Relationships between Aggregates and Individual Behaviour: The Nature, Direction and Size of Aggregation Bias

John Ermisch

Nuffield College, Oxford

6 July 2024

Abstract

Estimation of relationships between a dependent variable constructed by the aggregation of individual behaviour and aggregate independent variables such as mean income is common. The aim and contribution of the paper is to clarify when and how parameter estimates based on aggregates leads to bias and the likely degree of such bias. It demonstrates that use of aggregate data to estimate parameters associated with a model of individual behaviour when the outcome variable is binary (e.g. a birth) is not advisable. It only 'works' when the independent variables do not vary at the individual level (e.g. prices or the unemployment rate). Even then it requires prior distributional knowledge or assumptions. When the individual model also contains variables that vary across individuals, then the analysis in the paper suggests that all parameter estimates based solely on variation in the aggregates usually understate the size of their true value, even ones associated with variables which do not vary over individuals. Indeed, it is often the case that the 95% confidence interval of these latter parameter estimates never contains the parameter's true value.

It is common to estimate relationships between a dependent variable constructed by the aggregation of individual behaviour and aggregate independent variables, such as mean income or the proportion with a particular attribute (e.g. achieved a university degree), as well as variables that do not vary over individuals (e.g. prices). The aggregate variables may be for countries or regions/states/provinces within a country, and the relationship may be estimated using variation among the geographic groups or within them over time (e.g. a time series of country aggregates). Fertility is a particularly problematic dependent variable because at the individual level its relation to the variables that influence it (e.g. age, education, family income, housing cost) is inherently non-linear. Thus, this study focuses on a binary outcome variable at the individual level. It's aim and contribution is to clarify when and how parameter estimates based on aggregates leads to bias and the likely degree of such bias. It shows that estimates of parameters from models formulated at the individual level with independent variables which vary over individuals are usually unreliable when estimated with aggregate data in contrast to individual data.[1]

1. Foundations

A good example is an equation relating a metropolitan area's fertility rate to area house prices and the proportion who are homeowners, such as in Aksoy (2016) and Dettling and Kearney (2014), and a more recent example is Kearney and Levine (2022), which relates a US state's birth rate to the state's unemployment rate and an aggregate measure of household spending. Such equations are often interpreted as representing responses of individual fertility to independent variables, some of which are specific to the individual (e.g. homeownership

[1] The issue is related to the connection between ecological correlations and individual correlations examined by Robinson (1950): 'there are a large number of individual correlations which might correspond to any given ecological correlation.' (p.354)

status) rather than the area. But it is long-established that parameters estimated using aggregates do not generally correspond to the parameters at the individual level. In summarising earlier research, Green (1964) showed that if for the *i*-th individual $y_i = f_i(x_{1i}, \dots x_{mi})$, then consistent aggregation to an aggregate relationship $y_a = f(x_{1a}, \dots x_{ma})$, where $y_a = \sum_{i=1}^{n_a} y_i/n_a$, $x_{ka} = \sum_{i=1}^{n_a} x_{ki}/n_a$ $(k = 1, \dots, m)$ and n_a is the number of individuals in the aggregate unit *a* (e.g. metropolitan area), is only possible if the functions f_i are of the form

$$y_i = \alpha_i + \sum_{k=1}^{m} \beta_k \, x_{ki} \qquad (1)$$

Some of the x_{ki} may be constant across individuals and vary only by area. Examples are the area house price or unemployment rate; such variables are denoted by z_{ka}. Thus, we can rewrite (1) as

$$y_{ia} = \alpha_i + \sum_{k=1}^{m-K} \beta_k \, x_{ki} + \sum_{k=K}^{m} \beta_k \, z_{ka} \qquad (2)$$

