Department of Economics Discussion Paper Series

A two sample size estimator for large data sets

Martin O'Connell, Howard Smith and Øyvind Thomassen

Number 1001
February, 2023

Department of Economics
Manor Road Building, Manor Road
Oxford, OX1 3UQ

ISSN 1471-0498

A two sample size estimator for large data sets

Martin O'Connell, Howard Smith and Øyvind Thomassen*

February, 2023

Abstract

In GMM estimators moment conditions with additive error terms involve an observed component and a predicted component. If the predicted component is computationally costly to evaluate, it may not be feasible to estimate the model with all the available data. We propose an estimator that uses the full data set for the computationally cheap observed component, but a reduced sample size for the predicted component. We show consistency, asymptotic normality, and derive standard errors and a practical criterion for when our estimator is variance-reducing. We demonstrate the estimator's properties on a range of models through Monte Carlo studies and an empirical application to alcohol demand.

Keywords: GMM, estimation, micro data
JEL classification: C20, C51, C55
Acknowledgements: Financial support from the Center for National Competitiveness of the Institute of Economic Research of Seoul National University, from the Housing and Commercial Bank Economic Research Fund of the Institute of Economic Research of Seoul National University and from the Economic and Social Research Council (ESRC) under the Centre for the Microeconomic Analysis of Public Policy (CPP; grant number ES/T014334/1) are gratefully acknowledged. Data supplied by Kantar FMCG At-Home Purchase Panel. The use of Kantar FMCG At-Home Purchase Panel data in this work does not imply the endorsement of Kantar FMCG At-Home Purchase Panel in relation to the interpretation or analysis of the data. All errors and omissions remain the responsibility of the authors. We thank Richard Spady, Martin Weidner, Yoon-Jae Whang and Frank Windmeijer for comments.

*O'Connell is at the University of Wisconsin-Madison and Institute for Fiscal Studies, Smith is at Oxford University and Thomassen is at NHH Norwegian School of Economics. **Correspondence:** moconnell9@wisc.edu, howard.smith@economics.ox.ac.uk, oyvind.thomassen@nhh.no.

1 Introduction

The growing availability of large data sets, such as administrative tax records, household scanner data, and data collected by tech firms, in combination with the computational burden of complex models, can mean that researchers are forced to estimate their models with a subset of the available data. As stated by Lee and Ng (2020) in their recent review article: "analyzing large datasets is time consuming and sometimes beyond the limits of our computers. (...) One way to alleviate the bottleneck is to work with a sketch of the data. These are data sets of smaller dimensions and yet representative of the original data." Examples of this practice include Busse et al. (2013), who "draw a 10 percent random sample of all transactions (...) necessary to allow for estimation of specifications with multiple sets of high-dimensional fixed effects, including fixed effect interactions" (p. 228) and Varian (2014), who states that "it is often possible to select a subsample for statistical analysis. At Google, for example, I have found that random samples on the order of 0.1 percent work fine for analysis of business data."

For models with additive error terms, GMM moment conditions involve a difference between an observed and predicted component. In nonlinear settings the predicted part needs to be calculated for each trial value of the parameters during numerical minimization of the GMM objective function. This can be computationally costly if the model involves simulation, nested optimization or a fixed-point problem. Even in linear settings, if the model contains "multiple sets of high-dimensional fixed effects" computation of the predicted component can be burdensome. This means it can be necessary to use a subsample of the data for estimation. Typically one would use a subsample also for the observed part.

In this paper we propose an alternative estimator for cross-section or panel data that uses the full sample for the observed component, and a subsample only for the predicted component. The cost of computing the value of the observed part of the moment is usually negligible, because it needs to be done once, does not depend on the parameters to be estimated, and typically involves simple manipulations of a few variables at most. The calculation of the predicted part of the moment, on the other hand, may be costly in terms of CPU or memory demand. We show that the estimator is consistent and asymptotically normal, and how to estimate its asymptotic variance.

The rationale for using our estimator, instead of a GMM estimator with the subsample, is that it minimizes the sampling error of the observed component of the moment, which acts to lower the variance of the estimator. Yet there is an offsetting effect as, compared with small-sample GMM, our estimator lowers the correlation between the observed and predicted part of the moments, which acts to raise the variance of the estimator. We derive a condition, which can be straightforwardly checked with a subsample of the data, that, if satisfied, ensures our estimator will have lower variance than small-sample GMM. The condition requires the covariance between the observed and predicted component to be small relative to the variance of the observed component.

We illustrate our estimator's properties through two sets of simulations. The first is based on a Monte Carlo study in which we simulate a data generating process for a set of linear and non-linear models. In each case we fix a small sample size (used for the predicted component of the moment) and show how the

variance of our estimator changes with the size of the large sample (used for the observed component). The simulations illustrate the potential for our estimator to increase efficiency in linear and non-linear (namely a multinomial and random coefficient logit) settings.

Our second set of simulations is based on household scanner data (Kantar FMCG At-Home Purchase Panel) for the UK that covers alcohol purchases for 40,000 households over two decades. We specify an alcohol demand equation, with the purpose of estimating the slope of households' alcohol Engel curve. As the model is relatively simple (i.e., linear with a relatively small set of parameters) we are able to implement (both for a cross-sectional and panel data variant) GMM estimation on the full dataset. This acts as a benchmark against which we compare both GMM estimators implemented on sub-samples of the data and our estimator. We show that our estimator outperforms small-sample GMM, and that for the panel data model, the precision of our estimator is close to that of the GMM estimator implemented on the full dataset, even when the size of the sub-sample used to evaluate the predicted part of the moment represents a modest fraction (less than 15%) of the full sample.

Our estimator is related to that proposed by Imbens and Lancaster (1994), which consists of a GMM estimator that augments moments formed with micro or survey data with moments from census ("macro") data, which have no sampling error and act to improve the accuracy of estimation. It also is related to the two-sample 2SLS estimator of Angrist and Krueger (1992), Inoue and Solon (2010) and Pacini and Windmeijer (2016). Other than our more general nonlinear GMM setting, the main difference between our work and those papers is that they focus on a situation where only a subset of the required variables are observed in each of two data sets, so that both samples are needed for identification of parameters of interest. While they assume independence between the two samples, in our case one sample is a subset of the other.

The next section outlines our setting and estimator, and establish consistency, asymptotic normality and an asymptotic variance estimator. Section 3 discusses when our estimator is more efficient than small-sample GMM. Sections 4 and 5 present our Monte Carlo and data simulations and a final section concludes.

2 The estimator

Let I_N be a set of N individuals i, and let (y_i, w_i) for $i \in I_N$ be a random sample of observable variables y_i, which is a column vector, and w_i, whose dimensions are unspecified (but finite). The following population moment condition is assumed to hold at the true parameter value θ_0:

$$\mathbb{E}[y_i - h_i(\theta_0)] \quad = \quad 0, \tag{2.1}$$

where $h_i(\theta)$ is a vector valued function of θ and of w_i, $h_i(\theta) = h(w_i, \theta)$.

Define the observed part of the sample moment:

$$\bar{y}_N \quad = \quad \frac{1}{N} \sum_{i \in I_N} y_i, \tag{2.2}$$

Suppose it is costly to compute the predicted part of the sample moment using the full sample, $\bar{h}_N(\theta) = \frac{1}{N}\sum_{i\in I_N} h_i(\theta)$. To reduce the computational burden of estimating the model, we approximate the sample average $\bar{h}_N(\theta)$ by selecting (at random) a subset of households $I_n \subset I_N$ with

$$n = kN \tag{2.3}$$

for some "scaling-down" constant $k \leq 1$, where we think of n as the largest sample size for which the estimator is computationally feasible. In our asymptotic analysis, we assume that as n increases, N increases correspondingly, to ensure that condition (2.3) holds. We define the observed component of the sample moment as:

$$\bar{h}_n(\theta) = \frac{1}{n}\sum_{i\in I_n} h_i(\theta). \tag{2.4}$$

The large-small estimator is

$$\hat{\theta} = \arg\min_{\theta}[\bar{y}_N - \bar{h}_n(\theta)]'\hat{W}[\bar{y}_N - \bar{h}_n(\theta)]. \tag{2.5}$$

where θ is $K\times 1$, and y_i is $L\times 1$ for integers $L \geq K$, and $\hat{W} \xrightarrow{p} W$ for a positive semi-definite $L\times L$ matrix W. We make use of the following notation for the sample average of the moments:

$$\hat{g}_n(\theta) = \bar{y}_N - \bar{h}_n(\theta), \tag{2.6}$$

so that the estimator can be written $\hat{\theta} = \arg\min_\theta \hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$. Since, in general, $\hat{g}_n(\theta)$ is not a sample average over a single sample, this is not a traditional GMM estimator.[1] However, note that when $k = 1$ and so $n = N$, our estimator coincides with the standard GMM case.[2]

Example 1 – linear panel data model. We have a sample (X_{it}, Y_{it}, Z_{it}), $i \in I_N$, $t = 1,\ldots,T$, that is identically distributed, independent in the i-dimension and with unknown dependence in the t-dimension. The vector X_{it} is $1\times K$ for some integer $K > 1$, Z_{it} is $1\times L$ for some integer $L \geq K$, and Y_{it} is a scalar. The first entries of X_{it} and Z_{it} are both 1. The structural equation is $Y_{it} = X_{it}\theta + e_{it}$, where $\mathbb{E}(Z_{it}'e_{it}) = 0$ is assumed to hold in the population. Let $y_i = T^{-1}\sum_{t=1}^T Z_{it}'Y_{it}$ and $h_i(\theta) = T^{-1}\sum_{t=1}^T Z_i'X_i\theta$. Then condition (2.1) is $T^{-1}\sum_{t=1}^T \mathbb{E}(Z_{it}'e_{it}) = 0$. A pooled linear GMM estimator $\hat{\theta}$ is then given by equation (2.5) when $n = N$. Let X be the $NT\times K$ matrix that vertically stacks X_{it}, Z the corresponding $NT\times L$ matrix, and Y a stacked $NT\times 1$ vector. The GMM objective function is $[Z'Y/NT - (Z'X/NT)\theta]'\hat{W}[Z'Y/NT - (Z'X/NT)\theta]$. Solving the first-order condition gives the estimator $\hat{\theta} = [(Z'X/NT)'\hat{W}(Z'X/NT)]^{-1}(Z'X/NT)'\hat{W}(Z'Y/NT)$. ✓

Example 1 with large-small estimator. Let I_n be a subset of I_N, and pick at random a subset of size $\tau \leq T$ of the full set of T time periods. Let X_n be the $n\tau\times K$ matrix that vertically stacks X_{it} for i in I_n and $t = 1,\ldots,\tau$, and Z_n the corresponding $n\tau\times L$ matrix, so that $\bar{h}_n(\theta) = (Z_n'X_n/n\tau)\theta$. The objective function

[1] Newey and McFadden (1994) p. 2116) define a GMM estimator as $\arg\min_\theta \hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$ with $\hat{g}_n(\theta) = n^{-1}\sum_{i=1}^n g(z_i,\theta)$ for some function g of the data z_i.

