WORKING PAPER SERIES

Aggregated vs. Disaggregated Data in Regression Analysis: Implications for Inference

Thomas A. Garrett

Working Paper 2002-024B
http://research.stlouisfed.org/wp/2002/2002-024.pdf

November 2002

FEDERAL RESERVE BANK OF ST. LOUIS
Research Division
411 Locust Street
St. Louis, MO 63102

Aggregated vs. Disaggregated Data in Regression Analysis: Implications for Inference

Thomas A. Garrett
Federal Reserve Bank of St. Louis
Research Division
411 Locust Street
St. Louis, Missouri 63102

phone: 314-444-8601
email: tom.a.garrett@stls.frb.org

Abstract

This note demonstrates why regression coefficients and their statistical significance differ across degrees of data aggregation. Given the frequent use of aggregated data to explain individual behavior, data aggregation can result in misleading conclusions regarding the economic behavior of individuals.

Aggregated vs. Disaggregated Data in Regression Analysis: Implications for Inference

I. Introduction

Every field of economics uses aggregated data to test hypotheses about the behavior of individuals. Examples in macroeconomics include the use of aggregate consumption and income to test the permanent income hypothesis (Hall, 1978), and forecasting national personal consumption expenditures using consumer sentiment indices (Carroll, et al. 1994; Bram and Ludvigson, 1998). The use of aggregated data to explain individual behavior makes the assumption that the hypothesized relationship between the economic variables in question is homogenous across all individuals. When the behavior of economic agents is not the same, a regression analysis using aggregated data can provide conclusions regarding economic relationships that are different than if less aggregated data were used. Correcting for this 'aggregation bias' has received careful attention in the literature.[1]

This note develops a simple framework to show how coefficient estimates and their statistical significance can differ using aggregated versus less aggregated data. The analysis explores the effects of dependent variable data aggregation on coefficient estimates and standard errors. A classic example of the framework presented here is the empirical work of Carroll, et al. (1994) and Bram and Ludvigson (1998) that explores the ability of consumer sentiment (a measure of an individual's perception of economic conditions) to forecast national personal consumption expenditures (a highly aggregated measure of consumption). The analysis presented here is useful to both economics graduate students and applied economists as it provides general insights into the impact of data aggregation on regression estimates and conclusions made from statistical inference.

II. Analysis

Statistical inference on regression coefficients is traditionally done using a *t*-statistic. The *t*-statistic value for testing H_0: $\beta=0$ is the estimated coefficient, $\hat{\beta}_k$, divided by its standard error. Thus, for a chosen critical value, the significance of any coefficient depends upon the size of the coefficient and its variance. The difference in coefficient size and variance (and thus statistical significance) from regressions using various levels of data aggregation in the dependent variable is the focus of this analysis.

Consider *m* regression equations each having T observations and data matrix **X** that is assumed identical across equations:

$$\begin{aligned} \boldsymbol{Y_1} &= \boldsymbol{X\beta_1} + \boldsymbol{U_1} \\ \boldsymbol{Y_2} &= \boldsymbol{X\beta_2} + \boldsymbol{U_2} \\ &\cdots\cdots \\ &\cdots\cdots \\ \boldsymbol{Y_m} &= \boldsymbol{X\beta_m} + \boldsymbol{U_m} \end{aligned} \qquad \textbf{(1)}$$

where $\mathbf{Y}_m$ is a T×1 vector, **X** is T×K with K-1 explanatory variables (assume a constant term), $\boldsymbol{\beta}_m$ is the K×1 vector of estimated coefficients, and $\mathbf{U}_m$ is the T×1 residual vector. As an example of the above framework that follows the consumer sentiment literature, each $\mathbf{Y}_m$ could be personal consumption expenditures for each state (m = 50) and **X** consists of lagged consumer sentiment values.

