Manuscript submitted to *The Econometrics Journal*, pp. 1–27.

A two sample size estimator for large data sets

Martin O'Connell[†], Howard Smith[‡], and Øyvind Thomassen[⋆]

[†] *University of Wisconsin-Madison and Institute for Fiscal Studies*
E-mail: moconnell9@wisc.edu
[‡] *Oxford University*
E-mail: howard.smith@economics.ox.ac.uk
[⋆] *NHH Norwegian School of Economics*
E-mail: oyvind.thomassen@nhh.no

Summary 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 a simple two sample size "large-small" estimator that uses the full data set for the computationally cheap observed component, but a reduced sample size for the predicted component. We derive a practical criterion for when the large-small estimator has a lower variance than standard GMM with the reduced sample size. As an alternative, we show how the asymptotically efficient CEP-GMM estimator of Chen et al. (2005) and Chen et al. (2008) can also be used to reduce computational cost in our setting. We compare the performance of the estimators in a Monte Carlo study of a panel-data random coefficients logit model, and illustrate the use of our estimator in an empirical application to alcohol demand.

Keywords: *GMM, sample combination, estimation, micro data*

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 need 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" and Johansen and Munk-Nielsen (2022), who state that "[f]or computational reasons, we are forced to reduce the number of discrete choice observations to a random 0.2% subsample". Varian (2014) says 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 a predicted component. The cost of computing the value of the observed part of the moment is usually negligible, because it needs to be done only 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 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 demanding in terms of CPU or memory requirements, especially if the model involves simulation of integrals or a nested optimization or fixed-point problem.[1] The simplest way to overcome this challenge is to use a subsample of the data for estimation, with a resulting loss in precision.

In this paper we study how to reduce the computational burden at a smaller cost in terms of imprecision, by exploiting the fact that the observed part of the moment is computationally cheap to evaluate. We explore two ways of doing this. First, we propose a simple two sample size "large-small" GMM-style estimator for cross-section or panel data that uses the full sample for the observed component, and a subsample for the predicted component only. Second, we show how to apply methods from the sample combination literature, designed to deal with issues of missing data and measurement error, to the problem of reducing computation cost.

The large-small estimator has the advantage that it is straightforward to implement, since it uses essentially the same code as standard GMM. We show that the estimator can be cast as a standard minimum distance estimator, and hence is consistent and asymptotically normal. The rationale for using the large-small estimator, instead of a standard GMM estimator on the subsample only, is that it reduces 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: compared with small-sample GMM, the large-small estimator lowers the correlation between the observed and predicted part of the moments, and this acts to raise the variance of the estimator. We derive a condition for when the large-small estimator has a 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.

The problem we study is closely related to one studied in the sample combination literature (see Ridder and Moffitt (2007) for a survey): some components of (or variables in) the moment condition are available for a sample of size N (the *primary sample* in the sample combination terminology), while the full moment condition can only be computed for a sample of size $n < N$ (the *auxiliary sample*). Chen et al. (2005) and Chen et al. (2008) study this problem in the contexts of measurement error and missing data, respectively. We apply their conditional expectation projection (CEP) based GMM estimator to a novel setting in which it is computational constraints that prevent us from calculating the predicted part of the moment for the full size-N sample. Like the large-small estimator, CEP-GMM involves computing the predicted part of the moment only for the small sample. Relative to the large-small estimator, it has the advantage of being asymptotically efficient. However, it involves the additional calculation of a nonparametric estimate of the predicted moment for the large sample, so has greater computational cost.

Chen et al. (2005), Chen et al. (2008), Ridder and Moffitt (2007), and the papers cited there study the issue of sample combination in more generality, allowing, among other things, for the primary and auxiliary samples to be independent as well as dependent. In this paper we focus on the setting, commonly encountered by empirical researchers, where there is only one sample but it is too large for it to be computationally feasible to

[1] An example of models that involve a nested optimization problem are demand systems with binding non-negativity constraints (Wales and Woodland (1983), Lee and Pitt (1986)), and an example of a model with a nested fixed-point problem is the discrete choice demand model of Berry et al. (1995).

implement a standard GMM estimator. This corresponds to the special case where the small (auxiliary) sample is a subset of the large (primary) sample, called the "verify-in-sample" case in the sample combination literature. The general idea of using a subset of the data to speed up computation is related to the stochastic gradient descent method of Robbins and Monro (1951), reviewed in Bottou et al. (2018), and developed for a GMM setting in Chen et al. (2023).

We compare the performance of the estimators in a Monte Carlo study of a panel-data random-coefficients logit model, with a small sample size of five thousand and a large sample size of five hundred thousand. Both the large-small and the CEP-GMM estimators are substantially faster than large-sample GMM, with the large-small estimator a further four times faster than CEP-GMM. Both are substantially more precise than small-sample GMM, with the root mean squared error of the CEP-GMM estimates only 10-20 percent of the RMSE of the small-sample GMM estimates, while for the large-small estimates the RMSE is 10-70 percent (of small-sample GMM RMSE). Theoretical considerations as well as Monte Carlo results indicate that the difference in computation time between CEP-GMM and the large-small estimator increases approximately linearly in the size of the large sample (for a given size of the small sample).

In an empirical application of the large-small estimator we use household scanner data (Kantar FMCG At-Home Purchase Panel) for the UK that cover 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 the large-small estimator. We show that the large-small estimator outperforms small-sample GMM, and that for the panel data model, the precision of the large-small 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.

The next section defines the large-small estimator. Section 3 discusses when the large-small estimator is more efficient than small-sample GMM. Section 4 discusses the CEP-GMM estimator. Sections 5 and 6 present our Monte Carlo simulations and empirical application, and the final section concludes.