[2] The GMM estimator is consistent by Theorem 2.6 in Newey and McFadden (1994) with some additional assumptions, which we maintain throughout: (i) $W\mathbb{E}[y_i - h_i(\theta)] = 0$ only if $\theta = \theta_0$; (ii) $\theta_0 \in \Theta$ where Θ is compact; (iii) $h_i(\theta)$ is continuous at each $\theta\in\Theta$ with probability one; (iv) $\mathbb{E}[\sup_{\theta\in\Theta}\|h_i(\theta)\|] < \infty$ and $\mathbb{E}[\|y_i\|] < \infty$, which implies $\mathbb{E}[\sup_{\theta\in\Theta}\|y_i - h_i(\theta)\|] < \infty$ by the triangle inequality.

for the large-small estimator is $[Z'Y/NT - (Z_n'X_n/n\tau)\theta]'\hat{W}[Z'Y/NT - (Z_n'X_n/n\tau)\theta]$ and the estimator is $\hat{\theta} = [(Z_n'X_n/n\tau)'\hat{W}(Z_n'X_n/n\tau)]^{-1}(Z_n'X_n/n\tau)'\hat{W}(Z'Y/NT)$. ✓

We think of nonlinear models as the main application of the large-small estimator. But in Example 1, if N, L and K are very large (e.g. due to inclusion of high-dimensional fixed effects), the computational cost saving may be significant in these linear settings too.[3] We next consider two examples of nonlinear models, which highlight that our estimator can aid identification.

Example 2 – discrete-choice model. We have a random sample (X_i, Y_i, Z_i), $i \in I_N$. The setting is a discrete choice between J alternatives, with cross-sectional data for individual decision makers. J is large and some alternatives j are chosen only by a small fraction of consumers. The outcome variable Y_i is a $J \times 1$ vector of zeros except for the j-th entry which is 1 if i chooses alternative j. The $J \times K$ matrix X_i contains consumer- and alternative-specific attributes, and Z_i is a $J \times L$ matrix of instruments. A discrete-choice model gives the probability of choosing each alternative as a $J \times 1$ vector $\mathbb{P}(X_i, \theta)$ where θ includes coefficients on interactions between alternative-specific dummies and consumer attributes. The moment condition $\mathbb{E}[Z_i'(Y_i - \mathbb{P}(X_i, \theta_0))] = 0$ holds in the population. Define $y_i = Z_i'Y_i$ and $h_i(\theta) = Z_i'\mathbb{P}(X_i, \theta)$. Suppose $\mathbb{P}$ involves simulation, a nested optimization or fixed-point problem, or other features that are computationally costly, so that it is not feasible to compute the GMM estimator for the full data set. Instead we work with a subset $I_n \subset I_N$ of the data. If we compute a GMM estimator using the subsample I_n, some alternatives may not be chosen by any consumers in I_n, and the corresponding alternative-specific coefficients can no longer be estimated.[4] The large-small estimator does not have this problem, since predictions are formed for every alternative for every i and the full data set is used for the observed component of the moments. ✓

Example 3 – multi-product demand model with corner solutions. Let notation be as in Example 2, but now let Y_i be a $J \times 1$ vector of non-negative continuous choices, and $\mathbb{P}(X_i, \theta)$ a model of continuous choices subject to non-negativity constraints and permitting corner solutions. If utility is non-separable in products, and varies with continuous consumer attributes and simulated random shocks, J^2 combinations of interior/corner solutions must be checked and compared for each combination of household i and simulation draw r for each trial value of θ. This is an example of a nested optimization problem that makes estimation computationally costly. ✓

2.1 Consistency

Consistency follows from a standard result for extremum estimators (Newey and McFadden (1994), Theorem 2.1), where a key requirement is that the sample objective function converges uniformly in probability to its population counterpart. It is sufficient for $\hat{g}_n(\theta)$ to converge uniformly in probability to $g_0(\theta)$. Newey and McFadden (1994) provide a proof that this is the case for GMM estimators, which we adapt to show the same for our estimator.

[3] In Example 1, the large-small estimator involves $L(n\tau)K$ operations, instead of $L(NT)K$ operations, for the predicted part of the moment.

[4] See Lee and Ng (2020), Section 1.1, for a real data example of such "rank failure" in random subsampling.

Proposition 1. The large-small estimator (2.5) is consistent: $\hat{\theta} \xrightarrow{p} \theta_0$.

Proof. See Appendix A. □

2.2 Asymptotic normality

An estimator of the form $\hat{\theta} = \arg\min_\theta \hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$ is a *minimum distance estimator*[5] if at the true value θ_0, $\hat{g}_n(\theta_0) \xrightarrow{p} 0$. Minimum distance estimators for which $\sqrt{n}\hat{g}_n(\theta_0)$ is asymptotically normal are themselves asymptotically normal. We proceed by verifying these two requirements.

Proposition 2. The large-small estimator (2.5) is a minimum distance estimator.

Proof. By the law of large numbers, $\bar{h}_n(\theta_0) \xrightarrow{p} \mathbb{E}[h_i(\theta_0)]$ as $n \to \infty$. Since $N \geq n$ by condition (2.3), $\bar{y}_N \xrightarrow{p} \mathbb{E}[y_i]$ as $n \to \infty$. Then $\hat{g}_n(\theta_0) \xrightarrow{p} 0$ follows from the additivity of probability limits and condition (2.1): $\hat{g}_n(\theta_0) \xrightarrow{p} \mathbb{E}[y_i] - \mathbb{E}[h_i(\theta_0)] = 0$. □

Proposition 3. The moment in (2.6) satisfies:

$$\sqrt{n}\hat{g}_n(\theta_0) \xrightarrow{d} N(0, \Omega) \tag{2.7}$$

where $\Omega = k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma_{yh}')$, $\Sigma_y = \text{Var}[y_i]$, $\Sigma_h = \text{Var}[h_i(\theta_0)]$ and $\Sigma_{yh} = \text{Cov}[y_i, h_i(\theta_0)]$.

Proof. Define $\mu_y = \mathbb{E}[y_i]$ and $\mu_h = \mathbb{E}[h_i(\theta_0)]$. By the central limit theorem $\sqrt{n}(\bar{h}_n(\theta_0) - \mu_h) \xrightarrow{d} N(0, \Sigma_h)$ as $n \to \infty$ and $\sqrt{N}(\bar{y}_N - \mu_y) \xrightarrow{d} N(0, \Sigma_y)$ as $N \to \infty$. Given the fixed ratio $k = n/N$, it then follows that when $n \to \infty$, $\sqrt{n}(\bar{y}_N - \mu_y) = \frac{\sqrt{n}}{\sqrt{N}}\sqrt{N}(\bar{y}_N - \mu_y) \xrightarrow{d} N\left(0, \frac{n}{N}\Sigma_y\right)$. Online Appendix A shows that $\text{Cov}\left[\sqrt{n}(\bar{y}_N - \mu_y), \sqrt{n}(\bar{h}_n(\theta_0) - \mu_h)\right] = \frac{n}{N}\Sigma_{yh}$. Using the fact that $\mu_y = \mu_h$ by condition (2.1), we get $\sqrt{n}\hat{g}_n(\theta_0) = \sqrt{n}(\bar{y}_N - \bar{h}_n(\theta_0)) = \sqrt{n}(\bar{y}_N - \mu_y) - \sqrt{n}(\bar{h}_n(\theta_0) - \mu_h) \xrightarrow{d} N(0, k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma_{yh}'))$. □

We can now state the main result about the distribution of the large-small estimator, which follows from Theorem 3.2 in Newey and McFadden (1994).[6]

Proposition 4. The large-small estimator (2.5) satisfies

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, B\Omega B'), \tag{2.8}$$

where $B = (G'WG)^{-1}G'W$, $G = -\nabla_\theta \mathbb{E}[h_i(\theta_0)]$ (an $L \times K$ matrix), and $\Omega = k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma_{yh}')$.

[5] See Newey and McFadden (1994), p. 2116 - 2117.

[6] For simplicity, assume that (a) $\theta_0 \in \text{interior}(\Theta)$; (b) $h_i(\theta)$ is continuously differentiable on interior(Θ), and (c) $G'WG$ is nonsingular. But note we can relax (b) and (c) to the hypotheses in Newey and McFadden's Theorem 3.2: (d) $\hat{g}_n(\theta)$ is continuously differentiable in a neighbourhood $\mathcal{N}$ of θ_0; (e) there is a function $G(\theta)$ that is continuous at θ_0 and $\sup_{\theta \in \Theta} \|\nabla_\theta \hat{g}_n(\theta) - G(\theta)\| \xrightarrow{p} 0$; (f) for $G = G(\theta_0)$, $G'WG$ is nonsingular. Clearly (b) implies (d) and (e). Online Appendix B gives an alternative result for the case when condition (e) is not satisfied.

2.3 Asymptotic variance estimation

Given Proposition 4, we can obtain an estimator for the asymptotic variance in the usual way.

Proposition 5. The following estimator of the asymptotic covariance matrix is consistent:

$$\hat{B}\hat{\Omega}\hat{B}' \quad \xrightarrow{p} \quad B\Omega B', \tag{2.9}$$

where $\hat{B} = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}$, $\hat{G} = -n^{-1}\sum_{i\in I_n}\nabla_\theta h_i(\hat{\theta})$, and $\hat{\Omega} = k\hat{\Sigma}_y + \hat{\Sigma}_h - k(\hat{\Sigma}_{yh} + \hat{\Sigma}'_{yh})$, with

$$\hat{\Sigma}_y = N^{-1}\sum_{i\in I_N}[y_i - \bar{y}_N][y_i - \bar{y}_N]' \tag{2.10}$$

$$\hat{\Sigma}_h = n^{-1}\sum_{i\in I_n}[h_i(\hat{\theta}) - \bar{h}_n(\hat{\theta})][h_i(\hat{\theta}) - \bar{h}_n(\hat{\theta})]' \tag{2.11}$$

$$\hat{\Sigma}_{yh} = n^{-1}\sum_{i\in I_n}[y_i - \bar{y}_N][h_i(\hat{\theta}) - \bar{h}_n(\hat{\theta})]'. \tag{2.12}$$

Proof. By the law of large numbers, each of equations (2.10)-(2.12) converges in probability to the corresponding population object, so that $\hat{\Omega} \xrightarrow{p} \Omega$. The result then follows by Newey and McFadden (1994) Theorem 4.2. □

Standard errors are given by the square roots of the diagonal entries in $\hat{B}\hat{\Omega}\hat{B}'/n$. The covariance estimator is also valid for the special case of GMM, where $N = n$ and $k = 1$. It then corresponds to the centred covariance estimator that is "in general preferred" (Hansen (2022), p. 431) for GMM estimators.[7] Finally, the optimal weighting matrix is the usual choice:[8]

Proposition 6. The large-small estimator (2.5) is asymptotically efficient when $\hat{W}$ is chosen so that $W = \Omega^{-1}$.

