

Testing For Aggregation Bias in a Non-Linear Framework: Some Monte Carlo Results

Jac C. Heckelman

Department of Economics, Wake Forest University,
122 Carswell Hall, Winston-Salem, NC (USA);
Email : heckeljc@wfu.edu

ABSTRACT

Researchers modeling the behavior of individual people or firms are often unable to utilize micro-level data because such data are unavailable or unreliable. Faced with this dilemma, researchers often resort to using aggregate-level data. When the individual-level variable of interest is dichotomous, however, the aggregate-level model is subject to a special form of aggregation bias. Kelejian [1994] provides a methodology for testing for the presence of this form of bias. This study uses Monte Carlo analysis to evaluate the usefulness of Kelejian's test. Using 50 units, populated by, 100, 1000 and 10,000 individuals, we find that aggregation bias is almost universally present. Unfortunately, the Kelejian test identifies the bias in fewer than half of the cases.

Keywords: Aggregation bias, ecological regression, aggregate data, logistic transformation.

Mathematics Subject Classification: 62H10, 62P25

Journal of Economic Literature Classification: C35, C52

1. INTRODUCTION

Often, in social science and business applications, researchers empirically modeling the behavior of individual people or firms are unable to use individual, micro-level, data because such data are unavailable or unreliable. For example, individuals may be reluctant to disclose personal voting behavior or medical histories. When they do answer such questions, the data may be considered highly unreliable. Similarly, individual firms may be reluctant to supply data on hiring or production practices. When compelled by government agencies, the data are typically unavailable to non-government researchers due to their proprietary nature. Finally, historians of all types may face this problem when studying time periods prior to the development of large modern surveys.

Researchers facing this situation often resort to using aggregated data. Thus, for example, researchers may use state-level voting data to try to explain individual voting behavior, district-level graduation rates may be used to explain the behavior of individual students, or industry-level hiring data may be used to explain decisions made by individual firms.

In each case, the underlying theoretical probabilistic model can be expressed as:

$$y_{it} = f(x_i, z_{it}) + \varepsilon_{it},$$

where y_{it} is the endogenous outcome of member t ($t = 1, \dots, T_i$) of unit i ($i = 1, \dots, N$), (e.g., resident t of state i , or firm t in industry i), x_i is a vector of characteristics common to all members of the unit, z_{it} is a vector of characteristics that vary within the unit from member to member, and ε_{it} is a stochastic error term. For example, y_{it} could be an individual's income, x_i could be a vector of dummy variables indicating the presence of certain state laws, and z_{it} could be a vector of individual characteristics, such as the person's age, education level and union status. As another example, y_{it} might be a firm's profits, with x_i being a vector of variables describing the firm's industry, and z_{it} a vector of firm characteristics, such as the number of employees and number of patents received.

When the micro-level data, y_{it} and z_{it} , are not available but aggregate-level data are available, researchers often choose to estimate the parameters of an aggregate-level model:

$$\bar{y}_i = g(x_i, \bar{z}_i) + v_i,$$

where \bar{y}_i and \bar{z}_i represent the aggregate-level means (e.g., per-capita income, or average firm size in the industry), and v_i is a unit-level error term. The parameter estimates calculated using this aggregate-level model are often taken as estimates of the parameters of the underlying micro-level model. Unfortunately, there are some potential pitfalls to this strategy.

If the micro-level equation, $f()$, is specified as a linear function, as is often the case when the endogenous variable is continuous, ordinary least squares estimates of a linear aggregate model, $g()$, will yield unbiased estimates of the micro-level equation. However, since the aggregate error term, v_i , is, in general, heteroskedastic, researchers must use a generalized least squares estimator to obtain efficient estimates of the model's parameters. Since the form of the heteroskedasticity is well-known, and the weighting is straight-forward, these efficient estimates are easily calculated (see, for example Kmenta [1986], pp. 366-373).

Another consideration when using aggregate data is the possibility that the grouping of individuals will be correlated with y_{it} . Langbein and Lichtman [1978] show that this situation will lead to biased parameter estimates, and suggest procedures for obtaining unbiased estimates.