2. THE LARGE-SMALL ESTIMATOR

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

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

where $h_i(\theta)$ is a vector valued function of θ and of w_i, $h_i(\theta) = h(w_i, \theta)$, while y_i does not depend on θ.[2] Letting θ be $K \times 1$, and y_i and $h_i(\theta)$ $L \times 1$, we assume $L \geq K$. Note that (y_i, w_i) may contain multiple observations corresponding to different time periods

[2] Alternatively we could write $g(y_i)$ instead of y_i, where $g(y_i)$ is a vector valued function of (finite-dimensional) data y_i (but not of θ). In some settings, such as for the CEP-GMM estimator discussed in Section 4, it matters that the data y_i typically contain more information than $g(y_i)$ does – i.e. y_i cannot

for the same i, or more generally multiple observations belonging to the same cluster i. All our results allow for arbitrary dependence between observations within i in the case of panel or clustered data.

Define the observed part of the sample moment:

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

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 constant selection probability $k \in (0, 1]$, where we think of n as the largest sample size for which the estimator is computationally feasible. In the 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 think of nonlinear models as the main application of the large-small estimator, but to fix ideas Example 1 shows what the estimator would look like in the familiar setting of a linear model.[3]

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, and rank $\mathbb{E}(Z'_{it} X_{it}) = K$. 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 $[Z'Y/NT - (Z'X/NT)\theta]' \hat{W} [Z'Y/NT - (Z'X/NT)\theta]$. The first-order condition gives $\hat{\theta} = [(Z'X/NT)' \hat{W} (Z'X/NT)]^{-1} (Z'X/NT)' \hat{W} (Z'Y/NT)$. ✓

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

be inferred from $g(y_i)$. To keep the notation simple, however, we write y_i, and understand it to mean, depending on the context, either the observed part of the moment or the data used to form it.

[3] 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: the large-small estimator involves $L(n\tau)K$ term-by-term multiplications, instead of $L(NT)K$, for the predicted part of the moment.

$n\tau \times L$ matrix, so that $\bar{h}_n(\theta) = (Z_n' X_n / n\tau)\theta$. The objective function 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 $\hat{\theta} = [(Z_n' X_n/n\tau)'\hat{W}(Z_n' X_n/n\tau)]^{-1}(Z_n' X_n/n\tau)'\hat{W}(Z'Y/NT)$. ✓

Consistency and asymptotic normality. 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 $\hat{\theta} = \arg\min_\theta \hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta)$. When $k = 1$ and so $n = N$, the large-small estimator coincides with the standard GMM case.[4]

Consistency of the large-small estimator (2.5) follows from a standard result for extremum estimators (Newey and McFadden (1994), Theorem 2.1).[5] Furthermore, the estimator is a *minimum distance estimator* since $\hat{g}_n(\theta_0)$ has probability limit zero.[6] Minimum distance estimators for which $\sqrt{n}\hat{g}_n(\theta_0)$ is asymptotically normal are themselves asymptotically normal. 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 $n \to \infty$, $\sqrt{n}(\bar{y}_N - \mu_y) = \frac{\sqrt{n}}{\sqrt{N}}\sqrt{N}(\bar{y}_N - \mu_y) \xrightarrow{d} N(0, k\Sigma_y)$. Also, $\text{Cov}\left[\sqrt{n}(\bar{y}_N - \mu_y), \sqrt{n}(\bar{h}_n(\theta_0) - \mu_h)\right] = \frac{n}{N}\Sigma_{yh}$ (see Supporting Information). Using the fact that $\mu_y = \mu_h$ by condition (2.1), we get $\sqrt{n}\hat{g}_n(\theta_0) = \sqrt{n}(\bar{y}_N - \bar{h}_n(\theta_0)) = \sqrt{n}(\bar{y}_N - \mu_y) - \sqrt{n}(\bar{h}_n(\theta_0) - \mu_h) \xrightarrow{d} N(0, k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma_{yh}'))$.

It then follows from Theorem 3.2 in Newey and McFadden (1994)[7] that the large-small

[4] 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. The standard GMM estimator is consistent by Theorem 2.6 in Newey and McFadden (1994) with some additional assumptions, which we too 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.

[5] See the Supporting Information. The large-small estimator can also be cast as a GMM estimator, by defining the indicator function $s_i = 1(i \in I_n)$, so that $\sum_{i\in I_N} s_i = n$. We can then write $\bar{h}_n(\theta) = \frac{1}{N}\sum_{i\in I_N}\frac{s_i}{k}h_i(\theta)$, and $\hat{g}_n(\theta) = \frac{1}{N}\sum_{i\in I_N}\left[y_i - \frac{s_i}{k}h_i(\theta)\right]$. If the s_i are i.i.d. Bernoulli with $\Pr(s_i = 1) = k$, $\mathbb{E}\left[\frac{s_i}{k}h_i(\theta)\right] = \mathbb{E}[h_i(\theta)]$. We thank an anonymous referee for comments on this point.

[6] See Newey and McFadden (1994), p. 2116 - 2117 for this definition. To see that it holds for (2.5): 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): $\hat{g}_n(\theta_0) \xrightarrow{p} \mathbb{E}[y_i] - \mathbb{E}[h_i(\theta_0)] = 0$.

[7] For simplicity, assume that (a) $\theta_0 \in \text{interior}(\Theta)$; (b) $h_i(\theta)$ is continuously differentiable on $\text{interior}(\Theta)$, 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). In the Supporting Information we give an alternative result for the case when condition (e) is not satisfied.

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 $\Omega = k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma'_{yh})$.

