Econometrics Toolkit for Structural Estimation, with examples from disclosure models

AES Summer School, July 2020

Jeremy Bertomeu, Olin School of Business, Washington University in St Louis

Objectives of the session

- Generalized Method of Moments
- Weighting Matrix
- Standard Errors
- Maximum Likelihood

  $\longrightarrow$ central limit, delta method, influence functions, bootstrap.

Method of Moments (with an example)

- *Think first as a theorist*: state and solve a model where agents make optimal choices.
  - Example: Manager observes value of the firm $\tilde{v} \in [0, 1]$ and can disclose for a cost $c > 0$ or withhold to maximize price. Verrecchia (83) shows that firms disclose if and only if $v > \tau$, where

$$\tau - c = \mathbb{E}(\tilde{v}|\tilde{v} \leq \tau) = \frac{\tau}{2}, \tag{1}$$

    so $\tau = 2c$.
  - Compute one moment of this model that is *empirically* observable: for example, expected observed disclosure is

$$\mathbb{E}(\tilde{v}|\tilde{v} \geq \tau) = \frac{\tau + 1}{2} = c + \frac{1}{2}. \tag{2}$$

  - Match this moments to recover hidden parameter by, if expected disclosure in a sample is $\hat{d} = .6$, then, $\hat{d} = \hat{c} + \frac{1}{2}$, so $\hat{c} = .1$.

Method of Moments (2)

- Under method of moments, same nb. of parameters as number of moments. Formally, we are looking for a $n-$dimensional vector of parameters to estimate θ, and we have theoretical moments, a vector $g(\theta)$ (*to be solved for*) and corresponding empirical moments $\hat{m}$, then:

$$\hat{\theta} \in \text{Argmin}_{\theta}(g(\theta) - \hat{m})'(g(\theta) - \hat{m})$$

- Example (cont.): assume that $\tilde{v} \sim U[-b, b]$, so that $\tau = -b + 2c$.

$$\mathbb{E}(\tilde{v}|\tilde{v} \geq \tau) = \frac{2c - b + b}{2} = c, \quad (3)$$

$$Var(\tilde{v}|\tilde{v} \geq \tau) = \frac{1}{12}(b - \tau)^2 = \frac{(b - c)^2}{3}. \quad (4)$$

So set $\hat{c} = \hat{d}$ and denoting $\hat{d}_2 = 0.1$ as the empirical variance of disclosures, set $\hat{b} = \sqrt{3\hat{d}_2} + \hat{c} = 1.15$.
$\rightarrow$ Estimation is "as difficult" as being able to solve a model analytically or numerically.

Simulated Method of Moments

- Often, the theoretical moments $g(\theta)$ do not have an easy expression, that is, we can't explicitly write $g(\theta) = ...$ as a function of the parameters we want to estimate.
- Example (cont.): take previous model, but assume $\tilde{c} \sim U[0, \overline{c}]$ is random, observed by investors but not the econometrician. We want to estimate $\overline{c}$.
  - Pick b and $\overline{c}$, simulate K times v_i, c_i, compute disclosure threshold $-b + 2c_i$ and simulate disclosure $d_i = 1$ if $v_i > -b + 2c_i$ and $d_i = 0$ otherwise, so create simulated sample with $(d_i, d_i v_i)_{i=1}^{K}$. Compute the moments $g(\theta)$ from that simulation and match as in standard method of moments.
  - In practice: need to program a piece of code that conducts this simulation $g(\theta)$ and another that puts this function into the function to be minimized $(g(\theta) - \hat{m})'(g(\theta) - \hat{m})$, since no longer in closed-form.

Generalized Method of Moments

- Example (cont.): Suppose we return to model with $\tilde{v} \sim U[0,1]$ and c fixed, but try to match variance and mean:

$$\hat{c} \in \text{argmin}\, (\hat{d} - c - 1/2)^2 + (\hat{d}_2 - \underbrace{Var(\tilde{v}|\tilde{v} \geq 2c)}_{=\frac{1}{12}(1-2c)^2})^2$$

  Solution for $\hat{d} = 0.6$ and $\hat{d}_2 = 0.1$ is $\hat{c} = 0$. But should we weight the two moments equally?

- Generalized method moment: weight moments with suitably chosen matrix W,

$$\hat{\theta} \in \text{Argmin}_{\theta} (g(\theta) - \hat{m})' W (g(\theta) - \hat{m}),$$

  where W is the inverse variance-covariance matrix of the moments.