3 Conditions for efficiency gain

Holding the sample size for the predicted component of the moment condition in the estimator (2.5) fixed at n, we consider the effect on the variance of the estimator of increasing the sample size for the observed component, N.

Proposition 7. For the k-th element of the parameter vector θ, the difference between the asymptotic variance of the GMM estimator $\hat{\theta}_k^{n,n}$ that uses sample size n, and that of the large-small estimator $\hat{\theta}_k^{N,n}$ is

$$\text{Avar}(\hat{\theta}_k^{n,n}) - \text{Avar}(\hat{\theta}_k^{N,n}) = \left(\frac{1}{n} - \frac{1}{N}\right) b_k[\Sigma_y - (\Sigma_{yh} + \Sigma'_{yh})]b'_k, \tag{3.1}$$

where b_k is the k-th row of the matrix B.

[7] Then $\hat{\Omega} = n^{-1}\sum_{i\in I_n}[(y_i - h_i(\theta)) - (\bar{y}_n - \bar{h}_n(\theta))][(y_i - h_i(\theta)) - (\bar{y}_n - \bar{h}_n(\theta))]'$, which equals the $\hat{\Omega}$ above when we set $k = 1$ and replace N with n everywhere.

[8] See Theorem 5.2, Newey and McFadden (1994).

Proof. The change in $B\Omega B'/n$ is $B\left\{\left[\Sigma_y + \Sigma_h - (\Sigma_{yh} + \Sigma'_{yh})\right] - \left[k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma'_{yh})\right]\right\}B'/n$ $= (1-k)\,B[\Sigma_y - (\Sigma_{yh} + \Sigma'_{yh})]B'/n$. Element (k,k) of $B\Omega B'/n$ is $b_k \Omega b'_k/n$. □

Since the left-hand side of (3.1) is a quadratic form, the large-sample estimator with $N > n$ is variance-reducing for all $k = 1, \ldots, K$ if and only if

$$\Sigma_y - (\Sigma_{yh} + \Sigma'_{yh}) \tag{3.2}$$

is positive definite. Note that since equation (3.1) can be consistently estimated with the sample of size n, we can estimate the efficiency gain from the large-small estimator before actually implementing it.

Intuitively, potential efficiency gains are driven by a reduction in the asymptotic variance of the observed part of the moment, $\bar{y}_N$, since $\Sigma_y/N < \Sigma_y/n$. On the other hand, for given variances Σ_y/N and Σ_h/n of the observed and predicted components, respectively, the overall variance of the moment $\bar{y}_N - \bar{h}_n(\theta_0)$ increases if the covariance between the two components falls. Therefore, when the large-small estimator reduces the covariance term, $\Sigma_{yh}/N < \Sigma_{yh}/n$, it acts to increase the overall variance of the moment, everything else equal. It is the net effect of these two forces that determines whether the large-small estimator is variance-reducing. The following example explores these issues in the context of a simple linear regression model.

Example 4 – linear regression. Suppose we want to estimate the model $y = \theta x + e$, where $\mathrm{Cov}(x, e) = 0$ at the true value θ_0, so that $\theta_0 = \mathrm{Cov}(x, y)/\mathrm{Var}(x)$. Let $S_{xy}(n)$ and $S_{xy}(N)$ be the sample covariances between x and y in the small and large samples, respectively. Then the OLS estimator for the small sample is $\hat{\theta}_{n,n} = S_{xy}(n)/S_{xx}(n)$, while our estimator is $\hat{\theta}_{N,n} = S_{xy}(N)/S_{xx}(n)$. There are two opposite effects determining which of the two estimators has the smaller variance. One the one hand, $S_{xy}(N)$ is a less noisy estimate of the population covariance than $S_{xy}(n)$ is, which tends to make $\hat{\theta}_{N,n}$ less noisy. On the other hand, $S_{xx}(n)$ is more highly correlated with $S_{xy}(n)$ than with $S_{xy}(N)$, since $S_{xy}(n)$ is a function of exactly the same x_i (both directly for x and indirectly for y through $y_i = x_i\theta_0 + e_i$) that enter $S_{xx}(n)$. Therefore the noise in the numerator and denominator will cancel out to a larger extent in $\hat{\theta}_{n,n}$ than in $\hat{\theta}_{N,n}$, tending to make the ratio in the OLS estimator less noisy. The latter effect is particularly salient in the case of perfect fit, where $e_i = 0$. Then $S_{xy}(n) = \theta_0 S_{xx}(n)$, so that $\hat{\theta}_{n,n}$ has zero variance, while $\hat{\theta}_{N,n}$ still has sampling error from not using the same x_i in the observed and predicted parts of the moment. That is, sampling error in x is the only source of noise in the estimators, but it cancels between the numerator and denominator in the OLS estimator.

For a numerical illustration, suppose x and e are independent and both have a standard normal distribution. The positive-definiteness criterion is then satisfied if

$$0.7071 \approx 1/\sqrt{2} > |\theta|\,, \tag{3.3}$$

i.e. the error term e must be a relatively important source of variation in y. In this case, $G = \mathbb{E}(x^2) = \text{Var}(x) = 1$, so $B = 1$ for $W = 1$. From equation (3.1) we get a variance reduction of:[9]

$$\left(\frac{1}{n} - \frac{1}{N}\right)(1 - 2\theta^2), \tag{3.4}$$

so that if $n = 3,000$, $N = 100,000$ and $\theta = 0.15$, for instance, the variance reduction is 3.088e-4. The asymptotic variance of the moment with sample size n is $\text{Var}(xe)/n = 1/n = 3.333\text{e-}4$. It follows that the large-small estimator lowers the asymptotic standard deviation from 0.0183 for the GMM estimator to 0.0050, or by a factor of 3.7. ✓

4 Monte Carlo results

In this section we present Monte Carlo results for a linear regression, a multinomial logit model, and a random-coefficients logit model. Models 1–3 are linear regression models,

$$y_i \quad = \quad \theta x_i + e_i,$$

where both x_i and e_i are independent draws from a standard normal distribution.

Models 4–6 are multinomial logit discrete-choice models. Consumer i's conditional indirect utility u_{ij} from alternatives $j = 1, \ldots, J$, where $J = 20$, is given by

$$u_{ij} \quad = \quad \theta x_{ij} + \varepsilon_{ij}.$$

The variable x_{ij} is drawn independently from a standard normal distribution, and the shock ε_{ij} from a type-1 extreme value distribution. The variable x_{ij} mimics an observed consumer/alternative-specific variable, such as distance to a store from consumer i's home. Each observed choice is generated by one draw of the vector $(x_{i1}, \ldots, x_{iJ}, \varepsilon_{i1}, \ldots, \varepsilon_{iJ})$.

Model 7 is a panel data random-coefficients logit model. Consumer i at time $t = 1, \ldots, T$ gets conditional indirect utility u_{ijt} such that

$$u_{ijt} \quad = \quad (\beta + \sigma\nu_i + \gamma z_{it})x_{jt} + \varepsilon_{ijt}.$$

x_{jt} are independent draws from a standard normal distribution, and mimic observed product attributes. The shocks ε_{ijt} are independent draws from a type-1 extreme value distribution. The shocks ν_i, independent draws from a standard normal distribution, represent unobserved heterogeneity in the taste for the product attribute x_{jt}.[10] The variable z_{it}, independent across i, mimics an observed demographic variable with some dependence across t. It is generated as the sum of two standard uniform random variables, one of which is constant across time and one of which is independent across time. Each observed choice is generated by one

[9] See Online Appendix C for details.

[10] For simulation of the integral over the distribution of ν_i in the model's prediction, for each i we use five simulation draws.

draw of the vector $(z_{it}, \nu_i, \varepsilon_{i1t}, \ldots, \varepsilon_{iJt})$ while the product attributes $x_{1t}, \ldots, x_{Jt}$ are the same for all i. The number of alternatives is $J = 20$, time periods $T = 2$.

For the regression models 1–3, estimation is based on the moment condition $\mathbb{E}(x_i e_i) = 0$. For the multinomial logit models 4–6, the moment condition is $\mathbb{E}(\sum_j x_{ij} e_{ij}) = 0$, where $e_{ij} = \mathbb{I}_{ij} - \mathbb{P}_{ij}(\theta)$, $\mathbb{I}_{ij}$ is an indicator for whether i chose j, and $\mathbb{P}_{ij}$ is the model's choice probability for this outcome. For model 7, define the prediction error $e_{ijt} = \mathbb{I}_{ijt} - \mathbb{P}_{ijt}(\theta)$. We use the two moment conditions $\mathbb{E}[T^{-1} \sum_t \sum_j x_{jt} e_{ijt}] = 0$ and $\mathbb{E}[T^{-1} \sum_t \sum_j z_{it} x_{jt} e_{ijt}] = 0$. We add a third condition that exploits the panel structure to match the covariance of the purchase-weighted x in the two time periods:[11]

$$\mathbb{E}[(\sum_j x_{jt} \mathbb{I}_{ijt})(\sum_j x_{jt'} \mathbb{I}_{ijt'}) - (\sum_j x_{jt} \mathbb{P}_{ijt})(\sum_j x_{jt'} \mathbb{P}_{ijt'})] \quad = \quad 0,$$

We use 10,000 Monte Carlo samples of size n and $N \geq n$ respectively, where the small sample is a subset of the large sample.[12] Table 4.1 reports the sample means and sample standard deviations of estimates, as well as the sample mean of the standard errors. The results for model 1 correspond to the threshold derived in equation (3.3): no effect of changing N for a given n. Here the effects discussed in Section 3 exactly offset each other. The same is true for model 4. Models 2 and 5 have values of θ above the threshold, which means that the large sample increases the variance of the estimator. The model 2 mean standard errors for N=3,000 and N=100,000 match the asymptotic results based on equation (3.4) discussed in Section 3. In models 3, 6 and 7 the large sample reduces the variance of the estimator.

[11] The last condition is intended to identify the parameter σ. See Thomassen et al. (2017) and Berry and Haile (2021) for a discussion of this type of moment condition. As weighting matrix we use the diagonal matrix where entry (k, k) is the inverse of the square of the observed component of moment number k. This is a simple way to scale the moments in relation to each other without the need for two stages of estimation (see Low and Meghir (2017)).

[12] See Online Appendix D for details.