In this case, if (1) holds, then the relationship between area aggregates is

$$y_a = \alpha_a + \sum_{k=1}^{m-K} \beta_k \, x_{ka} + \sum_{k=K}^{m} \beta_k \, z_{ka} \qquad (3)$$

where $\alpha_a = \sum_{i=1}^{n_a} \alpha_i/n_a$ is the average intercept for individuals in area *a*. As long as α_a is not correlated with x_{ka} or z_{ka}, variation from aggregate cross-section data across areas would identify the β_k parameters, as would time series data if *a* indicated 'time' for a particular geographic area (e.g. a country) rather than area.[2]

[2] Estimation of equation (3) by least squares across areas would identify the mean α_a. Systematic sorting of individuals by areas could, however, induce a correlation between α_a and x_{ka} or z_{ka}; an example would be people with more favourable attitudes toward childbearing (higher α_i) moving to areas with lower housing costs.

2. Some special cases

2.1 Independent variables only vary by area

If the z_{ka} variables were the only regressors (i.e. $y_i = f_i(z_{1a}, \ldots z_{ma})$)[3], then even though f_i is not linear it may be possible to identify its parameters with aggregate data if we know the form of f_i; i.e. aggregation would be consistent. For example, let $y_i^* = \alpha + \sum_{k=1}^{m} \beta_k z_{ka} + u_i$ where u_i has a logistic distribution and $y_i = 0$ if $y_i^* \leq 0$ and $y_i = 1$ if $y_i^* > 0$. The probability that $y_i = 1$ is $p_i = \Pr(u_i > -\alpha - \sum_{k=1}^{m} \beta_k z_{ka}) = F(\alpha + \sum_{k=k}^{m} \beta_k z_{ka})$ where $F(.)$ is the logistic distribution function. Thus $p_i = \frac{\exp(\alpha + \sum_{k=1}^{m} \beta_k z_{ka})}{1 + \exp(\alpha + \sum_{k=1}^{m} \beta_k z_{ka})}$ implying $\ln(\frac{p_i}{1-p_i}) = \alpha + \sum_{k=1}^{m} \beta_k z_{ka}$. Because y_a is a consistent estimator of $p_a = \sum_{i=1}^{n_a} p_i / n_a$,

$$\ln\left(\frac{y_a}{1 - y_a}\right) = \alpha + \sum_{k=1}^{m} \beta_k z_{ka} \qquad (4)$$

Estimation of the parameters of equation (4) using variation over areas, or over time within areas, will yield consistent estimates of the parameters of the individual model.[4]

2.2 Normality of birth probabilities and of the linear latent variable index

This section presents an important result for understanding aggregation bias. For each area *a* denote $\alpha + \sum_{k=1}^{m-K} \beta_k x_{ki} + \sum_{k=K}^{m} \alpha_k z_{ka}$ as $\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i + \boldsymbol{\alpha}'\boldsymbol{Z}_a$ and assume that $\boldsymbol{\beta}'\boldsymbol{X}_i$ has a joint normal distribution with covariance matrix $\boldsymbol{\beta}'\boldsymbol{\Sigma}_{xx}\boldsymbol{\beta}$.[5] Also assume that the latent variable $y_i^* = \alpha + \boldsymbol{\alpha}'\boldsymbol{Z}_a + \boldsymbol{\beta}'\boldsymbol{X}_i + u_i$ where u_i has a standard Normal distribution. As before, $y_i = 0$ if $y_i^* \leq 0$ and $y_i = 1$ if $y_i^* > 0$. The probability that $y_i = 1$ is $\Pr(u_i > -\alpha - \boldsymbol{\beta}'\boldsymbol{X}_i -$

[3] This corresponds to Green's special case (ii) for consistent aggregation (Green 1964, p. 101) when the individual relationship is linear but the β_k vary over individuals.

[4] An approximation of such a model is Örsal and Goldstein (2018) who relate the annual change in the log of a country's total fertility rate to the change in a country's lagged unemployment rate, although they do not explicitly specify the underlying individual model.

[5] As Blundell and Stoker (2005, footnote 38)) point out, this is a weaker assumption than assuming that the $\boldsymbol{X}_i$ themselves are joint normal, because that would rule out dichotomous variables as regressors.