Summing equations provides the aggregated regression equation

$$\boldsymbol{Y_M} = \boldsymbol{X\beta_M} + \boldsymbol{U_M} \qquad \textbf{(2)}$$

where the aggregated regression coefficient vector $\boldsymbol{\beta}_M$ is[2]

$$\boldsymbol{\beta}_M = \sum_{m=1}^{M} \boldsymbol{\beta}_m \qquad \textbf{(3)}$$

It then follows that

$$\boldsymbol{Y}_M = \sum_{m=1}^{M} \boldsymbol{Y}_m ; \quad \boldsymbol{U}_M = \sum_{m=1}^{M} \boldsymbol{U}_m \qquad \textbf{(4)}$$

It is clear from (3) that any estimated coefficient from the aggregated regression is simply equal to the sum of the corresponding coefficients from the less aggregated regressions. Depending on the signs and magnitudes of each $\boldsymbol{\beta}_m$, the estimated impact of changes in each explanatory variable from the aggregated and less aggregated regressions can be quite different. For example, if all $\boldsymbol{\beta}_m$ are positive, then $\boldsymbol{\beta}_M$ will be larger than any $\boldsymbol{\beta}_m$. The estimated impact of changes in explanatory variable on the dependent variable is thus much greater in the aggregated regression than in the less aggregated regression.

Now consider the variance of a coefficient vector. The variance of a coefficient vector from any OLS regression is [see Greene (1990), page 184]

$$var[\boldsymbol{\beta}_m] = \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} \qquad \textbf{(5)}$$

where σ^2 is the sum of squared residuals (RSS). Because $\mathbf{X}$ is the same for each regression equation, the variance of a specific coefficient in the aggregated regression relative to the

variance of each corresponding coefficient in the less aggregated regressions thus depends solely on differences in the RSS between the less aggregated regressions and the aggregated regression.

Using the expression for $\mathbf{U}_M$ in (4), the aggregated regression RSS is

$$\boldsymbol{U}_{\boldsymbol{M}}^{\boldsymbol{2}} = (\sum_{m} \boldsymbol{U}_{\boldsymbol{m}})^2 \qquad \textbf{(6)}$$

The questioned relationship between the RSS from the aggregated regression and the summed RSS from the less aggregated regressions ($\sum_{m} \boldsymbol{U}_{\boldsymbol{m}}^{\boldsymbol{2}}$) can be written as:

$$(\sum_{m} \boldsymbol{U}_{\boldsymbol{m}})^2 \geq \sum_{m} \boldsymbol{U}_{\boldsymbol{m}}^{\boldsymbol{2}} \qquad \textbf{(7)}$$

or

$$(\sum_{m} \boldsymbol{U}_{\boldsymbol{m}}) \acute{} (\sum_{n} \boldsymbol{U}_{\boldsymbol{n}}) \geq \sum_{m} \boldsymbol{U}_{\boldsymbol{m}} \acute{} \boldsymbol{U}_{\boldsymbol{m}} \qquad \textbf{(8)}$$

Giving the final result

$$\sum_{m} \boldsymbol{U}_{\boldsymbol{m}} \acute{} \boldsymbol{U}_{\boldsymbol{m}} + 2 \cdot \sum_{m=1}^{N-1} \sum_{n=m+1}^{N} \boldsymbol{U}_{\boldsymbol{m}} \acute{} \boldsymbol{U}_{\boldsymbol{n}} \geq \sum_{m} \boldsymbol{U}_{\boldsymbol{m}} \acute{} \boldsymbol{U}_{\boldsymbol{m}} \qquad \textbf{(9)}$$

Let $Z = 2 \cdot \sum_{m=1}^{N-1} \sum_{n=m+1}^{N} \boldsymbol{U}_{\boldsymbol{m}} \acute{} \boldsymbol{U}_{\boldsymbol{n}}$. Z is the sum of the product of cross equation residuals. If $Z < 0$, then the RSS from the aggregated regression will be less than the sum of RSS from the less aggregated regressions. The sign of Z depends upon the covariance (and thus correlation) between the cross-equation residuals.[3] If $Z < 0$, then the cross-equation residuals have a negative covariance and are negatively correlated -- the residuals from the individual regression

equations have opposing signs (a result of differing signs on corresponding slope coefficients), and summing positive and negative residuals reduces the absolute magnitude of the aggregated regression residuals. Thus, the spread of the residuals in the aggregate regression is less than the sum of the residual spreads from the less aggregated regressions. Conversely, a $Z > 0$ implies that the summed residuals from the less aggregated regressions result in aggregated residuals that are more than the sum of the individual residual spreads -- residuals from the individual regression equations have the same signs (a result of same signs on corresponding slope coefficients), and summing same sign residuals increases the absolute magnitude of the aggregated regression residuals.