Another concern, discussed by Stoker [1993], is that there may be important differences in the parameters across individuals. In other words, the true micro-level is

$$y_{it} = f_{it}(x_i, z_{it}) + \varepsilon_{it},$$

giving each individual their own micro-level model.

Although each of these problems is often present when using aggregate data, and the latter two problems are often referred to as "aggregation bias," we are concerned here with a form of aggregation bias that has only recently been addressed. The form we focus on arises when $f()$ is specified as a non-linear function, in particular the case when y_{it} is dichotomous. In this case, \bar{y}_i will represent the percentage of observations in the unit that satisfy the condition of interest. For example, y_{it} may be a dummy variable indicating whether a person voted in an election. The variable \bar{y}_i would then be the state's voter turnout rate. Another example would be where y_{it} is a dummy variable indicating whether a person graduated from high school, and \bar{y}_i is the district's graduation rate.

In this case, using a *linear* specification at the aggregate level and treating it as a summation of a *nonlinear* micro-level model is inappropriate for several reasons. First, the linear functional form allows for predicted values of \bar{y}_i that are negative or exceed one. Second, the aggregate-level error term will be heteroskedastic. Finally these models implicitly assume, incorrectly, that within any unit,

$$E[g(x_i, z_{it})] = g(x_i, E(z_{it})).$$

Researchers have utilized several techniques to account for the first two of these issues. But there is no developed technique for addressing all three. Many researchers address the functional form problem by using a logistic transformation of \bar{y}_i .¹ The dependent variable then is the log of the odds that a member of the unit satisfied the condition of interest. Some researchers take this a step further by combining this transformation with methodologies that correct for heteroskedasticity.² This, however, still does not solve the problem of averaging over a nonlinear function of the explanatory variables.³

¹ This methodology is used, for example, by Geys and Heyndels [2006].

² See, for example, Guadalupe [2003].

³ A special case that can be addressed, relatively easily, occurs when all of the individuals in each unit have identical explanatory characteristics (i.e., $z_{it} = \bar{z}_i$ for all observations in the unit). Here, the problem can be addressed with minimum chi-square logit methods [Madalla, 1983]. In the previous example involving graduation rates, if all of the explanatory variables are determined at the district level (e.g., teacher salaries and course requirements for graduation) the minimum chi-square logit technique, using the aggregate data, would provide the researcher with appropriate estimates of the effects of these variables on an individual's probability of graduating. For an example involving historical bank failure rates, see Chung and Richardson [2006]. If,

Faced with this dilemma, the question turns from correcting the problem to simply determining when it is present. Kelejian [1994] provides a framework for testing for the presence of this type of aggregation bias. The purpose of the current study is to examine the properties of his suggested test. The remainder of this paper is organized as follows: the next section describes the Kelejian test, Section 3 describes the Monte Carlo experiment and presents the results, and Section 4 provides a summary and suggestions for researchers facing this form of bias.

2. THE KELEJIAN TEST

Kelejian considers the case where y_{it} is a dichotomous variable that is determined by the sign of

$$y_{it}^* = h(x_i, z_{it}) + \varepsilon_{it}.$$

If $y_{it}^* \geq 0$, then $y_{it} = 1$, otherwise $y_{it} = 0$. Suppose, for example, that y_{it} indicates whether an individual voted in an election. The vector x_i will include, perhaps, dummy variables indicating the presence of state-level laws regarding residency requirements. The vector z_{it} will include such individual-level descriptors as income and race. If the micro-level data are available, and the distribution of ε_{it} is known, the parameters of $h()$ can be estimated using maximum likelihood techniques. One of the most common specifications, and the case Kelejian addresses, is to treat $h()$ as linear, and to assume that ε_{it} is distributed logistic. Although $h()$ is linear, the equation determining the micro-level variable, y_{it} , is not. This implies that:

$$p_{it} = \text{Prob}(y_{it=1} | x_i, z_{it}) = \frac{e^{a+x_i b+z_{it} c}}{1 + e^{a+x_i b+z_{it} c}}, \quad (1)$$

where a is an intercept, and b and c are parameter vectors to be estimated. If the micro-level data are available, the parameters are estimated using the logit technique (Madalla [1983]). Kelejian addresses the case where the micro-level data are unavailable, but \bar{y}_i and \bar{z}_i (the unit-level means) are available. The framework treats z_{it} as being stochastic, and is based upon the following assumptions summarized below:

- The z_{it} values are independent across units. For example, individual voters' incomes cannot be correlated across states. There may, however, be dependence within a state.

however, some explanatory variables differ from individual-to-individual within the district (e.g., race and household income) minimum chi-square logit techniques are inappropriate. The researcher may have access to

- The distribution function of z_{it} varies (potentially) from state-to-state, but can be described as $F_z(\cdot | x_i, \lambda_i)$. Thus (for example) the distribution of income is allowed to vary from state-to-state as some function of residency requirement and a random coefficient vector, λ_i .
- The mean value of z_{it} in each state may be a function of x_i , but not λ_i . Thus λ_i may affect the dispersion of the z_{it} 's, but not the mean. In symbolic form, $E(z_{it} | x_i, \lambda_i) = E(z_{it} | x_i) = \mu_i$. For example, it is permissible for a state's mean income to be correlated with the residency requirement.
- The dispersion in the state, which involves λ_i , does not involve x_i . Thus, the level of dispersion in income is not dependent upon the state's residency requirements.
- Large samples are available, and $N/T_i \rightarrow 0$. Thus, the number of voters in each state must be large, relative to the number of states.

Note that equation (1) can be transformed so that

$$\ln \left[\frac{p_{it}}{1 - p_{it}} \right] = a + x_i b + z_{it} c. \quad (2)$$

However, the aggregate-level data only contains \bar{y}_i and \bar{z}_i (the unit-level means). Because this equation is non-linear it is inappropriate to simply substitute the aggregate-level mean values for the micro-level values. Kelejian shows that the equation can be transformed to

$$\ln \left[\frac{\bar{y}_i}{1 - \bar{y}_i} \right] = a + x_i b + \bar{z}_i c + g(a + x_i b + \bar{z}_i c) + \psi_i,$$

where ψ_i is an error term that, under the above assumptions, will have a mean of zero. Unfortunately, unless the distribution of z_{it} is known, the function $g()$ is unknown. Kelejian discusses approximating $g()$ with a polynomial of degree J in $x_i b + \bar{z}_i c$, yielding

$$\ln \left[\frac{\bar{y}_i}{1 - \bar{y}_i} \right] = a + x_i b + \bar{z}_i c + \sum_{r=0}^J (x_i b + \bar{z}_i c)^r \alpha_r + \psi_i.$$

An unfortunate drawback of this approach is that the underlying micro parameters, a , b , and c , cannot be identified. The estimation equation then becomes:

district-means for the explanatory variables (e.g., percent non-white, average household income). However, unlike in the linear case, using these aggregate-level measures is inappropriate.

$$\ln \left[\frac{\bar{y}_i}{1 - \bar{y}_i} \right] = (a + \alpha_0) + (x_i b + \bar{z}_i c)(1 + \alpha_1) + \sum_{r=2}^J (x_i b + \bar{z}_i c)^r \alpha_r + \psi_i$$

or

$$\ln \left[\frac{\bar{y}_i}{1 - \bar{y}_i} \right] = \gamma_0 + x_i \gamma_b + \bar{z}_i \gamma_c + \sum_{r=2}^J (x_i \gamma_b + \bar{z}_i \gamma_c)^r \alpha_r + \psi_i.$$

The parameters that can be identified are $\gamma_0 = a + \alpha_0$, $\gamma_b = b(1 + \alpha_1)$, $\gamma_c = c(1 + \alpha_1)$, and α_r ($r = 2, 3, \dots, J$). The absence of aggregation bias is the equivalent to the null hypothesis that $\alpha_2 = \alpha_3 = \dots = \alpha_J = 0$.⁴ This amounts to a relatively easy restriction test, and can be tested using any standard method.⁵ The micro-level coefficients cannot be directly recovered without additional restrictions⁶ although hypothesis tests that the variable(s) belong in the model can be performed. The following section presents results of Monte Carlo experiments that examine the performance of the Kelejian test for aggregation bias.