Approaches to recovering optimal weighting matrix

- Calculate $Var(\hat{m})$ in closed-form (cumbersome; sometimes can be done with delta method, see later).
- If we can write the empirical moment as mean of independent observations, i.e., using "influence" based method:

$$\hat{m} = \sum_{i=1}^{n} \frac{h(x_i; \theta)}{n},$$

then

$$Var(\hat{m}) = \frac{Var(h(\tilde{x}, \theta))}{n}.$$

So if $h(.)$ is a vector of moments with components h_1, ..., $h_J(.)$, build a matrix

$$V = \begin{pmatrix} h_1(x_1; \theta) & \dots & h_J(x_1; \theta) \\ \vdots & \dots & \vdots \\ h_1(x_n; \theta) & \dots & h_J(x_n; \theta) \end{pmatrix} \tag{5}$$

and compute the covariance of dataset V to get $\hat{Var}(h(\tilde{x}, \theta))$.

Estimating variances with influence functions

What if $\hat{m} \neq \sum_{i=1}^{n} h(x_i, \theta)/n$ is not additively separable in x_i? Suppose (more generally) that the moment m is defined by:

$$\mathbb{E}(g(\tilde{x}, m)) = 0.$$

Can use influence functions, i.e., same idea as (5):

$$V = \begin{pmatrix} \psi_1(x_1; \theta) & \dots & \psi_J(x_1; \theta) \\ \vdots & \dots & \vdots \\ \psi_1(x_n; \theta) & \dots & \psi_J(x_n; \theta) \end{pmatrix}$$

compute the covariance of dataset V and divide by n to get $Var(\hat{m})$.

How do we recover $\psi_j(.)$?

Calculating influence functions

Intuitively, $\psi_j(.)$ is the marginal effect of an observation x on $\hat{m}$: it can be shown (with a functional Taylor expansion) that:

$$\psi(x) = -\mathbb{E}(\nabla_m g(\tilde{x}, m))^{-1} g(x, m); \tag{6}$$

if you know $g(.)$, this expression can be evaluated.

An example: $x = (y, z)$ moments are mean of z and coefficient $m_2 = \beta$ in a linear regression $y = \alpha + \beta z$. Step 1, find the $g(.) = (g_1(.), g_2(.))$ function:

$$g_1 = m_1 - z \tag{7}$$

$$g_2 = z(y - \alpha - m_2 z) \tag{8}$$

Step 2, differentiate to get $\psi(.)$:

$$\psi(x) = -\begin{pmatrix} 1 & 0 \\ 0 & -\mathbb{E}(\tilde{z}^2) \end{pmatrix}^{-1} \begin{pmatrix} m_1 - z \\ z(y - \alpha - m_2 z) \end{pmatrix} = \begin{pmatrix} z - m_1 \\ \frac{y - \alpha - m_2 z}{\mathbb{E}(\tilde{z}^2)} \end{pmatrix} \tag{9}$$

$\rightarrow$ influence function methods work for any estimator θ defined by $\mathbb{E}(g(\tilde{x}, \theta)) = 0$, not just estimating the covariance matrix of moments.

Issues with optimal weighting matrix: dependence on parameter θ

- If $h(.,\theta)$ depends on θ. In "principle," no asymptotic efficiency loss in using $\hat{\theta}_0$ from a first-step estimate with identity matrix (most common method in applied work)

  $\rightarrow$ of course, not necessarily true in practice, so alternatives are *iterated* (replace first-step estimate by estimate, and repeat until convergence) or *one-step* (write $\hat{W}(\theta)$ as the part of the estimation).

Issues with optimal weighting matrix: Complicated models

What if we can't easily figure out $g(.)$ (*multi-step estimation, non-parametrics*); cumbersome analytical derivatives for $g(.)$ or unreliable numerical differentiation for $\nabla_m g$?

- Philosophy of bootstrap: variance of a moment is *also* its variability for a new empirical sample.

- But we only have one sample??
  $\longrightarrow$ we can rebuild a new sample by resampling (with replacement) a new dataset from the empirical sample. Do it M times, compute the moment each time, stack these estimates and calculate the variance-covariance of this moments.

- If $\hat{m}$ does not depend on θ, then procedure is very fast, safe (no risk of analytical error) and can accommodate any complexity in $\hat{m}$ (e.g., non-parametric first steps, plug-ins, etc.).