Table 4.1: *Monte Carlo results*

	n=3,000				
	N=3,000	N=10,000	N=30,000	N=100,000	$\frac{S.E.(\hat{\theta}_{3k,3k})}{S.E.(\hat{\theta}_{100k,3k})}$
Model 1: Linear regression ($\theta = 0.7071$)					
$\hat{\theta}$	0.7069	0.7073	0.7077	0.7077	
$S.E.(\hat{\theta})$	0.0183	0.0183	0.0183	0.0183	1.0
$S.D.(\hat{\theta})$	0.0183	0.0183	0.0184	0.0184	
Model 2: Linear regression ($\theta = 4$)					
$\hat{\theta}$	4.0001	4.0028	4.0022	4.0034	
$S.E.(\hat{\theta})$	0.0182	0.0871	0.0981	0.1018	0.2
$S.D.(\hat{\theta})$	0.0185	0.0875	0.0992	0.1009	
Model 3: Linear regression ($\theta = 0.15$)					
$\hat{\theta}$	0.1499	0.1501	0.1501	0.1501	
$S.E.(\hat{\theta})$	0.0182	0.0105	0.0068	0.0050	3.6
$S.D.(\hat{\theta})$	0.0182	0.0105	0.0069	0.0050	
Model 4: Multinomial logit ($\theta = 2.061$)					
$\hat{\theta}$	2.0615	2.0613	2.0620	2.0610	
$S.E.(\hat{\theta})$	0.0322	0.0321	0.0321	0.0321	1.0
$S.D.(\hat{\theta})$	0.0317	0.0320	0.0322	0.0322	
Model 5: Multinomial logit ($\theta = 4$)					
$\hat{\theta}$	4.0012	4.0048	4.0060	4.0103	
$S.E.(\hat{\theta})$	0.0731	0.1460	0.1610	0.1664	0.4
$S.D.(\hat{\theta})$	0.0730	0.1457	0.1625	0.1706	
Model 6: Multinomial logit ($\theta = 0.15$)					
$\hat{\theta}$	0.1502	0.1500	0.1500	0.1500	
$S.E.(\hat{\theta})$	0.0188	0.0109	0.0073	0.0054	3.5
$S.D.(\hat{\theta})$	0.0188	0.0109	0.0071	0.0054	
Model 7: R.C. logit ($\beta = 0.3$, $\sigma = 0.15$, $\gamma = 0.2$)					
$\hat{\beta}$	0.3012	0.3004	0.2999	0.3003	
$S.E.(\hat{\beta})$	0.0371	0.0265	0.0227	0.0211	1.8
$S.D.(\hat{\beta})$	0.0371	0.0267	0.0233	0.0216	
$\hat{\sigma}$	0.1667	0.1531	0.1486	0.1494	
$S.E.(\hat{\sigma})$	0.0802	0.0449	0.0267	0.0144	5.6
$S.D.(\hat{\sigma})$	0.0556	0.0374	0.0261	0.0147	
$\hat{\gamma}$	0.1998	0.1996	0.2002	0.1998	
$S.E.(\hat{\gamma})$	0.0344	0.0261	0.0232	0.0221	1.6
$S.D.(\hat{\gamma})$	0.0344	0.0264	0.0240	0.0226	

Notes: Numbers are mean parameter estimates and standard errors, and estimate standard deviations, over 10,000 simulations. The final column shows the ratio of the mean standard error when N=3,000 to the mean standard error when N=100,000.

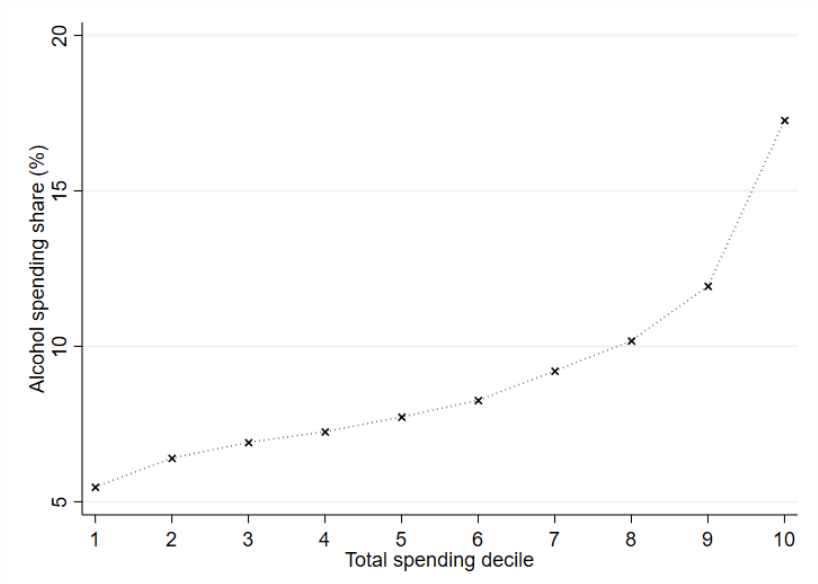
5 Empirical application

In this section we present an empirical application of the large-small estimator. We focus on a setting in which it is feasible to estimate model parameters using the full data set, and we show how the precision of estimates obtained using subsamples of different size, both using the large-small estimator and a conventional GMM estimator, compare to GMM estimates based on the full sample.

Our application entails estimating the relationship between a household's demand for alcohol and their total spending on food and drinks (both alcoholic and non-alcoholic). Figure 5.1 shows that households

with higher spending levels systematically allocate a higher spending share to alcohol. This pattern could reflect a causal income effect, i.e. exogenously raising a household's total spending leads them to allocate a higher share of spending to alcohol, or preference heterogeneity, i.e. households with higher spending have stronger permanent tastes for alcohol, or a combination of both. Understanding which is important for forecasting how alcohol consumption will change with fluctuations in income, and is relevant for the optimal design of alcohol taxation.[13]

Figure 5.1: *Alcohol budget share by spending deciles*



Notes: Based on authors' calculations using Kantar FMCG At-Home Purchase Panel. We group household-year observations into deciles based on the annual total food and drink (including alcohol) equivalized spending distribution. The markers show the average share of food and drink spending allocated to alcohol for each decile. Total spending is equivalized using the OECD-modified equivalence scale.

Data. We use household scanner data collected by the market research firm Kantar FMCG At-Home Purchase Panel. The data cover purchases of fast-moving consumer goods (food, drinks and household supplies - such as toiletries, non-prescription drugs, cleaning products and pet foods) brought into the home by a sample of households living in Great Britain. Households record purchases using handheld scanners or mobile phone apps. We aggregate the data to the annual level. Our sample covers 41,982 households (indexed $h = 1, \ldots, H$) and the period 2002-2021 (indexed $t = 1, \ldots, T$). The sample is an unbalanced panel, we use T_h to denote the number of periods household h is present in the sample. The total number of observations ($\sum_{h=1}^{H} T_h$) is 299,078.

Demand model. We specify a simple alcohol demand model in which the alcohol budget share of household h in year t, y_{ht}, is linear in the log of total (food and drink) expenditure x_{ht} (expressed in 2021 £s).

[13]For instance, if the cross-sectional pattern depicted in Figure 5.1 is driven by preference heterogeneity there is a redistributive rationale for taxing alcohol. On the other hand, if the pattern is entirely driven by income effects the case for alcohol taxation rests only on externality or internality correcting grounds (see Saez (2002) and Allcott et al. (2019)).

We control for price changes, through inclusion of time dummies, τ_t, and we control for a vector of demographics, D_{ht}, which include indicator variables for whether the household has 2 adults, 3 or more adults, 1 child, 2 children, or 3 or more children and for whether the household is located in Scotland and the year is 2018 or later (designed to capture the impact of the introduction of a minimum unit price for alcohol in Scotland).[14]

We estimate two alternative models:

1. Cross-sectional: The first ignores the panel structure of the data. The model is:

$$y_{ht} = \alpha + \alpha^D D_{ht} + \tau_t + \beta \ln(x_{ht}) + \epsilon_{ht}$$

In this case the size of the large sample is $N = \sum_{h=1}^{H} T_h (= 299,078)$ and we construct small samples by randomly drawing observations (i.e., household-periods) from the large sample.

2. Panel data: In the second model we include household fixed effects, a_h:

$$y_{ht} = a_h + \alpha^D D_{ht} + \tau_t + \beta \ln(x_{ht}) + \epsilon_{ht}$$

In this case the size of the large sample is $N = H (= 41,982)$ and we construct small samples by randomly drawing households from the large sample.

In each case we estimate the model i) including all explanatory variables in the instrument set and ii) allowing for the possibility of contemporaneous correlation between log total expenditure and the error term (by instrumenting log total expenditure with log expenditure on household supplies).

Simulations. Our aim to compare GMM estimates based on a sub- (or small) sample of a given size n with the large-small estimator (that uses large, size N, data to compute the observed component of the moment and the small sample to compute the predicted component of the moment). To do this we randomly draw with replacement a simulated large sample (of size N) from the full sample and from this we randomly draw a size n small sample. We repeat this procedure 10,000 times.

Results. We present simulation results in Table 5.1, focusing on estimates of the log total expenditure coefficient ($\hat{\beta}$).[15] The first two panels provide results for the cross-sectional model and the second two provide results for the panel data model. In each case we show estimates obtained under the assumption of (strict) exogeneity of the explanatory variables, and using log expenditure on household supplies as an instrument for log food and drink spending. Column (1) presents GMM estimates based on the full sample. Columns (2)-(6) present GMM estimates based on small samples of various sizes and columns (7)-(11) present corresponding large-small estimates. The first row of each panel shows the parameter estimates (for columns (2)-(11) means over 10,000 simulations are shown). For each model the estimate is positive,

[14] Alcohol taxes are the same throughout the UK. However in May 2018 the Scottish government introduced a price floor for alcohol which lead to price rises and falls in alcohol consumption in Scotland (see Griffith et al. (2022)).

[15] We present results for the other coefficients in the Online Appendix E. They exhibit similar patterns.

indicating that alcohol is a luxury good, though the implied total budget elasticity varies from 1.7 for the cross-sectional model estimated assuming regressor exogeneity to 1.1 for the panel data model (which controls for fixed household preference heterogeneity) estimated using the total spending instrument. The second row shows standard errors (for columns (2)-(11), means over simulations are shown) and the final row shows the standard deviation of the parameter estimates across 10,000 simulations.