Chang, Lipsitz, and Waternaux [2000] also note the presence of aggregation bias of the form considered here using simulations on (2) comparing the case of individual versus aggregated data. They consider three aggregate data estimators: simply replacing the individual-level data in (2) with aggregated data means \bar{y}_i and \bar{z}_i ; using Taylor series approximations; and applying Bayes' Theorem to derive a discriminate function estimator. They find all three estimators suffer from aggregation bias, and the bias was worse for the latter two more complicated methods compared to simply directly substituting \bar{y}_i and \bar{z}_i into (2). The tests here, following Kelejian's approach, will focus exclusively on the Kelejian test's ability to detect aggregation bias from the aggregate data substitution in (2).

3. MONTE CARLO RESULTS

To give some guidance regarding the usefulness of the Kelejian test, we begin by considering a hypothetical collection of 50 units, each populated by 100 individuals. Each unit is described by the scalar variable, x_i , and each individual is described by the scalar variable z_{it} .⁷ Values for these

⁴ Because α_0 and α_1 cannot be identified, the test is not able to detect aggregation bias that takes on a purely linear form.

⁵ In the following section we utilize a Wald Test.

⁶ See Heckelman [1997] for details.

⁷ Throughout the discussion, it is worth remembering that our results might depend upon our specification. For example, we select only one unit-level, and one individual-level variable, and make them both continuous. Other models could be tested in the future within this framework. Other decisions, such as the choice of parameters,

variables for the 5,000 members of our population are generated using a psuedo-random process. They are treated as fixed, and do not change throughout the experiment. Table 1 gives descriptive statistics for these variables.

Table 1: Descriptive statistics for explanatory variables ($T_i = 100$)

	x_i	z_{it}	\bar{z}_i
Mean	0.59510	-0.01222	-0.01222
Standard Deviation	0.26290	2.06251	0.64607
Kurtosis	-0.39731	-0.02465	-1.03275
Skewness	-0.34335	0.04165	-0.0033
Minimum	-0.10000	-7.69197	-1.24232
Maximum	0.97848	7.34361	1.31235
Count	50	5000	50

We next specify coefficients for our underlying micro model. Four different sets of parameters are used. These combinations are listed in the first row of Table 2, in the Appendix. Using these parameters, and randomly generated logistic random error terms, ε_{it} , we generate a value for y^* for each member of our population. As described in the previous section, population members with $y_{it}^* \geq 0$ are assigned $y_{it} = 1$, otherwise $y_{it} = 0$. In this way we generate a complete micro data set for our population. This process is repeated 500 times for each combination of parameters.

Table 2: Monte Carlo results for $T_i = 100$

	Parameter Values			
	$a = 1.0$	$a = 1.0$	$a = 1.0$	$a = 1.0$
	$b = 0.5$	$b = 1.0$	$b = 0.5$	$b = 1.0$
	$c = 1.0$	$c = 1.0$	$c = 0.5$	$c = 0.5$
<i>Aggregation Bias Present</i>	100%	100%	95.2%	83.2%
	<i>Aggregation Bias Detected*</i>			
$J = 2$	7.6%	9.6%	7.1%	8.9%
$J = 3$	15.8%	15.6%	11.3%	13.5%
$J = 4$	35.0%	36.4%	34.6%	42.9%
$J = 5$	47.2%	57.0%	51.5%	57.7%

*percent of cases where aggregation bias was detected, conditional on being present

Prior to aggregating our results, we first, as an experiment, estimate a logit model using the micro-level data, and perform a joint (Wald) test of the null hypothesis that the parameters are equal to the

the number of trials run, or even the choice of a random seed for the psuedo-random number generator could, at least in theory, influence the results. All estimations used TSP Version 4.3a.

actual values, using a significance level of 0.05.⁸ As would be expected, with a significance level of 0.05, in roughly five percent of the cases a Type I Error is committed.⁹

