Lecture 4: Linear GMM

January 24, 2025

Linear GMM Estimation

- We now turn to a general treatment with (potential) overidentification, that is, $L > K$.
- Though we assume the population moment conditions $E\left(\mathbf{Z}_i'\mathbf{X}_i\right)\boldsymbol{\beta} = E\left(\mathbf{Z}_i'\mathbf{y}_i\right)$ uniquely determine $\boldsymbol{\beta}$, the sample analog,

$$\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{X}_i\right)\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{y}_i\right), \tag{1.1}$$

  generally has no solution when $L > K$ (L equations in K unknowns).

Linear GMM Estimation

- We could choose $\hat{\boldsymbol{\beta}}$ to make the Euclidean length of the vector $\sum_{i=1}^{N} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}\right)$ as small as possible, that is, choose $\hat{\boldsymbol{\beta}}$ to solve

$$\min_{\mathbf{b}\in\mathbb{R}^K} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right)\right]' \left[\sum_{i=1}^{N} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i\mathbf{b}\right)\right]. \tag{1.2}$$

- This estimator is consistent and is sometimes used as an initial estimator, but it is essentially never efficient.

- Consider a general class of estimators that use a weighted Euclidean length.

Linear GMM Estimation

- Let $\hat{\mathbf{W}}$ be an $L \times L$ symmetric, positive semi-definite matrix, which can be random (and usually depends on the same random sample of data). Consider now the problem

$$\min_{\mathbf{b} \in \mathbb{R}^K} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i \mathbf{b} \right) \right]' \hat{\mathbf{W}} \left[\sum_{i=1}^{N} \mathbf{Z}_i' \left(\mathbf{y}_i - \mathbf{X}_i \mathbf{b} \right) \right]. \tag{1.3}$$

- The solution to (1.3) is called a **generalized method of moments (GMM)** estimator.
- Can solve this problem using multivariable calculus.
- Let $\mathbf{Z}$ be the $NG \times L$ matrix of instruments stacked by observation, and similarly for $\mathbf{X}(NG \times K)$ and $\mathbf{Y}(NG \times 1)$. Can show that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}\mathbf{Z}'\mathbf{X} \right)^{-1} \left(\mathbf{X}'\mathbf{Z}\hat{\mathbf{W}}\mathbf{Z}'\mathbf{Y} \right), \tag{1.4}$$

where

$$\mathbf{Z}'\mathbf{X} = \sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{X}_i, \mathbf{Z}'\mathbf{Y} = \sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{y}_i \tag{1.5}$$

Linear GMM Estimation

Assumption SIV. 3 (Positive Definite Limit): For an $L \times L$ nonrandom positive definite matrix $\mathbf{W}$,

$$\hat{\mathbf{W}} \xrightarrow{p} \mathbf{W} \text{ as } N \to \infty. \tag{1.6}$$

Positive definiteness is stonger than needed. As will be clear, having $\mathbf{C}'\mathbf{W}\mathbf{C}$ full rank (nonsigular) is sufficient, where $\mathbf{C} = E\left(\mathbf{Z}_i'\mathbf{X}_i\right)$.

Theorem (Consistency): Under SIV. 1 to SIV.3, $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$.

Proof: Write

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + &\left[\left(N^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Z}_i\right)\hat{\mathbf{W}}\left(N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1} \\ &\cdot\left(N^{-1}\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Z}_i\right)\hat{\mathbf{W}}\left(N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i\right) \end{aligned} \tag{1.7}$$

$$\begin{aligned} \text{plim}_{N\to\infty}(\hat{\boldsymbol{\beta}}) =& \boldsymbol{\beta} + \left(\mathbf{C}'\mathbf{W}\mathbf{C}\right)^{-1}\mathbf{C}'\mathbf{W}\left(\text{plim}_{N\to\infty} N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i\right) \\ =& \boldsymbol{\beta} + \left(\mathbf{C}'\mathbf{W}\mathbf{C}\right)^{-1}\mathbf{C}'\mathbf{W}\cdot\mathbf{0} = \boldsymbol{\beta}. \end{aligned} \tag{1.8}$$

