

EC241a  
Spring 2004

Econometric Theory  
UC Berkeley Department of Economics

## GENERALIZED METHOD OF MOMENTS ESTIMATION I:

### MOTIVATION

In generalized method of moments estimation we are going to look at estimation problems of the following type. We are given a function  $\psi(\cdot; \cdot)$ , such that

$$\mathbb{E}\psi(Z, \theta) = 0,$$

for some unknown value  $\theta_0$ , and this expectation differs from zero for all other values of  $\theta$ . The vector  $\psi(Z, \theta)$  is a known function of the random variable  $Z$  and the parameter  $\theta$ . Let us first look at some examples. The dimension of the vector of moment functions  $\psi$  is equal to  $M$ , and the dimension of the parameter vector  $\theta$  is equal to  $K$ .

#### Example I: Earnings Dynamics

Abowd and Card (1989) study the dynamics of earnings. Let  $y_{it}$  be log earnings in period  $t$  for individual  $i$ . Let  $x_i$  be a vector of individual, time-invariant, characteristics affecting earnings, such as education and age. Abowd and Card specify a model for earnings of the form:

$$y_{it} = x_i' \beta + \omega_i + \varepsilon_{it} + \eta_{it},$$

where

$$\varepsilon_{it} = \alpha \cdot \varepsilon_{it-1} + \nu_{it},$$

and

$$\begin{pmatrix} \omega_i \\ \nu_{it} \\ \eta_{it} \end{pmatrix} \Big| X \sim \mathcal{N} \left( \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\omega^2 & 0 & 0 \\ 0 & \sigma_\nu^2 & 0 \\ 0 & 0 & \sigma_\eta^2 \end{pmatrix} \right).$$

The earnings residual has a permanent component  $\omega_i$ , a transitory component  $\varepsilon_{it}$  following a first order autoregressive process, and measurement error component  $\eta_{it}$ . Given this process,

the covariance matrix of earnings is some function of the variances of the three components of the residual and the autoregressive parameter  $\alpha$ . These restrictions can then be used as moment restrictions in a GMM framework. Consider for example, the (1, 1) element of the covariance matrix.

$$\mathbb{E}[(y_{i1} - x'_i\beta)^2] = \sigma_\omega^2 + \frac{\sigma_\nu^2}{1 - \alpha^2} + \sigma_\eta^2,$$

so that a moment restriction is

$$\psi_1(y, x, \beta, \sigma_\omega^2, \sigma_\eta^2, \sigma_\nu^2, \alpha) = (y_{i1} - x'_i\beta)^2 - \sigma_\omega^2 - \frac{\sigma_\nu^2}{1 - \alpha^2} - \sigma_\eta^2.$$

Similarly we can calculate the expected value of all cross-moments of the form

$$\mathbb{E}[(y_{it} - x'_i\beta) \cdot (y_{it+s} - x_i\beta)] = \sigma_\omega^2 + \frac{\alpha^s \sigma_\nu^2}{1 - \alpha^2} + \sigma_\eta^2,$$

and use that to generate additional moment restrictions.  $\square$

### Example II: Euler Equations

Consider a rational expectations model as in Hansen and Singleton (1982). Agents choose consumption and the level of an asset in every period to maximize life-time discounted utility:

$$\max_{c_0, c_1, \dots, c_\infty, q_0, \dots, q_\infty} \mathbb{E} \left[ \sum_{t=0}^{\infty} \beta^t \cdot U(c_t; \gamma) \right],$$

where  $\beta$  is the discount factor, and  $U(c_t; \gamma)$  is the contemporaneous utility function indexed by parameter  $\gamma$ , subject to the budget constraint in each period

$$c_t + p_t \cdot q_t \leq (p_t + d_t) \cdot q_{t-1} + w_t,$$

where  $p_t$  is the price of the asset in period  $t$ ,  $d_t$  its dividend, and  $w_t$  is labor income in period  $t$ . We are interested in the discount factor and the parameters of the utility function. The first order condition is

$$p_t \cdot \frac{\partial U}{\partial c}(c_t; \gamma) = \beta \cdot \mathbb{E}_t \left[ (p_{t+1} + d_{t+1}) \cdot \frac{\partial U}{\partial c}(c_{t+1}; \gamma) \right].$$

This can be exploited in the form of a moment restriction

$$\mathbb{E}_t \left[ \beta \cdot \frac{\frac{\partial U}{\partial c}(c_{t+1}, \gamma)}{\frac{\partial U}{\partial c}(c_t, \gamma)} \frac{p_{t+1} + d_{t+1}}{p_t} - 1 \right] = 0.$$

Such moment restrictions are known as Euler equations.  $\square$ .

