The Challenges of Measuring Substitution Elasticities in Food Scanner Data

Hao Lan, Tim Lloyd, Steve McCorriston and Cesar Revoredo-Giha.

Abstract

Constant elasticity of substitution (CES)-based price indices for food inflation and consumer welfare are becoming increasingly popular for their use with scanner data where high product churn is observed. However, their use depend primarily on robust estimates of the substitution elasticity (σ). In the Feenstra–Broda–Weinstein (FBW) estimation, σ is jointly identified with the inverse supply elasticity (ω), which helps interpreting supply-side responses. To account for potential endogeneity, the parameters are estimated using the generalised method of moments (GMM) with instrumental variables. The only instruments available for the estimation (no other data are available besides prices and quantities) are product dummies; however, there is more than one option for defining those dummies. The purpose of this paper is to compare different strategies related to the use of product dummies as instrumental variables. Using Monte Carlo simulations calibrated to three UK food categories show that standard implementations that use dummies for the all products as instruments tend to severely inflate σ value, in contrast, considering dummies only for common-product yields more plausible elasticities, which bring clearer inflation and welfare implications.

Keywords: Elasticity of substitution, inverse supply elasticity, scanner data, instrumental variables estimation, Monte Carlo simulation
JEL Codes: C13, C26, C36, C81

Draft February 2026

1 Introduction

Recent advances in the measurement of retail price inflation and welfare—building on theoretical index-number results from the trade literature (notably Feenstra, 1994) and prominent empirical applications (e.g., Broda and Weinstein, 2010; Redding and Weinstein, 2020)—increasingly rely on demand-based cost-of-living indices derived from CES preferences. In this framework, the elasticity of substitution (σ) governs how consumers reallocate expenditure across varieties as relative prices change and therefore directly shapes inferred substitution effects, entry/exit (variety) effects, taste shocks, and welfare changes. In the canonical FBW system, σ is identified jointly with a (reduced-form) supply-side parameter, the inverse supply elasticity (ω), which is also of independent interest for incidence and pass-through analyses (the supply elasticity is $1/\omega$). Reliable estimation of both parameters is thus central for credible measurement of food inflation and the welfare consequences of food price movements.

Implementing these estimators in modern scanner data environments is challenging. Scanner datasets differ sharply from the trade panels for which the FBW approach was originally developed. They typically contain thousands to tens of thousands of UPCs per category observed at monthly or quarterly frequency, and they exhibit pronounced product turnover and highly unbalanced panels. A key implication is instrument proliferation: the standard FBW identification strategy uses product indicators as instruments under an assumption that demand and supply shocks are independent across products. In a trade setting with modest cross-sectional dimension, using all cross-sectional indicators as instruments is feasible. In scanner data, the same logic implies using thousands (or more) product instruments—often dominated by short-lived items with limited time-series support—making full-system GMM or LIML estimation computationally impractical and, as we show, empirically unreliable.

As a result, most scanner-data applications rely on the product-level weighted least squares implementation (FBW-WLS), which is asymptotically equivalent to full-instrument GMM under the maintained assumptions but is far cheaper computationally. In practice, however, FBW-WLS in scanner settings faces three related problems. First, it can generate implausible or "imaginary" estimates, motivating ad hoc corrections such as grid search. Second, it remains sensitive to heteroskedasticity and other deviations from the assumed moment structure, with limited consensus remedies in

applied work. Third, and crucially for current econometric practice, WLS is not well aligned with recent advances that are naturally implemented in an IV/GMM/LIML framework—such as LIML-based bias reductions (Soderbery, 2010, 2015) and post-estimation diagnostics for weak instruments and exclusion-restriction violations (Grant and Soderbery, 2024). These issues are amplified when the instrument set is dominated by short-lived products: many instruments may be weak, and some may be correlated with idiosyncratic pricing shocks, undermining the exclusion restriction and inflating elasticity estimates.

This paper re-examines CES elasticity estimation in high-frequency food scanner data and proposes a practical refinement that preserves the FBW identification logic while adapting it to the empirical structure of scanner panels. Our key observation is that scanner data typically contain only a small subset of "common" products—often around five percent—that are purchased in every period of the sample. We propose a common-product IV strategy that restricts the instrument set to these long-lived products. The refinement is transparent and objective: it requires no additional exogenous information and no subjective screening beyond the availability criterion. Economically, long-lived products tend to be more stable and better measured; econometrically, restricting to common products sharply reduces instrument dimensionality and shifts identifying variation away from short-lived, thinly observed items that are most likely to behave as outliers. This motivation is consistent with recent evidence that low-market-share products can disproportionately affect index construction and related inference (Ehrlich et al. 2023).

We implement the common-product IV strategy using both GMM and LIML, denoted comGMM and comLIML. We evaluate performance using large-scale Monte Carlo simulations that follow Soderbery's design but are calibrated to key scanner features—large cross-sectional dimension, short average product lives, and realistic entry/exit patterns. We consider five simulated scanner datasets, including stylised designs and template designs that replicate the unbalanced panel structure observed in UK Kantar food categories. Across these environments, the standard FBW-WLS implementation produces substantial upward bias in σ—often inflating substitution elasticities by an order of magnitude relative to the truth—and also distorts ω, though typically to a lesser degree. In contrast, comGMM and comLIML substantially reduce the bias in σ, and they improve the stability of ω estimates; comGMM is particularly robust in avoiding

estimation failures, while comLIML often yields σ estimates closest to the truth but can be less stable for ω in some designs.

Empirically, applying the approach to three UK food categories (chilled and frozen cakes, meat, and chilled and frozen ready meals) yields the same qualitative message: standard implementations generate implausibly large substitution elasticities, whereas common-product IV estimates are much more moderate and economically plausible. These differences are not merely statistical. Because σ directly governs substitution in demand-based cost-of-living indices, overstating σ mechanically attenuates the inferred importance of variety and taste-shock components, potentially distorting measured food inflation and the welfare assessment of price changes. Improving elasticity estimation is therefore an input into more credible measurement of inflation dynamics and consumer welfare in food markets.

The paper also reports robustness checks and extensions. We examine post-estimation diagnostics emphasised by Grant and Soderbery (2024), highlighting how inference can depend on the assumed dependence structure of the moment conditions. We also discuss a practical identification issue for ω: the mapping from reduced-form parameters to ω can become unstable near theoretical boundary conditions, producing extreme implied values even when σ is estimated reasonably. We then explore the method in a trade-data calibration and show that the strict "common-product" definition can be too restrictive over long horizons; a rolling-window version (e.g., common over 5–10 years) may be more appropriate in that context. Finally, we assess how two common implementation choices—iterated GMM and Fuller-modified LIML—affect estimation in scanner-data applications, and show that they can change point estimates and precision without altering the paper's main qualitative conclusions.

Overall, this paper contributes in three ways. First, it documents that scanner-data implementations that effectively use all products as instruments can generate severe upward bias in σ and distort ω, even when the underlying CES structure is correct. Second, it proposes a simple, implementable refinement—common-product IV—that materially improves finite-sample performance while enabling IV/GMM/LIML estimation and associated diagnostics in large scanner panels. Third, by stabilising the key elasticities used in demand-based index-number analysis, it strengthens the

empirical foundations of applied work on food inflation measurement, welfare analysis, and related questions in empirical industrial organisation.

2 Related Literature

The structural estimation of the CES elasticity of substitution (σ) and the (inverse) supply elasticity (ω) used in demand-based price indices originates in the international trade literature. Building on the identification logic in Leamer (1981), Feenstra (1994) develops a GMM framework that uses price and quantity data to recover import demand and export supply elasticities. Broda and Weinstein (2006) apply this approach to quantify welfare gains from trade and address extreme or implausible ("imaginary") estimates via a grid-search procedure. Soderbery (2010, 2015) shows that standard estimators can exhibit important small-sample bias—typically overstating demand elasticities—and proposes a LIML-based estimator to mitigate these problems. More recently, Feenstra et al. (2018) refine Armington-type elasticity estimation by adding moment conditions derived from the macro-demand equation, strengthening identification and robustness.

A key maintained assumption in the Feenstra–Broda–Weinstein (FBW) approach is that demand and supply shocks are uncorrelated within an exporter, which justifies using the full set of import–export pair dummies as instruments in GMM estimation. Under this instrument structure, the full-instrument GMM estimator is asymptotically equivalent to a country-level weighted least squares (WLS) implementation: one can estimate country-specific "hyperbolas" from time-series variation and identify demand and supply elasticities from their intersection across exporters. In trade applications, moving between the full instrument IV/GMM representation and the WLS representation is typically feasible. In scanner datasets, however, the same logic implies using tens of thousands of product dummies as instruments, making direct IV/GMM or LIML implementations computationally difficult and, as we show, empirically fragile. Consequently, most scanner-data applications adopt the WLS implementation (e.g., Broda and Weinstein, 2006, 2010; Jaravel, 2019; Redding and Weinstein, 2020).