$\boldsymbol{\alpha}'\boldsymbol{Z}_a) = \Phi(\alpha + \boldsymbol{\beta}'\boldsymbol{X}_i + \boldsymbol{\alpha}'\boldsymbol{Z}_a)$ where $\Phi(.)$ is the Normal cumulative distribution function. Blundell and Stoker (2005; pp. 376-377) show that under these assumptions that the aggregate probability function for an area is

$$p_a = E_a(y_i) = E_a(p_i) = \Phi\left(\frac{\alpha + \boldsymbol{\alpha}'\boldsymbol{Z}_a + \boldsymbol{\beta}'E_a(\boldsymbol{X}_i)}{\sqrt{1 + \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}}}\right) \qquad (5)$$

were $E_a(\boldsymbol{X}_i)$ is the area mean of $\boldsymbol{X}_i$. This is an important result because it clarifies the source of aggregation bias and what influences its direction and size, albeit under special distributional assumptions. $\boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}$ reflects the heterogeneity in the individual independent variables ($\boldsymbol{X}_i$) and is the source of bias. Equation (5) indicates that it is 'as if' the latent variable equation was $y_i^* = \frac{\alpha+\boldsymbol{\alpha}'\boldsymbol{Z}_a+\boldsymbol{\beta}'E_a(\boldsymbol{X}_i)}{\sqrt{1+\boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}}} + u_i$, which is non-linear in the parameters. An immediate corollary is that the absence of individual variables from the model removes aggregation bias in the estimates of $\boldsymbol{\alpha}$ (as in special case 2.1)**.**

If the analyst only knows that u_i has a standard Normal distribution, they are likely to assume that the latent variable is linear in the parameters. Then they would estimate the following equation by non-linear least squares or maximum likelihood from the aggregate data:

$$p_a = \Phi(c + \boldsymbol{c}'\boldsymbol{Z}_a + \boldsymbol{b}'E_a(\boldsymbol{X}_i)) \qquad (5E)$$

It is clear from equation (5) that the size of the estimates of the parameters $\boldsymbol{c}$ and $\boldsymbol{b}$ will under-state $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ by the factor $1/\sqrt{1 + \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}}$, and more heterogeneity reduces this fraction. The simulations below also consider estimation in the unlikely event the analyst also knows that $\boldsymbol{\beta}'\boldsymbol{X}_i$ has a joint Normal distribution with covariance matrix $\boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}$**.**

In general, however, if the individual relationship does not take the form of (1), estimation using aggregate data will not return consistent estimates of the parameters of the individual relationship $y_i = f_i(x_{1i}, \ldots x_{mi})$. In the first instance, the seriousness of the aggregation bias issue was explored using simulations of parameter estimation from data generated by a

known mechanism and results reported in the next section, and then in section 4 analysis with real data on individuals' fertility behaviour embedded within regions is reported.

3. Monte Carlo Simulations

Assume that y_i is a dichotomous variable indicating whether a woman has a birth or not. Define $price_a$ as a variable that only varies over area, not individuals; age_i is a continuous variable for a woman's age; Q_i is a binary variable identifying two groups who differ in their intercept and possibly also their response to price (e.g. differentiating owners and tenants). Define a latent variable equation for a woman *i* living in area *a*:

$$y_{ia}^{*} = \beta_0 + \beta_p \ln(price_a) + \gamma_1(age_i - 25) + \delta_1 Q_i$$

$$+\delta_2 Q_i \ln(price_a) + u_i = \boldsymbol{\beta}' \boldsymbol{X}_{ia} + u_i \qquad (6)$$

where u_i has a logistic distribution. Thus $p_{ia} = \frac{\exp(\boldsymbol{\beta}' \boldsymbol{X}_{ia})}{1+\exp(\boldsymbol{\beta}' \boldsymbol{X}_{ia})}$ implying

$$\ln\left(\frac{p_{ia}}{1-p_{ia}}\right) = \beta_0 + \beta_p \ln(price_a) + \gamma_1(age_i - 25) + \delta_1 Q_i$$

$$+\delta_2 Q_i \ln(price_a) \qquad (7)$$

This model is a simple version of the relationship for individual birth probabilities estimated in Ermisch (2023). Here it represents the true data generation mechanism, and the objective is to estimate its parameters with both individual data and aggregate data at the area level.