### Example III: Aggregate Information

In the Hellerstein and Imbens paper the interest is in running a wage regression of the form

$$\ln(\text{wage}_i) = \beta_0 + \beta_1 \text{educ}_i + \beta_2 \text{exper}_i + \beta_3 \text{exper}_i^2 + \beta_4 \text{iq}_i + \beta_5 \text{kww}_i + \varepsilon_i.$$

The hope is that by including ability measures such as iq and the test score kww we get estimates of the returns to education (estimates of  $\beta$ ) that are more closely corresponding to the causal effect of an additional year of education. To be able to run this regression you have to find a data set that contains these ability measures. The National Longitudinal Survey is one of those data sets. This is, however a small data set and therefore the estimates are not very precise. Hellerstein and Imbens use census data to get more precision. Specifically, they use census data to get some restrictions. For example one of the restrictions is that the expectation of  $\ln(\text{wage})$  is equal to 2.0647. If all the information we had was from the NLS, we could estimate  $\beta$  by least squares. This estimator can be put in the GMM framework using the moment restriction

$$\mathbb{E}[X \cdot (Y - X'\beta)] = 0,$$

where  $Y = \ln(\text{wage})$ , and  $X = (1, \text{educ}, \text{exper}, \text{exper}^2, \text{kww}, \text{iq})'$ . The second data set can be put in the GMM framework by adding a moment of the form

$$\mathbb{E}[\ln(\text{wage}) - 2.0647] = 0.$$

How is such a piece of information useful? Let us look at a simple example.

Suppose we wish to estimate the mean of a variable  $x$ . Given a random sample from its distribution, and without additional information, we would efficiently estimate its mean as

$$\hat{\mu}_x = \bar{x}.$$

This has a normalized variance of  $\sigma_x^2$ . Now suppose we also have a random sample of  $y$ , of the same size. This does not directly help us to improve on the estimate of  $\mu_x$ . However,

now suppose we know the mean of  $y$ ,  $\mu_y = 0$ . Then we can improve our estimate of  $\mu_x$  to:

$$\hat{\mu}_x = \bar{x} - \frac{\sigma_x}{\sigma_y} \rho_{xy} \sigma_x \sigma_y \bar{y}.$$

Although we do not know these variances, we can plug in estimates to get

$$\hat{\mu}_x = \bar{x} - \frac{S_{xy}}{S_{yy}} \bar{y},$$

where

$$S_{vw} = \frac{1}{N} \sum_{i=1}^N (v_i - \bar{v})(w_i - \bar{w}).$$

(Estimating these variances does not affect the large sample distribution of the estimator).

This has in large samples a normalized variance of  $\sigma_x^2(1 - \rho_{xy}^2) \leq \sigma_x^2$ , and hence it is more efficient than  $\bar{x}$ .

To understand why the extra information is useful, think back to seemingly unrelated regression, SUR:

$$Y_1 = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \varepsilon_1,$$

$$Y_2 = \alpha_0 + \alpha_1 \cdot X_1 + \varepsilon_2.$$

The fact that the regressors are not the same, combined with a non-zero correlation between  $\varepsilon_1$  and  $\varepsilon_2$  makes SUR more efficient than OLS (equation by equation). Here the first equation has only a constant,  $\mu_x$ , and the second had no unknown regression coefficients, yet the structure is the same, and thus the correlation and restriction combine to make OLS less than fully efficient in general.  $\square$

#### Example IV: Maximum Likelihood

Maximum likelihood estimation: we typically define the mle as maximizing the log likelihood function

$$\sum_{i=1}^N \ln f(y_i | x_i, \theta).$$

It can also be viewed as a Generalized Method of Moments estimator with moments equal to the score function  $\mathcal{S}$ :

$$\psi(y, x, \theta) = \mathcal{S}(y, x, \theta) = \frac{\partial \ln f}{\partial \theta}(y|x, \theta).$$