In scanner settings, reliance on WLS raises two related concerns. First, the empirical literature frequently reports very large substitution elasticities. Broda and Weinstein (2010), in their scanner-data application, use grid search to re-estimate imaginary values and impose a "reasonable range" based on economic judgement; however,

Soderbery (2015) argues that grid search can itself introduce substantial upward bias in small samples. Second, the WLS implementation is poorly aligned with more recent econometric advances that are naturally framed in IV/GMM/LIML terms. Soderbery (2010, 2015) emphasizes LIML as a practical bias-reduction device, and Grant and Soderbery (2024) advocate post-estimation diagnostics—such as weak-instrument checks and tests for violations of exclusion restrictions—that can be implemented after IV/GMM or LIML estimation but not after WLS. Moreover, available corrections applied to WLS have limited practical success, as discussed in Mohler (2009).[1]

Consistent with these concerns, recent scanner-based studies often report substitution elasticities that appear high relative to common theoretical priors. Broda and Weinstein (2010) report a median σ of 11.5, while Redding and Weinstein (2018) estimate 6.48, though both find wide dispersion across product groups. Jaravel (2018) reports values around 5.7 at the median and 9.3 at the 90th percentile, and Argente and Lee (2021) report average estimates around 20 for the United States. Such large σ estimates matter for the motivating objects in this literature: in demand-based cost-of-living indices, the contributions of product entry/exit and (in extensions) taste shocks are scaled by $(1/(\sigma-1))$, so inflated σ mechanically compresses these components and can distort inflation and welfare decompositions. This concern is particularly acute in frameworks that use σ to discipline multiple components simultaneously, such as the taste-shock decomposition in Redding and Weinstein (2018).

In response, some studies calibrate σ rather than estimate it. For example, Chen et al. (2024) adopt calibrated values typically ranging between 2 and 5. Others pursue alternative empirical strategies that rely on richer product characteristics and exogenous instruments. Colicev et al. (2024), for instance, estimate demand via fixed-effects regressions with external instruments, in the spirit of hedonic-type approaches discussed in Ehrlich et al. (2023); however, these approaches typically require highly

[1] Mohler (2009) applies generalized least squares (GLS) to correct for heteroskedasticity in the standard GMM framework but hardly found sufficient improvement in their empirical applications. Besies that, De Haan and Krsinich (2024) propose an algebraic method based on matched-model Jevons and price–quantity indices as an alternative to regression-based estimation. But this method is very preliminary, and they don't show a clear improvement from it. von Brasch and Raknerud (2021) and von Brasch and Vigtel (2024) explore advanced panel estimators, such as constrained GMM approaches. While they look promising but it seems less practical and too complex for standard empirical application and fails to meaningfully reduce the upward bias in the estimates.

detailed scanner attributes (manufacturer, brand, retailer identifiers) and, in some cases, machine-learning tools to construct the necessary covariates.

Overall, the literature highlights a tension: the FBW identification logic is appealing and widely used, but scanner-data environments exacerbate feasibility constraints and can amplify upward bias in σ under conventional implementations. This paper contributes by proposing a feasible and transparent instrument-design refinement—our common-product IV strategy—that is tailored to high-frequency scanner panels while remaining closely connected to recent IV/GMM/LIML developments in the trade-based elasticity literature.

3 FBW Approach

Following Feenstra (1994) and Broda and Weinstein (2010), consider a product category in which a representative household has CES preferences over UPCs. The associated demand system can be written in expenditure shares as

$$s_{it} = b_{it}\left(\frac{p_{it}}{P_t}\right)^{1-\sigma} \tag{1}$$

where $s_{it} \equiv \frac{p_{it}x_{it}}{\sum_{i\in I_t} p_{it}x_{it}}$ denotes the expenditure share of UPC i at time t, p_{it} is its price, b_{it} is an unobserved taste (quality) shifter, and P_t is the unit-expenditure function for the category. The elasticity of substitution σ is constant within the category and shared by all UPCs. We assume $\sigma > 1$, so $1 - \sigma < 0$.

Supply is modelled in reduced form. Using the share-based representation (Soderbery, 2024),

$$s_{it} = e^{-\frac{A_{it}}{\omega}} p_{it}^{\frac{1+\omega}{\omega}} E_t \tag{2}$$

where A_{it} is a supply shock, E_t is aggregate expenditure on the category, and ω is the inverse supply elasticity, assumed constant across UPCs with $\omega > 0$. (Equivalently, the supply elasticity is $1/\omega$.)

Equations (1)–(2) form the demand–supply system used to estimate the elasticity of substitution and the inverse supply elasticity jointly. To implement the estimation, we work in logs and use a double-differencing strategy to eliminate time-varying common terms. In particular, the differenced log system contains random terms $\phi_t \equiv (1-\sigma)\Delta \ln P_t$ and $\psi_t \equiv -\frac{\omega}{1+\omega}\Delta \ln(E_t)$, which are removed by differencing relative to a reference UPC K (typically the largest-selling product). The double-differenced system can then be written as

$$\Delta^K \ln s_{it} \equiv \ln s_{it} - \ln s_{Kt} = (1-\sigma)\Delta^K \ln p_{it} + \varepsilon_{it}^K$$

$$\Delta^K \ln p_{it} \equiv \ln p_{it} - \ln p_{Kt} = \frac{\omega}{1+\omega}\Delta^K \ln s_{it} + \delta_{it}^K \tag{3}$$

where $\varepsilon_{it} \equiv \Delta \ln(b_{it})$ and $\delta_{it} \equiv \frac{\Delta A_{it}}{1+\omega}$ are demand and supply shocks, respectively, and $\varepsilon_{it}^K \equiv \varepsilon_{it} - \varepsilon_{Kt}$, $\delta_{it}^K \equiv \delta_{it} - \delta_{Kt}$. Here Δ^Kdenotes differencing relative to the K-th product.

Identification relies on two maintained assumptions: (i) demand and supply shocks are independent across UPCs, and (ii) elasticities are constant across UPCs. Multiplying the two equations in (3) yields the single estimating equation

$$(\Delta^K \ln p_{it})^2 = \theta_1(\Delta^K \ln s_{it})^2 + \theta_2(\Delta^K \ln s_{it})(\Delta^K \ln p_{it}) + \nu_{it} \tag{4}$$

where $\theta_1 \equiv \frac{\omega}{(1+\omega)(\sigma-1)}$, $\theta_2 \equiv \frac{\omega(\sigma-2)-1}{(1+\omega)(\sigma-1)}$, and $\nu_{it} \equiv \frac{(1+\omega\sigma)}{(1+\omega)^2(\sigma-1)}\varepsilon_{it}^K\delta_{it}^K$ is the composite error term. Equation (4) cannot be consistently estimated by OLS because prices and shares are correlated with ν_{it}. Under the cross-UPC independence assumption, an IV approach can be implemented using dummy variables for other UPCs ($i \neq l$) as instruments. This yields consistent estimates of θ_1 and θ_2, and more efficient estimates can be obtained via GMM with a heteroskedasticity-robust weighting matrix. A constant term θ_0 is also commonly included to absorb measurement error associated with using unit values. Finally, Feenstra (1994) shows how the estimated coefficients can be mapped back into consistent estimates of σ and ω.

Formally, the GMM estimator minimizes

$$\hat{\beta} = \arg\min_{\beta} G^*(\beta)' W G^*(\beta) \tag{5}$$

where $G^*(\beta)$ stacks the sample moment conditions across UPCs, β collects the parameters to be estimated (typicallyθ_0, θ_1 and θ_2 in the implementation), and W is a positive definite weighting matrix. The cross-UPC independence assumption is the exclusion restriction in this overidentified setting and can be assessed using overidentification tests such as Sargan's/Hansen's J-test; rejection indicates that the proposed instruments are correlated with the composite error term, contrary to the maintained assumption.

In practice, the FBW method is often implemented using weighted least squares (WLS), which is asymptotically equivalent to the full-instrument GMM estimator but far faster when the instrument set is very large. Averaging equation (4) over time within each UPC yields

$$\overline{(\Delta^K \ln p_t)^2} = \theta_1 \overline{(\Delta^K \ln s_t)^2} + \theta_2 \overline{(\Delta^K \ln s_t)(\Delta^K \ln p_t)} + \bar{v}_i \tag{6}$$

where bars denote within-UPC sample means. Under the maintained exclusion restriction, $\mathbb{E}[\bar{v}_i \mid Z_i] = 0$, so estimating (6) by WLS is consistent; weighting each UPC by its number of observed periods is efficient because $\bar{v}_i$ is estimated more precisely for longer-lived products (its variance scales approximately with $1/T_i$). Intuitively, WLS uses the same population moments as full-instrument GMM but in a "collapsed" UPC-level representation, avoiding explicit use of a very high-dimensional instrument set.

4 Scanner dataset

This section describes the UK food scanner datasets used to motivate the estimation challenges in scanner environments and to calibrate the Monte Carlo simulations. We

use data from Kantar Worldpanel for three food categories that are representative in terms of product turnover, panel unbalancedness, and the scale of the cross section. Two categories are observed quarterly and one monthly, which also allows us to illustrate how frequency interacts with product churn and the effective amount of time-series information available for each product.