In all simulations $price_a$ is generated by a lognormal distribution at the area level; age_i is generated by a uniform distribution over the interval 18-44 at the individual level; and Q_i is generated by a latent variable with a standard Normal distribution at the individual level such that $E(Q_i) = 0.3$. 100 samples of 10,100 individuals from this model were created and they were randomly assigned to 101 areas of 100 persons each. In each sample, mean values of y_{ia}, Q_i and age_i were computed for each area (a) and are denoted y_a, Q_a and age_a,

respectively. These averages enter the aggregate equations. Simulations differ by the parameter constraints imposed on the latent variable equation (6).

Simulation 1: $\beta_0 = -0.5, \beta_p = -1, \gamma_1 = 0, \delta_1 = 0, \delta_2 = 0$. This is the simplest model considered. It corresponds with the special case 1.1 in which there is no systematic individual heterogeneity in the birth rate within areas. The model based on the aggregate data was assumed to take the form of equation (4), and Table 1 shows the parameter estimates based on individual data and aggregate data. We expect the estimates from both individual and aggregate data to be consistent estimators. The mean estimate of $\hat{\beta}_p$ based on aggregate data is within one standard error of its true value. The confidence interval of $\hat{\beta}_p$ from the aggregate data includes its true value much less than 95% of the time, but most of these breaches of the confidence interval are small. These are, however, ideal circumstances where we know the form of the individual birth rate equation and it does not depend on individual variables. Analogous results emerge when the aggregate birth probability is given by equation (5) and $\ln(\ price_a)$ is the only regressor (Simulation 5A below in which the aggregate equation (5) was estimated by used non-linear least squares).

Simulation 2 allows for binary heterogeneity in the intercept of the latent variable equation based on Q_i: $\beta_0 = -1, \beta_p = -1, \delta_1 = -1.5, \delta_2 = 0$ and $\gamma_1 = 0$.[6] The estimates from the individual data are, as expected, close to their true values, as Table 1 shows. In the estimates using aggregate data, Q_a replaces Q_i : $\ln\left(\frac{y_a}{1-y_a}\right) = \beta_0 + \beta_p \ln(price_a) + \delta_1 Q_a$. The size of the mean estimate of β_p from the aggregate data is slightly biased downward. The confidence interval of $\hat{\beta}_p$ includes its true value in only 71% of the samples, which occurs mainly (in all

[6] This model is equivalent to one with two groups defined by Q_i each of which has a different intercept but the same slope parameter for $\ln(price_a)$. This is similar to the demand model of equations (2.3) to (2.5) in Stoker (1993).

but one sample) because the lower bound of the confidence interval exceeds -1(i. e. $|\beta_p| < 1$). The parameter δ_1 is imprecisely estimated, and the mean value of its estimate is below its true value. Thus, estimation based on area aggregates in this simple model in which heterogeneity only involves a shifting intercept produces estimates of β_p which is only slightly biased, but its coverage of the true parameter is relatively poor.

Simulation 3 differs from simulation 2 in allowing the two groups also to differ in their response to price: $\beta_0 = -1$, $\beta_p = -0.5$, $\delta_1 = -1.5$, $\delta_2 = -0.5$ and $\gamma_1 = 0$. Examples of such a form are Aksoy (2016) and Dettling and Kearney (2014) in which Q_i represents housing tenure. Table 1 shows that when the price response is affected by individual heterogeneity the aggregate model produces poor estimates of the underlying individual model. Although the estimates are not significantly different from their true values, they are too imprecise to be useful.

