilon^2}{N}\right]}{(1-R^2_{gr})[\boldsymbol{Z}'\boldsymbol{Z}]}$$

where R^2_{gr} is the uncentered R-square from the group level auxiliary regression of ***Z*** on the grouped data, $\overline{\boldsymbol{X}}$.

Relative Efficiency

The relative efficiency of $\hat{\Gamma}_{gr}$ is given by

$$(7) \quad RE \equiv \frac{V(\hat{\Gamma}_m)}{V(\hat{\Gamma}_{gr})} = \frac{(1-R^2_{gr})[\boldsymbol{Z}'\boldsymbol{Z}]N}{(1-R^2_m)[\boldsymbol{Z}'\boldsymbol{Z}]N} = \frac{(1-R^2_{gr})}{(1-R^2_m)} \leq 1$$

This simple expression indicates that the loss in efficiency from estimating coefficients of group level variables with grouped data depends only on the two uncentered R-squares. Both can be computed directly with available data since they do not involve ***Y***, enabling one to evaluate ex ante the potential benefit from gathering micro data.

The loss in efficiency hinges on the correlations between the ***X***s and ***Z***. To the extent that these variables are orthogonal, there is no loss in efficiency from using grouped data. When the ***X***s and *Z* are not orthogonal, the difference between R^2_{gr} and R^2_m, and hence, relative efficiency, depends on whether the within-group variation in the ***X***s is large relative to the between-group variation. If the *X* values within each group are similar, the two R-squares will be similar, and the

loss in efficiency will be modest. If, on the other hand, replacing micro X values with group means removes a sizable portion of the variation in the Xs, the grouped regression will fit considerably better than the micro regression, and the loss in efficiency will be larger.

As long as R^2_m and R^2_{gr} are small or are of roughly equal size, there is little loss from using grouped data. The gain in efficiency may not warrant trade-offs with time and monetary costs involved in gathering micro data. Stated alternatively, studies with a large number of persons per group may not add much to the precision of estimated parameters of group specific variables unless R^2_m is very different from R^2_{gr}.[4]

III. Empirical Example

We now apply the preceding expression for relative efficiency in an empirical application in which the researcher has access to grouped data on the dependent variable $\boldsymbol{Y}$, and micro data on $\boldsymbol{X}$. Both R-squares can be computed directly and relative efficiency estimated using (7).

The state workers' compensation programs require that employers contribute a fixed percent of payroll for insurance for workers' compensation claims. The required rate of payroll, which varies by state, occupation, and year, is based on the expected value of the cost of claims.[5] If the incidence of workers' compensation insurance is shared between employer and employee, compensation should be lower for employees in jurisdictions with more generous workers' compensation rates.

In examining the impact of the state workers' compensation programs, researchers have used hourly wage as a proxy for hourly compensation because of a lack of micro level non-wage compensation data.[6] Non-wage compensation comprises about 30 percent of the compensation package. Instead of ignoring this component or gathering micro data, a researcher could use state level data on average non-wage compensation in a particular occupation to estimate a state level model.[7]

Consider a standard, log linear compensation model for a person in a specific occupation

$$(8) \qquad Y_{ig} = \beta_o + \beta_1 Z_g + \beta_2 \boldsymbol{X}_{ig} + \varepsilon_{ig},$$

where Y_{ig} refers to the natural log of hourly compensation of person i in state g, Z_g is the state level workers' compensation insurance rate, $\boldsymbol{X}_{ig}$ is a vector of individual characteristics, and ε_{ig} is an i.i.d. stochastic error. A value for β_1 equal to -1 would indicate that a one percentage point increase in the workers' compensation rate is met by a one-percent reduction in employee compensation. The loss

---

[4] This is consistent with Amemiya's (1978) result that efficient estimation of coefficients of group level variates in a model with random group effects is accomplished through a two step procedure where the second step is a regression using only grouped data to estimate coefficients on the group specific variates.

[5] See Burton (1980) for a full description of the program.

[6] See, for example, Gruber and Krueger (1991).

[7] Ignoring non-wage compensation may also introduce measurement error bias if it is correlated with model regressors.

in efficiency from estimating a state level version of (8) is readily estimated since the auxiliary regressions of $\boldsymbol{DZ}$ on $\boldsymbol{X}$ and $\boldsymbol{Z}$ on $\overline{\boldsymbol{X}}$ can be computed with existing data.