Our empirical applications use Chilled and frozen cakes (a disaggregated 4-digit category, quarterly), Meat (an aggregated 3-digit category, quarterly), and Chilled and frozen ready meals (monthly). These choices reflect a typical range of scanner panels: disaggregated categories contain fewer UPCs but substantial churn, whereas aggregated categories contain many more UPCs and often stronger unbalancedness. The monthly panel provides a complementary setting in which higher frequency can amplify entry and exit and shorten observed product lives in calendar time.

For each category, we construct a UPC-by-time panel with prices and expenditure shares defined consistently with the FBW framework. Prices are unit values at the UPC–period level and shares are expenditure shares within the category. We apply standard cleaning steps to remove observations that cannot be used in the FBW transformation, including periods with insufficient product overlap and UPCs with too few observations to form reliable within-product averages.

Table 1 summarizes the key features of the three category panels. Even at the category level, the number of UPCs is large, ranging from 2,381 in the disaggregated Cakes category to 18,705 in the aggregated Meat category (with 6,726 UPCs in Ready Meals). Product turnover is also substantial. Average UPC lifetimes are 15.3–15.9 quarters in the quarterly panels (Cakes and Meat) and 46.9 months in the monthly panel (Ready Meals), with wide dispersion: the standard deviation of UPC length is about 11 quarters in the quarterly data and 33 months in the monthly data. The lifetime distribution is highly skewed, with many UPCs appearing briefly (minimum length 4 quarters in the quarterly panels and 13 months in the monthly panel) and a smaller set of long-lived products persisting for much longer (up to 44 quarters and 130 months, respectively).

Most importantly for instrument construction, the panels are severely unbalanced. Only a small subset of UPCs is observed in every period of the sample. Across the three categories, the share of such common products is consistently low — 3.5% to 4.2% — despite the large number of UPCs overall. This feature is central for our approach:

conventional FBW implementations effectively use all products as instruments, but in scanner environments this means that the bulk of the instrument set is composed of short-lived products with limited time-series support.

These data characteristics create two practical problems for standard FBW estimation in scanner applications. First, product indicators as instruments generate instrument dimensionality that grows with the number of UPCs, making direct IV/GMM or LIML estimation computationally burdensome. Second, because many UPCs are short-lived, the effective instrument set is dominated by products with thin time-series information, which can weaken identification and exacerbate finite-sample bias in σ, and may contribute to unstable estimates of ω. These concerns motivate the Monte Carlo designs below, where we consider both stylised simulations that match scanner scale and template-based simulations that replicate the observed unbalanced panel structure of the Kantar categories.

Finally, the small share of common products suggests a simple and transparent instrument restriction. In later sections we define instruments using only UPCs observed in every period, concentrating identification on long-lived items that are measured repeatedly over time and sharply reducing the dimension of the instrument set. We refer to this as the common-product IV strategy; when implemented using GMM and LIML, we denote the resulting estimators comGMM and comLIML, respectively.

[Insert Table 1]

5 Simulation Analysis

This section evaluates the finite-sample performance of alternative FBW estimators in scanner-like environments. The objective is not to test the CES model per se, but to assess how conventional implementations behave when the data resemble modern food scanner panels: very large cross sections, substantial product turnover, and severe unbalancedness. The simulation design therefore follows the FBW identification structure and the Monte Carlo logic in Soderbery's work, while calibrating the panel structure to the empirical features documented in Section 4.

We use five simulated scanner datasets. sim_data1 and sim_data2 are stylised designs intended to mimic scanner scale and churn, with the maximum number of UPCs set to 5,000 and 10,000, respectively. The remaining designs are template-based simulations that replicate the unbalanced panel structure (UPC entry/exit and lifetime distribution) observed in the UK Kantar categories: sim_discat4 (Cakes, quarterly), sim_aggcat2 (Meat, quarterly), and sim_RMdata (Ready Meals, monthly). In the template designs, the panel structure is taken from the data, while shocks are simulated.

Following Grant and Soderbery (2024), we work with the differenced representation in which differenced prices and expenditure shares can be expressed solely in terms of the differenced shocks. Specifically, for UPC i in period t,

$$\Delta^K \ln p_{it} = \frac{\omega}{1+\omega\sigma}\varepsilon_{it}^K + \frac{1}{1+\omega\sigma}\xi_{it}^K$$

$$\Delta^K \ln s_{it} = \frac{\omega+1}{1+\omega\sigma}\varepsilon_{it}^K + \frac{1-\sigma}{1+\omega\sigma}\xi_{it}^K$$

where ε_{it}^K and ξ_{it}^K denote the K-differenced demand- and supply-side shocks. Given values of the underlying elasticities (σ, ω), we generate $\Delta^K \ln p_{it}$ and $\Delta^K \ln s_{it}$ from these equations and then apply the estimators exactly as in the empirical implementation.[2]

We set the true parameters at $\sigma = 3$ and $\omega = 0.5$, which are standard in the literature and empirically plausible for food categories. To mirror scanner-data noise, we allow for heteroskedasticity by drawing the error variances from Uniform distributions and then sampling shocks conditionally on those variances:

$$\varepsilon_{it}^K \sim N(0, \sigma_\varepsilon^2) \text{ where } \sigma_\varepsilon^2 \sim U(0, b_\varepsilon)$$

$$\xi_{it}^K \sim N(0, \sigma_\delta^2) \text{ where } \sigma_\xi^2 \sim U(0, b_\delta)$$

with b_ε and b_ξ chosen to range from 0 to 100 while imposing the scanner-relevant restriction $b_\varepsilon \ll b_\xi$. This feature differs from Soderbery (2010, 2015), who sets $b_\varepsilon =$

[2] In the simulation analysis, equation (4) follows Grant and Soderbery (2024), where we define the supply shock as $\xi_{it} \equiv \Delta A_{it}$. This differs slightly from equation (3), where the supply shock enters as $\delta_{it} \equiv \Delta A_{it}/(1+\omega)$. This rescaling has little effect on the estimation of the elasticity of substitution σ, but it can help the simulations better match the intended data-generating process on the supply side by modelling the primitive supply shock directly rather than a transformation that depends on ω.

b_ξ, a calibration that tends to generate relatively small (rather than inflated) estimates of σ and is therefore less representative of scanner environments; nevertheless, our qualitative conclusions are robust to that alternative setting.

Finally, we impose measurement error to reflect the use of unit values in place of true transaction prices. Observed prices in the Monte Carlo are replaced by

$$\Delta^K \ln UV_{it} = \Delta^K \ln p_{it} + \mu_{it} \; where \; \mu_{it} \sim N(0, 0.15)$$

so estimation is carried out using $\Delta^K \ln UV_{it}$ and $\Delta^K \ln s_{it}$, matching the empirical implementation.

To mimic the panel structure of UK scanner data, the stylised designs generate highly unbalanced panels with large cross sections. The number of UPCs N is drawn from a Uniform distribution over a scanner-relevant range (from 500 up to 5,000 or 10,000, depending on the dataset). The number of time periods per UPC, T_i, follows a Weibull distribution with shape parameter 1.5 and scale parameter 20, producing a decreasing lifetime profile in which short-lived UPCs dominate. We also impose a realistic fraction of common products—UPCs observed in every period—by drawing the common-product share from $U(0.03, 0.07)$ and constructing the panel to satisfy this proportion. Together, these rules reproduce the key empirical feature documented in Section 4: scanner datasets contain many short-lived products and only a small stable core of common products.

In addition to stylised simulations, we extract the panel structure directly from the three real-world scanner categories summarised in Table 1. In these template-based simulations, the entry/exit pattern and UPC-length distribution are taken from the data, and only the shocks are simulated. This hybrid design captures scanner-data complexities more accurately than purely stylised simulations and allows us to evaluate estimation strategies under empirically realistic unbalancedness.

Across replications, we compare the conventional scanner implementation (FBW-WLS) with our common-product IV strategy implemented as comGMM and comLIML. Performance is assessed by the distribution of $\hat{\sigma}$ and $\hat{\omega}$ relative to the true values, together with numerical stability (including imaginary or failed estimates). This environment is deliberately demanding—many products, substantial churn, heteroskedastic shocks, and noisy unit values—so that the consequences of instrument

design for elasticity estimation can be assessed under conditions that closely match empirical practice.

6 Results

The estimation procedure follows FBW, in which an additional reduced-form parameter is introduced to improve numerical behaviour. We define $\rho \equiv \frac{\omega(\sigma-1)}{1+\omega\sigma}$, so that $\omega > 0$ implies $\rho > 0$. In practice, we estimate (σ, ρ) first and then back out the inverse supply elasticity ω (details are provided in the online appendix). In the Monte Carlo analysis we compare the standard scanner implementation (FBW-WLS) with GMM and LIML estimators that employ our common-product IV instrument set, denoted comGMM and comLIML. Using the refined instrument set substantially reduces computation time and, more importantly, improves the recovery of elasticities. The simulation consists of 100 replications; summary results are reported below, with additional tables (including moments for $\theta_0, \theta_1, \theta_2$ and other distributional statistics) provided upon requests.

