2
Model Specification and Estimation Strategies

A DAPM MAY: (1) provide a complete characterization of the joint distribution of all of the variables being studied; or (2) imply restrictions on some moments of these variables, but not reveal the form of their joint distribution. A third possibility is that there is not a well-developed theory for the joint distribution of the variables being studied. Which of these cases obtains for the particular DAPM being studied determines the feasible estimation strategies; that is, the feasible choices of $\mathcal{D}$ in the definition of an *estimation strategy*. This chapter introduces the maximum likelihood (ML), generalized method of moments (GMM), and linear least-squares projection (LLP) estimators and begins our development of the interplay between model formulation and the choice of an estimation strategy discussed in Chapter 1.

2.1. Full Information about Distributions

Suppose that a DAPM yields a complete characterization of the joint distribution of a sample of size T on a vector of variables y_t, $\vec{y}_T \equiv \{y_1, \ldots, y_T\}$. Let $L_T(\beta) = L(\vec{y}_T; \beta)$ denote the family of joint density functions of $\vec{y}_T$ implied by the DAPM and indexed by the K-dimensional parameter vector β. Suppose further that the admissible parameter space associated with this DAPM is $\Theta \subseteq \mathbb{R}^K$ and that there is a unique $\beta_0 \in \Theta$ that describes the true probability model generating the asset price data.

In this case, we can take $L_T(\beta)$ to be our sample criterion function—called the *likelihood function* of the data—and obtain the *maximum likelihood* (ML) estimator b_T^{ML} by maximizing $L_T(\beta)$. In ML estimation, we start with the joint density function of $\vec{y}_T$, evaluate the random variable $\vec{y}_T$ at the realization comprising the observed historical sample, and then maximize the value of this density over the choice of $\beta \in \Theta$. This amounts to maximizing,

over all admissible β, the "likelihood" that the realized sample was drawn from the density $L_T(\beta)$. ML estimation, when feasible, is the most econometrically efficient estimator within a large class of consistent estimators (Chapter 3).

In practice, it turns out that studying L_T is less convenient than working with a closely related objective function based on the conditional density function of y_t. Many of the DAPMs that we examine in later chapters, for which ML estimation is feasible, lead directly to knowledge of the density function of y_t conditioned on $\vec{y}_{t-1}$, $f_t(y_t|\vec{y}_{t-1};\beta)$ and imply that

$$f_t(y_t|\vec{y}_{t-1};\beta) = f\big(y_t|\vec{y}_{t-1}^{\,J};\beta\big), \tag{2.1}$$

where $\vec{y}_t^{\,J} \equiv (y_t, y_{t-1}, \ldots, y_{t-J+1})$, a J-history of y_t. The right-hand side of (2.1) is *not* indexed by t, implying that the conditional density function does not change with time.[1] In such cases, the likelihood function L_T becomes

$$L_T(\beta) = \prod_{t=J+1}^{T} f\big(y_t|\vec{y}_{t-1}^{\,J};\beta\big) \times f_m(\vec{y}_J;\beta), \tag{2.2}$$

where $f_m(\vec{y}_J)$ is the marginal, joint density function of $\vec{y}_J$. Taking logarithms gives the *log-likelihood* function $l_T \equiv T^{-1}\log L_T$,

$$l_T(\beta) = \frac{1}{T}\sum_{t=J+1}^{T} \log f\big(y_t|\vec{y}_{t-1}^{\,J};\beta\big) + \frac{1}{T}\log f_m(\vec{y}_J;\beta). \tag{2.3}$$

Since the logarithm is a monotonic transformation, maximizing l_T gives the same ML estimator b_T^{ML} as maximizing L_T.

The first-order conditions for the sample criterion function (2.3) are

$$\frac{\partial l_T}{\partial\beta}\big(b_T^{\mathrm{ML}}\big) = \frac{1}{T}\sum_{t=J+1}^{T} \frac{\partial \log f}{\partial\beta}\big(y_t|\vec{y}_{t-1}^{\,J};\, b_T^{\mathrm{ML}}\big) + \frac{1}{T}\frac{\partial \log f_m}{\partial\beta}\big(\vec{y}_J;\, b_T^{\mathrm{ML}}\big) = 0, \tag{2.4}$$

where it is presumed that, among all estimators satisfying (2.4), b_T^{ML} is the one that maximizes l_T.[2] Choosing $z_t' = (y_t', \vec{y}_{t-1}^{\,J\prime})$ and

[1] A sufficient condition for this to be true is that the time series $\{y_t\}$ is a strictly stationary process. Stationarity does not preclude time-varying conditional densities, but rather just that the functional form of these densities does not change over time.

[2] It turns out that b_T^{ML} need not be unique for fixed T, even though β_0 is the unique minimizer of the population objective function Q_0. However, this technical complication need not concern us in this introductory discussion.

$$\mathcal{D}(z_t;\beta) \equiv \frac{\partial \log f}{\partial \beta}\left(y_t \middle| \vec{y}_{t-1}^{\,J};\beta\right) \tag{2.5}$$

as the function defining the moment conditions to be used in estimation, it is seen that (2.4) gives first-order conditions of the form (1.12), except for the last term in (2.4).[3] For the purposes of large-sample arguments developed more formally in Chapter 3, we can safely ignore the last term in (2.3) since this term converges to zero as $T \to \infty$.[4] When the last term is omitted from (2.3), this objective function is referred to as the *approximate* log-likelihood function, whereas (2.3) is the *exact* log-likelihood function. Typically, there is no ambiguity as to which likelihood is being discussed and we refer simply to the log-likelihood function l.