Variance-Covariance Matrix: An Example

Let $\tilde{x}_i \sim N(0,1)$ with $i = 1, \ldots, n$ iid observations. Let $\hat{m} = (\sum \frac{x_i}{n}, \sum \frac{x_i^2}{n})$ be a vector of two moments.

1. Analytical method (exact):

$$Var(\hat{m}) = \begin{pmatrix} \frac{E(\tilde{x}^2)}{n} & \frac{E(\tilde{x}^3)}{n} \\ \frac{E(\tilde{x}^3)}{n} & \frac{E(\tilde{x}^4) - E(\tilde{x}^2)^2}{n} \end{pmatrix} = \begin{pmatrix} \frac{1}{n} & 0 \\ 0 & \frac{2}{n} \end{pmatrix}$$

2. Influence method, write:

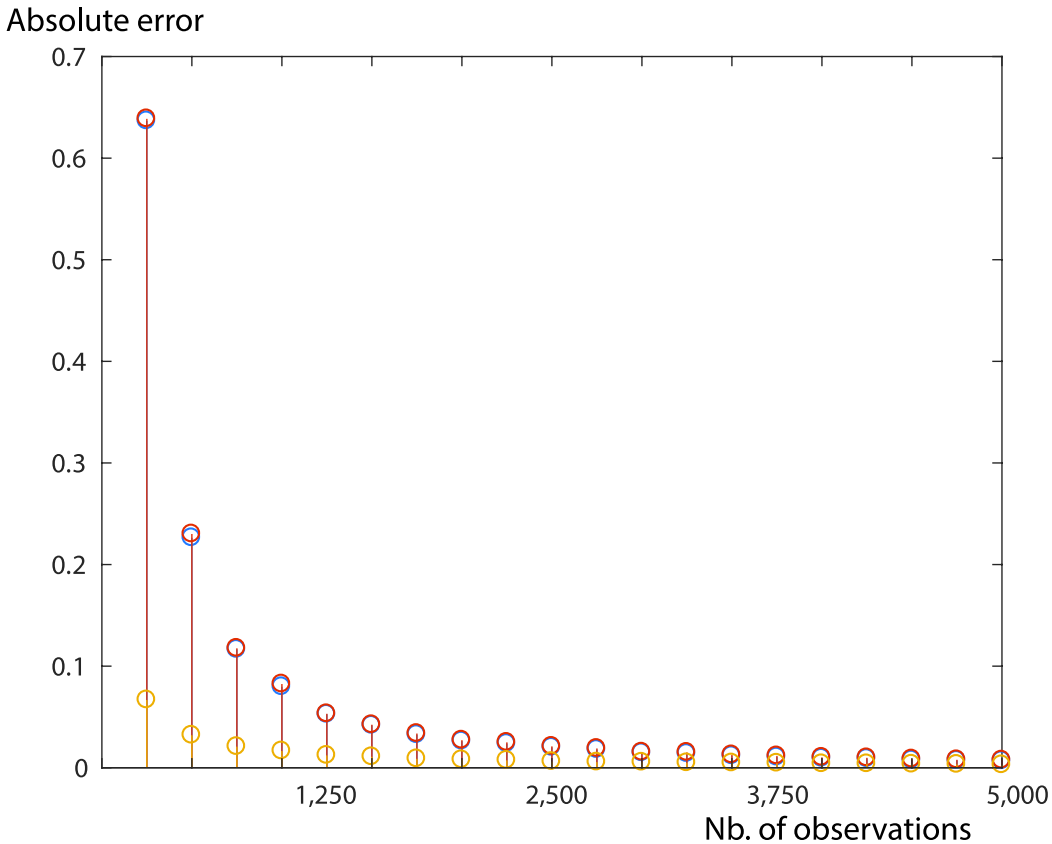
$$V = \begin{pmatrix} x_1 & x_1^2 \\ \vdots & \vdots \\ x_n & x_n^2 \end{pmatrix}$$

   Taking covariance matrix of this and multiply by $1/n$.

3. Bootstrap. Sample for the dataset $(x_i)_{i=1}^n$ to create new bootstrapped samples $(x_i^j)_{i=1}^n$ for $j = 1, \ldots, 100$. Compute the means of $(x_i^j, (x_i^j)^2)_{i=1}^n$, as a vector m_j. Compute the covariance of the matrix $(m_j)_{j=1}^{500}$. Do not multiply by $1/n$!