The asymptotic variance can be consistently estimated in the usual way as

$$\hat{B}\hat{\Omega}\hat{B}' \tag{2.9}$$

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 $\hat{\Sigma}_y = N^{-1}\sum_{i \in I_N}[y_i - \bar{y}_N][y_i - \bar{y}_N]'$, $\hat{\Sigma}_h = n^{-1}\sum_{i \in I_n}[h_i(\hat{\theta}) - \bar{h}_n(\hat{\theta})][h_i(\hat{\theta}) - \bar{h}_n(\hat{\theta})]'$, and $\hat{\Sigma}_{yh} = n^{-1}\sum_{i \in I_n}[y_i - \bar{y}_N][h_i(\hat{\theta}) - \bar{h}_n(\hat{\theta})]'$.[8] 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 standard 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.[9] Finally, the optimal weighting matrix is the usual choice:[10] $W = \Omega^{-1}$.

3. CONDITIONS FOR EFFICIENCY GAIN

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

Proposition 3.1. *For the j-th element of the parameter vector θ, the difference between the $\sqrt{n}$-asymptotic variance of the GMM estimator $\hat{\theta}_j^{n,n}$ that uses sample size n, and that of the large-small estimator $\hat{\theta}_j^{N,n}$ is*

$$\text{Avar}(\hat{\theta}_j^{n,n}) - \text{Avar}(\hat{\theta}_j^{N,n}) = (1-k)b_j[\Sigma_y - (\Sigma_{yh} + \Sigma'_{yh})]b'_j, \tag{3.10}$$

where b_j is the j-th row of the matrix B.

Proof. The change in $B\Omega B'$ is $B\left\{\left[\Sigma_y + \Sigma_h - (\Sigma_{yh} + \Sigma'_{yh})\right] - \left[k\Sigma_y + \Sigma_h - k(\Sigma_{yh} + \Sigma'_{yh})\right]\right\}B'$ or $(1-k)\,B[\Sigma_y - (\Sigma_{yh} + \Sigma'_{yh})]B'$. Element (j,j) of $B\Omega B'$ is $b_j \Omega b'_j$.

Since the right-hand side of (3.10) is a quadratic form, the large-small estimator with $N > n$ is variance-reducing for all elements of θ if and only if

$$\Sigma_y - (\Sigma_{yh} + \Sigma'_{yh}) \tag{3.11}$$

is positive definite.

Intuitively, potential efficiency gains are driven by a reduction in the asymptotic variance of the observed part of the moment, $\bar{y}_N$, since $k\Sigma_y < \Sigma_y$. On the other hand, for given variances $k\Sigma_y$ and Σ_h of the observed and predicted components, respectively, the

[8] By the law of large numbers, $\hat{\Sigma}_y$, $\hat{\Sigma}_h$ and $\hat{\Sigma}_{yh}$ converge in probability to the corresponding population objects, so that $\hat{\Omega} \xrightarrow{p} \Omega$. The result then follows by Newey and McFadden (1994) Theorem 4.2.

[9] 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.

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

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, $k\Sigma_{yh} < \Sigma_{yh}$, 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 2 – linear regression. Suppose we want to estimate the model $y = \theta x + e$, where $\text{Cov}(x, e) = 0$ at the true value θ_0, so that $\theta_0 = \text{Cov}(x, y)/\text{Var}(x)$. 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 the large-small 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 small-small 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

$$0.7071 \approx 1/\sqrt{2} > |\theta|, \tag{3.12}$$

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 (3.10) we get a variance reduction of:[11]

$$(1-k)(1-2\theta^2), \tag{3.13}$$

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

4. AN ASYMPTOTICALLY EFFICIENT ESTIMATOR

Chen et al. (2008) derive asymptotic variance lower bounds for estimators in a number of settings involving missing data. One of these settings is closely related to that studied in this paper: the moment can only be computed for a subset (the auxiliary sample) of the full (primary) sample, but some variables that enter the moments are available for the entire primary sample. The probability of an observation not being in the auxiliary sample, called the propensity score, and corresponding to $1 - k = 1 - n/N$ in our case, is assumed known. Chen et al. (2008) show that the CEP-GMM estimator of Chen et al.

[11] See Supporting Information for details.

(2005) achieves the relevant lower bound in this case. Furthermore, they show that the small-sample GMM estimator, which corresponds to their inverse probability weighting (IPW) based GMM, is asymptotically inefficient in this setting.

In the preceding section we show that the large-small estimator may be less efficient than small-sample GMM. It follows that the large-small estimator is not asymptotically efficient, and that the CEP-GMM is preferable to the large-small estimator on theoretical grounds. However, since the motivation for not using standard GMM on the large (size-N) sample is a computational constraint, computational concerns are also relevant when choosing between the large-small and CEP-GMM estimators.

Here we briefly describe the CEP-GMM estimator, in order to show how it differs from the large-small estimator in terms of computation. Let $\hat{h}_n(\theta, y_i)$ be a nonparametric estimate of $\mathbb{E}[h(w_i, \theta)|y_i]$ obtained using only the small (auxiliary) sample I_n. The CEP-GMM estimator is then:[12]

$$\hat{\theta}^{\text{CEP}} = \arg\min_{\theta} \left\{ \frac{1}{N} \sum_{i \in I_N} [y_i - \hat{h}_n(\theta, y_i)] \right\}' \hat{W} \left\{ \frac{1}{N} \sum_{i \in I_N} [y_i - \hat{h}_n(\theta, y_i)] \right\}. \tag{4.14}$$