Focusing on the approximate log-likelihood function, fixing $\beta \in \Theta$, and taking the limit as $T \to \infty$ gives, under the assumption that sample moments converge to their population counterparts, the associated population criterion function

$$Q_0(\beta) = E\left[\log f\left(y_t \middle| \vec{y}_{t-1}^{\,J};\beta\right)\right]. \tag{2.6}$$

To see that the β_0 generating the observed data is a maximizer of (2.6), and hence that this choice of Q_0 underlies a sensible estimation strategy, we observe that since the conditional density integrates to 1,

$$\begin{aligned} 0 &= \frac{\partial}{\partial \beta}\int_{-\infty}^{\infty} f\left(y_t \middle| \vec{y}_{t-1}^{\,J};\beta_0\right) dy_t \\ &= \int_{-\infty}^{\infty} \frac{\partial \log f}{\partial \beta}\left(y_t \middle| \vec{y}_{t-1}^{\,J};\beta_0\right) f\left(y_t \middle| \vec{y}_{t-1}^{\,J};\beta_0\right) dy_t \\ &= E\left[\frac{\partial \log f}{\partial \beta}(y_t \middle| \vec{y}_{t-1}^{\,J};\beta_0) \;\middle|\; \vec{y}_{t-1}^{\,J}\right], \end{aligned} \tag{2.7}$$

which, by the law of iterated expectations, implies that

$$\frac{\partial Q_0}{\partial \beta}(\beta_0) = E\left[\frac{\partial \log f}{\partial \beta}(y_t \middle| \vec{y}_{t-1}^{\,J};\beta_0)\right] = E\left[\mathcal{D}(z_t;\beta_0)\right] = 0. \tag{2.8}$$

Thus, for ML estimation, (2.8) is the set of constraints on the joint distribution of $\vec{y}_T$ used in estimation, the ML version of (1.10). Critical to (2.8)

[3] The fact that the sum in (2.4) begins at $J+1$ is inconsequential, because we are focusing on the properties of b_T^{ML} (or θ_T) for large T, and J is fixed a priori by the asset pricing theory.

[4] There are circumstances where the small-sample properties of b_T^{ML} may be substantially affected by inclusion or omission of the term $\log f_m(\vec{y}_J;\beta)$ from the likelihood function. Some of these are explored in later chapters.

being satisfied by β_0 is the assumption that the conditional density f implied by the DAPM is in fact the density from which the data are drawn.

An important special case of this estimation problem is where $\{y_t\}$ is an independently and identically distributed (i.i.d.) process. In this case, if $f_m(y_t; \beta)$ denotes the density function of the vector y_t evaluated at β, then the log-likelihood function takes the simple form

$$l_T(\beta) \equiv T^{-1} \log L_T(\beta) = \frac{1}{T} \sum_{t=1}^{T} \log f_m(y_t; \beta). \tag{2.9}$$

This is an immediate implication of the independence assumption, since the joint density function of $\vec{y}_T$ factors into the product of the marginal densities of the y_t. The ML estimator of β_0 is obtained by maximizing (2.9) over $\beta \in \Theta$. The corresponding population criterion function is $Q_0(\beta) = E[\log f_m(y_t; \beta)]$.

Though the simplicity of (2.9) is convenient, most dynamic asset pricing theories imply that at least some of the observed variables y are not independently distributed over time. Dependence might arise, for example, because of mean reversion in an asset return or persistence in the volatility of one or more variables (see the next example). Such time variation in conditional moments is accommodated in the formulation (2.1) of the conditional density of y_t, but not by (2.9).

Example 2.1. *Cox, Ingersoll, and Ross [Cox et al., 1985b] (CIR) developed a theory of the term structure of interest rates in which the instantaneous short-term rate of interest, r, follows the mean reverting diffusion*

$$dr = \kappa(\bar{r} - r)\, dt + \sigma \sqrt{r}\, dB. \tag{2.10}$$

An implication of (2.10) is that the conditional density of r_{t+1} given r_t is

$$f(r_{t+1}|r_t; \beta_0) = ce^{-u_t - v_{t+1}} \left(\frac{v_{t+1}}{u_t}\right)^{\frac{q}{2}} I_q\left(2(u_t v_{t+1})^{\frac{1}{2}}\right), \tag{2.11}$$

where

$$c = \frac{2\kappa}{\sigma^2(1 - e^{-\kappa})}, \tag{2.12}$$

$$u_t = \frac{2\kappa}{\sigma^2(1 - e^{-\kappa})} e^{-\kappa} r_t, \tag{2.13}$$

$$v_{t+1} = \frac{2\kappa}{\sigma^2(1 - e^{-\kappa})} r_{t+1}, \tag{2.14}$$

$q = 2\kappa\bar{r}/\sigma^2 - 1$, and I_q is the modified Bessel function of the first kind of order q. This is the density function of a noncentral χ^2 with $2q + 2$ degrees of freedom and noncentrality parameter $2u_t$. For this example, ML estimation would proceed by substituting (1.11) into (2.4) and solving for b_T^{ML}. The short-rate process (2.10) is the continuous time version of an interest-rate process that is mean reverting to a long-run mean of $\bar{r}$ and that has a conditional volatility of $\sigma\sqrt{r}$. This process is Markovian and, therefore, $\vec{y}_t^{\,J} = y_t$, which explains the single lag in the conditioning information in (1.11).

Though desirable for its efficiency, ML may not be, and indeed typically is not, a feasible estimation strategy for DAPMs, as often they do not provide us with complete knowledge of the relevant conditional distributions. Moreover, in some cases, even when these distributions are known, the computational burdens may be so great that one may want to choose an estimation strategy that uses only a portion of the available information. This is a consideration in the preceding example given the presence of the modified Bessel function in the conditional density of r. Later in this chapter we consider the case where only limited information about the conditional distribution is known or, for computational or other reasons, is used in estimation.