Linear GMM Estimation

Theorem (Asymptotic Normality): Under SIV.1, SIV.2, and SIV.3, $\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ is asymptotically normal with mean zero and variance matrix

$$Avar\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\mathbf{C}'\mathbf{W}\mathbf{C}\right)^{-1}\mathbf{C}'\mathbf{W}\boldsymbol{\Lambda}\mathbf{W}\mathbf{C}\left(\mathbf{C}'\mathbf{W}\mathbf{C}\right)^{-1}. \tag{1.9}$$

where

$$\boldsymbol{\Lambda} = \mathrm{Var}\left(\mathbf{Z}_i'\mathbf{u}_i\right) = E\left(\mathbf{Z}_i'\mathbf{u}_i\mathbf{u}_i'\mathbf{Z}_i\right). \tag{1.10}$$

- Consistent estimation of $Avar\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ uses $\hat{\mathbf{C}} = N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{X}_i$, $\hat{\mathbf{W}}$ and a consistent estimator $\hat{\boldsymbol{\Lambda}}$ of $\boldsymbol{\Lambda}$ (more later).

System 2SLS

- The **System 2SLS** estimator uses weight matrix

$$\hat{\mathbf{W}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \mathbf{Z}_i \right)^{-1}, \tag{2.1}$$

  and the estimator can be written as

$$\hat{\boldsymbol{\beta}}_{\text{S2SLS}} = \left[\mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{X} \right]^{-1} \mathbf{X}'\mathbf{Z} \left(\mathbf{Z}'\mathbf{Z} \right)^{-1} \mathbf{Z}'\mathbf{Y} \tag{2.2}$$

- Inference with S2SLS is possible without further assumptions.
- In the SUR case, the S2SLS estimator is **2SLS equation-by-equation**.

Optimal Weighting Matrix

- Given an infinite number of choices for $\mathbf{W}$, can we choose the best? Yes. (And then it is obvious how to estimate the optimal $\mathbf{W}$.)
- Recall that $\boldsymbol{\beta}$ is defined by

$$E\left[\mathbf{Z}_i'\left(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta}\right)\right] = E\left(\mathbf{Z}_i'\mathbf{u}_i\right) = \mathbf{0}. \tag{3.1}$$

- The optimal weighting matrix is the inverse of the variance matrix of $\mathbf{Z}_i'\mathbf{u}_i, \boldsymbol{\Lambda} = \mathrm{Var}\left(\mathbf{Z}_i'\mathbf{u}_i\right)$

Optimal Weighting Matrix

Assumption SIV. 4 (Optimal Weighting Matrix):

$$\mathbf{W} = \mathbf{\Lambda}^{-1}. \tag{3.2}$$

- With this choice of $\mathbf{W}$, the asymptotic variance collapses to

$$\left(\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{\Lambda}^{-1}\mathbf{C}\left(\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{C}\right)^{-1} = \left(\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{C}\right)^{-1}. \tag{3.3}$$

- It can be shown that $\left(\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{C}\right)$ is the "smallest" possible by showing

$$\left(\mathbf{C}'\mathbf{\Lambda}^{-1}\mathbf{C}\right) - \left(\mathbf{C}'\mathbf{W}\mathbf{C}\right)\left(\mathbf{C}'\mathbf{W}\mathbf{\Lambda}\mathbf{W}\mathbf{C}\right)^{-1}\left(\mathbf{C}'\mathbf{W}\mathbf{C}\right) \tag{3.4}$$

  is positive semi-definite for any $L \times L$ positive definite matrix $\mathbf{W}$.

Optimal Weighting Matrix

To obtain an actual GMM estimator using an efficient weighting matrix, use a two-step procedure.