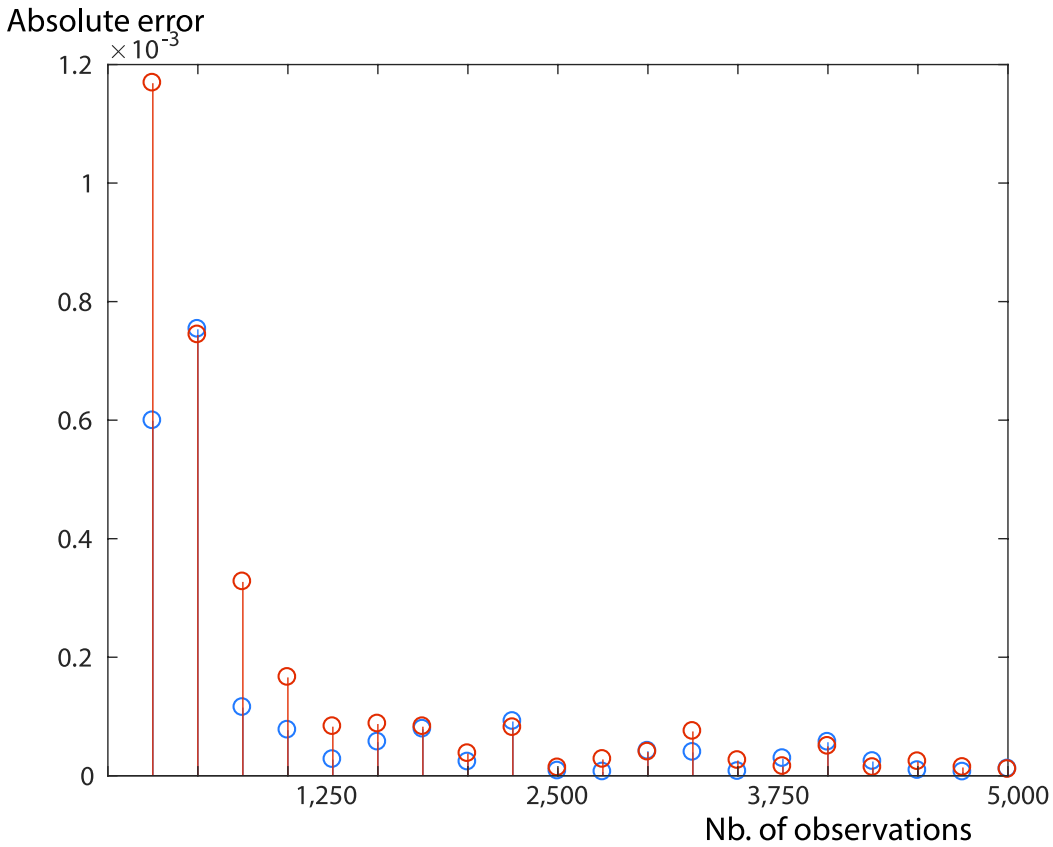
Absolute error: mean and variance of normal

Compute absolute error as $|Var(\hat{m}) - \hat{Var}(\hat{m})|$ where $Var(\hat{m})$ is the analytical method (exact) and $\hat{Var}(\hat{m})$ is computed with three methods: (i) influence (blue), (ii) bootstrap over the sample dataset $(x_i)_{i=1}^n$, (iii) model bootstrap, i.e., resample $\tilde{x}_i$ from $N(0, 1)$ (yellow).



Absolute error: mean and OLS coefficient

Same exercise but with mean and OLS coefficient in examples (7) and (8). Compare influence to bootstrap as before when estimating the error (relative to numerical model bootstrap).



Interesting properties of GMM

- If moments can be satisfied only by one parameter (identification), then the estimator is consistent.
- Estimates are asymptotically normal with, given $G = \mathbb{E}(\nabla g(\theta))$,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, (G'WG)^{-1}) \tag{10}$$

  $\rightarrow$ but can also avoid calculations or numerical differentiation by bootstrapping standard-errors (more on this later).
- Hansen J-test. Suppose we have k parameters to estimates, l moments and sample size n, with $k > l$. Under the hypothesis that the moments are satisfied,

$$J \equiv n(g(\hat{\theta}) - \hat{m})'\hat{W}(g(\hat{\theta}) - \hat{m}) \xrightarrow{d} \chi^2_{k-l} \tag{11}$$