Chen et al. (2005) propose a sieve least squares estimator of $\hat{h}_n(\theta, y_i)$.[13] Let $q_1(y_i), q_2(y_i), \ldots$ be a sequence of known basis functions that can approximate any square-measurable function of y_i arbitrarily well. Typically splines or polynomials are used. Define the $1 \times l(n)$ vector consisting of the first $l(n)$ terms of the sequence: $q^{l(n)}(y_i) = (q_1(y_i), \ldots, q_{l(n)}(y_i))$, where $l(n)$ is some integer such that $l(n) \to \infty$ and $l(n)/n \to 0$ when $n \to \infty$. Stack $q^{l(n)}(y_j)$ for each $j \in I_n$ in the $n \times l(n)$ matrix Q_n. Also stack $h(w_j, \theta)'$ for $j \in I_n$ in the $n \times L$ matrix $H_n(\theta)$.[14] Then $\hat{h}_n(\theta, y_i)$ is the i-th column of the $L \times N$ matrix

$$H_n(\theta)' Q_n (Q_n' Q_n)^{-1} Q_N' \tag{4.15}$$

where the $N \times l(n)$ matrix Q_N stacks $q^{l(n)}(y_i)$ for all $i \in I_N$. While the large-small estimator requires the evaluation of $H_n(\theta)$ for each trial value of the parameter vector, the CEP-GMM estimator involves the additional multiplication at each value of θ of $H_n(\theta)'$ with the $n \times N$ matrix $Q_n(Q_n'Q_n)^{-1}Q_N'$ (which only needs to be computed once). In practice it will usually be faster to first multiply $H_n(\theta)'$, which is $L \times n$, and $Q_n(Q_n'Q_n)^{-1}$, which is $n \times l(n)$, to get an $L \times l(n)$ matrix, and then multiply by Q_N', which is $l(n) \times N$. This gives $Lnl(n) + Ll(n)N$ term-by-term multiplications instead of LnN, and suggests that the difference in computation time between the CEP-GMM and large-small estimators increases close to linearly in N for fixed L and n, at a rate proportional to $Ll(n)$.[15] The next section compares the two estimators in a Monte Carlo study.

5. MONTE CARLO RESULTS

In this section we present Monte Carlo results for a panel-data random-coefficients logit model. Consumer i at time t gets conditional indirect utility u_{ijt} from choosing alternative

[12] See the Supporting Information for an estimator for the asymptotic variance of CEP-GMM.

[13] See Chen (2007) for a survey of sieve estimation.

[14] For a given value of θ, we can calculate $h(w_j, \theta)$ for $j \in I_n$, but not for the $N - n$ observations that are in $I_N \setminus I_n$ (i.e. in the primary sample but not in the auxiliary sample).

[15] In the Monte Carlo experiments presented in the next section, $N = 500,000$ and $n = 5,000$, which means that an $n \times N$ matrix requires 20GB of memory.

j such that

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

The 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}.[16] 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. The true values of the parameters are $\beta = 0.3$, $\sigma = 0.15$, and $\gamma = 0.2$. Each observed choice is generated by one draw of the vector $(z_{it}, \nu_i, \varepsilon_{i1t}, \dots, \varepsilon_{iJt})$ while the product attributes $x_{1t}, \dots, x_{Jt}$ are the same for all i. The number of time periods is $T = 2$ and the number of alternatives is $J = 100$.

Let $\mathbb{I}_{ijt} = 1$ if i chooses j in t and $\mathbb{I}_{ijt} = 0$ otherwise, and let $\mathbb{P}_{ijt}(\theta)$ be the corresponding choice probability. 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:[17]

$$\mathbb{E}[(\sum_j x_{jt}\mathbb{I}_{ijt})(\sum_j x_{jt'}\mathbb{I}_{ijt'}) - (\sum_j x_{jt}\mathbb{P}_{ijt})(\sum_j x_{jt'}\mathbb{P}_{ijt'})] = 0 \quad \text{for } t \neq t'.$$

Each replication (Monte Carlo sample) involves drawing a large ($N = 500,000$) random sample, and a small ($n = 5,000$) random subsample of the large sample. We use 1000 replications, with the exception of the full-sample GMM estimator (denoted "L-L"), where we use 100 replications.[18] In the reported results we include only runs where the GMM objective ends up below a threshold value of 5E−9, to prevent misleading results from instances where the minimum of the objective function has not been found. (This happens in 8.3 percent of replications for the standard GMM estimator on the small sample, and for one replication, i.e. 0.1 percent, for the large-small estimator.)

Table 1 reports results for four different estimators: CEP-GMM ("CEP"),[19] small-sample GMM ("S-S"), the large-small estimator ("L-S"), and large-sample GMM ("L-L"). S-S and L-L use the same standard GMM estimator and differ only in the sample sizes, n and N, respectively. The first panel shows the mean of estimates for the three parameters across replications. The second panel shows the bias (difference between mean estimate and true parameter value). The third panel shows the root mean squared error, and the fourth panel the mean standard error for each sample. The row denoted

[16] For simulation of the integral over the distribution of ν_i in the model's prediction, for each i we use one simulation draw.

[17] 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. Since the model is exactly identified, we use the identity matrix as weighting matrix.

[18] The objective functions are minimized with Matlab's quasi-Newton fminunc with objective and step tolerances set to 1e-6. Starting values for all parameters were 0.5. The simulations were run on a Lenovo ThinkPad T14s Gen 3. Processor: 12th Gen Intel(R) Core(TM) i7-1260P, 2100 Mhz, 12 Cores, 16 Logical Processors. Installed Physical Memory (RAM): 32.0 GB.

[19] For the CEP-GMM estimator, the vector $q^{l(n)}(y_i)$ of basis functions consists of a constant, and for each of $t = 1, 2$, of z_{it}, $(z_{it})^2$, $(\max\{0, z_{it} - z^l\})^2$ for equidistant knots $\underline{z} < z^1 < z^2 < z^3 < z^4 < \bar{z}$ where $\underline{z}$ and $\bar{z}$ are the min and max across all z_{it}, respectively, as well as $z_{it} \cdot [\mathbb{I}_{i2t}, \dots, \mathbb{I}_{iJt}]$.

'mean seconds' shows the average computation time for each estimator across replications. The rows 'no. replications', i.e. number of replications, and the row below it, number of replications with objective below the threshold value, differ only in the case of S-S because the threshold value for the objective function is reached in all replications for the other estimators. The mean values of the objective function are shown in the last row. Finally, for the large-small (L-S) estimator, we check for each replication whether the variance-reduction criterion (3.11) is satisfied. This is the case in every replication, as can be seen in the second to last row.