To quantify how strongly alternative estimators correct the WLS result toward the true elasticity, we report a "correction power" metric defined relative to WLS. Let $e_m = \widehat{\omega}_m - \omega$ denote the estimation error for method m. We define correction power as $\frac{|e_{wls}|-|e_m|}{|e_{wls}|} \times 100$, which equals 100 if a method eliminates the WLS absolute error, equals 0 if it offers no improvement over WLS, and is negative if it increases the absolute error. For example, with $\omega = 0.5$, WLS yields $\widehat{\omega} = 0.436$ (absolute error 0.064), while GMM yields $\widehat{\omega} = 0.489$ (absolute error 0.011), implying a correction power of about 82.8%. By contrast, LIML yields $\widehat{\omega} = 0.567$ (absolute error 0.067), corresponding to a slightly negative correction power (about −4.7%) in that case.

Table 2 reports mean estimates across simulation scenarios. The main pattern is stark: the elasticity of substitution estimated using the standard FBW implementation (WLS) is severely upward biased. Across scanner-like designs, the mean $\hat{\sigma}$ under WLS typically falls between about 40 and 100, despite the true σ being 3. In contrast, both comGMM and comLIML deliver $\hat{\sigma}$ values close to the truth. The improvement is also large for the reduced-form parameter ρ: WLS tends to produce values around 0.8–0.9

when the true value is 0.4, while the common-product estimators move $\hat{\rho}$ much closer to the target. For the inverse supply elasticity ω, bias under WLS is smaller than for σ and ρ (WLS often yields values around 0.4 when the true value is 0.5), but comGMM still exhibits strong correction power relative to WLS. Overall, comGMM typically removes around 90% of the WLS absolute error for σ, around 70–80% for ρ, and around 70–80% for ω, with these gains broadly robust across simulated datasets. comLIML often corrects σ even more strongly—frequently approaching 90–100% correction—but its correction for ω is less stable, and in some cases the correction power becomes negative, indicating that it can marginally increase absolute error relative to WLS.

[Insert Table 2]

Figures 1–4 report the distributions of the estimated substitution elasticity and inverse supply elasticity, along with their associated standard errors. These figures reinforce the message from Table 2 while revealing important distributional details. Under WLS, $\hat{\sigma}$ can take extremely large values in some designs—reaching several hundred in the most distorted replications—whereas in the monthly template design this problem is substantially attenuated, with the maximum $\hat{\sigma}$ much smaller. The standard errors help interpret these extremes: the most distorted WLS estimates are typically statistically insignificant, while the higher-frequency monthly design shows a noticeably tighter and more significant distribution of $\hat{\sigma}$. Under comGMM and comLIML, the distribution of $\hat{\sigma}$ tightens sharply around the true value, but the two estimators differ in shape. comGMM estimates tend to lie slightly above 3 —suggesting a small remaining upward bias — whereas comLIML estimates are more symmetric around 3. At the same time, comLIML occasionally produces values close to the theoretical boundary (just above 1), which raises two concerns: it can generate economically implausible substitution behaviour and it can create near-boundary instability when mapping reduced-form parameters into ω. This is visible in the standard-error distributions: comLIML exhibits a small number of extremely large standard errors (sometimes very large relative to the bulk), and these outliers are associated with the near-boundary $\hat{\sigma}$ draws. comGMM, by contrast, produces consistently small standard errors across designs, implying tight confidence intervals and stable inference.

[Insert Figures 1-4]

For the inverse supply elasticity ω, the distributions are generally tighter than for σ, and both WLS and comGMM concentrate around the true value with relatively well-behaved standard errors. The most distorted outcomes occur under comLIML, which exhibits a pronounced right tail: some replications yield very large $\hat{\omega}$ values, often accompanied by extremely large standard errors. This pattern mirrors the instability noted above and suggests that, while comLIML can be attractive for recovering σ in many cases, it carries a larger risk of producing extreme but statistically uninformative supply-side estimates. By contrast, comGMM remains robust, with a tight distribution of $\hat{\omega}$ and uniformly small standard errors.

A further practical concern in this literature is estimator failure. Broda and Weinstein (2010) correct imaginary values using grid search, and Soderbery proposes hybrid procedures in which constrained LIML is invoked when unconstrained LIML fails. Table 3 reports failure frequencies across Monte Carlo designs. Consistent with this concern, FBW-WLS exhibits a non-trivial number of failures, particularly for ω; we discuss the ω-specific instability separately below. Notably, failures are less prevalent in the large-scale high-frequency designs in our calibration, a pattern that differs from some earlier studies. Applying the common-product instrument restriction eliminates failures entirely for comGMM, which records zero failures across all simulated datasets. In contrast, comLIML still exhibits a sizeable number of failures even with the refined IV set, and in several designs it does not outperform WLS in this respect. This pattern is consistent with the motivation for constrained or hybrid LIML procedures in related work.

[Insert Table 3]

Taken together, the Monte Carlo evidence implies that the standard scanner implementation (FBW-WLS) can deliver severely biased substitution elasticities in scanner-like panels, while bias in ω is smaller but still meaningful. The common-product IV strategy substantially improves performance. comGMM emerges as the most stable and reliable estimator for joint recovery of (σ, ω) in large-scale, unbalanced scanner panels: it achieves large error reductions relative to WLS, yields

tight inference, and eliminates failures. comLIML frequently recovers σ very well, but it is less reliable for ω and exhibits more failures and near-boundary behaviour.

Finally, we apply the estimators to the real-world UK scanner datasets summarised in Table 1. The empirical results mirror the simulation findings (Table 4). Standard methods produce implausibly large substitution elasticities (e.g., 76.7 for Cakes and 94.8 for Meat), whereas the common-product IV strategy yields estimates in a much narrower and more plausible range—around 2–3 across all three categories—consistent with economic intuition about substitution behaviour in food markets.

Finally, we apply the estimators to the three real-world UK scanner datasets described in Section 4 and report the results in Table 4. Under the standard FBW implementation, substitution elasticities are implausibly large in Cakes and Meat ($\hat{\sigma} = 76.7$ and 94.9), and remain high in Ready Meals ($\hat{\sigma} = 13.4$). Restricting instruments to common products reduces these estimates substantially: comGMM delivers $\hat{\sigma} = 13.8$ (Cakes), 14.9 (Meat), and 5.3 (Ready Meals), and these estimates are precisely estimated and strongly significant. By contrast, comLIML produces smaller point estimates of σ (1.6, 2.2, and 3.7), but the quarterly-category estimates are statistically insignificant with very large standard errors, indicating materially lower stability than comGMM; only in the monthly Ready Meals case is $\hat{\sigma}$ estimated precisely. The Cakes comLIML estimate ($\hat{\sigma} = 1.618$) is also close to the theoretical lower bound $\sigma > 1$, which may help explain its imprecision and associated instability. Overall, the empirical results are consistent with the Monte Carlo evidence and, in practice, point to comGMM as the most reliable and economically interpretable strategy for scanner-data applications.

[Insert Table 4]

7 Robustness and Extensions

This section provides additional evidence on robustness and clarifies several practical issues that arise in FBW-style implementations. We first report post-estimation diagnostics for weak instruments and overidentifying restrictions following Grant and Soderbery (2024). We then explain why the implied inverse supply elasticity can become unstable near theoretical boundaries, generating extreme values in a small number of simulations. Next, we examine the feasibility and performance of the

common-product IV idea in a trade-data Monte Carlo design, where a strict "common-product" definition can be too restrictive and rolling-window variants become more appropriate. Finally, we assess how two implementation choices that are common in applied work—iterated GMM and Fuller-modified LIML—affect estimation performance in scanner-data applications.

7.1 Diagnostic

We assess robustness to weak-instrument and exclusion-restriction concerns emphasised by Grant and Soderbery (2024). Motivated by their discussion, we compute (i) the first-stage Kleibergen–Paap rk Wald F-statistic and (ii) Hansen's J-test after comGMM and comLIML across simulated datasets, as robust standard errors are used in typical applications. Table 5 reports the mean and median of the F-statistic, the J-statistic, and the corresponding p-value.

The weak-instrument diagnostics indicate limited first-stage strength. Across simulated datasets, both mean and median Kleibergen–Paap F-statistics are around 1.7–2. Stock–Yogo critical values are not available when the number of instruments is very large, so we cannot apply the standard critical-value approach. Relative to the usual "rule-of-thumb" threshold of 10, however, these magnitudes suggest that weak-instrument concerns are not eliminated by restricting to common-product instruments. Compared with Soderbery (2024), our values are of similar order but slightly smaller. Under the rejection rules used by Grant and Soderbery (2024), this pattern implies that passing the weak-instrument diagnostic is difficult in this environment, particularly for GMM; LIML tends to be more tolerant in practice even when F-statistics remain low.

The overidentification diagnostics differ sharply between comGMM and comLIML. Under comGMM, Hansen's J-statistics are typically large and p-values small, implying frequent rejection of valid overidentifying restrictions. Under comLIML, J-statistics are generally smaller and p-values often indicate that the overidentification test is passed. We also find that, for comGMM, J-test conclusions are sensitive to the assumed dependence structure of moment conditions. Allowing for within-cluster correlation via cluster-robust covariance estimation increases p-values relative to heteroskedasticity-robust (non-clustered) inference, suggesting that ignoring within-cluster dependence can lead to over-rejection. Finally, running the full battery of tests for comLIML can be computationally burdensome: for example, 100 replications of the template

simulation based on sim_aggcat2 may require close to a week to complete in Stata on a moderately performing machine.