Common questions with GMM - How do we choose moments? Good properties to seek

- Moments that seem to (intuitively) identify the parameters we're looking for.
  $\longrightarrow$ expected disclosure identifies truncation, variance of disclosure identifies of untruncated information.
- Moments we're interested in explaining
  $\longrightarrow$ expected disclosure captures tendency to hide bad news if higher than unconditional information
- Moments that can be precisely estimated
  $\longrightarrow$ Variance is easier to estimate than kurtosis.
- Moments that don't over-rely in places the model is not designed to work well $\longrightarrow$ (with uniform) flat density of disclosures above the threshold.
  $\longrightarrow$ sometimes can be seen ex-post, as a diagnostics on what the model cannot do well: what if using that moment yields nonsensical result.

  (with uniform) flat density of disclosures above the threshold.
- Moments that capture different things and are not too redundant.
  $\longrightarrow$ usually visible when the eigenvalues of the weighting matrix are too large.
- Moments that are more robust to variations on details of the model, e.g., functional forms, additional noise terms, etc.

Bootstrap: A simple approach for SEs

- Generic term for obtaining estimates (of bias or SE) via resampling. Consider an estimator $\hat{\theta}$ given data set $X_n = (x_i)_{i=1}^n$.
  - Resample j (large enough) times to create M resampled data sets X_n^j for $j = 1, \ldots M$.
  - Compute estimator $\hat{\theta}_j$ in each sample.
  - Empirical distribution of $\hat{\theta}_j$ is an estimate of distribution of $\hat{\theta}$.
- Two types of bootstrap:
  - Parametric: if fully-specified model, resample X_n^j from the model given estimate $\hat{\theta}$.
  - Non-parametric: resample X_n^j by uniformly drawing (with repetition) n times in X_n.
    $\rightarrow$ in panels, randomly draw clusters of observations (block bootstrap).
- Examples:
  - Estimate finite-sample bias $\frac{1}{M} \sum \hat{\theta}_j - \hat{\theta}$.
  - Construct asymmetric confidence intervals.
  - SE for set identified models.
  - Multi-step estimation procedure with non-standard SE.
- Notes: (i) usually better finite-sample but not panacea (same small-sample problems), (ii) can be infeasible if estimator does not compute fast (M slower), (iii) bootstrap the SE of a bootstrapped estimator? (M^2 slower)

Maximum likelihood estimation

Why maximum likelihood?

- for certain models, likelihoods can be easier to write (*closed-form*) even when moments can only be simulated.
- avoid headaches over which moments to choose, i.e., implicitly selects moments efficiently.
- Flexible tool: full likelihood, partial likelihood (if only part of the likelihood is easy to use), simulated likelihood.
- purist perspective: fully write the assumed data-generating process, model the noise/disturbance terms if any, and estimate.
- econometrics are usually more straightforward.

$\rightarrow$ downside (?): *MLE won't work if theoretical model is too misspecified, w/o modelling noise formally*. In many dynamic models with state variables, writing likelihood can be very difficult.

Maximum likelihood Definition

Let $f(x|\theta)$ be a family of densities indexed by θ. For any sample $X = (x_i)_{i=1}^n$ drawn from $f(x|\theta_0)$, define the average log likelihood as

$$\mathcal{L}(X|\theta) = \frac{1}{n} \sum \ln f(x_i|\theta). \tag{12}$$

The maximum likelihood estimator (MLE) for θ_0 is such that

$$\hat{\theta}_{MLE} \in argmax_\theta \quad \mathcal{L}(X_n|\theta). \tag{13}$$

Under mild regularity conditions, $plim \quad \hat{\theta}_{MLE} = \theta_0$ and attains the Cramér-Rao lower bound, i.e., no other consistent estimator has lower asymptotic mean squared error.

Maximum likelihood estimation - standard errors

Under certain regularity conditions (see Hayashi Chapter 7),

$$\sqrt{n}(\hat{\theta}_{mle} - \theta_0) \xrightarrow{d} N(0, I^{-1}) \tag{14}$$

where $I = \mathbb{E}(-\frac{\delta^2 \ln f_{\theta_0}(x_i)}{\delta \theta^2})$ is Fisher's information matrix.