□

### Example V: Instrumental Variables

Suppose we have a linear model

$$Y_i = X_i' \beta + \varepsilon_i,$$

but we are concerned about endogeneity of the covariates (correlation between  $\varepsilon_i$  and  $X_i$ ). In that case we cannot estimate  $\beta$  by ordinary least squares methods. If we have a set of instruments  $Z_i$  that are uncorrelated with the error terms  $\varepsilon_i$ , and which are correlated with the endogenous regressors  $X_i$ , we can exploit the moment conditions

$$\mathbb{E}[Z(Y - X'\beta)] = 0.$$

The number of instruments (the dimension of  $Z_i$ ) is not necessarily the same as the number of covariates (the dimension of  $X_i$ ). □

### Example VI: Panel Data Models with Fixed Effects

Suppose we have a two period panel, with

$$Y_{it} = X_{it}' \beta + \alpha_i + \varepsilon_{it},$$

for  $i = 1, \dots, N$  and  $t = 1, 2$ . If  $\alpha_i$  is correlated with  $X_{it}$ , we cannot estimate  $\beta$  by regressing  $Y_{it}$  on  $X_{it}$ . If  $\varepsilon_{it}$  is independent of  $X_{is}$  for all  $i, t$ , and  $s$ , differencing leads to moment conditions of the form

$$\mathbb{E}[h(X_{i1}, X_{i2}) \cdot (Y_{i2} - Y_{i1} - (X_{i2} - X_{i1})' \beta)],$$

for arbitrary functions  $h(X_{i1}, X_{i2})$ . □

From this list of examples it is clear that most if not all estimators commonly used in econometrics fit into this framework. It is therefore extremely useful to be able to deal with the general case; we can then go back to the special cases we knew already how to deal with and see how things work out there.

First let us consider estimation. In general Hansen (1982) considers estimating  $\theta$  by minimizing

$$Q_C(\theta) = \frac{1}{N} \left[ \sum_{i=1}^N \psi(z_i, \theta) \right]' \cdot C \cdot \left[ \sum_{i=1}^N \psi(z_i, \theta) \right],$$

for some positive semi-definite and symmetric matrix  $C$ . This is one of the places where the distinction between the just-identified, where  $M = K$ , and over-identified, where  $M > K$ , case is important. In the just-identified case the choice of  $C$  in principle does not matter: the solution satisfies the estimating equations:

$$\sum_{i=1}^N \psi(z_i, \hat{\theta}) = 0,$$

and therefore

$$Q_C(\hat{\theta}) = 0,$$

irrespective of the choice of  $C$ , subject to positive definiteness. Consider the ols example (Example 3). In that case solving the estimating leads to the standard ols solution

$$\hat{\beta} = \left( \sum_{i=1}^N x_i x_i' \right)^{-1} \left( \sum_{i=1}^N x_i y_i \right).$$

In the mle case (Example 4) it leads to setting the average score function equal to zero, which if the second derivative of the likelihood function is negative, corresponds the maximum likelihood estimator.

Next, consider the over-identified case. Here the choice of  $C$  is going to matter. In the remainder of these notes we will go through a simple example to give some intuition for the optimal choice. Although this optimal choice depends on things we do not actually observe,

it turns out we can then estimate the optimal weight matrix without any loss of precision (up to first order). Suppose someone repeatedly draws objects from a population and obtains two independent measurements of their height. Call the first measurement of the  $i$ th object  $x_i$  and the second measurement  $y_i$ . The parameter of interest is the population average height. We assume that both measurements are unbiased, but they may have different variances and may even be correlated if the height of the objects differ. The moment functions are

$$\psi(x, y, \theta) = \begin{pmatrix} x - \theta \\ y - \theta \end{pmatrix}.$$

Now suppose that the  $y$  measurements are really bad (i.e., imprecise) relative to the  $x$  measurements. In that case it seems obvious that one would like to only use the  $x$  measurements and the efficient solution should be close to  $\bar{x}$ . This corresponds to a choice for  $C$  close to

$$C = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}.$$

(This matrix is obviously not strictly positive definite, but it could be made so by making the second diagonal element close to zero but positive.) A similar argument suggests that if the  $x$  measurement is really imprecise the estimate should be close to  $\bar{y}$  and the optimal weight matrix should be close to

$$C = \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}.$$

Now let us look at the general case where

$$\mathbb{E} \begin{pmatrix} x - \theta \\ y - \theta \end{pmatrix} \begin{pmatrix} x - \theta \\ y - \theta \end{pmatrix}' = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}.$$