The mean estimated response to $\ln(price_a)$ in this model is $\hat{\beta}_p + Q_a\hat{\delta}_2$. Q_a varies between 0.18 and 0.39 across the areas, with a mean value of 0.3. The aggregate estimates from Table 1 suggest that the mean response is $-0.34 - (0.3) \cdot 0.74 = -0.56$ ($SE = 0.03$), which is significantly below its true value of $-0.50 - (0.3) \cdot 0.50 = -0.65$.

Table 1: Parameter estimates from simulated data (100 sample replications, N=10,100 individuals, 101 regions**)**

Parameter [Variable]	Simulation 1 Individual[1]	Simulation 1 Aggregate[1]	Simulation 2 Individual[2]	Simulation 2 Aggregate[2]
	Mean par. Est. (Mean SE) *Coverage**	Mean par. Est. (Mean SE) *Coverage**	Mean par. Est. (Mean SE) *Coverage**	Mean par. Est. (Mean SE) *Coverage**
β_p $[\ln(price_a)]$	-1.00 (0.03) *0.96*	-1.02 (0.03) *0.87*	-1.00 (0.07) *0.94*	-0.97 (0.04) *0.71*
δ_1 $[Q_i]$	0	0	-1.51 (0.07) *0.97*	-1.23 (0.62) *0.90*
δ_2 $[Q_i \ln(price_a)]$	0	0	0	0
γ_1 $[age_i - 25]$	0	0	0	0
β_0	-0.51 (0.02)	-0.51 (0.02)	-1.00 (0.03)	-0.99 (0.20)
	Simulation 3 Individual[3]	Simulation 3 Aggregate[3]	Simulation 4 Individual[4]	Simulation 4 Aggregate[4]
β_p $[\ln(price_a)]$	-0.51 (0.03) *0.92*	-0.51 (0.19) *0.88*	-1.00 (0.05) *0.96*	-0.46 (0.02) *0*
δ_1 $[Q_i]$	-1.50 (0.08) *0.93*	-1.13 (0.57) *0.89*	0	0
δ_2 $[Q_i \ln(price_a)]$	-0.49 (0.08) *0.95*	-0.17 (0.62) *0.91*	0	0
γ_1 $[age_i - 25]$	0	0	0.50 (0.01) *0.96*	0.21 (0.03) *0*
β_0	-1.00 (0.03)	-0.99 (0.18)	-7.00 (0.17)	-2.74 (0.18)

* Coverage is the proportion of 95% confidence intervals that contain the true parameter value.

[1] True parameter values: $\beta_0 = -0.5, \beta_p = -1, \gamma_1 = 0, \delta_1 = 0, \delta_2 = 0$.

[2] True parameter values: $\beta_0 = -1, \beta_p = -1, \delta_1 = -1.5, \delta_2 = 0, \gamma_1 = 0,$

[3] True parameter values: $\beta_0 = -1, \beta_p = -0.5, \delta_1 = -1.5, \delta_2 = -0.5, \gamma_1 = 0.$

[4] True parameter values: $\beta_0 = -7\ \beta_p = -1, \gamma_1 = 0.5, \delta_1 = 0, \delta_2 = 0$

Simulation 4 introduces individual heterogeneity with a continuous variable, age, but no interaction with $price_a$: $\beta_0 = -7, \beta_p = -1, \gamma_1 = 0.5, \delta_1 = \delta_2 = 0$. What is most striking about the aggregate estimates is that the 95% confidence interval of neither $\hat{\beta}_p$ nor $\hat{\gamma}_1$ ever contains its true value. There is substantial downward bias in the size of both, even though $price_a$ does not vary over individuals (i.e. is not itself subject to aggregation) and age_a and $price_a$ are uncorrelated. Dropping age_a from the aggregate equation does not solve the estimation problem: in the model here, the point estimate $\hat{\beta}$ =0.46 (SE=0.03) is the same and its 95% confidence interval never contains its true value.