We then aggregate the data to the unit level and proceed as if only aggregate data were available. For each of these 500 trials we estimate the parameters of the log odds (or logistic transformation) model discussed in Section 2, substituting \bar{z}_i for z_{it} , and using a robust method that accounts for heteroskedasticity (White [1980b]). We then perform a Wald test of the null hypothesis that the parameters are equal to the actual values. The second row of Table 2 gives the percentage of the time that we reject this hypothesis (indicating the presence of aggregation bias). As seen, we are able to reject the null hypothesis in the vast majority of cases. In fact, for the first two sets of parameters, every experiment resulted in parameter estimates that were statistically different from the true parameters. This indicates that, indeed, the use of unit-level averages is a poor method of estimating micro-level coefficients. These results corroborate the simulations by Chang, Lipsitz, and Waternaux [2000] who did not use robust techniques for their aggregate data estimators.

The remainder of Table 2 summarizes the results when Kelejian's suggested approach is used to test for the presence of aggregation bias, using a polynomial to approximate the form of the bias. Parameters are estimated by Non-linear Least Squares using White's [1980a] covariance adjustment for non-linear regressions subject to heteroskedasticity. Separate results are presented for polynomials of degree two through five. For the second-degree polynomial, the test successfully detected aggregation bias in only 7.1 to 9.6 percent of the cases when it was present.

Scanning the results for higher-level polynomials, we find that increasing the degree of the polynomial increases the probability that bias will be detected. Fifth-degree polynomials, for example, detect bias in about half of the cases when it exists. When the degree of the polynomial was increased beyond five, the detection continued to improve (along with the percentage of false indications). However, estimation of the parameters began to become problematic due to non-convergence. This is likely to be of even greater concern to researchers using more elaborate models that include more variables.

As a further experiment, we increased the population size from 100 individuals per unit to 1,000, and then 10,000 individuals per unit. This largest population size would more closely reflect the population, for example, of a United States county. It also more closely reflects Kelejian's final assumption that $N/T_i \rightarrow 0$. With the larger populations, aggregation bias was present in all trials.¹⁰

Tables 3 and 4 summarize the results from this experiment for $T_i = 1,000$ and $T_i = 10,000$. The lower-order polynomials are much more effective at identifying aggregation bias than with the smaller population sizes. Unfortunately, even with the higher-order polynomials, the Kelejian test was only able to detect the bias in about half of the trials.

⁸ All hypothesis tests discussed in this study are two-tailed, and use a significance level of 0.05.

⁹ Specific results are available upon request.

¹⁰ This is primarily because of the small standard errors that result from the large population. A similar phenomenon has been observed with minimum logit chi-square methods (Greene [1997, pg. 896]).

Table 3: Descriptive statistics for explanatory variables for $T_i = 100$ and $T_i = 1000$

	$T_i=1,000$			$T_i=10,000$		
	x_i ,	z_{it}	\bar{z}_i	x_i ,	z_{it}	\bar{z}_i
Mean	0.59510	0.00655	0.00655	0.59510	0.01640	0.01640
Standard Deviation	0.26290	2.07482	0.57233	0.26290	2.08396	0.58439
Kurtosis	-0.39731	0.00841	-1.14784	-0.39731	-0.00332	-1.22417
Skewness	-0.34335	0.00162	0.00431	-0.34335	-0.00384	-0.03461
Minimum	-0.10000	-8.31795	-0.92974	-0.10000	-10.6119	-0.93233
Maximum	0.97848	8.63594	0.98233	0.97848	9.9626	0.97096
Count	50	5,0000	50	50	500,000	50