(1) Let $\check{\boldsymbol{\beta}}$ be an initial consistent estimator of $\boldsymbol{\beta}$, usually the system 2SLS estimator or the estimator using $\hat{\mathbf{W}} = \mathbf{I}_L$. Obtain the $G \times 1$ residual vectors, $\check{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i \check{\boldsymbol{\beta}}, i = 1, \ldots, N$, and compute

$$\hat{\boldsymbol{\Lambda}} = N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \check{\mathbf{u}}_i \check{\mathbf{u}}_i' \mathbf{Z}_i \xrightarrow{p} \boldsymbol{\Lambda}. \tag{3.5}$$

(2) Choose

$$\hat{\mathbf{W}} = \left(N^{-1} \sum_{i=1}^{N} \mathbf{Z}_i' \check{\mathbf{u}}_i \check{\mathbf{u}}_i' \mathbf{Z}_i \right)^{-1}. \tag{3.6}$$

Optimal Weighting Matrix

- We call such an estimator an **optimal GMM** estimator. It is sometimes called a **minimum chi-square** estimator (reason will be clear later).
- Important: the optimal weighting matrix provides an asymptotically efficient estimator in the class of estimators based on

$$E\left[\mathbf{Z}_i'(\mathbf{y}_i - \mathbf{X}_i\boldsymbol{\beta})\right] = \mathbf{0}. \tag{3.7}$$

  In other words, the estimator is asymptotically efficient for the given set of moment conditions (instruments).
- It is possible we can find additional moment conditions that can enhance efficiency.

Optimal Weighting Matrix

- When $L = K$ the weighting matrix is irrelevant. There is only one estimator consistent under (3.7), and that is the system IV estimator.
- If $\hat{\boldsymbol{\beta}}$ is now an optimal GMM estimator and $\hat{\mathbf{u}}_i = \mathbf{y}_i - \mathbf{X}_i\hat{\boldsymbol{\beta}}$ are the optimal GMM residuals,

$$\begin{aligned}\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}) &= N^{-1}\left[(\mathbf{X}'\mathbf{Z}/N)\left(N^{-1}\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}(\mathbf{Z}'\mathbf{X}/N)\right]^{-1} \\ &= \left[(\mathbf{X}'\mathbf{Z})\left(\sum_{i=1}^{N}\mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}(\mathbf{Z}'\mathbf{X})\right]^{-1}\end{aligned} \tag{3.8}$$

- Recent research has focused on the small-sample properties of optimal GMM estimators. While replacing $\boldsymbol{\Lambda}$ with $\hat{\boldsymbol{\Lambda}}$ does not affect the $\sqrt{N}$-asymptotic distribution of $\hat{\boldsymbol{\beta}}$, it can have deleterious effects on the actual (finite sample) distribution. (More sophisticated asymptotic analysis picks this up.)

Hausman, Leonard, and Zona (1994)

Demand Estimation

Consumer Preference of Differentiated Products

- Store level data of differentiated products with brand names: cereal, beer, soda, etc.
- Two challenges:
  - A large number of brands and products
  - Traditional cost shifters common across brands
- Solutions
  - Model a multi-level demand system
  - Instruments: use spatial variation at city-level
- Use scanner data of beer industry for competitive analysis

Multi-level Demand System

- Gorman (1971)
  - Top level: overall demand for beer
  - Middle level: segments of beer - premium, light, imported, non-premium
  - Bottom level: brands given a segment - Bud, Miller, etc
- Bottom level takes a Almost Ideal Demand System (AIDS)
  - Deaton and Mellbauer

$$s_{int} = \alpha_{in} + \beta_i \log \left(y_{Gnt} / \mathrm{P}_{nt} \right) + \sum_{j=1}^{J} \gamma_{ij} \log p_{jnt} + \varepsilon_{int}, \quad i = 1, \ldots, J, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T \tag{1}$$

  - i: brands, n: city, t: time,
    where s_{int} is the revenues share of total segment expenditure of the ith brand in city n in period t, y_{Gnt} is overall segment expenditure, P_{nt} is a price index, and p_{jnt} is the price of the jth brand in city n.

