Chapter 7

Estimation and Testing

7.1 The Dynamic Regression Model

Chapter 6 demonstrated the style of argument required to establish the asymptotic properties of regression, by working through a simple case. Now we consider the generalization of these results to practical cases. Necessarily, this will require 'higher-level' assumptions. Properties that before could be proved from primitive conditions on the DGP (like stability conditions on the autoregressive roots) must now be assumed directly. In fact, all the results can be proved at the same level as before by assuming, for example, that the equation of interest is embedded in a stationary Gaussian VAR, as in Chapter 4. Apart from the complexity of this exercise, it would not be the most helpful approach because the results actually apply much more widely. We exhibit a sufficient collection of conditions, and these can be put to practical use by showing that they hold in a given application; or more truthfully, devoutly hoping that they do. As the reader knows, economists rarely hesitate to assume things that they would like to be true, and we shall follow this honourable tradition.

7.1.1 The Setup

The model to be studied corresponds to sampling model C, in the categorization of §1.3. To cope with dynamic elements of the specification, and the fact that variables are typically related to their own lags in the sequence of observations, it is necessary to introduce conditioning assumptions less stringent that we were able to use in sampling model B. Let $\mathcal{I}_t$ represent a 'set of conditioning variables'. The mathematically correct way to express this representation is to say that $\mathcal{I}_t$ is the smallest σ-field of events containing the σ-fields generated by the conditioning variables, but this is rather a mouthful. We commit a mild abuse of notation by using the inclusion symbol $\in$ to indicate that a random vector is measurable with respect to the indicated σ-field, and so write things like $\boldsymbol{x}_t \in \mathcal{I}_t$. Although there is an important distinction between the set of random variables and the set of events with respect to which the variables are measurable, for the present purpose

it is convenient to use one as shorthand for the other. See §B.10 for the relevant background.

The models studied take the basic form[1]

$$y_t = \boldsymbol{x}_t'\boldsymbol{\beta} + u_t \tag{7.1.1}$$

where $\boldsymbol{x}_t$ $(k \times 1) \in \mathcal{I}_t$. The fundamental assumption is

Assumption 7.1.1 $E(u_t|\mathcal{I}_t) = 0$ a.s. □

A subclass of models enjoying certain extra desirable properties satisfy a further condition,

Assumption 7.1.2 $E(u_t^2|\mathcal{I}_t) = \sigma^2$ a.s. □

The parameter σ^2 is defined, in either case, to equal $E(u_t^2)$. Initially we analyse the models that satisfy both assumptions, and subsequently note the consequences of dropping Assumption 7.1.2.

The most important feature of the setup is the specification of the set $\mathcal{I}_t$ of valid conditioning variables. This is the complement of the set of variables specifically excluded, by the model, from the explanation or prediction of y_t. The latter typically include any dated $t + j$ for $j > 0$, and any regarded as being determined jointly with y_t. A very large collection remains, generally including all the variables in the following categories:

- deterministic variables, such as the intercept, seasonal dummies, etc.:
- lagged variables, dated $t - j$ for $j > 0$;
- current dated variables that are weakly exogenous for $(\boldsymbol{\beta}, \sigma^2)$.

Moreover, any Borel-measurable function of variables in $\mathcal{I}_t$ is also in $\mathcal{I}_t$. Thus, $u_{t-j} \in \mathcal{I}_t$ since $u_{t-j} = y_{t-j} - \boldsymbol{\beta}'\boldsymbol{x}_{t-j}$. One implication of Assumption 7.1.1 is that the disturbances must be serially uncorrelated.

To take a specific example, suppose that $(y_t, \boldsymbol{z}_t)$ is a vector of variables generated by a dynamic data generation process, represented by the density with factorization

$$D_t(y_t, \boldsymbol{z}_t|\mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}; \boldsymbol{\psi}) = D_t(y_t|\boldsymbol{z}_t, \mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}; \boldsymbol{\psi}_1) D_t(\boldsymbol{z}_t|\mathcal{Y}_{t-1}, \mathcal{Z}_{t-1}; \boldsymbol{\psi}_2) \tag{7.1.2}$$

and that $\boldsymbol{x}_t'\boldsymbol{\beta}$ is the mean of y_t under the first conditional factor where $\boldsymbol{x}_t$ is composed of elements of $\boldsymbol{z}_{t-j}$ for $j \geq 0$ and y_{t-j} for $j > 0$, and $\boldsymbol{\beta}$ and σ^2 are elements of $\boldsymbol{\psi}_1$. In this case $\mathcal{I}_t = \sigma(\boldsymbol{z}_t) \vee \mathcal{Z}_{t-1} \vee \mathcal{Y}_{t-1}$ and the decomposition corresponds to (4.5.3), specialized here by taking y_t to be a scalar. If the conditional factor in (7.1.2) is Gaussian then $\boldsymbol{\psi}_1 = (\boldsymbol{\beta}, \sigma^2)$, but note that the Gaussianity assumption is not needed for the results.