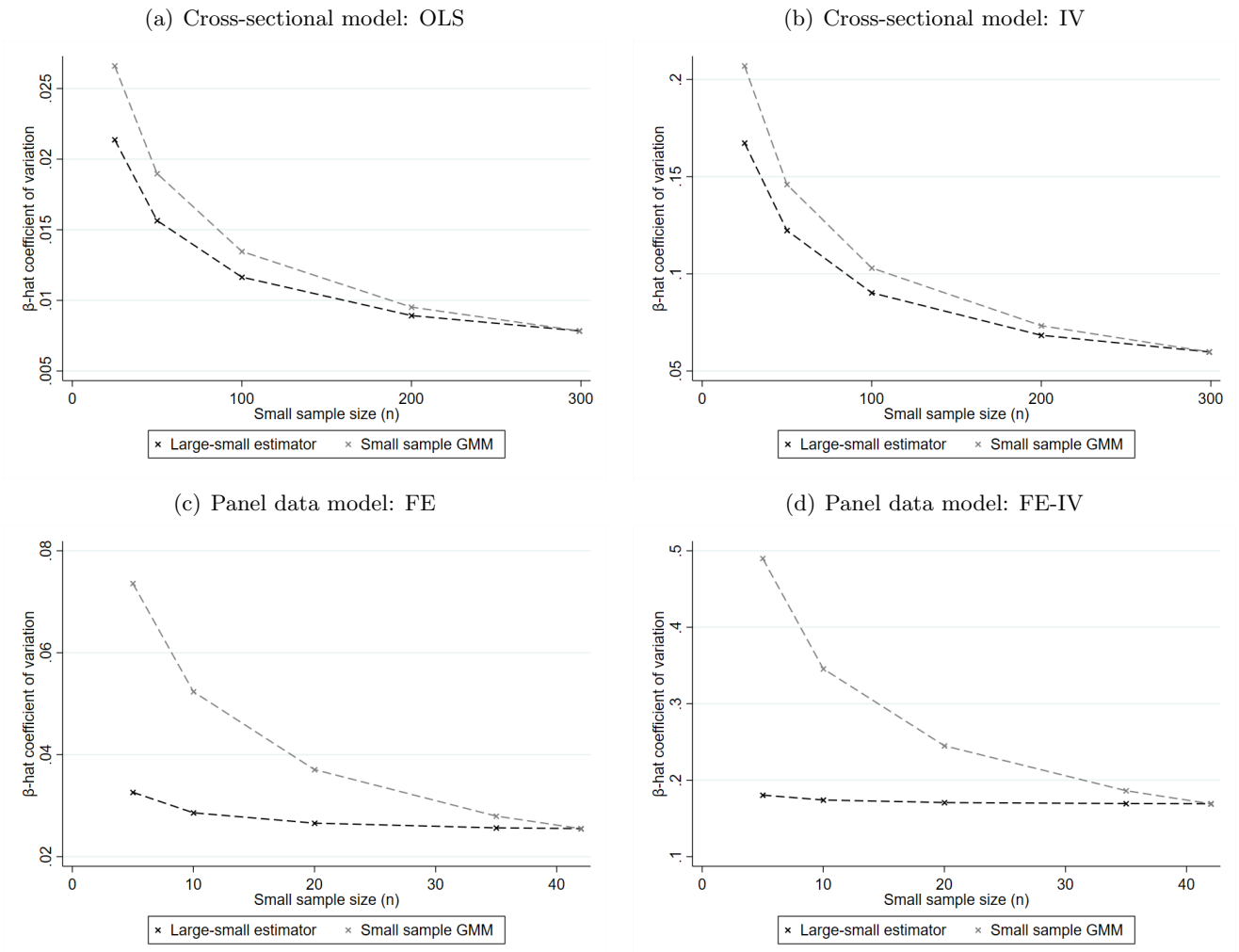
In all cases the large-small estimates are more precise than the corresponding small sample GMM estimates. We illustrate this graphically in Figure 5.2, which show how the coefficient of variation for $\hat{\beta}$ across simulations, for both GMM and large-small estimates, varies with the size of the small sample. The efficiency gains from the large-small estimator are larger the smaller the size of the small sample. For a given small sample size (aside from the limiting case in which the small and large samples coincide), the efficiency gains from the large-small estimator are largest for the panel data model. For instance, for the panel data model estimated using the spending IV, when the small size is 5,000 (under 15% the size of the large sample), the coefficient of variation of the large-small estimate is only 7% higher than that for the GMM estimate on the large sample (whereas the small sample GMM estimate has a coefficient of variation that is 2.9 times larger than the large sample GMM estimate).

Table 5.1: *Simulation results*

	$\hat{\beta}_{LL}$	$\hat{\beta}_{SS}$					$\hat{\beta}_{LS}$				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
Cross-sectional model											
	$N =$ 299k	$n =$ 25k	$n =$ 50k	$n =$ 100k	$n =$ 200k	$n =$ 294k	$n =$ 25k	$n =$ 50k	$n =$ 100k	$n =$ 200k	$n =$ 299k
OLS											
$\hat{\beta}$	6.385	6.385	6.385	6.385	6.385	6.385	6.393	6.388	6.386	6.386	6.385
$S.E.(\hat{\beta})$	0.049	0.171	0.121	0.085	0.060	0.049	0.137	0.099	0.073	0.056	0.049
$S.D.(\hat{\beta})$		0.170	0.121	0.086	0.061	0.050	0.137	0.100	0.074	0.057	0.050
IV											
$\hat{\beta}$	1.415	1.416	1.417	1.416	1.416	1.416	1.418	1.414	1.415	1.417	1.416
$S.E.(\hat{\beta})$	0.084	0.291	0.206	0.145	0.103	0.084	0.237	0.171	0.126	0.096	0.084
$S.D.(\hat{\beta})$		0.293	0.207	0.146	0.104	0.085	0.237	0.173	0.128	0.097	0.085
Panel data model											
	$N =$ 42k	$n =$ 5k	$n =$ 10k	$n =$ 20k	$n =$ 35k	$n =$ 42k	$n =$ 5k	$n =$ 10k	$n =$ 20k	$n =$ 35k	$n =$ 42k
FE											
$\hat{\beta}$	3.777	3.779	3.777	3.776	3.776	3.777	3.785	3.780	3.778	3.777	3.777
$S.E.(\hat{\beta})$	0.096	0.278	0.197	0.139	0.105	0.096	0.138	0.116	0.103	0.097	0.096
$S.D.(\hat{\beta})$		0.278	0.198	0.140	0.106	0.096	0.123	0.108	0.100	0.097	0.096
FE-IV											
$\hat{\beta}$	0.892	0.897	0.892	0.892	0.891	0.891	0.894	0.893	0.892	0.892	0.891
$S.E.(\hat{\beta})$	0.150	0.434	0.307	0.217	0.164	0.150	0.161	0.155	0.152	0.150	0.150
$S.D.(\hat{\beta})$		0.439	0.308	0.218	0.166	0.151	0.161	0.155	0.152	0.151	0.151

Notes: $\hat{\beta}_{LL}$ denotes GMM estimates with the full sample. $\hat{\beta}_{SS}$ denotes GMM estimates with a sub-sample (i.e. the small sample). $\hat{\beta}_{LS}$ denotes denotes large-small estimates. The small sample size is displayed in the second row. For columns (2)-(11), rows 1 and 2 of each panel report means across 10,000 simulations. In the panel data model standard errors are clustered at the household level.

Figure 5.2: *Precision of* $\hat{\beta}$



Notes: Each marker shows the ratio of the standard deviation to the mean of the estimate across simulations. The standard deviations and means are reported in Tables 5.1.

6 Conclusion

In this paper we derive the asymptotic properties of a large/small-sample GMM-type estimator. The estimator will be of use in situations in which researchers have access to large datasets and wish to estimate a computationally intensive model (involving, for instance, nested optimization or fixed point problems). Standard practice entails using a subset of the data for estimation. In contrast, rather than discarding completely the data not in the feasible subsample, our estimator brings the information in the full data to bear by using a large sample for the observed part of the moment (which only needs to be computed once). We show that if a simple criterion, that can be checked with the small sample only, is satisfied then using the large-small estimator will be more efficient than GMM using the small sample. We verify the estimator's properties for different models with Monte Carlo studies and simulations with real data.

Appendix

A Proof of Proposition 1

We maintain the assumptions in footnote 2, here referred to as Assumption 1. We first prove two lemmas that we use for the main proof.

Lemma 1. Let f_n and f be vector-valued functions such that $\sup_{\theta\in\Theta}\|f_n(\theta)-f(\theta)\| \xrightarrow{p} 0$ and x_n and μ vectors such that $\|x_n-\mu\| \xrightarrow{p} 0$. Suppose $N \geq n$. Then as $n\to\infty$,

$$\sup_{\theta\in\Theta}\|x_N - f_n(\theta) - [\mu - f(\theta)]\| \xrightarrow{p} 0$$

Proof. First note that for all N and n,

$$\begin{aligned}\|x_N-\mu\| + \sup_\theta\|f_n(\theta)-f(\theta)\| &= \sup_\theta\{\|x_N-\mu\| + \|f_n(\theta)-f(\theta)\|\} \\ &\geq \sup_\theta\{\|x_N-\mu-[f_n(\theta)-f(\theta)]\|\}, \end{aligned} \tag{A.1}$$

where the last line follows from the triangle inequality. By assumption, for any $\varepsilon>0$ and any $\delta>0$ there exists n_1 such that for all $n\geq n_1$,

$$\Pr\left(\sup_{\theta\in\Theta}\|f_n(\theta)-f(\theta)\| > \frac{\varepsilon}{2}\right) < \frac{\delta}{2}, \tag{A.2}$$

and n_2 such that for all $N\geq n_2$,

$$\Pr\left(\|x_N-\mu\| > \frac{\varepsilon}{2}\right) < \frac{\delta}{2}. \tag{A.3}$$

Then for $n_0=\max\{n_1,n_2\}$, $n>n_0$ implies

$$\begin{aligned}\Pr\left(\sup_\theta\|x_N-\mu-[f_n(\theta)-f(\theta)]\|>\varepsilon\right) &\leq \Pr\left(\|x_N-\mu\| + \sup_{\theta\in\Theta}\|f_n(\theta)-f(\theta)\|>\varepsilon\right) \\ &\leq \Pr\left(\|x_N-\mu\|>\frac{\varepsilon}{2} \text{ or } \sup_{\theta\in\Theta}\|f_n(\theta)-f(\theta)\|>\frac{\varepsilon}{2}\right) \\ &\leq \Pr\left(\|x_N-\mu\|>\frac{\varepsilon}{2}\right) + \Pr\left(\sup_{\theta\in\Theta}\|f_n(\theta)-f(\theta)\|>\frac{\varepsilon}{2}\right) \\ &< \delta.\end{aligned}$$

□

Lemma 2. Let $\hat{g}_n(\theta)=\bar{y}_N-\bar{h}_n(\theta)$ and $g_0(\theta)=\mathbb{E}[y_i-h_i(\theta)]$. Then (i) $g_0(\theta)$ is continuous and (ii) $\sup_{\theta\in\Theta}\|\hat{g}_n(\theta)-g_0(\theta)\| \xrightarrow{p} 0$.