We are interested in choosing the weight matrix  $C$  that minimizes the asymptotic variance of the estimator. Minimizing  $Q_C(\theta)$  leads to the first order condition

$$(C_{11} + C_{21}) \cdot \sum_{i=1}^N (x_i - \theta) + (C_{12} + C_{22}) \cdot \sum_{i=1}^N (y_i - \theta) = 0.$$

We can therefore write the estimator for a given choice of  $C$  as

$$\hat{\theta}_\lambda = \lambda \cdot \bar{x} + (1 - \lambda) \cdot \bar{y},$$

where  $\lambda = (C_{11} + C_{12}) / (C_{11} + C_{21} + C_{12} + C_{22})$ . The variance of  $\hat{\theta}_\lambda$  is

$$\lambda^2 \sigma_{xx} + (1 - \lambda)^2 \sigma_{yy} + 2\lambda(1 - \lambda)\sigma_{xy}.$$

Let us minimize this over  $\lambda$ . The first order condition is

$$2\lambda\sigma_{xx} - 2(1 - \lambda)\sigma_{yy} - 2\lambda(1 - \lambda)\sigma_{xy} = 0,$$

leading to the solution

$$\lambda = \frac{\sigma_{yy} - \sigma_{xy}}{\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy}}.$$

To translate this into a solution for  $C$  remember that with  $C$  symmetric  $C_{12} = C_{21}$ . Also clearly we can normalize  $C$  because multiplying it by a constant does not change the solution for  $\theta$ . A solution for  $C$  based on a particular normalization is the inverse of the variance/covariance matrix of the moments:

$$C = \Delta^{-1} = \begin{pmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{xy} & \sigma_{yy} \end{pmatrix}^{-1} = \frac{1}{\sigma_{xx}\sigma_{yy} - \sigma_{xy}^2} \cdot \begin{pmatrix} \sigma_{yy} & -\sigma_{xy} \\ -\sigma_{xy} & \sigma_{xx} \end{pmatrix},$$

and

$$\lambda = \frac{C_{11} + C_{12}}{C_{11} + C_{12} + C_{21} + C_{22}} = \frac{\sigma_{yy} - \sigma_{xy}}{\sigma_{yy} + \sigma_{xx} - 2\sigma_{xy}}.$$

This turns out to be true in general (see Hansen (1982)), and we therefore in general look at estimators for  $\theta$  that minimize

$$Q_{\mathbb{E}[\Delta]^{-1}}(\theta).$$

Of course we typically do not know what the variance covariance matrix  $\Delta$  is. Hansen's solution is the following 2-step procedure:

1. Estimate  $\theta$  by minimizing  $Q_C(\theta)$  for an arbitrary positive definite symmetric matrix  $C$  to get a consistent but not necessarily efficient estimate  $\hat{\theta}_{\text{init}}$ .

2. Use this initial estimate to get an estimate of the optimal weight matrix:

$$\hat{\Delta}^{-1} = \left[ \sum_{i=1}^N \psi(z_i, \hat{\theta}_{\text{init}}) \psi(z_i, \hat{\theta}_{\text{init}})' / N \right]^{-1},$$

and use this estimate in the second round by minimizing  $Q_{\hat{\Delta}^{-1}}(\theta)$ .

We shall see that the large sample properties of this procedure are not affected by the estimation of the weight matrix, that is, by using  $\hat{\Delta}^{-1}$  instead of  $\Delta^{-1}$ .

#### ADDITIONAL REFERENCES

HANSEN, L.-P., AND K. SINGLETON, (1982), "Generalized Instrumental Variable Estimation of Nonlinear Rational Expectations Models", *Econometrica*, Vol 50, 1269–1286.

HELLERSTEIN, J., AND G. IMBENS, (1999), "Imposing Moment Restrictions by Weighting", *Review of Economics and Statistics*.