Note that $\hat{I}$ can be consistently estimated as the hessian of the average log-likelihood

$$\hat{I} = \frac{\delta^2 \mathcal{L}(X|\theta)}{\delta \theta^2}\Big|_{\theta=\hat{\theta}_{mle}}. \tag{15}$$

Disclosure theory with MLE

Consider example model in (1) where $\tilde{v}_i \sim U[0,1]$ with pdf g_v and cdf G_v, let $d_i \in \{0,1\}$ indicate disclosure and x_i be the disclosure (if any). Recall that the disclosure threshold is $\tau = 2c$ so that the likelihood of this model is:

$$f(d_i, x_i|c) = 1_{d_i=0} \times \underbrace{2c}_{G_v(\tau)} + 1_{d_i=1} \times \underbrace{1_{x_i \in [2c,1]}}_{g_v(x_i)}. \tag{16}$$

Note that $\mathcal{L}(c) = -\infty$ if there exists at least one disclosure x_i below $2c$. To be estimated, this model needs noise! For example, as in Bertomeu, Ma and Marinovic (TAR, forth), can assume that cost is random $c_i \sim H(c;\theta)$ with pdf $h(.)$ and parameter θ to be estimated.

$$\begin{aligned} f(d_i, x_i|\theta) &= 1_{d_i=0} \times \int \min(1, 2c) h(c;\theta) dc \\ &\quad + 1_{d_i=1} \times \int 1_{x_i \in [2c,1]} h(c;\theta) dc. \end{aligned}$$

Two more useful theorems

Theorem (Lindeberg-Lévy central limit theorem)

Let $(X_i)_{i=1}^{\infty}$ be a sequence of i.i.d. random variables with finite mean M and variance Σ. Define $S_n = \sum_{i=1}^{n} X_i/n$. Then,

$$\sqrt{n}(S_n - M) \xrightarrow{d} N(0, \Sigma) \tag{17}$$

That's good but structural estimation need not be in terms of means!! So generalize to any function of means (or functions of known estimator) using Delta Method:

Theorem (Delta Method)

Let S_n be a sequence of random variables such that $\sqrt{n}(S_n - M) \xrightarrow{d} N(0, \Sigma)$ with Σ positive definite. Let f be a function $\nabla f(M) \neq 0$. Then,

$$\sqrt{n}(f(S_n) - f(M)) \xrightarrow{d} N(0, \nabla f(M)' \Sigma \nabla f(M)) \tag{18}$$

Note: for the case $f : \mathbb{R}^J \to \mathbb{R}$, the gradient is $\nabla f(x)' = (f_1'(x_1), \ldots, f_J'(x_J))'$ where x_i is the i^{th} component.

Standard errors with Delta Method: Cheynel and Liu-Watts, RASt

- With probability p, the manager is uninformed about x, drawn from a distribution with p.d.f. $f(.)$, c.d.f. $F(.)$, which we normalize to a mean zero.

- If informed, the manager can disclose for a cost c (i.e., $d(x) = x$) or stay silent (i.e., $d(x) = ND$) implying market price $P(d)$, where $P(x) = x$ and $P(ND) = \mathbb{E}(x|d(x) = ND)$.

- Manager discloses to maximize price, implying disclosure threshold τ with:

$$\tau - c = \mathbb{E}(x|ND). \tag{19}$$

Recovering implied disclosure costs

Lemma

Let q be the probability of disclosure,

$$c = \tau + \frac{q}{1-q}\mathbb{E}(\tilde{x}|\tilde{x} \geq \tau). \tag{20}$$

Proof. Start from disclosure equation:

$$\tau - c = \mathbb{E}(x|ND). \tag{21}$$

Rewrite $\mathbb{E}(x|ND)$ using the law of total expectation:

$$0 = \mathbb{E}(x) = q\mathbb{E}(x|x \geq \tau) + (1-q)\mathbb{E}(x|ND),$$

where q is the probability of disclosure. That is,

$$\mathbb{E}(x|ND) = -\frac{q}{1-q}\mathbb{E}(\tilde{x}|\tilde{x} \geq \tau).$$

and (20) follows by reinjecting this in (21). □

Estimation of Disclosure Cost

- From (3), the following is a *consistent* estimator for any c,

$$\hat{c} = \hat{\tau} + \frac{\hat{q}}{1 - \hat{q}} \hat{m},$$

  where $\hat{\tau}$ is the minimum disclosure, $\hat{q}$ is the sample disclosure frequency and $\hat{m}$ is the average disclosure.