[1] The symbol $\boldsymbol{x}_t$ is being used here to denote the vector of regressors, to match the conventional usage in Chapters 1-3. We avoid the usage $\boldsymbol{x}_t = (y_t, \boldsymbol{z}_t')'$, as employed in Chapter 4.

Additional regularity conditions are needed to validate the asymptotic analysis, of which the most important is

$$\operatorname*{plim}_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t' = \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}E(\boldsymbol{x}_t\boldsymbol{x}_t') = \boldsymbol{M}_{XX} < \infty \text{ (positive definite).} \quad (7.1.3)$$

This is of course identical to (3.5.2), but the underlying conditions would be quite different in the two cases. All that is needed with independent data is to verify that the moment conditions for the law of large numbers is obeyed. Here, we should need either a generalization of Theorem 6.3.2, or an invocation of mixing or NED properties. If the data are stationary then $\boldsymbol{M}_{XX} = E(\boldsymbol{x}_t\boldsymbol{x}_t')$, but this isn't necessary.

7.1.2 Consistency

The least squares estimator, reproducing equations (3.5.4) here for the sake of emphasis, is

$$\hat{\boldsymbol{\beta}} = \left(\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t y_t = \boldsymbol{\beta} + \left(\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\sum_{t=1}^{n}\boldsymbol{x}_t u_t. \quad (7.1.4)$$

As before, the analysis of consistency begins with the vector $\boldsymbol{x}_t u_t$ $(k \times 1)$. Since $\boldsymbol{x}_t \in \mathcal{I}_t$, the properties of conditional expectations imply

$$E(\boldsymbol{x}_t u_t|\mathcal{I}_t) = \boldsymbol{x}_t E(u_t|\mathcal{I}_t) = \boldsymbol{0} \text{ a.s.} \quad (7.1.5)$$

Define $\mathcal{I}_t^* = \sigma(y_t) \vee \mathcal{I}_t$, or in words, the σ-field with respect to which all the variables contained in the model are measurable. Since $y_{t-1} \in \mathcal{I}_t$, note that $\mathcal{I}_{t-1}^* \subseteq \mathcal{I}_t$. Hence, (7.1.5) further implies, using the LIE, that

$$E(\boldsymbol{x}_t u_t|\mathcal{I}_{t-1}^*) = E[E(\boldsymbol{x}_t u_t|\mathcal{I}_t)|\mathcal{I}_{t-1}^*] = \boldsymbol{0} \text{ a.s.} \quad (7.1.6)$$

Since each element of the vector $\boldsymbol{x}_t u_t$ is $\mathcal{I}_t^*$-measurable, it accordingly forms a martingale difference sequence with respect to this information set. In fact a stronger property obtains, which is that the sequence $\{\boldsymbol{x}_t u_t, \mathcal{I}_t^*\}$ is a *vector martingale difference* (v.m.d.).

One way to define a v.m.d. is as a vector sequence, say $\{\boldsymbol{y}_t, \mathcal{F}_t\}$, such that for any fixed, finite vector $\boldsymbol{\lambda}$, $\{\boldsymbol{\lambda}'\boldsymbol{y}_t, \mathcal{F}_t\}$ is an m.d. sequence. This is not the same as a vector of m.d. sequences. Consider the vector $\boldsymbol{y}_t = (x_t, x_{t-1})'$ where x_t is a m.d. sequence. In general $\boldsymbol{\lambda}'\boldsymbol{y}_t = \lambda_1 x_t + \lambda_2 x_{t-1}$ is a serially correlated sequence, and so cannot be a m.d.