[Insert Table 5]

7.2 Boundary behaviour in ω

Although scanner-data applications typically focus on σ, the FBW framework also identifies the inverse supply elasticity ω. In a small number of Monte Carlo replications we obtain extremely large—or even negative—$\hat{\omega}$. This behaviour reflects the nonlinear mapping from reduced-form parameters into ω, rather than a breakdown of the CES structure.

Using $\omega = \frac{\rho}{\sigma(1-\rho)-1}$, the mapping becomes unstable when the denominator $\sigma(1-\rho)-1$ is close to zero. For $\omega > 0$, the model implies $0 < \rho < 1 - 1/\sigma$, and as $\omega \to \infty$, ρ approaches the upper bound $1 - 1/\sigma$. Hence, when $(\hat{\sigma}, \hat{\rho})$ lies near (or slightly above) $1 - 1/\hat{\sigma}$, even small sampling error in $\hat{\sigma}$ or $\hat{\rho}$ can produce an extremely large $\hat{\omega}$, or flip its sign if the bound is crossed. For example, one replication yields $\hat{\sigma} = 2.3169$ and $\hat{\rho} = 0.5663$, yet $\hat{\omega} \approx 117.98$ because $1 - 1/\hat{\sigma} \approx 0.5684$, making $\hat{\sigma}(1-\hat{\rho}) - 1 \approx 0.0048$. Such outcomes should be interpreted as near-boundary artefacts rather than reliable supply estimates.

The same nonlinearity also explains why $\hat{\omega}$ can sometimes appear stable even when $\hat{\sigma}$ and $\hat{\rho}$ are badly biased. If estimation errors move $(\hat{\sigma}, \hat{\rho})$ in a way that keeps $\hat{\sigma}(1-\hat{\rho}) - 1$ away from zero, the implied $\hat{\omega}$ can remain moderate. For instance, a replication with $\hat{\sigma} = 97.6534$ and $\hat{\rho} = 0.9671$ yields $\hat{\omega} \approx 0.437$, close to the truth, because $\hat{\rho}$ is not sufficiently close to the upper bound $1 - 1/\hat{\sigma} \approx 0.9898$ to trigger the asymptote. The practical implication is that inference for ω can be fragile near the boundary even when σ appears well recovered, and conversely that moderate $\hat{\omega}$ values do not guarantee accurate recovery of σ and ρ.

7.3 Trade setting

We also examine the common-product IV idea in a trade-data context using the Monte Carlo design (MC4) provided by Grant and Soderbery (2024), calibrated to SITC 8947 and constructed to allow for weak instruments and potential exclusion-restriction failures. A key difference from scanner data is feasibility: requiring a supplying country

to be present in every year of a multi-decade panel can leave too few (or no) "common" units. In such cases, the strict common-product strategy fails mechanically, weakening direct applicability in long-horizon trade settings.

As a robustness check, we therefore relax the criterion and construct common-product instrument sets using shorter survival windows that are interpretable as "stable" trade relationships (e.g., 5-year or 10-year windows). We compare these estimates with standard WLS and with LIML using all instruments. Table 6 reports mean estimates and failure rates across methods and window definitions. When the strict full-horizon rule is imposed, failure rates exceed 60%, indicating that the strictest version of the strategy is often infeasible in typical trade panels. Using shorter windows substantially reduces failures, bringing them close to the rates observed under standard WLS or LIML with all instruments. In terms of point estimates, the trade-data environment differs from scanner data: standard WLS and LIML do not display the same extreme inflation in σ, and LIML tends to improve upon WLS, consistent with Grant and Soderbery (2024). Even so, the window-based common-product variants still tend to improve estimation of σ and ω, suggesting that the instrument-design idea remains valuable but must be adapted to context.

[Insert Table 6]

7.4 Using Iterated GMM and LIML with Fuller's modification

An additional practical point is that the iterated GMM option (iGMM) and Fuller's modification (F-LIML) can materially affect the performance of the corresponding GMM and LIML estimators in scanner-data settings. Fuller's modification is designed to reduce small-sample bias and improve behaviour under weak or near-weak instruments relative to plain LIML/2SLS. We also experimented with the continuously-updated GMM estimator or CUE of Hansen, Heaton and Yaron (1996). In our scanner-data applications, CUE behaved similarly to LIML but was less stable, so we leave the results for interested readers upon requests.

Table 7 reports estimates for the real scanner datasets using iterated comGMM and Fuller-modified comLIML. Relative to plain comGMM, iterated comGMM further reduces the estimated substitution elasticity in Cakes and Ready Meals (e.g., $\hat{\sigma} = 8.589$ for Cakes and $\hat{\sigma} = 4.260$ for Ready Meals), while leaving the overall pattern

unchanged: σ remains far below the FBW estimates and is precisely estimated. For comLIML, Fuller's modification increases σ in the quarterly categories and yields more moderate implied ω values than plain comLIML, particularly for Cakes and Meat. However, σ under Fuller-LIML remains statistically insignificant in these two quarterly categories, as reflected in the large standard errors. In contrast, in the monthly Ready Meals category, Fuller-LIML produces a precisely estimated $\hat{\sigma}$ and a significant $\hat{\omega}$.

Overall, these extensions do not alter our main qualitative conclusion that comGMM is the most stable and reliable estimator in practice, while LIML-based estimates can be sensitive and, in quarterly panels, often imprecise. A practical drawback is computation time: both iGMM and Fuller-LIML are noticeably slower than plain GMM/LIML in large panels, and Fuller-LIML can take several hours in high-UPC categories such as Meat.

[Insert Table 7]

8 Conclusion

Reliable estimation of CES elasticities is essential for demand-based price indices that aim to separate pure price movements from substitution, product entry/exit, and taste shocks. In food markets—where product turnover is high and household budgets are sensitive to price changes—these parameters feed directly into the measurement of food inflation and the welfare analysis surrounding inflation. When the elasticity of substitution (σ) is overstated, demand-based cost-of-living indices mechanically attribute too much adjustment to substitution and too little to variety and taste components. In the FBW system, the inverse supply elasticity (ω) is identified jointly with σ and shapes the interpretation of supply-side forces in observed price–quantity movements.

This paper shows that widely used scanner-data implementations of the FBW approach can perform poorly in the environments where scanner data are most valuable: large numbers of products, high frequency, and severe panel unbalancedness. The evidence from Monte Carlo experiments calibrated to UK scanner data indicates that the conventional FBW-WLS implementation can generate extreme upward bias in σ and

meaningful distortions in ω, driven largely by instrument design when short-lived items dominate the effective instrument set.

We propose a transparent alternative: a common-product IV strategy that restricts instruments to long-lived products observed in every period. Implemented as comGMM and comLIML, this refinement reduces instrument dimensionality, improves finite-sample performance, and yields more plausible elasticity estimates in UK scanner-data applications, while retaining the underlying identification logic and enabling direct IV/GMM/LIML estimation and associated diagnostics. A practical implication is that very large σ estimates from scanner data should be treated with caution unless instrument construction and the role of short-lived products have been carefully assessed; conversely, common-product instruments provide a simple, replicable default that can be complemented with robustness checks based on alternative longevity thresholds.

Several extensions are natural. One is to develop data-driven rules for defining "long-lived" instruments when the strict common-product criterion is too restrictive (for example, in long-horizon trade panels), and to characterise the associated bias–variance trade-off across alternative thresholds or rolling-window definitions. Another is to extend the analysis from elasticity estimation to full index construction by implementing demand-based inflation and welfare decompositions under alternative instrument strategies, thereby quantifying the sensitivity of those decompositions to σ (and ω) in scanner environments.

References

Argente, D., & Lee, M. (2021). Cost of living inequality during the great recession. *Journal of the European Economic Association*, *19*(2), 913-952.

Broda, C., & Weinstein, D. E. (2006). Globalization and the Gains from Variety. *The Quarterly journal of economics*, *121*(2), 541-585.

Broda, C., & Weinstein, D. E. (2010). Product creation and destruction: Evidence and price implications. *American Economic Review*, *100*(3), 691-723.

Chen, T., Levell, P., & O'Connell, M. (2024). Cheapflation and the rise of inflation inequality. *CEPR Discussion Papers*, *1*, 22.

Colicev, A., Hoste, J., & Konings, J. (2024). The impact of a large depreciation on the cost of living of rich and poor consumers. *International Economic Review*, *65*(4), 1625-1656.

De Haan, J., & Krsinich, F. (2024, May). Product Churn and the GEKS-Törnqvist Price Index: The "Feenstra Adjustment". In *18th Meeting of the Ottawa Group, Ottawa, Canada. website: https://stats. unece. org/ottawagroup/meeting/18*.

Diewert, W. E., & Feenstra, R. C. (2019). *Estimating the benefits of new products* (No. c14281). National Bureau of Economic Research.