2.2. No Information about the Distribution

At the opposite end of the knowledge spectrum about the distribution of $\vec{y}_T$ is the case where we do not have a well-developed DAPM to describe the relationships among the variables of interest. In such circumstances, we may be interested in learning something about the joint distribution of the vector of variables z_t (which is presumed to include some asset prices or returns). For instance, we are often in a situation of wondering whether certain variables are correlated with each other or if one variable can predict another. Without knowledge of the joint distribution of the variables of interest, researchers typically proceed by *projecting* one variable onto another to see if they are related. The properties of the estimators in such projections are examined under this case of no information.[5] Additionally, there are occasions when we reject a theory and a replacement theory that explains the rejection has yet to be developed. On such occasions, many have resorted to projections of one variable onto others with the hope of learning more about the source of the initial rejection. Following is an example of this second situation.

[5] Projections, and in particular linear projections, are a simple and often informative first approach to examining statistical dependencies among variables. More complex, nonlinear relations can be explored with nonparametric statistical methods. The applications of nonparametric methods to asset pricing problems are explored in subsequent chapters.

Example 2.2. *Several scholars writing in the 1970s argued that, if foreign currency markets are informationally efficient, then the forward price for delivery of foreign exchange one period hence (F_t^1) should equal the market's best forecast of the spot exchange rate next period (S_{t+1}):*

$$F_t^1 = E[S_{t+1}|I_t], \tag{2.15}$$

where I_t denotes the market's information at date t. This theory of exchange rate determination was often evaluated by projecting $S_{t+1} - F_t^1$ onto a vector x_t and testing whether the coefficients on x_t are zero (e.g., Hansen and Hodrick, 1980). The evidence suggested that these coefficients are not zero, which was interpreted as evidence of a time-varying market risk premium $\lambda_t \equiv E[S_{t+1}|I_t] - F_t^1$ (see, e.g., Grauer et al., 1976, and Stockman, 1978). Theory has provided limited guidance as to which variables determine the risk premiums or the functional forms of premiums. Therefore, researchers have projected the spread $S_{t+1} - F_t^1$ onto a variety of variables known at date t and thought to potentially explain variation in the risk premium. The objective of the latter studies was to test for dependence of λ_t on the explanatory variables, say x_t.

To be more precise about what is meant by a *projection*, let L^2 denote the set of (scalar) random variables that have finite second moments:

$$L^2 = \{\text{random variables } x \text{ such that } Ex^2 < \infty\}. \tag{2.16}$$

We define an inner product on L^2 by

$$\langle x \,|\, y \rangle \equiv E(xy), \quad x, y \in L^2, \tag{2.17}$$

and a norm by

$$\| x \| = [\langle x \,|\, x \rangle]^{\frac{1}{2}} = \sqrt{E(x^2)}. \tag{2.18}$$

We say that two random variables x and y in L^2 are *orthogonal* to each other if $E(xy) = 0$. Note that being orthogonal is not equivalent to being uncorrelated as the means of the random variables may be nonzero.

Let A be the closed linear subspace of L^2 generated by all linear combinations of the K random variables $\{x_1, x_2, \ldots, x_K\}$. Suppose that we want to project the random variable $y \in L^2$ onto A in order to obtain its best linear predictor. Letting $\delta' \equiv (\delta_1, \ldots, \delta_K)$, the best linear predictor is that element of A that minimizes the distance between y and the linear space A:

$$\min_{z \in A} \| y - z \| \quad \Leftrightarrow \quad \min_{\delta \in \mathbb{R}^K} \| y - \delta_1 x_1 - \ldots - \delta_K x_K \|. \tag{2.19}$$

The *orthogonal projection theorem*[6] tells us that the *unique* solution to (2.19) is given by the $\delta_0 \in \mathbb{R}^K$ satisfying

$$E\left[(y - x'\delta_0)x\right] = 0, \quad x' = (x_1, \ldots, x_K); \tag{2.20}$$

that is, the forecast error $u \equiv (y - x'\delta_0)$ is orthogonal to all linear combinations of x. The solution to the first-order condition (2.20) is

$$\delta_0 = E[xx']^{-1}E[xy]. \tag{2.21}$$

In terms of our notation for criterion functions, the population criterion function associated with least-squares projection is

$$Q_0(\delta) = E\left[(y_t - x_t'\delta)^2\right], \tag{2.22}$$

and this choice is equivalent to choosing $z_t' = (y_t, x_t')$ and the function $\mathcal{D}$ as

$$\mathcal{D}(z_t; \delta) = (y_t - x_t'\delta)x_t. \tag{2.23}$$

The interpretation of this choice is a bit different than in most estimation problems, because our presumption is that one is proceeding with estimation in the absence of a DAPM from which restrictions on the distribution of (y_t, x_t) can be deduced. In the case of a least-squares projection, we view the moment equation

$$E\left[\mathcal{D}(y_t, x_t; \delta_0)\right] = E\left[(y_t - x_t'\delta_0)x_t\right] = 0 \tag{2.24}$$

as the moment restriction that *defines* δ_0.

The sample least-squares objective function is

$$Q_T(\delta) = \frac{1}{T}\sum_{t=1}^{T}(y_t - x_t'\delta)^2, \tag{2.25}$$

with minimizer

$$\delta_T = \left[\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right]^{-1} \frac{1}{T}\sum_{t=1}^{T} x_t y_t. \tag{2.26}$$

[6] The orthogonal projection theorem says that if L is an inner product space, M is a closed linear subspace of L, and y is an element of L, then $z^* \in M$ is the unique solution to

$$\min_{z \in M} \| y - z \|$$

if and only if $y - z^*$ is orthogonal to all elements of M. See, e.g., Luenberger (1969).