Multi-level Demand System

- How to calculate segment price index?
  - Most empirical literature use a linear approximation
  - $\log(P) = \sum_j s_j \log(p_j)$
- Middle level: Log-Log demand system

$$\log q_{mnt} = \beta_m \log y_{Bnt} + \sum_{k=1}^{n} \delta_k \log \pi_{knt} + \alpha_{mn} + \varepsilon_{mnt} \tag{2}$$
$$m = 1, \ldots, M, \quad n = 1, \ldots, N, \quad t = 1, \ldots, T$$

where the left hand side variable q_{mnt} is log quantity of the mth segment in city n in period t, the expenditure variable y_{Bnt} is total beer expenditure, and the Π_{knt} are the segment price indices for city n.

Multi-level Demand System

- Top level (i.e. demand for beer)
- This is more or less similar to the broiler example we covered

$$\log u_t = \beta_0 + \beta_1 \log y_t + \beta_2 \log \Pi_t + \mathrm{Z}_t \delta + \varepsilon_t \tag{3}$$

where u_t is overall consumption of beer, y_t is deflated disposable income, Π_t is the deflated price index for beer, and Z_t are variables which account for changes in demographics, monthly (seasonal) factors, and minimum age for purchasing beer.

To estimate equation (3) we use national (BLS) monthly data over a sixteen year period with instrumental variables. We have found that a longer time period than may be available from store level data is often useful to estimate the top level demand elasticity. The instruments we use in estimation of equation (3), are factors which shift costs such as different ingredients, packaging, and labor. We estimate the overall price elasticity of beer to be -1.36 with an estimated standard error of 0.21.

Deaton and Muellbauer (1980)

- AIDS cost (expenditure) function

$$\log c(u,p) = \alpha_0 + \sum_k \alpha_k \log p_k + \frac{1}{2}\sum_k \sum_j \gamma^*_{kj} \log p_k \log p_j + u\beta_0 \prod_k p_k^{\beta_k} \tag{DM.4}$$

- Demand function (Shepard's Lemma)

$$\frac{\partial \log c(u,p)}{\partial \log p_i} = \frac{p_i q_i}{c(u,p)} = w_i \tag{DM.5}$$

$$w_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i u \beta_0 \Pi p_k^{\beta_k} \tag{DM.6}$$

where

$$\gamma_{ij} = \frac{1}{2}\left(\gamma^*_{ij} + \gamma^*_{ji}\right) \tag{DM.7}$$

Deaton and Muellbauer (1980)

For a utility-maximizing consumer, total expenditure x is equal to $c(u, p)$ and this equality can be inverted to give u as a function of p and x, the indirect utility function. If we do this for (DM.4) and substitute the result into (DM.6) we have the budget shares as a function of p and x; these are the AIDS demand functions in budget share form:

$$w_i = \alpha_i + \sum_j \gamma_{ij} \log p_j + \beta_i \log (x/P) \tag{DM.8}$$

where P is a price index defined by

$$\log P = \alpha_0 + \sum_k \alpha_k \log p_k + \frac{1}{2} \sum_j \sum_k \gamma_{kj} \log p_k \log p_j \tag{DM.9}$$

Deaton and Muellbauer (1980)

The restrictions on the parameters of (DM.4) plus equation (DM.7) imply restrictions on the parameters of the AIDS equation (DM.8). We take these in three sets

$$\sum_{i=1}^{n} \alpha_i = 1 \quad \sum_{i=1}^{n} \gamma_{ij} = 0 \quad \sum_{i=1}^{n} \beta_i = 0 \tag{DM.10}$$

$$\sum_{j} \gamma_{ij} = 0 \tag{DM.11}$$

$$\gamma_{ij} = \gamma_{ji} \tag{DM.12}$$

Provided (DM.10), (DM.11), and (DM.12) hold, equation (DM.8) represents a system of demand functions which add up to total expenditure ($\sum w_i = 1$), are homogeneous of degree zero in prices and total expenditure taken together, and which satisfy Slutsky symmetry.