Simulation 5 address the estimation of the parameters the model in section 2.2 in which the Normality assumptions introduced at the outset of section 2.2 hold by construction. A new variable X_i was added. It was assumed to have a standard Normal distribution and X_a is denoted as its area mean. The true parameters were assumed to be $\alpha = -0.5, \alpha_p = -1$ and $\beta_x = 0.5$ *or* 0. As the main issue concerns the aggregate estimates, only the Monte Carlo simulations of these are reported in Table 2: simulation 5A sets $\beta_x = 0$ and 5B sets $\beta_x = 0.5$. In these two simulations we assume 'limited information' in the sense that the analyst assumes that the latent variable is linear in parameters and its error term is standard Normal, and so we estimate equation (5E). In simulation 5C there are the same two independent variables as in 5B, but it is assumed that the analyst also knows that the distribution of X_i is standard Normal, in which case we should be able to estimate the parameters α_p and β_x consistently by estimating the exact counterpart of equation (5) by non-linear least squares. This is, of course, a patently unreal situation.

Table 2: Aggregate estimates of equation (5E) or (5)
(Estimation used non-linear least squares**)**

Parameter [Variable]	Simulation 5A Aggregate[1]	Simulation 5B Aggregate[2]	Simulation 5C Aggregate[3]
	Mean par. Est. (Mean SE) *Coverage**	Mean par. Est. (Mean SE) *Coverage**	Mean par. Est. (Mean SE) *Coverage**
α_p [ln $(price_a)$]	-1.00 (0.02) *0.99*	-0.89 (0.02) *0*	-1.04 (0.12) *0.89*
β_x $[X_a]$	0	0.45 (0.16) *0.91*	0.55 (0.23) *0.95*
c	-0.50 (0.02)	-0.45 (0.02)	-0.52 (0.06)

[1] $\alpha = -0.5, \alpha_p = -1, \beta_x = 0$, 'limited information' in estimation, eq'n (5E)
[2] $\alpha = -0.5, \alpha_p = -1\,, \beta_x = 0.5$, 'limited information' in estimation, eq'n (5E).
[3] $\alpha = -0.5, \alpha_p = -1\,, \beta_x = 0.5$, 'full information' in estimation, eq'n (5).

When the independent variables only vary by area (no heterogeneity) equation (5) indicated that the parameter estimates from the aggregate data using equation (5E) should be consistent. Simulation 5A in Table 2 indicates that they are indeed unbiased and relatively precise, like in simulation 1.

In Simulation 5B there is one individual variable with $\Sigma_{xx} = var(X_i) = 1$ so that $= \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta} = \beta_x^2$. Given the assumed parameter values, $\sqrt{1 + \boldsymbol{\beta}'\Sigma_{xx}\boldsymbol{\beta}} = 1.118$. Thus, from equation (5), we would expect the estimates of c_p and b_x in equation (5E) to be -0.894 and 0.447 and, respectively. This is what we find in Table 2 when we estimate equation (5E) in which we have 'limited information' about the data information mechanism; thus, the aggregate estimates are biased and understate the size of both parameters α_p and β_x. As in simulation 4, aggregate estimation of a model with an independent variable which varies over individuals produces biased estimates of the price effect. Indeed, in this simulation the 95% confidence interval of $\hat{\alpha}_p$ never contains its true value.

Finally, when we assume 'full information' and estimate the exact counterpart of equation (5) the mean of estimate of α_p is slightly closer to its true value than in simulation 5B and its coverage of the true parameter value is much better, but the estimates of both parameters are not very precise.

3. Estimation with real data

Individual panel data from the UK Household Longitudinal Study (*Understanding Society*) used in Ermisch (2023) was used to generate estimates of the mean proportion having a birth by each of 10 standard regions in England and Wales for each year of the sample (2010-21) (denoted $\hat{y}_{rt}$). The region by year sample sizes range from 150 to 1,200. The mean $\hat{y}_{rt}$ (over the 120 year-region combinations) was 0.064 and its mean standard error was 0.011. The aggregate fixed regional effect regressions that follow were weighted by the inverse of a region's mean standard error of $\hat{y}_{rt}$ over the 12 years.