The estimator δ_T is also obtained directly by replacing the population moments in (2.21) by their sample counterparts.

In the context of the pricing model for foreign currency prices, researchers have projected $(S_{t+1} - F_t^1)$ onto a vector of explanatory variables x_t. The variable being predicted in such analyses, $(S_{t+1} - F_t^1)$, is not the risk premium, $\lambda_t = E[(S_{t+1} - F_t^1)|I_t]$. Nevertheless, the resulting predictor in the population, $x_t'\delta_0$, is the same regardless of whether λ_t or $(S_{t+1} - F_t^1)$ is the variable being forecast. To see this, we digress briefly to discuss the difference between *best linear* and *best* prediction.

The predictor $x_t'\delta_0$ is the best linear predictor, which is defined by the condition that the projection error $u_t = y_t - x_t'\delta_0$ is orthogonal to all linear combinations of x_t. Predicting y_t using linear combinations of x_t is only one of many possible approaches to prediction. In particular, we could also consider prediction based on both linear and nonlinear functions of the elements of x_t. Pursuing this idea, let V denote the closed linear subspace of L^2 generated by all random variables $g(x_t)$ with finite second moments:

$$V = \{g(x_t) : g : \mathbb{R}^K \to \mathbb{R}, \text{ and } g(x_t) \in L^2\}. \tag{2.27}$$

Consider the new minimization problem $\min_{z \in V} \| y_t - z_t \|$. By the orthogonal projection theorem, the unique solution z_t^* to this problem has the property that $(y_t - z_t^*)$ is orthogonal to all $z_t \in V$. One representation of z^* is the conditional expectation $E[y_t|x_t]$. This follows immediately from the properties of conditional expectations: the error $\epsilon_t = y_t - E[y_t|x_t]$ satisfies

$$E[\epsilon_t g(x_t)] = E\left[(y_t - E[y_t|x_t])g(x_t)\right] = 0, \tag{2.28}$$

for all $g(x_t) \in V$. Clearly, $A \subseteq V$ so the best predictor is at least as good as the best linear predictor. The precise sense in which best prediction is better is that, whereas ϵ_t is orthogonal to *all* functions of the conditioning information x_t, u_t is orthogonal to only linear combinations of x_t.

There are circumstances where best and best linear predictors coincide. This is true whenever the conditional expectation $E[y_t|x_t]$ is linear in x_t. One well-known case where this holds is when (y_t, x_t') is distributed as a multivariate normal random vector. However, normality is not necessary for best and best linear predictors to coincide. For instance, consider again Example 2.1. The conditional mean $E[r_{t+\Delta}|r_t]$ for positive time interval Δ is given by (Cox et al., 1985b)

$$\mu_{rt}(\Delta) \equiv E[r_{t+\Delta}|r_t] = r_t e^{-\Delta\kappa} + \bar{r}(1 - e^{-\Delta\kappa}), \tag{2.29}$$

which is linear in r_t, yet neither the joint distribution of $(r_t, r_{t-\Delta})$ nor the distribution of r_t conditioned on $r_{t-\Delta}$ is normal. (The latter is noncentral chi-square.)

With these observations in mind, we can now complete our argument that the properties of risk premiums can be studied by linearly projecting $(S_{t+1} - F_t^1)$ onto x_t. Letting $\text{Proj}[\cdot|x_t]$ denote linear least-squares projection onto x_t, we get

$$\begin{aligned}\text{Proj}[\lambda_t|x_t] &= \text{Proj}\left[\left(S_{t+1} - F_t^1\right) - \epsilon_{t+1}\middle|x_t\right] \\ &= \text{Proj}\left[\left(S_{t+1} - F_t^1\right)\middle|x_t\right], \end{aligned} \tag{2.30}$$

where $\epsilon_{t+1} \equiv (S_{t+1} - F_t^1) - \lambda_t$. The first equality follows from the definition of the risk premium as $E[S_{t+1} - F_t^1|I_t]$ and the second follows from the fact that ϵ_{t+1} is orthogonal to all functions of x_t including linear functions.

2.3. Limited Information: GMM Estimators

In between the cases of full information and no information about the joint distribution of $\vec{y}_T$ are all of the intermediate cases of *limited information*. Suppose that estimation of a parameter vector θ_0 in the admissible parameter space $\Phi \subset \mathbb{R}^K$ is to be based on a sample $\vec{z}_T$, where z_t is a subvector of the complete set of variables y_t appearing in a DAPM.[7] The restrictions on the distribution of $\vec{z}_T$ to be used in estimating θ_0 are summarized as a set of restrictions on the moments of functions of z_t. These moment restrictions may be either *conditional* or *unconditional*.

2.3.1. Unconditional Moment Restrictions

Consider first the case where a DAPM implies that the unconditional moment restriction

$$E[h(z_t; \theta_0)] = 0 \tag{2.31}$$

is satisfied uniquely by θ_0, where h is an M-dimensional vector with $M \geq K$. The function h may define standard central or noncentral moments of asset returns, the orthogonality of forecast errors to variables in agents' information sets, and so on. Illustrations based on Example 2.1 are presented later in this section.

To develop an estimator of θ_0 based on (2.31), consider first the case of $K = M$; the number of moment restrictions equals the number of parameters to be estimated. The function $H_0 : \Phi \to \mathbb{R}^M$ defined by $H_0(\theta) =$

[7] There is no requirement that the dimension of Φ be as large as the dimension of the parameter space Θ considered in full information estimation; often Φ is a lower-dimensional subspace of Θ, just as z_t may be a subvector of y_t. However, for notational convenience, we always set the dimension of the parameter vector of interest to K, whether it is θ_0 or β_0.