Proof. We verify the hypotheses of Lemma 2.4 in Newey and McFadden (1994).[16] Recall that $h_i(\theta)=h(w_i,\theta)$. We have assumed that w_i are i.i.d., Θ is compact, and $h(w_i,\theta)$ is continuous at each $\theta\in\Theta$ with probability one.

Define $d(w_i)=\sup_{\theta\in\Theta}\|h(w_i,\theta)\|$. Then we have $\|h(w_i,\theta)\|\leq d(w_i)$ for all $\theta\in\Theta$. By Assumption 1(iv), $\mathbb{E}[d(w_i)] = \mathbb{E}\left[\sup_{\theta\in\Theta}\|h(w_i,\theta)\|\right]<\infty$. Then by Lemma 2.4 in Newey and McFadden (1994), $\mathbb{E}[h(w_i,\theta)]$ is continuous and

$$\sup_{\theta\in\Theta}\left\|\bar{h}_n(\theta)-\mathbb{E}[h(w_i,\theta)]\right\| = \sup_{\theta\in\Theta}\left\|n^{-1}\sum_{i\in I_n}h(w_i,\theta)-\mathbb{E}[h(w_i,\theta)]\right\| \xrightarrow{p} 0.$$

Continuity of $\mathbb{E}[h(w_i,\theta)]$ implies continuity of $g_0(\theta)=\mathbb{E}[y_i]-\mathbb{E}[h(w_i,\theta)]$. This completes the proof of part (i) of the Lemma.

In Lemma 1, let $f_n(\theta)=\bar{h}_n(\theta)$, $f(\theta)=\mathbb{E}[h(w_i,\theta)]$, $x_N=\bar{y}_N$ and $\mu=\mathbb{E}[y_i]$. Since the assumptions of Lemma 1 are satisfied, it follows that

$$\sup_{\theta\in\Theta}\|\hat{g}_n(\theta)-g_0(\theta)\| = \sup_{\theta\in\Theta}\left\|\bar{y}_N-\bar{h}_n(\theta)-\mathbb{E}[y_i-h(w_i,\theta)]\right\| \xrightarrow{p} 0.$$

[16] Our h is their a, and our w_i is their z.

This completes the proof of part (ii) of the lemma. □

We can now prove Proposition 1.

Proof. We let $\hat{g}_n(\theta) = \bar{y}_N - \bar{h}_n(\theta)$ and $g_0(\theta) = \mathbb{E}[y_i - h(w_i, \theta)]$. We also define $Q_0(\theta) = -g_0(\theta)' W g_0(\theta)$ and $\hat{Q}_n(\theta) = -\hat{g}_n(\theta)' \hat{W} \hat{g}_n(\theta)$.

We proceed by verifying the hypotheses of Theorem 2.1 in Newey and McFadden (1994).[17] Their condition (i), that $Q_0(\theta)$ be uniquely maximized at θ_0, follows by our Assumption 1(i) and Lemma 2.3 (GMM identification) in Newey and McFadden (1994). Their condition (ii), that Θ be compact, is our Assumption 1(ii).

By Lemma 2, $g_0(\theta)$ is continuous and $\sup_{\theta \in \Theta} \|\hat{g}_n(\theta) - g_0(\theta)\| \xrightarrow{p} 0$. We use these facts to verify the remaining conditions in Newey and McFadden's Theorem 2.1. Their condition (iii), that $Q_0(\theta) = g_0(\theta)' W g(\theta)$ be continuous, follows from the continuity of $g_0(\theta)$.

By Θ compact, $g_0(\theta)$ continuous and the extreme value theorem, $g_0(\theta)$ is bounded on Θ. By the triangle and Cauchy-Schwartz inequalities,

$$\begin{aligned} |\hat{Q}_n(\theta) - Q_0(\theta)| &\leq |[\hat{g}_n(\theta) - g_0(\theta)]' \hat{W} [\hat{g}_n(\theta) - g_0(\theta)]| + |g_0(\theta)' (\hat{W} + \hat{W}') [\hat{g}_n(\theta) - g_0(\theta)]| \\ &\quad + |g_0(\theta)' (\hat{W} - W) g_0(\theta)| \\ &\leq \|\hat{g}_n(\theta) - g_0(\theta)\|^2 \left\|\hat{W}\right\| + 2 \|g_0(\theta)\| \|\hat{g}_n(\theta) - g_0(\theta)\| \left\|\hat{W}\right\| \\ &\quad + \|g_0(\theta)\|^2 \left\|\hat{W} - W\right\|. \end{aligned}$$

Then condition (iv), that $\sup_{\theta \in \Theta} |\hat{Q}_n(\theta) - Q_0(\theta)| \xrightarrow{p} 0$, holds by $\sup_{\theta \in \Theta} \|\hat{g}_n(\theta) - g_0(\theta)\| \xrightarrow{p} 0$ and our assumption that $\hat{W} \xrightarrow{p} W$. □

References

Allcott, H., B. B. Lockwood, and D. Taubinsky (2019). Regressive Sin Taxes, with an Application to the Optimal Soda Tax. *Quarterly Journal of Economics 134*(3), 1557–1626.

Angrist, J. D. and A. B. Krueger (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association 87*(418), 328–336.

Berry, S. T. and P. A. Haile (2021). Chapter 1 - foundations of demand estimation. In K. Ho, A. Hortaçsu, and A. Lizzeri (Eds.), *Handbook of Industrial Organization, Volume 4*, Volume 4 of *Handbook of Industrial Organization*, pp. 1–62. Elsevier.

Busse, M. R., C. R. Knittel, and F. Zettelmeyer (2013). Are consumers myopic? evidence from new and used car purchases. *The American Economic Review 103*(1), 220–256.

Griffith, R., M. O'Connell, and K. Smith (2022). Price floors and externality correction. *The Economic Journal 132*, 2273–2289.

Hansen, B. E. (2022). *Econometrics*. Princeton University Press.

Imbens, G. and T. Lancaster (1994). Combining micro and macro data in microeconometric models. *Review of Economic Studies 61*(4), 655–680.

Inoue, A. and G. Solon (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics 92*(3), 557–561.

Lee, S. and S. Ng (2020). An econometric perspective on algorithmic subsampling. *Annual Review of Economics 12*(1), 45–80.

Low, H. and C. Meghir (2017). The use of structural models in econometrics. *Journal of Economic Perspectives 31*(2), 33–58.

[17] We follow the proof of Theorem 2.6 in Newey and McFadden (1994), but adapt it to the case of a minimum distance estimator.

Newey, W. K. and D. McFadden (1994). Chapter 36 large sample estimation and hypothesis testing. Volume 4 of *Handbook of Econometrics*, pp. 2111–2245. Elsevier.

Pacini, D. and F. Windmeijer (2016). Robust inference for the two-sample 2sls estimator. *Economics Letters 146*, 50–54.

Saez, E. (2002). The desirability of commodity taxation under non-linear income taxation and heterogeneous tastes. *Journal of Public Economics 83*(2), 217–230.

Thomassen, Ø., H. Smith, S. Seiler, and P. Schiraldi (2017). Multi-category competition and market power: A model of supermarket pricing. *American Economic Review 107*(8), 2308–51.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Appendix: For online publication

A two sample size estimator for large data sets

Martin O'Connell, Howard Smith and Øyvind Thomassen
February, 2023

A The covariance between the observed and predicted moments

We make use of the shorthand notation $a_i = y_i - \mu_y$ and $b_i = h_i(\theta_0) - \mu_h$. Then $\mathbb{E}(a_i) = 0$, $\mathbb{E}(b_i) = 0$, and, because of independence between households i, j, $\mathbb{E}(a_i b_j') = 0$ for $i \neq j$. We have $\bar{y}_N - \mu_y = \frac{1}{N}\sum_{i\in I_N} y_i - \mu_y = \frac{1}{N}\sum_{i\in I_N} a_i$ and $\bar{h}_n - \mu_h = \frac{1}{n}\sum_{i\in I_n} h_i(\theta_0) - \mu_h = \frac{1}{n}\sum_{i\in I_n} b_i$. Then:

$$
\begin{aligned}
&\text{Cov}\left[\sqrt{n}(\bar{y}_N - \mu_y), \sqrt{n}(\bar{h}_n - \mu_h)\right] \\
&= n\mathbb{E}\left[\left(\frac{1}{N}\sum_{i\in I_N} a_i\right)\left(\frac{1}{n}\sum_{i\in I_n} b_i\right)'\right] = \frac{1}{N}\mathbb{E}\left[\left(\sum_{i\in I_N} a_i\right)\left(\sum_{i\in I_n} b_i'\right)\right] \\
&= \frac{1}{N}\mathbb{E}\left[\left(\sum_{i\in I_n} a_i + \sum_{i\in I_N\setminus I_n} a_i\right)\left(\sum_{i\in I_n} b_i'\right)\right] = \frac{1}{N}\mathbb{E}\left(\sum_{i\in I_n} a_i b_i' + \sum_{i\in I_n}\sum_{j\in I_n\setminus\{i\}} a_i b_j' + \sum_{i\in I_N\setminus I_n}\sum_{j\in I_n} a_i b_j'\right) \\
&= \frac{1}{N}\left(\sum_{i\in I_n}\mathbb{E}(a_i b_i') + \sum_{i\in I_n}\sum_{j\in I_n\setminus\{i\}}\mathbb{E}(a_i b_j') + \sum_{i\in I_N\setminus I_n}\sum_{j\in I_n}\mathbb{E}(a_i b_j')\right) = \frac{1}{N}\sum_{i\in I_n}\mathbb{E}(a_i b_i') \\
&= \frac{1}{N} n\mathbb{E}(a_i b_i') = \frac{n}{N}\mathbb{E}\left[(y_i - \mu_y)\left(h_i(\theta_0) - \mu_h\right)'\right] = \frac{n}{N}\text{Cov}\left(y_i, h_i(\theta_0)\right).
\end{aligned}
$$

B Non-smoothness

In some cases assumption (vii) in footnote 6 may not hold. The following set of alternative assumptions are also sufficient for asymptotic normality.

Assumption 1. Let $\hat{g}_n(\theta)$ be defined by equation (2.6) and let $g_0(\theta) = \mathbb{E}[y_{it} - h(w_{it}, \theta)]$. Then

(a) $\theta_0 \in \text{interior}(\Theta)$

(b) $\hat{g}_n(\hat{\theta})'\hat{W}\hat{g}_n(\hat{\theta}) \leq \hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta) + o_p(n^{-1})$

(c) $g_0(\theta)$ is differentiable at θ_0 with derivative G such that $G'WG$ is nonsingular.

(d) For any $\delta_n \to 0$, $\sup_{\|\theta-\theta_0\|<\delta_n} \frac{\sqrt{n}\|\hat{g}_n(\theta)-\hat{g}_n(\theta_0)-g_0(\theta)\|}{1+\sqrt{n}\|\theta-\theta_0\|} \xrightarrow{p} 0$.