Identification

What is potential IV at bottom/middle level?
Think about a pricing model

$$\log p_{jnt} = \delta_j \log c_{jt} + \alpha_{jn} + w_{jnt} \tag{4}$$

> *where p_{jnt} is the price for brand j in city n in period t. The determinants of the brand price for brand j are c_{it}, the cost which is assumed not to have a city specific time shifting component which is consistent with the national shipments and advertising of most differentiated products, α_{jn}, which is a city specific brand differential which accounts for transportation costs or local wage differentials, and w_{jnt}, which is a mean zero stochastic disturbance which accounts of sales promotion run for brand j in city n*

Key assumption: promotions are independent across cities

> *The idea is that prices in one city (atter elimination of city and brand specific effects) are driven by underlying costs, c_{jt}, which provide instrumental variables which are correlated with prices but are uncorrelated with stochastic disturbances in the demand equations, e.g., w_{jnt} from equation* (4) *is uncorrelated with ε_{ilt} from equation* (1) *when the cities are different, $n \neq 1$. Thus, the availability of panel data is a crucial factor which allows for estimation of the all the own price and cross price brand elasticities.*

Middle-level Result

- Quite Strong cross-segment substitution patterns

Beer Segment Conditional Demand Equations.

	Premium	Popular	Light
Constant	0.501	−4.021	−1.183
	(0.283)	(0.560)	(0.377)
log (Beer Exp)	0.978	0.943	1.067
	(0.011)	(0.022)	(0.015)
log ($P_{PREMIUM}$)	−2.671	2.704	0.424
	(0.123)	(0.244)	(0.166)
log ($P_{POPULAR}$)	0.510	−2.707	0.747
	(0.097)	(0.193)	(0.127)
log (P_{LIGHT})	0.701	0.518	−2.424
	(0.070)	(0.140)	(0.092)
Time	−0.001	−0.000	0.002
	(0.000)	(0.001)	(0.000)
log (# of Stores)	−0.035	0.253	−0.176
	(0.016)	(0.034)	(0.023)

Number of Observations = 101.

- Light vs. Popular vs. Premium brands can not be treated as separate markets

Bottom-level Results: Premium Brands

- The authors impose symmetry

Brand Share Equations: Premium.

	1 Budweiser	2 Molson	3 Labatts	4 Miller	5 Coors
Constant	0.393	0.377	0.230	−0.104	–
	(0.062)	(0.078)	(0.056)	(0.031)	–
Time	0.001	−0.000	0.001	0.000	–
	(0.000)	(0.000)	(0.000)	(0.000)	–
log (Y/P)	−0.004	−0.011	−0.006	0.017	–
	(0.006)	(0.007)	(0.005)	(0.003)	–
log ($P_{Budweiser}$)	−0.936	0.372	0.243	0.150	–
	(0.041)	(0.231)	(0.034)	(0.018)	–
log (P_{Molson})	0.372	−0.804	0.183	0.130	–
	(0.231)	(0.031)	(0.022)	(0.012)	–
log ($P_{Labatts}$)	0.243	0.183	−0.588	0.028	–
	(0.034)	(0.022)	(0.044)	(0.019)	–
log (P_{Miller})	0.150	0.130	0.028	−0.377	–
	(0.018)	(0.012)	(0.019)	(0.017)	–
log (# of Stores)	−0.010	0.005	−0.036	0.022	–
	(0.009)	(0.012)	(0.008)	(0.005)	–
Conditional Own	−3.527	−5.049	−4.277	−4.201	−4.641
Price Elasticity	(0.113)	(0.152)	(0.245)	(0.147)	(0.203)

Bottom-level Results: Economy Brands

It is important to note that these results are conditional on expenditure in a given category, y_{Gnt} from equation (1); the overall brand price elasticities arise from a combination of the estimates from all three levels.

Overall Elasticities.

	Elasticity	Standard Error
Budweiser	−4.196	0.127
Molson	−5.390	0.154
Labatts	−4.592	0.247
Miller	−4.446	0.149
Coors	−4.897	0.205
Old Milwaukee	−5.277	0.118
Genesee	−4.236	0.129
Milwaukee's Best	−6.205	0.170
Busch	−6.051	0.332
Piels	−4.117	0.469
Genesee Light	−3.763	0.072
Coors Light	−4.598	0.115
Old Milwaukee Light	−6.097	0.140
Lite	−5.039	0.141
Molson Light	−5.841	0.148