$E[h(z_t;\theta)]$ satisfies $H_0(\theta_0) = 0$. Therefore, a natural estimation strategy for θ_0 is to replace H_0 by its sample counterpart,

$$H_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} h(z_t;\theta), \tag{2.32}$$

and choose the estimator θ_T to set (2.32) to zero. If H_T converges to its population counterpart as T gets large, $H_T(\theta) \to H_0(\theta)$, for all $\theta \in \Phi$, then under regularity conditions we should expect that $\theta_T \to \theta_0$. The estimator θ_T is an example of what Hansen (1982b) refers to as a generalized method-of-moments, or GMM, estimator of θ_0.

Next suppose that $M > K$. Then there is not in general a unique way of solving for the K unknowns using the M equations $H_T(\theta) = 0$, and our strategy for choosing θ_T must be modified. We proceed to form K linear combinations of the M moment equations to end up with K equations in the K unknown parameters. That is, letting $\bar{\mathcal{A}}$ denote the set of $K \times M$ (constant) matrices of rank K, we select an $A \in \bar{\mathcal{A}}$ and set

$$\mathcal{D}^A(z_t;\theta) = Ah(z_t;\theta), \tag{2.33}$$

with this choice of $\mathcal{D}^A$ determining the estimation strategy. Different choices of $A \in \bar{\mathcal{A}}$ index (lead to) different estimation strategies. To arrive at a sample counterpart to (2.33), we select a possibly sample-dependent matrix A_T with the property that $A_T \to A$ (almost surely) as sample size gets large. Then the $K \times 1$ vector θ_T^A (the superscript A indicating that the estimator is A-dependent) is chosen to satisfy the K equations $\sum_t \mathcal{D}_T(z_t, \theta_T^A) = 0$, where $\mathcal{D}_T(z_t, \theta_T^A) = A_T h(z_t; \theta_T^A)$. Note that we are now allowing $\mathcal{D}_T$ to be sample dependent directly, and not only through its dependence on θ_T^A. This will frequently be the case in subsequent applications.

The construction of GMM estimators using this choice of $\mathcal{D}_T$ can be related to the approach to estimation involving a criterion function as follows: Let $\{a_T : T \geq 1\}$ be a sequence of $s \times M$ matrices of rank s, $K \leq s \leq M$, and consider the function

$$Q_T(\theta) = |a_T H_T(\theta)|, \tag{2.34}$$

where $|\cdot|$ denotes the Euclidean norm. Then

$$\operatorname*{argmin}_\theta |a_T H_T(\theta)| = \operatorname*{argmin}_\theta |a_T H_T(\theta)|^2 = \operatorname*{argmin}_\theta H_T(\theta)' a_T' a_T H_T(\theta), \tag{2.35}$$

and we can think of our criterion function Q_T as being the quadratic form

$$Q_T(\theta) = H_T'(\theta) W_T H_T(\theta), \tag{2.36}$$

where $W_T \equiv a_T' a_T$ is often referred to as the *distance matrix*. This is the GMM criterion function studied by Hansen (1982b). The first-order conditions for this minimization problem are

$$\frac{\partial H_T}{\partial \theta}(\theta_T)' W_T H_T(\theta_T) = 0. \tag{2.37}$$

By setting

$$A_T = [\partial H_T(\theta_T)'/\partial \theta] W_T, \tag{2.38}$$

we obtain the $\mathcal{D}_T(z_t; \theta)$ associated with Hansen's GMM estimator.

The population counterpart to Q_T in (2.36) is

$$Q_0(\theta) = E[h(z_t; \theta)]' W_0 E[h(z_t; \theta)]. \tag{2.39}$$

The corresponding population $\mathcal{D}_0(z_t, \theta)$ is given by

$$\mathcal{D}_0(z_t, \theta) = E\left[\frac{\partial h}{\partial \theta}(z_t; \theta_0)'\right] W_0 h(z_t; \theta) \equiv A_0 h(z_t; \theta), \tag{2.40}$$

where W_0 is the (almost sure) limit of W_T as T gets large. Here $\mathcal{D}_0$ is not sample dependent, possibly in contrast to $\mathcal{D}_T$.

Whereas the first-order conditions to (2.36) give an estimator in the class $\bar{\mathcal{A}}$ [with A defined by (2.40)], not all GMM estimators in $\bar{\mathcal{A}}$ are the first-order conditions from minimizing an objective function of the form (2.36). Nevertheless, it turns out that the *optimal* GMM estimators in $\bar{\mathcal{A}}$, in the sense of being asymptotically most efficient (see Chapter 3), can be represented as the solution to (2.36) for appropriate choice of W_T. Therefore, the large-sample properties of GMM estimators are henceforth discussed relative to the sequence of objective functions $\{Q_T(\cdot) : T \geq 1\}$ in (2.36).

2.3.2. Conditional Moment Restrictions

In some cases, a DAPM implies the stronger, conditional moment restrictions

$$E[h(z_{t+n}; \theta_0)|I_t] = 0, \qquad \text{for given } n \geq 1, \tag{2.41}$$

where the possibility of $n > 1$ is introduced to allow the conditional moment restrictions to apply to asset prices or other variables more than one period in the future. Again, the dimension of h is M, and the information set I_t may be generated by variables other than the history of z_t.