The parameter estimates using the individual data are of the following model:[7]

$$\ln\left(\frac{p_{itr}}{1-p_{itr}}\right) = \alpha_0 + \alpha_1 age_{it} + \alpha_2 age_{it}^2 + \mu_1 year1316 + \mu_2 year1721 + \delta_1 year1316 * age_{it} + \delta_2 year1721 * age_{it} + \sum_r \gamma_{r0}\, region_t + \beta_1 \ln(realhouseprice)_{rt-1} + \beta_2\, unempr_{rt-1} + \beta_3 degree_i + age_{it} * (\gamma_{London} London_{it} + \gamma_{SEE} SEE_{it}) \quad (8)$$

where *t*=year, *r*=region, *i*=individual woman, age_{it} is age *minus* 31, $year1316$ is a dummy variable for years 2013-16, $year1721$ is a dummy variable for years 2017-21, $region_{rt}$ indicates a dummy variable for region *r*, $London_{it}$ and SEE_{it} are dummy variables for residence in London and the Southeast or East, respectively, $degree_i$ indicates whether the woman has a university degree, $realhouseprice_{rt-1}$ and $unempr_{rt-1}$ indicate lagged regional real house price and unemployment rate, respectively. Estimates of the key

[7] The small proportion of women who change regions between waves are excluded.

parameters of this model are shown in Table 3 in the Individual column. Provided the logistic assumption is correct, these are consistent estimates.

When using the regional aggregates to estimate the model it is not possible to obtain any reasonable degree of precision if we include all the age interactions with the year variables and the indicators of the London region or the South or East regions (SSE). Thus, these were excluded. The aggregate regression relates $\ln(\frac{\hat{y}_{rt}}{1-\hat{y}_{rt}})$ to the lagged regional house price, lagged regional unemployment rate, mean age, mean age squared, proportion who have a degree, dummy variables for years 2013-16 and 2017-21 and a fixed effect for each region, as in the individual analysis.

Table 3: Selected parameter estimates from real data (standard errors in parentheses)

Parameter [Variable]	Individual Data	Aggregate $\hat{y}_{rt}$[a]	Aggregate $\hat{p}_{rt}$[b]
β_1 $\ln(realhouseprice)_{rt-1}$	-0.50 (0.25)	-0.29 (0.22)	-0.38 (0.05)
β_2 $unempr_{rt-1}$	0.053 (0.021)	0.054 (0.031)	0.066 (0.005)
β_3 $degree_i$	-0.05 (0.03)	-0.17 (0.97)	-0.33 (0.15)
α_1 age_{it}	-0.073[c] (0.006)	-0.099 (0.066)	-0.046 (0.008)
α_2 age_{it}^2	-0.016[c] (0.0004)	0.016 (0.016)	0.004 (0.003)

[a] $\hat{y}_{rt}$ = mean of binary birth variable by region and year. Regression weighted by the inverse of the mean standard error of the estimate of $\hat{y}_{rt}$ for each region.
[b] $\hat{p}_{rt}$ is the mean predicted birth probability by region and year.
[c] Parameters reported are relevant to women living outside London, the East and Southeast during 2010-12.

The aggregate estimates are shown in the Aggregate column of Table 3. Compared to the estimates based on individual data, the estimated size of the house price effect (β_1) is smaller in the regression using the aggregates, although both estimates have relatively large standard errors, making their estimated difference statistically imprecise. Both estimators produce similar point estimates of the unemployment effect (β_2). The estimated impact of having a degree (β_3) is too imprecise to be useful.