Proposition 8. The estimator (2.5) satisfies:

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, B\Omega B'), \tag{B.1}$$

where $B = (G'WG)^{-1}G'W$ and $\Omega = k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma_{yh}')$.

Proof. Newey and McFadden (1994) Theorem 7.2. □

C Proof of condition (3.3)

We have

$$\begin{aligned}\Sigma_y &= \text{Var}(xy) = \text{Var}(x(\theta x + e)) = \theta^2 \text{Var}(x^2) + \text{Var}(xe) + 2\theta \text{Cov}(x^2, xe) \\ \Sigma_{yh} &= \text{Cov}(xy, \theta x^2) = \text{Cov}(x(\theta x + e), \theta x^2) = \theta^2 \text{Var}(x^2) + \theta \text{Cov}(x^2, xe)\end{aligned}$$

Under the mean zero and independence assumptions, $\text{Var}(xe) = \text{Var}(x)\text{Var}(e)$. This can be seen from the general formula

$$\text{Var}(xe) = \text{Cov}(x^2, e^2) + [\text{Var}(x) + (\mathbb{E}x)^2][\text{Var}(e) + (\mathbb{E}e)^2] - [\text{Cov}(x, e) + \mathbb{E}x\mathbb{E}e]^2.$$

By equation (3.2) the variance (for a given n) will be reduced by using a larger sample for the observed component of the moment, if and only if

$$0 > 2\Sigma_{yh} - \Sigma_y = \theta^2 \text{Var}(x^2) - \text{Var}(x)\text{Var}(e) \Leftrightarrow \text{Var}(e) > \theta^2 \frac{\text{Var}(x^2)}{\text{Var}(x)}.$$

Since x is $N(0,1)$, $x^2 \sim \chi^2(1)$ (variance 2). The criterion for the large/small-sample estimator to reduce variance (relative to the OLS estimator using sample size n for both the observed and predicted components) is then

$$1 > \theta^2 2 \Leftrightarrow 0.7071 \approx \frac{1}{\sqrt{2}} > |\theta|.$$

D Details about Monte Carlo simulations

For all models objective functions were minimized numerically with Matlab's quasi-Newton `fminunc` with objective and step tolerances set to 1e-6. Starting values for all parameters were drawn from $U(1,2)$. For model 7, if the resulting minimum of the objective was greater than 1e-9, we did a second optimization run with the true parameter values as starting values. We then kept the results that gave the lowest value for the objective. To avoid misleading results from optimization failures, we dropped cases where the objective function was greater than 1e-5 at both optimization runs. This happened in 14.7 percent of simulations for $N = 3,000$, 4.5 percent for $N = 10,000$, 0.72 percent for $N = 50,000$, and 0.45 percent for $N = 100,000$. To see that the discarded runs were outliers: the 95th, 75th, 50th and 25th percentiles of the objective at the minimum are, respectively, 5.887e-10, 6.064e-11, 7.019e-12, and 5.116e-13. Among the remaining simulation runs, for each of the three parameters, there were outliers among the standard errors, which might have been caused by numerical error in the (finite-difference) calculation of moment derivatives. We therefore dropped standard errors (and the corresponding parameter estimates) that were above the 95th percentile of standard errors in each of the six parameter / sample-size combinations where N=3,000 or N=10,000; and above the 99th percentile of standard errors in each of the six parameter / sample-size combinations where N=30,000 or N=100,000. Table D.1 shows the Monte Carlo simulations for model 7 without this last step. Standard deviations of estimates are broadly unchanged, while mean standard errors are now very different because of the outliers. Table D.2 shows results with the same 95th and 99th percentile cutoffs as in the paper, but where we have used 100 simulation draws to simulate the distribution of ν_i instead of 5 as in the main table. On the other hand, we used only 100 simulated sample to generate Table D.2.

Table D.1: *Monte Carlo results - with outlier standard errors*

	n=3,000			
	N=3,000	N=10,000	N=30,000	N=100,000
Model 7: R.C. logit ($\beta = 0.3$, $\sigma = 0.15$, $\gamma = 0.2$)				
$\hat{\beta}$	0.3012	0.3004	0.3000	0.3003
$S.E.(\hat{\beta})$	1.0125	0.4178	0.1497	0.0213
$S.D.(\hat{\beta})$	0.0375	0.0274	0.0237	0.0220
$\hat{\sigma}$	0.1594	0.1474	0.1475	0.1489
$S.E.(\hat{\sigma})$	56.7494	26.6544	9.3956	0.0254
$S.D.(\hat{\sigma})$	0.0631	0.0446	0.0283	0.0155
$\hat{\gamma}$	0.1999	0.1999	0.2001	0.1998
$S.E.(\hat{\gamma})$	0.9350	0.4137	0.1448	0.0223
$S.D.(\hat{\gamma})$	0.0346	0.0271	0.0243	0.0229

Notes: Numbers are mean parameter estimates and standard errors, and estimate standard deviations, over 10,000 simulations.

Table D.2: *Monte Carlo results - simulation draws increased from 5 to 100*

	n=3,000			
	N=3,000	N=10,000	N=30,000	N=100,000
Model 7: R.C. logit ($\beta = 0.3$, $\sigma = 0.15$, $\gamma = 0.2$)				
$\hat{\beta}$	0.2919	0.3004	0.3004	0.2967
$S.E.(\hat{\beta})$	0.0376	0.0267	0.0229	0.0208
$S.D.(\hat{\beta})$	0.0399	0.0274	0.0223	0.0205
$\hat{\sigma}$	0.1594	0.1568	0.1496	0.1486
$S.E.(\hat{\sigma})$	0.0908	0.0433	0.0261	0.0139
$S.D.(\hat{\sigma})$	0.0538	0.0330	0.0264	0.0165
$\hat{\gamma}$	0.2091	0.1996	0.2000	0.2038
$S.E.(\hat{\gamma})$	0.0350	0.0264	0.0235	0.0218
$S.D.(\hat{\gamma})$	0.0369	0.0276	0.0226	0.0213

Notes: Numbers are mean parameter estimates and standard errors, and estimate standard deviations, over 100 simulations, using 100 simulation draws per i to simulate the distribution of ν_i.

E Empirical application: full simulation results

Table E.1: *Cross-sectional model: OLS*

	$\hat{\theta}_{LL}$	$\hat{\theta}_{SS}$					$\hat{\theta}_{LS}$				
	$N =$ 299k	$n =$ 25k	$n =$ 50k	$n =$ 100k	$n =$ 200k	$n =$ 294k	$n =$ 25k	$n =$ 50k	$n =$ 100k	$n =$ 200k	$n =$ 299k
$\bar{\hat{\theta}}$											
Constant	-38.285	-38.285	-38.286	-38.287	-38.282	-38.284	-38.326	-38.300	-38.291	-38.288	-38.284
No. adults=2	-1.593	-1.590	-1.591	-1.592	-1.593	-1.594	-1.600	-1.596	-1.595	-1.594	-1.594
No. adults=3	-2.837	-2.836	-2.837	-2.837	-2.838	-2.838	-2.842	-2.840	-2.839	-2.839	-2.838
No. adults>3	-4.267	-4.267	-4.267	-4.266	-4.268	-4.268	-4.274	-4.270	-4.268	-4.269	-4.268
No. kids=1	-3.384	-3.383	-3.382	-3.383	-3.384	-3.384	-3.383	-3.383	-3.384	-3.384	-3.384
No. kids=2	-5.167	-5.168	-5.165	-5.167	-5.167	-5.167	-5.169	-5.169	-5.167	-5.167	-5.167
No. kids>2	-6.863	-6.862	-6.863	-6.864	-6.863	-6.863	-6.865	-6.864	-6.863	-6.863	-6.863
Scotland-post 2017	0.932	0.934	0.929	0.932	0.933	0.933	0.962	0.947	0.936	0.935	0.933
Log expenditure	6.385	6.385	6.385	6.385	6.385	6.385	6.393	6.388	6.386	6.386	6.385
$S.E.(\hat{\theta})$											
Constant	0.378	1.308	0.925	0.654	0.462	0.378	1.075	0.776	0.571	0.434	0.378
No. adults=2	0.065	0.223	0.158	0.112	0.079	0.065	0.184	0.133	0.098	0.074	0.065
No. adults=3	0.081	0.280	0.198	0.140	0.099	0.081	0.236	0.170	0.124	0.094	0.081
No. adults>3	0.087	0.300	0.212	0.150	0.106	0.087	0.252	0.182	0.133	0.100	0.087
No. kids=1	0.055	0.192	0.136	0.096	0.068	0.055	0.172	0.123	0.089	0.065	0.055
No. kids=2	0.051	0.178	0.126	0.089	0.063	0.051	0.158	0.113	0.082	0.060	0.051
No. kids>2	0.062	0.214	0.151	0.107	0.076	0.062	0.185	0.133	0.097	0.072	0.062
Scotland-post 2017	0.164	0.568	0.401	0.284	0.201	0.164	0.519	0.370	0.266	0.195	0.164
Log expenditure	0.049	0.171	0.121	0.085	0.060	0.049	0.137	0.099	0.073	0.056	0.049
$S.D.(\hat{\theta})$											
Constant		1.299	0.926	0.660	0.468	0.383	1.070	0.782	0.578	0.440	0.383
No. adults=2		0.226	0.157	0.111	0.078	0.065	0.186	0.134	0.100	0.075	0.065
No. adults=3		0.281	0.197	0.139	0.099	0.082	0.238	0.171	0.126	0.095	0.082
No. adults>3		0.302	0.213	0.149	0.106	0.087	0.252	0.184	0.134	0.101	0.087
No. kids=1		0.191	0.136	0.095	0.068	0.055	0.174	0.122	0.088	0.065	0.055
No. kids=2		0.176	0.125	0.089	0.062	0.051	0.158	0.113	0.081	0.060	0.051
No. kids>2		0.216	0.152	0.108	0.077	0.063	0.184	0.133	0.097	0.073	0.063
Scotland-post 2017		0.565	0.399	0.283	0.199	0.162	0.517	0.367	0.262	0.191	0.162
Log expenditure		0.170	0.121	0.086	0.061	0.050	0.137	0.100	0.074	0.057	0.050