To construct an estimator of θ_0 based on (2.41), we proceed as in the case of unconditional moment restrictions and choose K sample moment

equations in the K unknowns θ. However, because $h(z_{t+n}; \theta_0)$ is orthogonal to any random variable in the information set I_t, we have much more flexibility in choosing these moment equations than in the preceding case. Specifically, we introduce a class of $K \times M$ full-rank "instrument" matrices $\mathcal{A}_t$ with each $A_t \in \mathcal{A}_t$ having elements in I_t. For any $A_t \in \mathcal{A}_t$, (2.41) implies that

$$E[A_t h(z_{t+n}; \theta_0)] = 0 \tag{2.42}$$

at $\theta = \theta_0$. Therefore, we can define a family of GMM estimators indexed by $A \in \mathcal{A}$, θ_T^A, as the solutions to the corresponding sample moment equations,

$$\frac{1}{T} \sum_t A_t h\big(z_{t+n}; \theta_T^A\big) = 0. \tag{2.43}$$

If the sample mean of $A_t h(z_{t+n}; \theta)$ in (2.43) converges to its population counterpart in (2.42), for all $\theta \in \Phi$, and A_t and h are chosen so that θ_0 is the unique element of Φ satisfying (2.42), then we might reasonably expect θ_T^A to converge to θ_0 as T gets large. The large-sample distribution of θ_T^A depends, in general, on the choice of A_t.[8]

The GMM estimator, as just defined, is not the extreme value of a specific criterion function. Rather, (2.42) defines θ_0 as the solution to K moment equations in K unknowns, and θ_T solves the sample counterpart of these equations. In this case, $\mathcal{D}_0$ is chosen directly as

$$\mathcal{D}_0(z_{t+n}, A_t; \theta) = \mathcal{D}_T(z_{t+n}, A_t; \theta) = A_t h(z_{t+n}; \theta). \tag{2.44}$$

Once we have chosen an A_t in $\mathcal{A}_t$, we can view a GMM estimator constructed from (2.41) as, trivially, a special case of an estimator based on unconditional moment restrictions. Expression (2.42) is taken to be the basic K moment equations that we start with. However, the important distinguishing feature of the class of estimators $\mathcal{A}_t$, compared to the class $\bar{\mathcal{A}}$, is that the former class offers much more flexibility in choosing the weights on h. We will see in Chapter 3 that the most efficient estimator in the class $\mathcal{A}$ is often more efficient than its counterpart in $\bar{\mathcal{A}}$. That is, (2.41) allows one to exploit more information about the distribution of z_t than (2.31) in the estimation of θ_0.

[8] As is discussed more extensively in the context of subsequent applications, this GMM estimation strategy is a generalization of the instrumental variables estimators proposed for classical simultaneous equations models by Amemiya (1974) and Jorgenson and Laffont (1974), among others.

2.3.3. Linear Projection as a GMM Estimator

Perhaps the simplest example of a GMM estimator based on the moment restriction (2.31) is linear least-squares projection. Suppose that we project y_t onto x_t. Then the best linear predictor is defined by the moment equation (2.20). Thus, if we define

$$h(y_t, x_t; \delta) = (y_t - x_t'\delta)\, x_t, \tag{2.45}$$

then by construction δ_0 satisfies $E[h(y_t, x_t; \delta_0)] = 0$.

One might be tempted to view linear projection as special case of a GMM estimator in $\mathcal{A}_t$ by choosing $n = 0$,

$$A_t = x_t \quad \text{and} \quad h(y_t, x_t; \delta) = (y_t - x_t'\delta). \tag{2.46}$$

However, importantly, we are not free to select among other choices of $A_t \in \mathcal{A}_t$ in constructing a GMM estimator of the linear predictor $x_t'\delta_0$. Therefore, least-squares projection is appropriately viewed as a GMM estimator in $\bar{\mathcal{A}}$.

Circumstances change if a DAPM implies the stronger moment restriction

$$E\big[(y_t - x_t'\delta_0)\big|x_t\big] = 0. \tag{2.47}$$

Now we are no longer in an environment of complete ignorance about the distribution of (y_t, x_t), as it is being assumed that $x_t'\delta_0$ is the best, not just the best linear, predictor of y_t. In this case, we are free to choose

$$A_t = g(x_t) \quad \text{and} \quad h(y_t, x_t; \delta) = (y_t - x_t'\delta), \tag{2.48}$$

for any $g : \mathbb{R}^K \to \mathbb{R}^K$. Thus, the assumption that the best predictor is linear puts us in the case of conditional moment restrictions and opens up the possibility of selecting estimators in $\mathcal{A}$ defined by the functions g.

2.3.4. Quasi-Maximum Likelihood Estimation

Another important example of a limited information estimator that is a special case of a GMM estimator is the *quasi-maximum likelihood* (QML) estimator. Suppose that $n = 1$ and that I_t is generated by the J-history $\vec{y}_t^{\,J}$ of a vector of observed variables y_t.[9] Further, suppose that the functional

[9] We employ the usual, informal notation of letting I_t or $\vec{y}_t^{\,J}$ denote the σ-algebra (information set) used to construct conditional moments and distributions.

forms of the population mean and variance of y_{t+1}, conditioned on I_t, are known and let θ denote the vector of parameters governing these first two conditional moments. Then ML estimation of θ_0 based on the classical normal conditional likelihood function gives an estimator that converges to θ_0 and is normally distributed in large samples (see, e.g., Bollerslev and Wooldridge, 1992).