In closing, the analysis finds the following connection between data aggregation and statistical inference: the size of the RSS from the aggregated regression relative to the sum of the RSS from less aggregated regressions depends on the correlation of residuals across the less aggregated regressions. The difference in the RSS then translates into differences in coefficient standard errors as seen in (5). This, combined with the additive relationship between the less aggregated and aggregated coefficients shown in (3), gives different conclusions from statistical inference across levels of data aggregation.

III. Summary

This analysis has shown why the sign and significance of coefficient estimates from regressions using aggregated data can differ from regressions using less aggregated data. The size of a coefficient estimate from aggregated data is shown to be the sum of each coefficient from the less aggregated regressions. More importantly, it is shown that the RSS from the

aggregated regression can be larger or smaller than the sum of RSS from less aggregated regressions depending upon the covariance and correlation between cross-equation residuals. Because, given a chosen critical value, statistical significance of a coefficient is a positively related to coefficient size and negatively related to its standard error, it is likely that coefficients from an aggregated regression are statistically significant whereas identical coefficients from less aggregated regressions are statistically insignificant, and vice versa. This results in different conclusions regarding economic behavior depending upon the level of data aggregation.

Endnotes

1. See Goodfriend (1992), Thomas and Tauer (1994), Mittlehammer, et. al (1996), Davis (1997), and Cherry and List (2002).

2. This is derived from the fact that

$$\begin{aligned} \boldsymbol{\beta}_m &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}_m \\ \sum_m \boldsymbol{\beta}_m &= \sum_m (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}_m \\ \sum_m \boldsymbol{\beta}_m &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sum_m \boldsymbol{Y}_m \end{aligned}$$

3. This is seen by taking the expectation of (9). $\text{Cov}(\mathbf{U}_m,\mathbf{U}_n) = \rho_{U_m,U_n} \cdot \sigma_{U_m} \cdot \sigma_{U_n}$ where ρ is the correlation between residual vectors and σ is the standard error of each residual vector.

References

Bram, Jason and Sydney Ludvigson, "Does Consumer Confidence Forecast Household Expenditures? A Sentiment Index Horse Race." *Federal Reserve Bank of New York Economic Policy Review*, vol. 4 (June 1998): 59-78.

Carroll, Christopher, Jeffrey Fuhrer, and David Wilcox, "Does Consumer Sentiment Forecast Household Spending? If So, Why?" *American Economic Review*, vol. 84, no. 5 (December 1994): 1397-1408.

Cherry, Todd and John List, "Aggregation Bias in the Economic Model of Crime." *Economics Letters*, vol. 75, no. 1 (March 2002): 81-86.

Davis, George, "Product Aggregation Bias as a Specification Error in Demand Systems," *American Journal of Agricultural Economics*, vol. 79, no. 1 (February 1997): 100-109.

Goodfriend, Marvin, "Information-Aggregation Bias." *American Economic Review*, vol. 82, no. 3 (June 1992): 508-519.

Greene, William H. Econometric Analysis. MacMillan Publishing, New York, 1990.

Hall, Robert E. "Intertemporal Substitution and Consumption." *Journal of Political Economy*, vol. 96 (April 1998): 339-357.

Mittlehammer, Ron, Hongqi Shi and Thomas Wahl, "Accounting for Aggregation Bias in Almost Ideal Demand Systems." *Journal of Agricultural and Resource Economics*, vol. 21, no. 2 (December 1996): 247-262.

Thomas, Arthur and Loren Tauer, "Linear Input Aggregation Bias in Nonparametric Technical Efficiency Measurement." *Canadian Journal of Agricultural Economics*, vol. 42, no. 1 (March 1994): 77-76.