Table E.2: *Cross-sectional model: IV*

	$\hat{\theta}_{LL}$	$\hat{\theta}_{SS}$					$\hat{\theta}_{LS}$				
	$N =$ 299k	$n =$ 25k	$n =$ 50k	$n =$ 100k	$n =$ 200k	$n =$ 294k	$n =$ 25k	$n =$ 50k	$n =$ 100k	$n =$ 200k	$n =$ 299k
$\bar{\hat{\theta}}$											
Constant	-1.726	-1.734	-1.740	-1.736	-1.729	-1.732	-1.729	-1.713	-1.725	-1.737	-1.732
No. adults=2	0.736	0.739	0.737	0.737	0.736	0.735	0.732	0.735	0.735	0.734	0.735
No. adults=3	0.113	0.113	0.112	0.113	0.111	0.111	0.111	0.113	0.111	0.111	0.111
No. adults>3	-1.021	-1.021	-1.022	-1.020	-1.022	-1.022	-1.024	-1.021	-1.022	-1.023	-1.022
No. kids=1	-2.842	-2.842	-2.841	-2.842	-2.842	-2.842	-2.841	-2.841	-2.842	-2.843	-2.842
No. kids=2	-4.010	-4.011	-4.008	-4.010	-4.010	-4.010	-4.010	-4.010	-4.010	-4.010	-4.010
No. kids>2	-5.299	-5.299	-5.300	-5.301	-5.301	-5.300	-5.301	-5.300	-5.300	-5.300	-5.300
Scotland-post 2017	0.979	0.980	0.976	0.979	0.981	0.980	1.008	0.994	0.983	0.982	0.980
Log expenditure	1.415	1.416	1.417	1.416	1.416	1.416	1.418	1.414	1.415	1.417	1.416
$S.E.(\hat{\theta})$											
Constant	0.636	2.198	1.554	1.099	0.777	0.635	1.789	1.293	0.954	0.728	0.635
No. adults=2	0.072	0.248	0.175	0.124	0.088	0.072	0.198	0.144	0.106	0.082	0.072
No. adults=3	0.090	0.310	0.219	0.155	0.110	0.090	0.254	0.184	0.135	0.103	0.090
No. adults>3	0.096	0.330	0.234	0.165	0.117	0.096	0.272	0.196	0.145	0.110	0.096
No. kids=1	0.056	0.194	0.137	0.097	0.069	0.056	0.165	0.119	0.087	0.065	0.056
No. kids=2	0.053	0.182	0.129	0.091	0.064	0.053	0.156	0.112	0.082	0.061	0.053
No. kids>2	0.064	0.220	0.156	0.110	0.078	0.064	0.181	0.130	0.096	0.073	0.064
Scotland-post 2017	0.170	0.586	0.414	0.293	0.207	0.170	0.507	0.363	0.264	0.197	0.170
Log expenditure	0.084	0.291	0.206	0.145	0.103	0.084	0.237	0.171	0.126	0.096	0.084
$S.D.(\hat{\theta})$											
Constant		2.215	1.567	1.103	0.784	0.639	1.788	1.305	0.965	0.732	0.639
No. adults=2		0.251	0.174	0.122	0.087	0.071	0.200	0.145	0.108	0.082	0.071
No. adults=3		0.313	0.218	0.153	0.109	0.090	0.255	0.185	0.136	0.104	0.090
No. adults>3		0.336	0.235	0.164	0.117	0.096	0.271	0.199	0.146	0.111	0.096
No. kids=1		0.193	0.137	0.096	0.069	0.056	0.168	0.119	0.087	0.065	0.056
No. kids=2		0.182	0.129	0.091	0.064	0.052	0.155	0.111	0.081	0.061	0.052
No. kids>2		0.223	0.157	0.111	0.079	0.065	0.179	0.130	0.097	0.074	0.065
Scotland-post 2017		0.584	0.412	0.292	0.206	0.168	0.504	0.360	0.261	0.194	0.168
Log expenditure		0.293	0.207	0.146	0.104	0.085	0.237	0.173	0.128	0.097	0.085

Table E.3: *Panel data model: Fixed effects*

	$\hat{\theta}_{LL}$	$\hat{\theta}_{SS}$					$\hat{\theta}_{LS}$				
	$N =$ 42k	$n =$ 5k	$n =$ 10k	$n =$ 20k	$n =$ 35k	$n =$ 42k	$n =$ 5k	$n =$ 10k	$n =$ 20k	$n =$ 35k	$n =$ 42k
$\bar{\hat{\theta}}$											
No. adults=2	-0.514	-0.511	-0.511	-0.511	-0.512	-0.513	-0.512	-0.513	-0.512	-0.512	-0.513
No. adults=3	-0.550	-0.545	-0.545	-0.547	-0.548	-0.549	-0.549	-0.550	-0.549	-0.549	-0.549
No. adults>3	-0.855	-0.852	-0.855	-0.855	-0.856	-0.856	-0.857	-0.856	-0.855	-0.856	-0.856
No. kids=1	-1.587	-1.589	-1.589	-1.590	-1.590	-1.589	-1.595	-1.592	-1.590	-1.590	-1.589
No. kids=2	-2.470	-2.470	-2.472	-2.471	-2.472	-2.471	-2.482	-2.476	-2.473	-2.472	-2.471
No. kids>2	-3.217	-3.221	-3.221	-3.217	-3.219	-3.218	-3.236	-3.226	-3.220	-3.219	-3.218
Scotland-post 2017	-0.157	-0.160	-0.160	-0.159	-0.157	-0.157	-0.154	-0.156	-0.157	-0.157	-0.157
Log expenditure	3.777	3.779	3.777	3.776	3.776	3.777	3.785	3.780	3.778	3.777	3.777
$S.E.(\hat{\theta})$											
No. adults=2	0.156	0.449	0.318	0.226	0.171	0.156	0.207	0.180	0.165	0.158	0.156
No. adults=3	0.199	0.575	0.407	0.288	0.218	0.199	0.271	0.233	0.211	0.201	0.199
No. adults>3	0.257	0.743	0.527	0.373	0.282	0.257	0.352	0.302	0.273	0.260	0.257
No. kids=1	0.103	0.297	0.211	0.149	0.113	0.103	0.142	0.122	0.110	0.104	0.103
No. kids=2	0.141	0.407	0.289	0.204	0.155	0.141	0.197	0.168	0.151	0.143	0.141
No. kids>2	0.207	0.596	0.423	0.299	0.227	0.207	0.299	0.251	0.223	0.210	0.207
Scotland-post 2017	0.162	0.466	0.331	0.234	0.177	0.162	0.194	0.176	0.167	0.163	0.162
Log expenditure	0.096	0.278	0.197	0.139	0.105	0.096	0.138	0.116	0.103	0.097	0.096
$S.D.(\hat{\theta})$											
No. adults=2		0.453	0.320	0.227	0.171	0.156	0.209	0.181	0.165	0.158	0.156
No. adults=3		0.577	0.408	0.289	0.218	0.199	0.271	0.232	0.211	0.202	0.199
No. adults>3		0.739	0.522	0.372	0.281	0.257	0.354	0.301	0.273	0.260	0.257
No. kids=1		0.300	0.214	0.152	0.113	0.103	0.140	0.121	0.110	0.104	0.103
No. kids=2		0.411	0.290	0.205	0.154	0.140	0.192	0.164	0.149	0.142	0.140
No. kids>2		0.605	0.424	0.299	0.226	0.206	0.291	0.247	0.221	0.209	0.206
Scotland-post 2017		0.470	0.335	0.235	0.179	0.163	0.193	0.176	0.168	0.164	0.163
Log expenditure		0.278	0.198	0.140	0.106	0.096	0.123	0.108	0.100	0.097	0.096

Table E.4: *Panel data model: Fixed effects-IV*

	$\hat{\theta}_{LL}$	$\hat{\theta}_{SS}$					$\hat{\theta}_{LS}$				
	$N =$ 42k	$n =$ 5k	$n =$ 10k	$n =$ 20k	$n =$ 35k	$n =$ 42k	$n =$ 5k	$n =$ 10k	$n =$ 20k	$n =$ 35k	$n =$ 42k
$\bar{\hat{\theta}}$											
No. adults=2	-0.120	-0.117	-0.117	-0.117	-0.118	-0.119	-0.117	-0.118	-0.119	-0.119	-0.119
No. adults=3	0.067	0.073	0.072	0.070	0.069	0.068	0.070	0.068	0.068	0.068	0.068
No. adults>3	-0.021	-0.016	-0.019	-0.021	-0.022	-0.022	-0.019	-0.020	-0.021	-0.022	-0.022
No. kids=1	-1.121	-1.122	-1.122	-1.123	-1.122	-1.122	-1.127	-1.124	-1.123	-1.122	-1.122
No. kids=2	-1.681	-1.679	-1.681	-1.681	-1.682	-1.681	-1.689	-1.685	-1.682	-1.681	-1.681
No. kids>2	-2.132	-2.132	-2.133	-2.131	-2.132	-2.131	-2.144	-2.137	-2.133	-2.132	-2.131
Scotland-post 2017	-0.179	-0.181	-0.181	-0.180	-0.178	-0.179	-0.175	-0.177	-0.178	-0.179	-0.179
Log expenditure	0.892	0.897	0.892	0.892	0.891	0.891	0.894	0.893	0.892	0.892	0.891
$S.E.(\hat{\theta})$											
No. adults=2	0.155	0.446	0.316	0.224	0.170	0.155	0.186	0.169	0.160	0.156	0.155
No. adults=3	0.200	0.576	0.408	0.289	0.219	0.200	0.244	0.220	0.207	0.201	0.200
No. adults>3	0.258	0.746	0.528	0.374	0.283	0.258	0.314	0.283	0.267	0.260	0.258
No. kids=1	0.104	0.300	0.213	0.151	0.114	0.104	0.126	0.114	0.108	0.105	0.104
No. kids=2	0.144	0.415	0.294	0.208	0.158	0.144	0.173	0.157	0.148	0.145	0.144
No. kids>2	0.210	0.606	0.430	0.304	0.230	0.210	0.259	0.233	0.218	0.212	0.210
Scotland-post 2017	0.165	0.475	0.338	0.239	0.181	0.165	0.180	0.172	0.168	0.166	0.165
Log expenditure	0.150	0.434	0.307	0.217	0.164	0.150	0.161	0.155	0.152	0.150	0.150
$S.D.(\hat{\theta})$											
No. adults=2		0.450	0.318	0.225	0.170	0.155	0.187	0.169	0.160	0.156	0.155
No. adults=3		0.581	0.409	0.290	0.218	0.200	0.243	0.220	0.207	0.201	0.200
No. adults>3		0.744	0.525	0.374	0.283	0.258	0.314	0.284	0.267	0.260	0.258
No. kids=1		0.304	0.216	0.153	0.115	0.105	0.125	0.114	0.108	0.105	0.105
No. kids=2		0.422	0.296	0.209	0.157	0.143	0.169	0.154	0.147	0.143	0.143
No. kids>2		0.618	0.431	0.304	0.230	0.209	0.253	0.229	0.217	0.210	0.209
Scotland-post 2017		0.480	0.342	0.241	0.183	0.167	0.181	0.173	0.169	0.167	0.167
Log expenditure		0.439	0.308	0.218	0.166	0.151	0.161	0.155	0.152	0.151	0.151