The analysis of these data concludes with an artificial exercise. According to the individual model there are two sources of variation in the individual birth probability: (1) that arising from variation in the independent variables over time and area (age, house price, etc.) and (2) variation from the random error term in the latent variable equation, which was assumed to have a logistic distribution. The latter source can be excluded by only considering variation in the predicted birth probability $\hat{p}_{itr}$, which is generated by the estimated parameters and variables in equation (8). $\hat{p}_{itr}$ was aggregated by region and year to produce $\hat{p}_{rt}$. The final

regression in Table 3 uses $\hat{p}_{rt}$ as the dependent variable.[8] Unsurprisingly, the parameter estimates are much more precise than earlier columns because variation in $\hat{p}_{tr}$ is generated solely by the independent variables. Yet aggregation continues to produce lower estimates of the size of β_1, but now higher estimates of the size of β_2 and β_3 compared to the individual model, particularly the last: the effect of a degree from the aggregate estimates is much larger. Even in these favourable circumstances (by construction), the aggregate estimates of the parameters associated with the variables which vary over individuals are not reliable as estimates of influences on individual behaviour when compared with estimates from the individual data.

4. Conclusions

Use of aggregate data to estimate parameters associated with a model of individual behaviour when the outcome variable is binary (e.g. a birth) is not advisable. It only 'works' when the independent variables do not vary at the individual level (e.g. prices or the unemployment rate). Even then it requires prior distributional knowledge (e.g. that the logistic or normal distribution is a good approximation for the mechanism generating the binary outcome variable). When the model also contains variables that vary across individuals (i.e. there is individual heterogeneity), the Monte Carlo simulations indicate that all parameter estimates based solely on variation in the aggregates usually understate the size of their true value and have poor coverage of the true parameter, a finding that was anticipated by the theoretical arguments leading to equation (5). This also applies to parameters of variables that are associated with variables which do not vary over individuals. Indeed, it is often the case that the 95% confidence interval of these latter parameter estimates never contains the parameter's true value. Yet it is when there is particular interest in the impact of variables

[8] Having estimates of $\hat{p}_{tr}$ is of course an unreal situation because we would not need the aggregate estimates if the individual data were available.

which do not vary over individuals that applied researchers often call upon aggregate data, either by geographic area, time or both.

For some purposes it may be sufficient to have information on the relationship between aggregates, but not for estimating how heterogeneous individuals respond to their environment (e.g. prices) and their individual circumstances and attributes (e.g. education). As Stoker (1993, p.1871) cautions:

> The problem is that for any equation connecting aggregates, there are a plethora of behaviorally different "stories" that could generate the equation, which are observationally equivalent from the vantage point of aggregate data alone.

References

Aksoy, C.G. (2016). Short-term effects of house prices on birth rates. European Bank for Reconstruction and Development Working Paper No. 192.

Blundell, R. and Stoker, T.M. (2005). Heterogeneity and aggregation. *Journal of Economic Literature* 43 (June):347-391.

Ermisch, J. 2023. The recent decline in period fertility in England and Wales: Differences associated with family background and intergenerational educational mobility. *Population Studies*, https://doi.org/10.1080/00324728.2023.2215224).

Dettling, L.J. and Kearney. M.S. (2014). House prices and birth rates: The impact of the real estate market on the decision to have a baby. *Journal of Public Economics* 110: 82-100.

Green, H.A. J. (1964, republished 2015). *Aggregation in Economic Analysis: An Introductory Survey*. Princeton University Press.

Kearney, Melissa Schettini and Phillip B. Levine (2022). The US Covid-19 baby bust and rebound. NBER Working Paper 30000 http://www.nber.org/papers/w30000.

Örsal, D. D. K., & Goldstein, J. R. (2018). The changing relationship between unemployment and total fertility. *Population Studies*, *72*(1), 109–121.

W. S. Robinson 1950. Ecological Correlations and the Behavior of Individuals. *American Sociological Review*, Vol. 15, No. 3 pp. 351-357.

Stoker, T.M. (1993). Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature* 31(4): 1827-1874