Ehrlich, G., Haltiwanger, J. C., Jarmin, R. S., Johnson, D., Olivares, E., Pardue, L. W., ... & Zhao, L. (2023). *Quality adjustment at scale: Hedonic vs. exact demand-based price indices* (No. w31309). National Bureau of Economic Research.

Feenstra, R. C. (1994). New product varieties and the measurement of international prices. *The American Economic Review*, 157-177.

Feenstra, R. C., Luck, P., Obstfeld, M., & Russ, K. N. (2018). In search of the Armington elasticity. *Review of Economics and Statistics*, *100*(1), 135-150.

Grant, M., & Soderbery, A. (2024). Heteroskedastic supply and demand estimation: Analysis and testing. *Journal of International Economics*, *150*, 103817.

Hansen, L. P., Heaton, J., & Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics*, *14*(3), 262-280.

Jaravel, X. (2019). The unequal gains from product innovations: Evidence from the us retail sector. *The Quarterly Journal of Economics*, *134*(2), 715-783.

Leamer, E. E. (1981). Is it a demand curve, or is it a supply curve? Partial identification through inequality constraints. *The Review of Economics and Statistics*, 319-327.

Mohler, L. (2009). On the sensitivity of estimated elasticities of substitution. *FREIT Worker Paper*, (38).

Redding, S. J., & Weinstein, D. E. (2020). Measuring aggregate price indices with taste shocks: Theory and evidence for CES preferences. *The Quarterly Journal of Economics*, *135*(1), 503-560.

Soderbery, A. (2010). Investigating the asymptotic properties of import elasticity estimates. *Economics Letters*, *109*(2), 57-62.

Soderbery, A. (2015). Estimating import supply and demand elasticities: Analysis and implications. *Journal of International Economics*, *96*(1), 1-17.

von Brasch, T., & Raknerud, A. (2021). *A two-stage pooled panel data estimator of demand elasticities* (No. 951). Discussion Papers.

von Brasch, T., Raknerud, A., & Vigtel, T. C. Identifying Demand Elasticity via Heteroscedasticity.

Tables and Figures

Table 1. Summary statistics of UPCs per category and income group

Category	Chilled and frozen cakes (01.1.1.4)	Meat (01.1.2)	Ready meals
No. of UPCs	2,381	18,705	6,726
Common UPC share	0.042	0.035	0.036
UPC length (mean)	15.3	15.9	46.9
UPC length (SD)	11.4	11.3	32.9
UPC length (min)	4	4	13
UPC length (max)	44	44	130

Note: UPC length is the number of periods a UPC is observed in the panel (quarters for Cakes and Meat; months for Ready Meals). "Common UPC share" is the fraction of UPCs observed in every period of the sample. Income group is randomly selected for each category.

Table 2. Summary of estimation results across different methods

	Mean Estimates			Correction power to FBW (%)	
	FBW-WLS	comGMM	comLIML	comGMM	comLIML
			sim_data1 (Stylised–MaxN5k)		
$\rho = 0.4$	0.913	0.531	0.402	74.5	99.6
$\sigma = 3$	47.971	4.521	2.971	96.6	99.9
$\omega = 0.5$	0.436	0.489	0.567	82.8	-4.7
			sim_data2 (Stylised–MaxN10k)		
$\rho = 0.4$	0.914	0.527	0.391	75.3	98.2
$\sigma = 3$	40.512	4.43	2.98	96.2	99.9
$\omega = 0.5$	0.437	0.488	0.506	81.0	90.5
			sim_discat4 (Template–Cakes-Q)		
$\rho = 0.4$	0.945	0.511	0.419	79.6	96.5
$\sigma = 3$	98.673	4.312	3.06	98.6	99.9
$\omega = 0.5$	0.419	0.476	0.777	70.4	-242.0
			sim_aggcat2 (Template–Meat-Q)		
$\rho = 0.4$	0.93	0.506	0.399	80.0	99.8
$\sigma = 3$	47.354	4.207	3.008	97.3	100.0
$\omega = 0.5$	0.419	0.472	0.511	65.4	86.4
			sim_RMdata (Template–RM-M)		
$\rho = 0.4$	0.863	0.474	0.402	84.0	99.6
$\sigma = 3$	21.176	3.738	3.014	95.9	99.9
$\omega = 0.5$	0.463	0.493	0.509	81.1	75.7

Note: "sim_data1- sim_data2" are stylised simulations with $N \sim U(500, MaxN)$. "sim_discat4 – sim_RMdata" are template-based simulations that replicate the unbalanced panel structure (entry/exit and UPC length distribution) from the named Kantar categories at the stated frequency (see table 1); only shocks are simulated. Correction power$(m) = \frac{|e_{wls}| - |e_m|}{|e_{wls}|} \times 100$ where e_{wls} and e_m refer to errors between estimates and true values for WLS and $m = comGMM, comLIML$ respectively. If it's 100%: perfect correction (hits the truth); if it's 0: no improvement over WLS; and if it's negative: worse than WLS.

Table 3. Percentage of failed cases

Simulated dataset	FBW-WLS	comGMM	comLIML
	$\hat{\sigma}$		
sim_data1	0	0	0
sim_data2	0	0	2
sim_discat4	0	0	2
sim_aggcat2	11	0	4
sim_RMdata	0	0	0
	$\hat{\omega}$		
sim_data1	3	0	3
sim_data2	1	0	2
sim_discat4	0	0	2
sim_aggcat2	16	0	7
sim_RMdata	0	0	0

Table 4. Estimation based on real scanner datasets

Estimates	FBW-WLS	comGMM	comLIML
	Chilled and frozen cakes (01.1.1.4)		
ρ	0.969^{***}	0.773^{***}	0.025
	(0.017)	(0.029)	(0.188)
σ	76.677^{**}	13.77^{***}	1.618
	(34.892)	(1.237)	(3.555)
ω	0.704^{***}	0.364^{***}	0.042
	(0.153)	(0.037)	(0.094)
	Meat (01.1.2)		
ρ	0.980^{***}	0.827^{***}	0.072
	(0.007)	(0.009)	(0.118)
σ	94.890^{***}	14.939^{***}	2.211
	(29.667)	(0.507)	(1.386)
ω	1.096^{***}	0.520^{***}	0.069
	(0.134)	(0.021)	(0.045)
	Chilled and frozen ready meals		
ρ	0.766^{***}	0.321^{***}	0.140
	(0.037)	(0.015)	(0.103)
σ	13.385^{***}	5.323^{***}	3.682^{***}
	(1.334)	(0.147)	(1.221)
ω	0.3581^{***}	0.123^{***}	0.065^{***}
	(0.0502)	(0.005)	(0.028)

Note: Robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Table 5. Diagnostic Tests after comGMM and comLIML

	comGMM		comLIML	
	Mean	Median	Mean	Median
		sim_data1		
F^{KP}	2.129	2.092	2.129	2.092
J^{Stats}	231.114	217.085	92.322	94.776
$J^{\text{P-value}}$	0.014	0.000	0.912	0.996
		sim_data2		
F^{KP}	2.199	2.201	2.199	2.201
J^{Stats}	444.362	379.913	184.639	163.868
$J^{\text{P-value}}$	0.002	0.000	0.919	0.991
		sim_discat4		
F^{KP}	1.723	1.704	1.723	1.704
J^{Stats}	170.596	166.642	62.913	65.026
$J^{\text{P-value}}$	0.009	0.000	0.915	0.995
		sim_aggcat2		
F^{KP}	1.774	1.765	1.774	1.765
J^{Stats}	1212.422	1212.834	509.720	540.378
$J^{\text{P-value}}$	0.000	0.000	0.982	1.000
		sim_RMdata		
F^{KP}	1.905	1.907	1.905	1.907
J^{Stats}	329.227	330.730	217.011	217.132
$J^{\text{P-value}}$	0.028	0.000	0.764	0.863

Note: F^{KP} is Kleibergen–Paap rk Wald F-statistics, J^{Stats} and $J^{\text{P-value}}$ is the J-statistic and p-value of the estimated J-statistic.