- For simplicity, assume here that $\hat{\tau} = \tau$ is known (can be estimated as $\min x_i$)

Step 1: Use CLT to get Variance matrix for inputs to $\hat{c}$

Proposition

The estimator is consistent and asymptotically normal with

$$\sqrt{N}(\hat{c} - c) \rightarrow_d N(0, \sigma_c^2), \tag{22}$$

where $\sigma_c^2 = \frac{qm^2 + (1-q)qv_x}{(1-q)^3}$ *and* $v_x = Var(x|x \geq k)$.

Proof. In what follows, let us denote x as the random variable corresponding to the manager's private information and $d = 1$ if a forecast is issued $d = 0$ otherwise. Let the sample frequency of disclosure be denoted $\hat{q}$ and the "average" sample forecast (coding zeros for non-forecast periods) be denoted $\hat{w}$, then, from the central limit theorem,

$$\sqrt{N}\left(\begin{pmatrix}\hat{q}\\ \hat{w}\end{pmatrix} - \begin{pmatrix}q\\ qm\end{pmatrix}\right) \rightarrow_d N(\mathbf{0_2}, \underbrace{\begin{pmatrix}Var(d) & cov(d,dx)\\ cov(d,dx) & Var(dx)\end{pmatrix}}_{\mathbf{v_0}}).$$

□

Step 1: Use CLT to get Variance matrix for inputs to $\hat{c}$, (cont.)

Proof. Simplifying this variance-covariance matrix and denoting $m = \mathbb{E}(x|x \geq k)$ and $v_x = Var(x|x \geq k)$,

$$\mathbf{V_0} = \begin{pmatrix} q(1-q) & (1-q)qm \\ (1-q)qm & q(v_{x+}(1-q)m^2) \end{pmatrix} \tag{23}$$

$$\begin{aligned} \text{because} \quad Var(d) &= q(1-q); \\ cov(d, dx) &= \mathbb{E}(d^2 x) - \mathbb{E}(d)\mathbb{E}(dx) \\ &= (1-q)q\mathbb{E}(x|x \geq k) = (1-q)qm; \end{aligned}$$

$$\begin{aligned} \text{and} \quad Var(dx) &= \mathbb{E}(d^2(x)^2) - \mathbb{E}(dx)^2 \\ &= q\mathbb{E}(x^2|x \geq k) - (q\mathbb{E}(x|x \geq k))^2 \\ &= q(Var(x|x \geq k) + \mathbb{E}(x|x \geq k)^2) - (q\mathbb{E}(x|x \geq k))^2 \\ &= q(v_x + (1-p)m^2). \end{aligned}$$

□

Step 2: apply delta method to obtain covariance matrix for $\hat{c}$

Proof. Next, note that $\hat{c} = G(\hat{p}, \hat{w})$ where $G(y, z) = k + \frac{z}{1-y}$. Applying the delta method,

$$\sqrt{N}(\hat{c} - c) \to_d N(0, \underbrace{AV_0A}_{\sigma_c^2})$$

such that $A = \nabla G = (\frac{qm}{(1-q)^2}, \frac{1}{1-q})$. Therefore,

$$\begin{aligned}
\sigma_c^2 &= (\frac{qm}{(1-q)^2}, \frac{1}{1-q}) \begin{pmatrix} q(1-q) & (1-q)qm \\ (1-q)qm & q(v_x + (1-q)m^2) \end{pmatrix} \begin{pmatrix} \frac{qm}{(1-q)^2} \\ \frac{1}{1-q} \end{pmatrix} \\
&= \frac{q(qm + m(1-q))^2 + (1-q)qv_x}{(1-q)^3}.
\end{aligned}$$

□

Concluding Notes: Architecture of a Structural Estimation

1. Load dataset.
2. Write a single code to compute empirical moment; if using simulated moments (SMM), use the **same** code.
3. For SMM, write a program that simulates the same type of data that is available empirically. Keep random seed constant at the start of program.
4. Program to compute an optimal weighting matrix, if more moments than parameters.
5. If boostrapping, either (a) pre-draw M datasets and save them offline (*better*), or (b) write code that randomly samples M datasets when code is executed.
6. Write code to minimize objective method of moments objective function; avoid local search or unbounded search as much as possible.
7. Write code to compute SEs, either by re-estimating the bootstrap samples or using an equation.
8. Write code to simulate model for *any* parameters and use it to compute unobservables from the estimated parameters or unobservables/observables from counterfactuals.
9. Write code to compute other moments (*not used in the estimation*). Use this code on data vs. simulation to report distance.