Table 4: Monte Carlo Results for $T_i = 100$ and $T_i = 1000$

	Parameter Values			
	$a = 1.0$	$a = 1.0$	$a = 1.0$	$a = 1.0$
	$b = 0.5$	$b = 1.0$	$b = 0.5$	$b = 1.0$
	$c = 1.0$	$c = 1.0$	$c = 0.5$	$c = 0.5$
<i>Aggregation Bias Present</i>	100%	100%	100%	100%
	<i>Aggregation Bias Detected ($T_i=1,000$)</i>			
$J = 2$	13.8%	13.2%	9.8%	12.8%
$J = 3$	18.0%	21.0%	17.0%	20.2%
$J = 4$	26.4%	33.8%	31.0%	34.6%
$J = 5$	39.6%	45.4%	44.2%	47.4%
	<i>Aggregation Bias Detected ($T_i=10,000$)</i>			
$J = 2$	36.6%	33.4%	18.4%	23.2%
$J = 3$	37.2%	41.8%	25.0%	33.8%
$J = 4$	44.6%	48.8%	37.4%	48.0%
$J = 5$	53.4%	57.8%	48.2%	60.2%

4. SUMMARY AND CONCLUSION

Aggregation bias may arise when researchers, wishing to model the behavior of individuals, use aggregated, macro-level, data. This typically occurs when the micro-level data are not available. The appropriate procedure to address the bias depends upon the form of the bias.

This study addresses a particular form of bias that occurs when the individual-level variable being explained is dichotomous. Studies that use aggregate dichotomous data, such as the percent of registered voters in a state that vote, to study individual decisions are likely to suffer from such bias. We use a Monte Carlo framework, generating populations of first 5,000 individuals (then 50,000 and 500,000) distributed across 50 units. After aggregating the data to the unit level, we proceed to estimate the micro parameters using the aggregate data.

Our first result is the large percentage of cases in which the parameters estimated using the aggregate data are statistically significantly different from the actual parameter values. This indicates that the problem of aggregation bias is an important consideration for all researchers using aggregate-level data to explain categorical data. A similar conclusion was reached by Chang, Lipsitz, and Wateraux [2000].

We next examine a procedure, suggested by Kelejian [1994], that uses a semi-parametric approach to model the bias using a polynomial form. Unfortunately, in most cases, Kelejian's suggested test is unable to correctly identify the presence of bias. The likelihood of detecting the difference increases as the degree of the polynomial increases.

These results indicate that researchers should be very cautious when using aggregate dichotomous data, and in interpreting the coefficients as being estimates of the parameters of the individual-level model. The suggested test by Kelejian is fairly easy to implement, and interpret. Unfortunately, because of the low power of the test, its usefulness is suspect.

Acknowledgment

Timothy Sullivan contributed to an earlier version of this paper. Thanks are given to Harry Kelejian for clarifications and helpful comments on the earlier draft.

5. REFERENCES

Chang, H.-H., Lipsitz, S., Wateraux, C., 2000, Logistic regression in meta-analysis using aggregate data. *J. Appl. Stat.* **27**, 411-424.

Chung, C.-Y., Richardson, G., 2006, Deposit insurance altered the composition of bank suspensions during the 1920s: Evidence from the archives of the Board of Governors. *Contrib. to Econ. Anal. Pol.* **5**, Article 34.

Geys, B., Heyndels B., 2006, Disentangling the effects of political fragmentation on voter turnout: the Flemish municipal elections. *Econ. Polit.* **8**, 367-387.

Guadalupe, M., 2003, The hidden costs of fixed term contracts: the impact of work accidents. *Labour Econ.* **10**, 339-357.

Greene, W.H., 1997, *Econometric Analysis*, 3rd Ed., Prentice Hall.

Heckelman, J.C., 1997, Determining who voted in historical elections: an aggregated logit approach. *Soc. Sci. Res.* **26**, 121-134.

Kelejian, H.H., 1994, Aggregated heterogeneous dependent data and the logit model: a suggested approach. *Econ. Lett.* **47**, 243-248.

Kmenta, J., 1986, *Elements of Econometrics*, 2nd Ed. Macmillan Publishing Company.

Langbein, L.I., Lichtman, A.J., 1978, *Ecological Inference*, Sage Publications.

Madalla, G.S., 1983, *Limited-dependent and Qualitative Variables in Econometrics*, Cambridge University Press.

Stoker, T.M., 1993, Empirical approaches to the problem of aggregation over individuals. *J. Econ. Lit.* **31**, 1827-1874.

White, H., 1980a, Nonlinear regression on cross-section data. *Econometrica* **48**, 721-746.

White, H., 1980b, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817-838.