Table 6. Summary of estimation using trade data

	Mean Estimates			Failed (Proportion)	
	$\rho = 0.364$	$\sigma = 3$	$\omega = 0.4$	$\sigma = 3$	$\omega = 0.4$
			FBW		
WLS/GMM	0.442	4.877	0.449	0.015	0.013
LIML	0.362	2.781	0.558	0.017	0.01
		Common IV set (full window – 26 years)			
GMM	0.4	3.855	0.804	0.658	0.634
LIML	0.405	4.302	0.836	0.678	0.644
		Common IV set (5-year window)			
GMM	0.438	3.953	0.45	0.015	0.015
LIML	0.362	2.758	0.554	0.011	0.021
		Common IV set (10-year window)			
GMM	0.429	4.17	0.454	0.013	0.013
LIML	0.364	2.772	0.591	0.032	0.043

Table 7. Estimation based on iGMM and F-LIML

Estimates	FBW (WLS)	comGMM (iGMM)	comLIML (F-LIML)
	Chilled and frozen cakes (01.1.1.4)		
ρ	0.969***	0.566***	0.116
	(0.017)	(0.039)	(0.492)
σ	76.677**	8.589***	2.865
	(34.893)	(0.668)	(5.576)
ω	0.704***	0.207***	0.076
	(0.153)	(0.019)	(0.147)
	Meat (01.1.2)		
ρ	0.980***	0.800***	0.076
	(0.007)	(0.010)	(0.122)
σ	94.831***	13.393***	2.255
	(29.628)	(0.435)	(1.407)
ω	1.096***	0.477***	0.070
	(0.134)	(0.019)	(0.046)
	Ready Meals		
ρ	0.766***	0.209***	0.141
	(0.037)	(0.009)	(0.103)
σ	13.385***	4.260***	3.692***
	(1.334)	(0.074)	(1.223)
ω	0.358***	0.088***	0.065**
	(0.050)	(0.003)	(0.028)

Note: Robust standard errors are reported in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Figure 1. Distribution of $\hat{\sigma}$ by methods and datasets