The 1980 and 1990 Decennial Censuses provide micro data on the $\boldsymbol{X}$ variates for 18 to 64 year-old workers in a variety of occupations. The $\boldsymbol{X}$'s include years of education, potential experience, the square of potential experience, and indicators for male, black, other non-white, part-time, living in metropolitan area, married, and married male. Workers' compensation insurance rate data are from the National Council on Compensation Insurance (NCCI), the rating board that establishes rates for over 600 detailed occupations in 36 states, and from independent state councils for 7 states with rate-making procedures and occupational classifications similar to those of NCCI.[8] The state level models use 86 observations (43 states, 2 time periods).

The results are shown in Table 1. Columns (1) and (2) show the auxiliary regression uncentered R-squares. Equation (7) is used to calculate relative efficiency, shown in column (3), and the implied increase in the standard error of the workers' compensation insurance rate coefficient, shown in column (4). Column (5) provides the sample sizes for the micro data models. The results indicate a range of relative efficiency from a low of .40 for secretaries to .90 for street pavers, implying increases in the standard errors of the coefficients for the workers compensation rate variable in state level models ranging from 6 to 57 percent, with a median of 14 percent.[9]

Gains in efficiency will not alter test results if the variance of the grouped data estimator is already small enough to reject the null hypothesis. Therefore, the cost effectiveness of obtaining micro data will differ across studies. For example, Table 1 is based on a larger project which attempts to test whether increases in workers' compensation costs are fully passed on to workers.[10] The increases in standard errors from using grouped data affect the significance of the workers' compensation rate in this study in only two of the nineteen occupations examined. From this, we conclude that the gains from obtaining micro data on employee compensation would be small and likely would not warrant the cost in this study.

---

[8] Rate data were provided by Jon Gruber at MIT and Mike Curran at Liberty Mutual Insurance Company.

[9] The increase in the standard error on the workers' compensation rate coefficient is equal to $\mathbf{1} / \sqrt{\mathbf{RE}}$.

[10] Lang (1995).

REFERENCES

Amemiya, T., "A Note on the Random Coefficients Model," *International Economic Review,* 1978, 19:3, 793-796.

Burton, John F. Jr., 1980, "Workers' Compensation Costs for Employers" in Research Report of the Interdepartmental Workers' Compensation Task Force, Volume 3, U.S. Department of Labor, Washington D.C.

Census of Population and Housing, 1980: Public Use Microdata Samples [machine-readable data files] / prepared by the Bureau of the Census. Washington: The Bureau, 1982.

Census of Population and Housing, 1990: Public Use Microdata Samples [machine-readable data files] / prepared by the Bureau of the Census. Washington: The Bureau, 1992.

Gruber, J. and Alan Krueger, "The Incidence of Mandated Employer-Provided Insurance: Lessons from Workers' Compensation Insurance," in David Bradford, ed., Tax Policy and the Economy, National Bureau of Economic Research, Volume 5, Cambridge, MA: MIT Press, 1991.

Lang, Kathleen M. (1995). "The Incidence of an Employer Mandate: A Theoretical and Empirical Analysis." Ph.D. Dissertation. Boston College.

Moulton, B.R., "Random Group Effects and the Precision of Regression Estimates", *Journal of Business and Economic Statistics* 1987, 385-397.

Theil, H. (1971) Principles of Econometrics. New York: John Wiley & Sons.

Table 1: Relative Efficiency of Micro versus Grouped Data in the Estimation of the Workers' Compensation Coefficient In a Compensation Equation

Occupation	Uncentered R-square: Group Auxiliary Regression (1)	Uncentered R-square: Micro Auxiliary Regression (2)	Relative Efficiency RE (3)	Increase in Standard Error of Coefficient (4)
Truck Driver	0.9756	0.9725	0.89	1.06
Plumber	0.9839	0.9792	0.77	1.14
Carpenter	0.9822	0.9799	0.89	1.06
Electrician	0.9794	0.9733	0.77	1.14
Gas Station	0.9761	0.9715	0.84	1.09
Street Paving	0.9804	0.9781	0.90	1.06
Hospital Worker	0.9774	0.9706	0.77	1.14
Convalescent Home	0.9869	0.9785	0.61	1.28
Nurse	0.9716	0.9403	0.48	1.45
Clerical	0.9784	0.9723	0.78	1.13
Secretary	0.9884	0.9713	0.40	1.57
Department Store	0.9715	0.9588	0.69	1.20
Convenience Store	0.9814	0.9651	0.53	1.37
Apparel	0.9828	0.9788	0.81	1.11
Grocery Clerk	0.9858	0.9663	0.42	1.54
Hotel Bag/Bellhop	0.9768	0.9712	0.81	1.11
Furniture Mover	0.9710	0.9648	0.82	1.10
Office Machine Repair	0.9703	0.9643	0.83	1.10
Restaurant Worker	0.9887	0.9777	0.51	1.40