Table 1. Monte Carlo results

	CEP	S-S	L-S	L-L
mean $\hat{\beta}$	0.30008	0.29740	0.30011	0.29967
mean $\hat{\sigma}$	0.14983	0.14651	0.14976	0.14931
mean $\hat{\gamma}$	0.19998	0.20208	0.19986	0.20011
bias $\hat{\beta}$	0.00008	-0.00260	0.00011	-0.00033
bias $\hat{\sigma}$	-0.00017	-0.00349	-0.00024	-0.00069
bias $\hat{\gamma}$	-0.00002	0.00208	-0.00014	0.00011
RMSE $\hat{\beta}$	0.00577	0.02778	0.01601	0.00312
RMSE $\hat{\sigma}$	0.00541	0.05849	0.00584	0.00478
RMSE $\hat{\gamma}$	0.00531	0.02544	0.01683	0.00287
mean $s.e.(\hat{\beta})$	0.00562	0.02800	0.01560	0.00279
mean $s.e.(\hat{\sigma})$	0.00539	0.06228	0.00570	0.00519
mean $s.e.(\hat{\gamma})$	0.00508	0.02593	0.01636	0.00258
mean seconds	12.54	1.83	3.17	225.83
no. replications	1000	1000	1000	100
no. valid replications (GMM obj < 5E-9)	1000	917	999	100
no. repl. w/variance red. criterion satisf.			1000	
mean value of GMM objective	3.76E-11	2.88E-06	4.98E-11	6.61E-11

Notes: authors' simulations. True parameter values are $\beta = 0.3$, $\sigma = 0.15$, and $\gamma = 0.2$. Each replication draws a new large sample of size $N = 500,000$, and a new random subsample, of size $n = 5,000$, from the large sample. "CEP" is the CEP-GMM discussed in section 4 with primary sample of size $N = 500,000$ and auxiliary sample of size $n = 5,000$, "S-S" is the standard GMM estimator on the small ($n = 5,000$) sample, "L-S" is the large-small estimator, and "L-L" is the standard GMM estimator on the large ($N = 500,000$) sample.

Starting with computation times, we see that CEP, S-S, and L-S all entail significant reductions in computation speed relative to the full sample GMM, with the CEP-GMM and large-small estimators giving a speedup of 18.0 times and 71.2 times, respectively, relative to the large-sample GMM. The speedup of small-sample GMM is 123.4 times relative to large-sample GMM, which is broadly in line with the fact that these are identical estimators but with S-S using a sample that is 100 times smaller. In an additional Monte Carlo exercise, reported in the Supporting Information, we run the same experiments but for different values of N. We find that the difference between CEP computation time and large-small computation time increases at a nearly constant rate of about 2 seconds for each additional 100,000 observations in N.[20]

[20] That is, for $N = 0.5$, 0.75, 1, 1.25, 1.5 million, we find that $1e5 \cdot (time_{CEP} - time_{LS})/N \approx 2.0$. The exact numbers for the five values of N are 2.26, 1.88 2.00, 2.72, and 2.86, respectively. The ratio

Turning next to the precision of the estimates, both CEP-GMM and large-small outperform small-sample GMM, with the RMSE for the three parameters being, respectively, 20.8, 9.2, and 20.9 percent of the S-S ones for CEP-GMM, and 57.6, 10.0, and 66.2 percent of the S-S ones for the large-small estimator. For comparison, the large-sample estimator has an RMSE that, as expected, is around 10 percent of those for S-S, at 11.2, 8.2, and 11.3 percent, respectively.

6. 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 1 shows that households with higher spending levels systematically allocate a higher spending share to alcohol.[21] 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.[22]

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 $h = 1, \dots, H$) and the period 2002-2021 (indexed $t = 1, \dots, T$). 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). 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

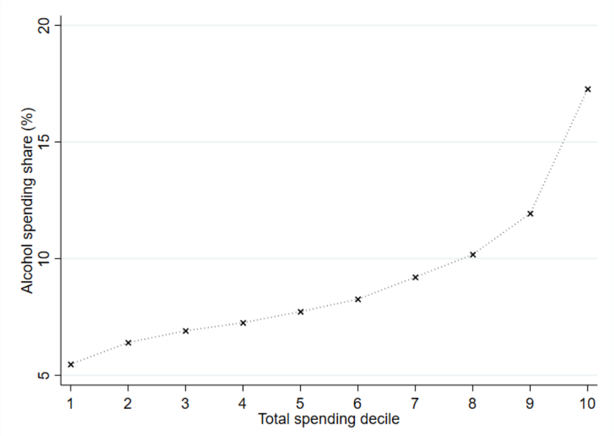
$time_{CEP}/times_{LS}$ increases in N, with 4.00, 4.49, 5.12, 5.36, and 6.18, for the five values of N. Here we do 100 replications per setting. See the Supporting Information for full results.

[21] This pattern holds for consumption, with the total alcohol units of the top and second top spending deciles being 14 and 7 times those for the bottom spending decile.

[22] For instance, if the cross-sectional pattern depicted in Figure 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)).

and the year is 2018 or later (designed to capture the impact of the introduction of a minimum unit price for alcohol in Scotland).[23]

Figure 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.

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 is to compare conventional GMM estimates based on a sub- (or small) sample of a given size n with the large-small estimator (that uses large, size

[23] 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)).

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 2, focusing on estimates of the log total expenditure coefficient ($\hat{\beta}$).[24] 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 (valid under separability of household supply and food and drink preferences). 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, 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.

[24] We present results for the other coefficients in the Supporting Information. They exhibit similar patterns.