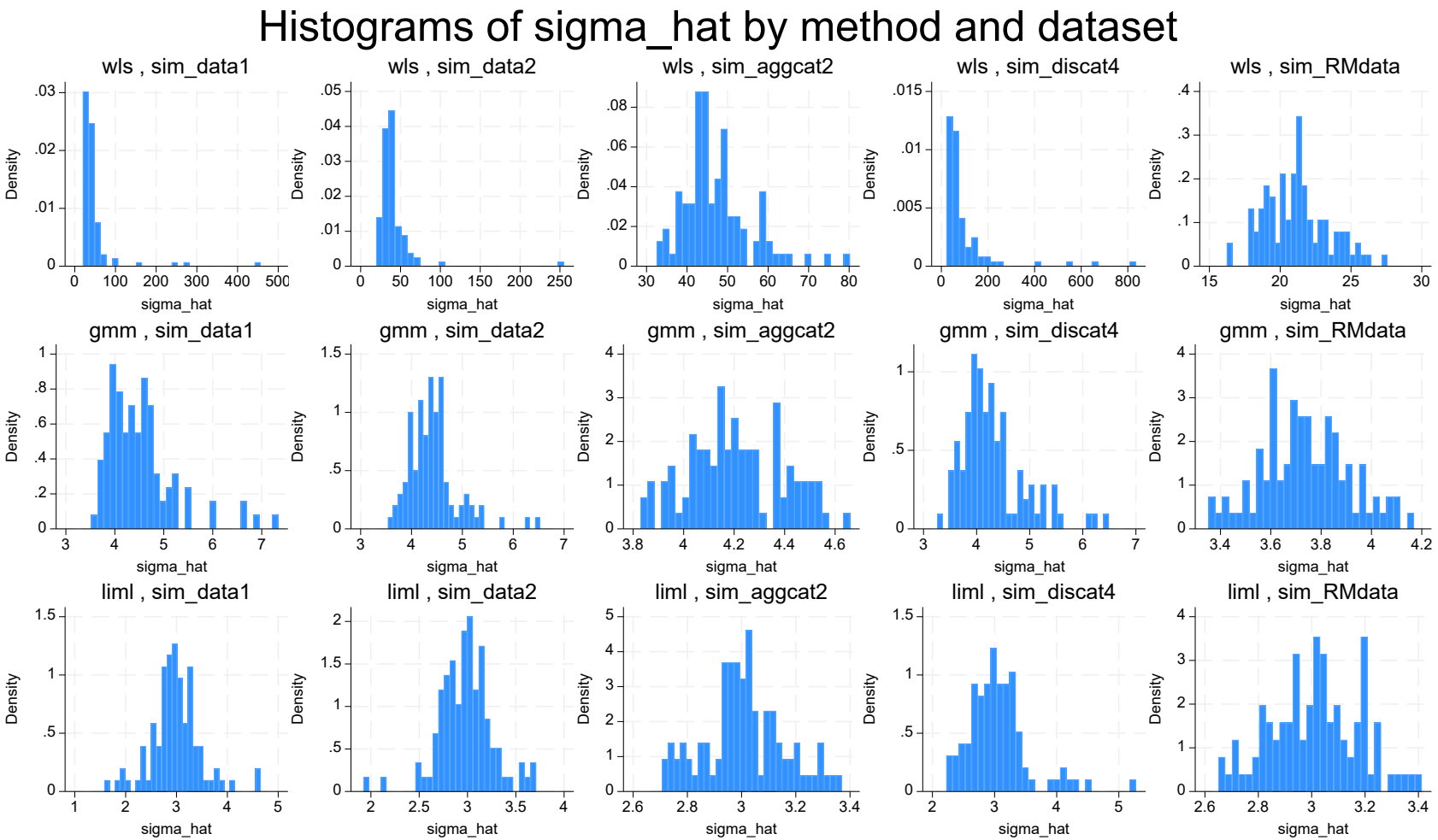


Figure 2. Distribution of $\hat{\omega}$ by methods and datasets

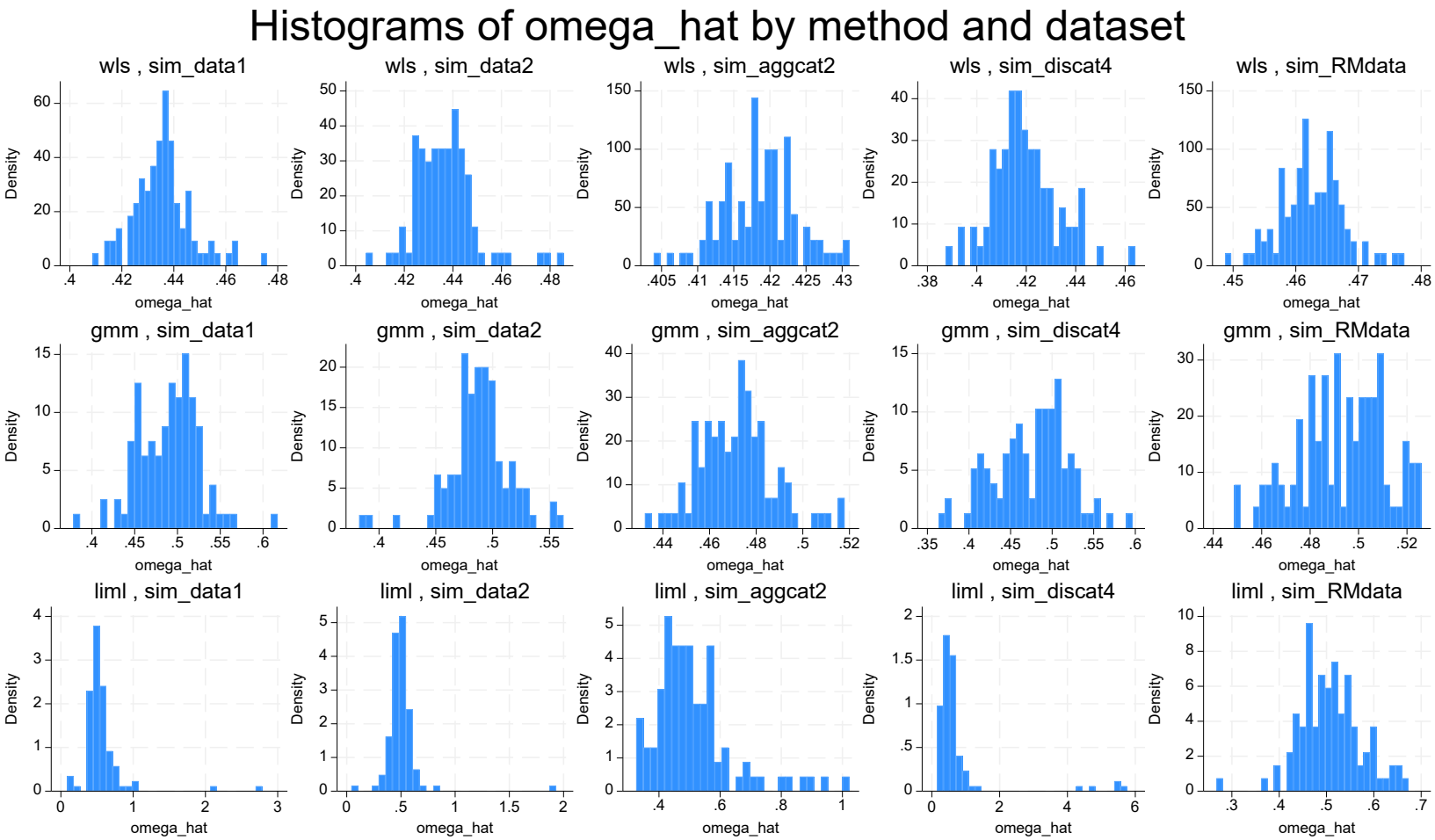


Figure 3. Distribution of S.E. of $\hat{\sigma}$ by methods and datasets

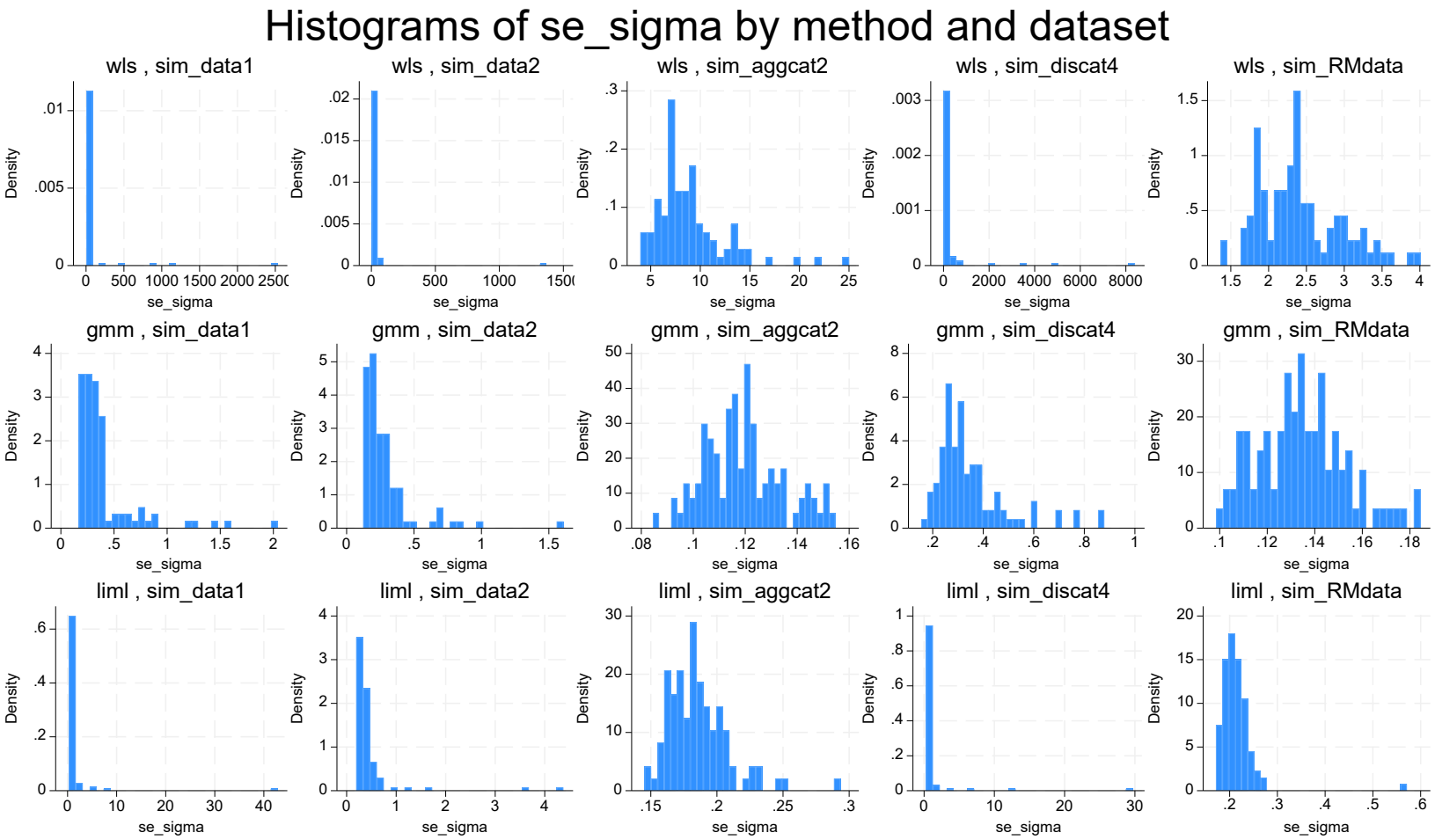
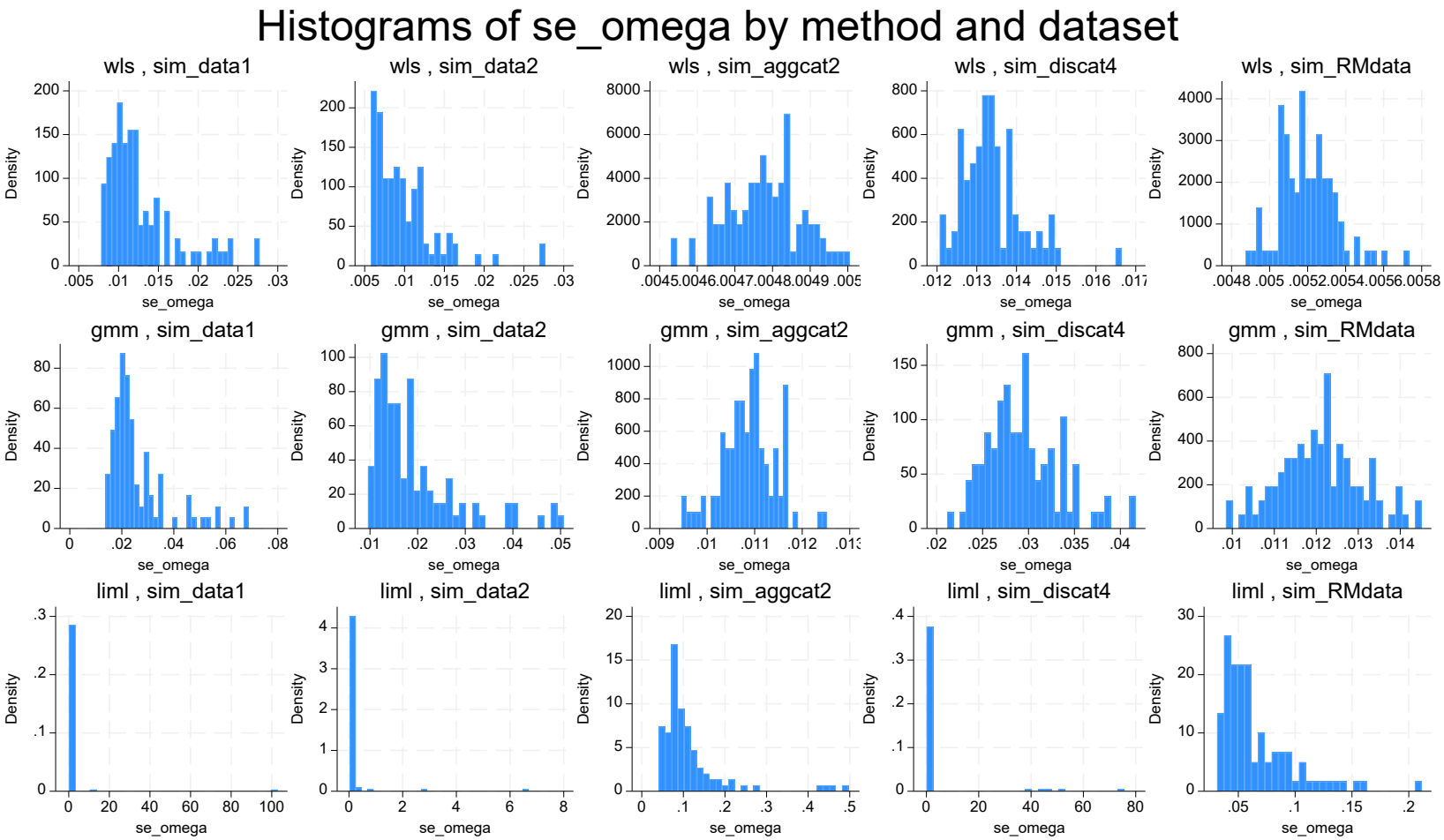


Figure 4. Distribution of S.E. of $\hat{\omega}$ by methods and datasets



Online Appendix

A1. Parameters derivation

Feenstra (1994) summarizes the possible outcomes for $\hat{\sigma}$ and $\hat{\rho}$:

If $\hat{\theta}_1 > 0$ and $\hat{\theta}_2 > 0$, then

$$\hat{\sigma} = 1 + \left(\frac{2\hat{\rho} - 1}{1 - \hat{\rho}}\right)\frac{1}{\hat{\theta}_2} \quad where \quad \hat{\rho} = \frac{1}{2} + \left(\frac{1}{4} - \frac{1}{4 + \left(\hat{\theta}_2^2/\hat{\theta}_1\right)}\right)^{\frac{1}{2}}$$

If $\hat{\theta}_1 > 0$ and $\hat{\theta}_2 < 0$, then

$$\hat{\sigma} = 1 + \left(\frac{2\hat{\rho} - 1}{1 - \hat{\rho}}\right)\frac{1}{\hat{\theta}_2} \quad where \quad \hat{\rho} = \frac{1}{2} - \left(\frac{1}{4} - \frac{1}{4 + \left(\hat{\theta}_2^2/\hat{\theta}_1\right)}\right)^{\frac{1}{2}}$$

If $\hat{\theta}_1 < 0$, then estimates of ρ and σ that are both in the ranges $\hat{\sigma} > 1$ and $0 \leq \hat{\rho} < 1$ cannot be obtained.

From the definition of ρ, we can finally recover the $\hat{\omega}$ as

$$\hat{\omega} = \frac{\hat{\rho}}{\hat{\sigma}(1 - \hat{\rho}) - 1}$$