Referring back to the introductory remarks in Chapter 1, we see that the function $\mathcal{D}$ $(= \mathcal{D}_0 = \mathcal{D}_T)$ determining the moments used in estimation in this case is

$$\mathcal{D}(z_t; \theta) = \frac{\partial \log f_N}{\partial \theta}\left(y_t \middle| \vec{y}_{t-1}^{\,J}; \theta\right), \tag{2.49}$$

where $z_t' = (y_t', \vec{y}_{t-1}^{\,J\prime})$ and f_N is the normal density function conditioned on $\vec{y}_{t-1}^{\,J}$. Thus, for QML to be an admissible estimation strategy for this DAPM it must be the case that θ_0 satisfies

$$E\left[\frac{\partial \log f_N}{\partial \theta}\left(y_t \middle| \vec{y}_{t-1}^{\,J}; \theta_0\right)\right] = 0. \tag{2.50}$$

The reason that θ_0 does in fact satisfy (2.50) is that the first two conditional moments of y_t are correctly specified and the normal distribution is fully characterized by its first two moments. This intuition is formalized in Chapter 3. The moment equation (2.50) defines a GMM estimator.

2.3.5. Illustrations Based on Interest Rate Models

Consider again the one-factor interest rate model presented in Example 2.1. Equation (2.29) implies that we can choose

$$h\left(\vec{z}_{t+1}^{\,1}; \theta_0\right) = \left[r_{t+1} - \bar{r}(1 - e^{-\kappa}) - e^{-\kappa} r_t\right], \tag{2.51}$$

where $\vec{z}_{t+1}^{\,2} = (r_{t+1}, r_t)'$. Furthermore, for any 2×1 vector function $g(r_t)$: $\mathbb{R} \to \mathbb{R}^2$, we can set $A_t = g(r_t)$ and

$$E\left[(r_{t+1} - \bar{r}(1 - e^{-\kappa}) - e^{-\kappa} r_t) g(r_t)\right] = 0. \tag{2.52}$$

Therefore, a GMM estimator $\theta_T^{A'} = (\bar{r}_T, \kappa_T)$ of $\theta_0' = (\bar{r}, \kappa)$ can be constructed from the sample moment equations

$$\frac{1}{T}\sum_t \left[r_{t+1} - \bar{r}_T(1 - e^{-\kappa_T}) - e^{-\kappa_T} r_t\right] g(r_t) = 0. \tag{2.53}$$

Each choice of $g(r_t) \in \mathcal{A}_t$ gives rise to a different GMM estimator that in general has a different large-sample distribution. Linear projection of r_t onto r_{t-1} is obtained as the special case with $g(r_{t-1})' = (1, r_{t-1})$, $M = K = 2$, and $\theta' = (\kappa, \bar{r})$.

Turning to the implementation of QML estimation in this example, the mean of $r_{t+\Delta}$ conditioned on r_t is given by (2.29) and the conditional variance is given by (Cox et al., 1985b)

$$\sigma_{rt}^2(\Delta) \equiv Var[r_{t+\Delta}|r_t] = r_t \frac{\sigma^2}{\kappa}(e^{-\Delta\kappa} - e^{-2\Delta\kappa}) + \bar{r}\frac{\sigma^2}{2\kappa}(1 - e^{-\Delta\kappa})^2. \quad (2.54)$$

If we set $\Delta = 1$, it follows that discretely sampled returns $(r_t, r_{t-1}, \ldots)$ follow the model

$$r_{t+1} = \bar{r}(1 - e^{-\kappa}) + e^{-\kappa} r_t + \sqrt{\sigma_{rt}^2}\epsilon_{t+1}, \quad (2.55)$$

where the error term ϵ_{t+1} in (2.55) has (conditional) mean zero and variance one. For this model, $\theta_0 = (\bar{r}, \kappa, \sigma^2)' = \beta_0$ (the parameter vector that describes the entire distribution of r_t), though this is often not true in other applications of QML.

The conditional distribution of r_t is a noncentral χ^2. However, suppose we ignore this fact and proceed to construct a likelihood function based on our knowledge of (2.29) and (2.54), assuming that the return r_t is distributed as a normal conditional on r_{t-1}. Then the log-likelihood function is (l^q to indicate that this is QML)

$$l_T^q(\theta) \equiv \frac{1}{T}\sum_{t=2}^{T}\left(-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\sigma_{rt-1}^2\right) - \frac{1}{2}\frac{(r_t - \mu_{rt-1})^2}{\sigma_{rt-1}^2}\right). \quad (2.56)$$

Computing first-order conditions gives

$$\frac{\partial l_T^q}{\partial \theta_j}\left(\theta_T^q\right) = \frac{1}{T}\sum_{t=2}^{T} -\frac{1}{2\hat{\sigma}_{rt-1}^2}\frac{\partial \hat{\sigma}_{rt-1}^2}{\partial \theta_j} + \frac{1}{2}\frac{(r_t - \hat{\mu}_{rt-1})^2}{\hat{\sigma}_{rt-1}^4}\frac{\partial \hat{\sigma}_{rt-1}^2}{\partial \theta_j}$$

$$+ \frac{(r_t - \hat{\mu}_{rt-1})}{\hat{\sigma}_{rt-1}^2}\frac{\partial \hat{\mu}_{rt-1}}{\partial \theta_j} = 0, \quad j = 1, 2, 3, \quad (2.57)$$

where θ_T^q denotes the QML estimator and $\hat{\mu}_{rt-1}$ and $\hat{\sigma}_{rt-1}^2$ are μ_{rt-1} and σ_{rt-1}^2 evaluated at θ_T^q. As suggested in the preceding section, this estimation strategy is admissible because the first two conditional moments are correctly specified.

Though one might want to pursue GMM or QML estimation for this interest rate example because of their computational simplicity, this is not

the best illustration of a limited information problem because the true likelihood function is known. However, a slight modification of the interest rate process places us in an environment where GMM is a natural estimation strategy.