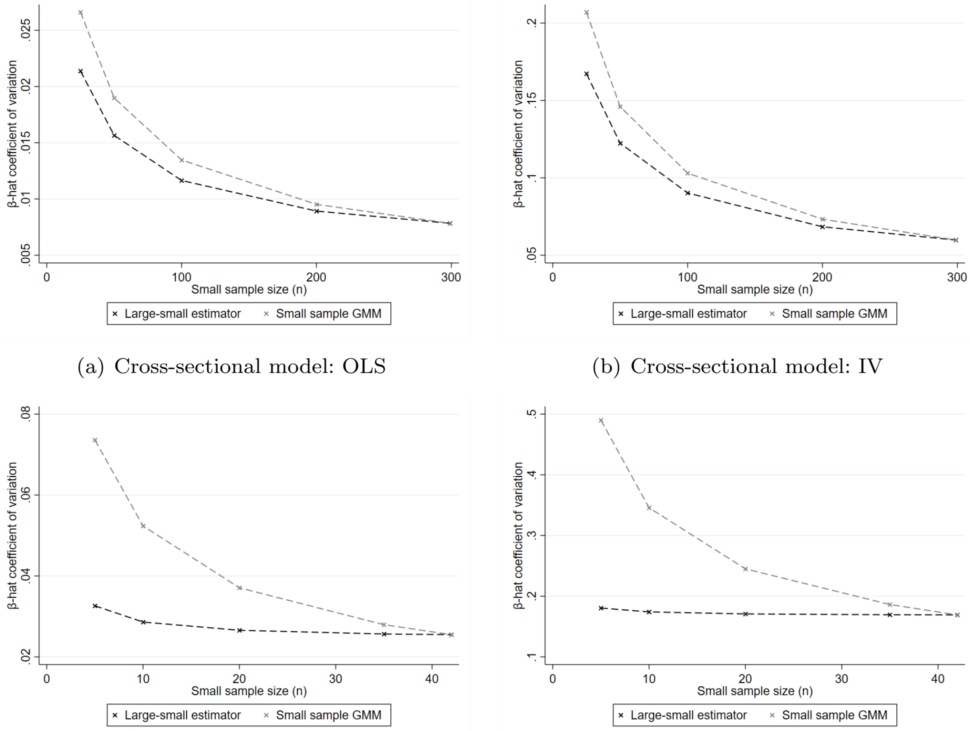
Table 2. Simulation results

	$\hat{\beta}_{LL}$	$\hat{\beta}_{SS}$					$\hat{\beta}_{LS}$				
	(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.

In all cases the large-small estimates are more precise than the corresponding small sample GMM estimates. We illustrate this graphically in Figure 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).

Figure 2. Precision of $\hat{\beta}$



(a) Cross-sectional model: OLS

(b) Cross-sectional model: IV

(c) Panel data model: FE

(d) Panel data model: FE-IV

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 Table 2.

7. CONCLUSION

We propose a two sample size "large-small" GMM-type estimator that 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, simulated integrals or 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, the large-small 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 is satisfied then using the large-small estimator will be more efficient than standard GMM using the small sample only. As an alternative, we show how to apply the asymptotically efficient, and therefore theoretically preferable, CEP-GMM estimator of Chen et al. (2005) to the problem of reducing the computational burden in a GMM setting with large data sets.

We compare the performance of the large-small and CEP-GMM estimators in a Monte Carlo study. We show that for given small and large sample sizes both estimators are substantially more precise than GMM estimation on the small sample, and that CEP-GMM results in more precise estimates than the large-small sample estimator but with higher computation time. An interesting avenue for future research is to explore the choice of estimator *and* sample sizes that maximize precision subject to not exceeding a given computational cost. We also undertake an empirical application to alcohol demand, which shows, even when the size of the small sample is modest relative to that of the large sample, it is possible for the large-small estimator to yield estimates with precision close to those of the asymptotically efficient GMM estimator applied to the large sample.

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 Jeff Smith, Richard Spady, Martin Weidner, Yoon-Jae Whang and Frank Windmeijer for comments.

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.

Berry, S., J. Levinsohn, and A. Pakes (1995). Automobile prices in market equilibrium. *Econometrica 63*(4), 841–890.

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.

Bottou, L., F. E. Curtis, and J. Nocedal (2018). Optimization methods for large-scale machine learning. *SIAM Review 60*(2), 223–311.

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

Chen, X. (2007). Chapter 76 large sample sieve estimation of semi-nonparametric models. Volume 6 of *Handbook of Econometrics*, pp. 5549–5632. Elsevier.

Chen, X., H. Hong, and E. Tamer (2005). Measurement Error Models with Auxiliary Data. *The Review of Economic Studies 72*(2), 343–366.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in GMM models with auxiliary data. *The Annals of Statistics 36*(2), 808 – 843.

Chen, X., S. Lee, Y. Liao, M. H. Seo, Y. Shin, and M. Song (2023). SGMM: Stochastic Approximation to Generalized Method of Moments. *Journal of Financial Econometrics*, nbad027.

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

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

Johansen, B. G. and A. Munk-Nielsen (2022). Portfolio complementarities and electric vehicle adoption.

Lee, L.-F. and M. M. Pitt (1986). Microeconometric demand system with binding non-negativity constraints: The dual approach. *Econometrica 54*(5), 1237–1242.

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

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.

Ridder, G. and R. Moffitt (2007). The econometrics of data combination. In J. Heckman and E. Leamer (Eds.), *Handbook of Econometrics* (1 ed.), Volume 6B, Chapter 75. Elsevier.

Robbins, H. and S. Monro (1951). A Stochastic Approximation Method. *The Annals of Mathematical Statistics 22*(3), 400 – 407.

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.

Wales, T. J. and A. Woodland (1983). Estimation of consumer demand systems with binding non-negativity constraints. *Journal of Econometrics 21*(3), 263–285.

Supporting Information

A. CONSISTENCY OF THE LARGE-SMALL ESTIMATOR

We maintain the assumptions in footnote 4, here referred to as Assumption 1. We start by proving two lemmas that we use for the main proof.

Lemma 1.1. *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,

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

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 $n \geq n_1$,

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

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

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

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

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

Lemma 1.2. *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 simply verify the hypotheses of Lemma 2.4 in Newey and McFadden (1994).[25] 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

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

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 1.1, 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 1.1 are satisfied, it follows that

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

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

We can now prove that the large-small estimator (2.5) is consistent.

Proof. We still 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).[26] 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 1.2, $g_0(\theta)$ 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 $\hat{W} \xrightarrow{p} W$.

B. 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 $\bar{h}_n - \mu_h = \frac{1}{n} \sum_{i \in I_n} h_i(\theta_0) - \mu_h =$

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

$\frac{1}{n}\sum_{i\in I_n} b_i$. Then:

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

C. NON-SMOOTHNESS

In some cases assumption (vii) in footnote 7 may not hold. The following set of alternative assumptions are also sufficient for asymptotic normality.

ASSUMPTION 3.1. *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*

1 $\theta_0 \in$ *interior*(Θ)
2 $\hat{g}_n(\hat{\theta})'\hat{W}\hat{g}_n(\hat{\theta}) \leq \hat{g}_n(\theta)'\hat{W}\hat{g}_n(\theta) + o_p(n^{-1})$
3 $g_0(\theta)$ *is differentiable at* θ_0 *with derivative* G *such that* $G'WG$ *is nonsingular.*
4 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 3.1. *The estimator (2.5) satisfies:*

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

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.

D. PROOF OF STATEMENT IN REGRESSION EXAMPLE

This is a proof of statement (3.12) in the paper. We have

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

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

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

By equation (3.11) 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

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

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

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

E. DETAILS OF THE CEP-GMM ESTIMATOR

Here we describe the CEP-GMM estimator in some more detail for our setting and in our notation, and show how to calculate standard errors. Our estimation problem is based on the moment condition (2.1). In our setting, y_i is the only part of the moment that can be computed for the full (primary) sample. In the missing data setting of Chen et al. (2008) this corresponds to the subset of variables that are not missing in the primary sample.[27] Our w_i corresponds to the variables that are observed only in the auxiliary sample in their setting.[28] Define $\mathcal{E}(y_i;\theta) = \mathbb{E}[y_i - h(w_i,\theta)|y_i] = y_i - \mathbb{E}[h(w_i,\theta)|y_i]$ and $\hat{\mathcal{E}}(y_i;\theta) = y_i - \hat{h}_n(\theta, y_i)$, where $\hat{h}_n(\theta, y_i)$ denotes a nonparametric estimate of $\mathbb{E}[h(w_i,\theta)|y_i]$ obtained using only the small (auxiliary) sample I_n. The CEP-GMM estimator is then

$$\hat{\theta}^{\text{CEP}} = \arg\min_{\theta} \left[\frac{1}{N}\sum_{i\in I_N} \hat{\mathcal{E}}(\theta;y_i)\right]' \hat{W} \left[\frac{1}{N}\sum_{i\in I_N} \hat{\mathcal{E}}(\theta;y_i)\right]. \tag{E.1}$$

Chen et al. (2005) propose a sieve least squares estimator of $\hat{h}_n(\theta, y_i)$. Let $q_1(y_i), q_2(y_i), \ldots$ be a sequence of known basis functions that can approximate any square-measurable function of y_i arbitrarily well. Typically splines or polynomials are used. Define the $l(n)\times 1$ vector consisting of the first $l(n)$ terms of the sequence:[29]

$$q^{l(n)}(y_i) = (q_1(y_i), \ldots, q_{l(n)}(y_i))', \tag{E.2}$$

where $l(n)$ is some integer such that $l(n)\to\infty$ and $l(n)/n \to 0$ when $n\to\infty$. We stack these vectors for each $i = 1,\ldots,n$ in the $n\times l(n)$ matrix

$$Q = (q^{l(n)}(y_1), \ldots, q^{l(n)}(y_n))'.$$

For a given value of θ, we can calculate the $L\times 1$ vector of predicted moments $h(w_j,\theta)$ for $j\in I_n$, but not for the $N-n$ observations that are in $I_N \setminus I_n$ (i.e. in the primary sample but not in the auxiliary sample). We stack $h(w_j,\theta)'$ for $j\in I_n$ in the $n\times L$ matrix

$$H(\theta) = (h(w_1,\theta), \ldots, h(w_n,\theta))'.$$

OLS regressions of each of the L entries of $h(w_j,\theta)$ on $q^{l(n)}(y_j)$ using the auxiliary sample I_n gives the $l(n)\times L$ coefficient matrix:[30]

$$\Gamma = (Q'Q)^{-1}Q'H(\theta)$$

[27] In their notation, these variables are denoted X_i, corresponding to our y_i.

[28] Our w_i corresponds to Y_i in their notation.

[29] In the body of the text we use $q^{l(n)}(y_i)$ to mean the transpose of its definition here.

[30] Consisting of one $l(n)\times 1$ vector of coefficients for each of the L moment conditions.

and, for any y_i, which may not be in the auxiliary sample, an $L \times 1$ fitted value

$$\begin{aligned}\hat{h}_n(\theta, y_i) &= [q^{l(n)}(y_i)'\Gamma]' \\ &= \Gamma' q^{l(n)}(y_i) \\ &= H(\theta)'Q(Q'Q)^{-1}q^{l(n)}(y_i).\end{aligned}$$

This is then used to get the CEP-GMM estimator.