Example 2.3. *Suppose we extend the one-factor model introduced in Example 2.1 to the following two-factor model:*

$$\begin{aligned} dr &= \kappa(\bar{r} - r)\,dt + \sigma_r\sqrt{v}\,dB_r, \\ dv &= \nu(\bar{v} - v)\,dt + \sigma_v\sqrt{v}\,dB_v. \end{aligned} \tag{2.58}$$

In this two-factor model of the short rate, v plays the role of a stochastic volatility for r. Similar models have been studied by Anderson and Lund (1997a) and Dai and Singleton (2000). The volatility shock in this model is unobserved, so estimation and inference must be based on the sample $\vec{r}_T$ and r_t is no longer a Markov process conditioned on its own past history.

An implication of the assumptions that r mean reverts to the long-run value of $\bar{r}$ and that the conditional mean of r does not depend on v is that (2.29) is still satisfied in this two-factor model. However, the variance of r_t conditioned on r_{t-1} is not known in closed form, nor is the form of the density of r_t conditioned on $\vec{r}_{t-1}^{\,J}$. Thus, neither ML nor QML estimation strategies are easily pursued.[10] Faced with this limited information, one convenient strategy for estimating $\theta_0' \equiv (\bar{r}, \kappa)$ is to use the moment equations (2.52) implied by (2.29).

This GMM estimator of θ_0 ignores entirely the known structure of the volatility process and, indeed, σ_r^2 is not an element of θ_0. Thus, not only are we unable to recover any information about the parameters of the volatility equation using (2.52), but knowledge of the functional form of the volatility equation is ignored. It turns out that substantially more information about $f(r_t|r_{t-1};\theta_0)$ can be used in estimation, but to accomplish this we have to extend the GMM estimation strategy to allow for unobserved state variables. This extension is explored in depth in Chapter 6.

2.3.6. GMM Estimation of Pricing Kernels

As a final illustration, suppose that the pricing kernel in a DAPM is a function of a state vector x_t and parameter vector θ_0. In preference-based DAPMs, the pricing kernel can be interpreted as an agent's intertemporal

[10] Asymptotically efficient estimation strategies based on approximations to the true conditional density function of r have been developed for this model. These are described in Chapter 7.

Table 2.1. Summary of Population and Sample Objective Functions for Various Estimators

	Maximum likelihood	GMM	Least-squares projection
Population objective function	$\max_{\beta\in\Theta} E\left[\log f\left(y_t \left\vert \vec{y}_{t-1}^{\,J}; \beta\right.\right)\right]$	$\min_{\theta\in\Theta} E[h(z_t;\theta)]' W_0 E[h(z_t;\theta)]$	$\min_{\delta\in\mathbb{R}^K} E\left[\left(y_t - x_t'\delta\right)^2\right]$
Sample objective function	$\max_{\beta\in\Theta} \frac{1}{T}\sum_{t=J+1}^{T} \log f\left(y_t \left\vert \vec{y}_{t-1}^{\,J}; \beta\right.\right)$	$\min_{\theta\in\Theta} H_T(\theta)' W_T H_T(\theta)$ $H_T(\theta) = \frac{1}{T}\sum_{t=1}^{T} h(z_t;\theta)$	$\min_{\delta\in\mathbb{R}^K} \frac{1}{T}\sum_{t=1}^{T}\left(y_t - x_t'\delta\right)^2$
Population F.O.C.	$E\left[\frac{\partial \log}{\partial\beta} f\left(y_t \left\vert \vec{y}_{t-1}^{\,J}; \beta_0\right.\right)\right] = 0$	$A_0 E[h(z_t;\theta_0)] = 0$	$E\left[\left(y_t - x_t'\delta_0\right) x_t\right] = 0$
Sample F.O.C.	$\frac{1}{T}\sum_{t=J+1}^{T} \frac{\partial \log}{\partial\beta} f\left(y_t \left\vert \vec{y}_{t-1}^{\,J}; b_T^{ML}\right.\right) = 0$	$A_T \frac{1}{T}\sum_{t=1}^{T} h(z_t;\theta_T) = 0$	$\frac{1}{T}\sum_{t=1}^{T}\left(y_t - x_t'\delta_T\right)x_t = 0$

marginal rate of substitution of consumption, in which case x_t might involve consumptions of goods and θ_0 is the vector of parameters describing the agent's preferences. Alternatively, q^* might simply be parameterized directly as a function of financial variables. In Chapter 1 it was noted that

$$E\left[\left(q^*_{t+n}(x_{t+n};\theta_0)r_{t+n}-1\right)\middle|I_t\right]=0, \tag{2.59}$$

for investment horizon n and the appropriate information set I_t. If r_{t+n} is chosen to be a vector of returns on M securities, $M \geq K$, then (2.59) represents M conditional moment restrictions that can be used to construct a GMM estimator of θ (Hansen and Singleton, 1982).

Typically, there are more than K securities at one's disposal for empirical work, in which case one may wish to select $M > K$. A $K \times M$ matrix $A_t \in \mathcal{A}_t$ can then be used to construct K unconditional moment equations to be used in estimation:

$$E\left[A_t\left(q^*_{t+n}(x_{t+n};\theta_0)r_{t+n}-1\right)\right]=0. \tag{2.60}$$

Any $A_t \in I_t$ is an admissible choice for constructing a GMM estimator (subject to minimal regularity conditions).

2.4. Summary of Estimators

The estimators introduced in this chapter are summarized in Table 2.1, along with their respective first-order conditions. The large-sample properties of ML, GMM, and LLP estimators are explored in Chapter 3.