By Theorem 2 in Chen et al. (2005), $\sqrt{n}(\hat{\theta}^{\text{CEP}} - \theta_0) \xrightarrow{\text{d}} N(0, V)$ as $n \to \infty$, where the $\sqrt{n}$-asymptotic variance V can be consistently estimated as (see Chen et al. (2005), section 3.3) $\hat{B}\hat{\Omega}\hat{B}'$, where $\hat{B} = (\hat{G}'\hat{W}\hat{G})^{-1}\hat{G}'\hat{W}$, with[31]

$$\begin{aligned}\hat{G} &= -N^{-1} \sum_{i \in I_N} \nabla_\theta \hat{h}_n(\theta, y_i) \\ \hat{\Omega} &= \frac{1}{n} \sum_{j=1}^{n} \hat{U}_j \hat{U}_j' + \frac{n}{N} \frac{1}{N} \sum_{i \in I_N} \hat{\mathcal{E}}(\theta; y_i) \hat{\mathcal{E}}(\theta; y_i)' \\ \hat{U}_j &= [y_j - h(w_j, \theta)] - \hat{\mathcal{E}}(\theta; y_j).\end{aligned}$$

F. ADDITIONAL MONTE CARLO RESULTS

Table 3 shows Monte Carlo results for the same model as in Table 1 in the paper, but now for different values of N. Note that we here use 100 replications per setting, compared to 1000 in the main table.

[31] The formula given in Chen et al. (2005) is

$$\begin{aligned}\hat{\Omega} &= \frac{1}{n} \sum_{j=1}^{n} (\hat{v}_j \hat{U}_j)(\hat{v}_j \hat{U}_j)' + \frac{n}{N} \frac{1}{N} \sum_{i \in I_N} \hat{\mathcal{E}}(\theta; y_i) \hat{\mathcal{E}}(\theta; y_i)' \\ \hat{v}_j &= \left[\frac{1}{N} \sum_{i \in I_N} q^{l(n)}(y_i) \right]' \left(\frac{Q'Q}{n} \right)^{-1} q^{l(n)}(y_j),\end{aligned}$$

but in our case the weighting by $\hat{v}_j$ can be replaced by a 1, since the auxiliary sample is a subset of the primary sample. The reason is that $\hat{v}_j$ is an estimate of the ratio of the density of y in the primary and the auxiliary data sets, evaluated at y_j. In our case these two densities are equal.

Table 3. Monte Carlo results, varying N

	CEP 500,000	L-S 500,000	CEP 750,000	L-S 750,000	CEP 1,000,000	L-S 1,000,000	CEP 1,250,000	L-S 1,250,000	CEP 1,500,000	L-S 1,500,000
mean $\hat{\beta}$	0.2999	0.3003	0.3005	0.3018	0.3002	0.2987	0.3001	0.3013	0.3004	0.2995
mean $\hat{\sigma}$	0.1498	0.1499	0.1499	0.1502	0.1500	0.1499	0.1502	0.1501	0.1503	0.1499
mean $\hat{\gamma}$	0.2004	0.1995	0.1993	0.1982	0.1997	0.2012	0.2000	0.1991	0.1993	0.2003
bias $\hat{\beta}$	-0.0001	0.0003	0.0005	0.0018	0.0002	-0.0013	0.0001	0.0013	0.0004	-0.0005
bias $\hat{\sigma}$	-0.0002	-0.0001	-0.0001	0.0002	-0.0000	-0.0001	0.0002	0.0001	0.0003	-0.0001
bias $\hat{\gamma}$	0.0004	-0.0005	-0.0007	-0.0018	-0.0003	0.0012	-0.0000	-0.0009	-0.0007	0.0003
RMSE $\hat{\beta}$	0.0057	0.0168	0.0054	0.0149	0.0058	0.0168	0.0048	0.0158	0.0054	0.0152
RMSE $\hat{\sigma}$	0.0051	0.0055	0.0042	0.0048	0.0035	0.0037	0.0038	0.0044	0.0034	0.0038
RMSE $\hat{\gamma}$	0.0053	0.0178	0.0049	0.0158	0.0052	0.0183	0.0047	0.0169	0.0047	0.0160
mean $s.e.(\hat{\beta})$	0.0056	0.0159	0.0054	0.0153	0.0053	0.0156	0.0052	0.0154	0.0052	0.0155
mean $s.e.(\hat{\sigma})$	0.0053	0.0056	0.0045	0.0049	0.0038	0.0043	0.0036	0.0040	0.0034	0.0038
mean $s.e.(\hat{\gamma})$	0.0051	0.0167	0.0049	0.0160	0.0047	0.0164	0.0047	0.0162	0.0047	0.0163
mean seconds	15.09	3.77	18.13	4.04	24.83	4.85	41.77	7.80	51.27	8.30
$time_{CEP} - time_{LS}$	11.31		14.09		19.98		33.97		42.97	
$time_{CEP}/time_{LS}$	4.00		4.49		5.12		5.36		6.18	
1E5$\cdot(time_{CEP} - time_{LS})/N$	2.26		1.88		2.00		2.72		2.86	
no. valid repl.	100	100	100	100	100	100	100	100	100	100
v. red. Crit.	0	100		100		100		100		100
GMM objective	3.98E-11	8.78E-11	4.89E-11	2.77E-11	2.87E-11	5.90E-11	3.52E-11	3.94E-11	3.14E-11	3.42E-11

Notes: authors' simulations. True parameter values are $\beta = 0.3$, $\sigma = 0.15$, and $\gamma = 0.2$. Each replication draws a new large sample of size N, and a new random subsample, of size $n = 5,000$, from the large sample. "CEP" is the CEP-GMM discussed in section 4 with primary sample of size N and auxiliary sample of size $n = 5,000$, and "L-S" is the large-small estimator.

G. EMPIRICAL APPLICATION: FULL SIMULATION RESULTS

Table 4. 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
$\overline{\hat{\theta}}$											
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 5. 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 6. 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 7. 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