Assumptions 7.1.1 and 7.1.2 also make $\{\operatorname{Vec}\boldsymbol{x}_t\boldsymbol{x}_t'(u_t^2 - \sigma^2), \mathcal{I}_t^*\}$ a v.m.d. The properties generalize naturally from the univariate case, and in particular, the following results hold.

$$E(\boldsymbol{x}_t u_t) = \boldsymbol{0} \quad (7.1.7)$$

$$\operatorname{Var}(\boldsymbol{x}_t u_t) = E(u_t^2\boldsymbol{x}_t\boldsymbol{x}_t') = E[E(u_t^2|\mathcal{I}_t)\boldsymbol{x}_t\boldsymbol{x}_t'] = \sigma^2 E(\boldsymbol{x}_t\boldsymbol{x}_t') \quad (7.1.8)$$

$$\operatorname{Cov}(\boldsymbol{x}_t u_t, \boldsymbol{x}_s u_s) = E(u_t u_s \boldsymbol{x}_t\boldsymbol{x}_s') = E[E(u_t|\mathcal{I}_t)u_s\boldsymbol{x}_t\boldsymbol{x}_s'] = \boldsymbol{0} \quad t > s. \quad (7.1.9)$$

(7.1.8) and (7.1.3) together imply

$$\frac{1}{n}\sum_{t=1}^{n}\text{Var}(\boldsymbol{x}_t u_t) = \sigma^2\frac{1}{n}\sum_{t=1}^{n}E(\boldsymbol{x}_t\boldsymbol{x}_t') \to \sigma^2\boldsymbol{M}_{XX} < \infty \text{ as } n \to \infty. \tag{7.1.10}$$

Together with conditions (7.1.7) and (7.1.9), (7.1.10) is sufficient for the Chebyshev WLLN to hold for each element of the vector, yielding

$$\text{plim}\,\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_t u_t = \boldsymbol{0}. \tag{7.1.11}$$

Applying the Slutsky Theorem and (7.1.3) now gives $\text{plim}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$, by the argument paralleling (3.5.7).

7.1.3 Asymptotic Normality

For asymptotic normality, either strict stationarity or the assumption

$$E|\boldsymbol{\lambda}'\boldsymbol{x}_t u_t|^{2+\delta} \leq B < \infty \quad \delta > 0,\ \forall\ t,\ \forall \text{ fixed } \boldsymbol{\lambda} \text{ with } \boldsymbol{\lambda}'\boldsymbol{\lambda} = 1 \tag{7.1.12}$$

is all that is needed. Since $\{\text{Vec}\,\boldsymbol{x}_t\boldsymbol{x}_t'(u_t^2 - \sigma^2), \mathcal{I}_t^*\}$ is a v.m.d., this condition is sufficient by Theorem 6.2.2 for

$$\frac{1}{n}\sum_{t=1}^{n}(\boldsymbol{\lambda}'\boldsymbol{x}_t)^2(u_t^2 - \sigma^2) \xrightarrow{\text{pr}} 0. \tag{7.1.13}$$

Hence, in view of (7.1.3) and (7.1.10) and arguments similar to §3.5.3, the conditions of Theorem 6.2.3 are also satisfied, and

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{\lambda}'\boldsymbol{x}_t u_t \xrightarrow{\text{D}} \text{N}(0,\ \sigma^2\boldsymbol{\lambda}'\boldsymbol{M}_{XX}\boldsymbol{\lambda}) \tag{7.1.14}$$

for each such $\boldsymbol{\lambda}$. A point to note is that (7.1.12) is not enough for (7.1.3), and (7.1.13) imposes no restrictions on the memory of the $\boldsymbol{x}_t$ process, as (7.1.3) does. Given the m.d. property, the moment conditions are enough here.

The argument leading to (3.5.11) now applies identically, leading to the conclusion

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \left(\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{x}_t\boldsymbol{x}_t'\right)^{-1}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{x}_t u_t \xrightarrow{\text{D}} \text{N}(0,\ \sigma^2\boldsymbol{M}_{XX}^{-1}). \tag{7.1.15}$$

This completes the proof of the following theorem.

Theorem 7.1.1 If Assumptions 7.1.1 and 7.1.2 plus conditions (7.1.3) and (7.1.12) hold, the OLS estimator of $\boldsymbol{\beta}$ is CAN, with the distribution given in (7.1.15). □

The practical application of these results is straightforward. They imply that in large samples, the interval estimates and test procedures described in §2.4 can be applied in the present context, basically without modification. A consistent

estimate of $\sigma^2 \boldsymbol{M}_{XX}^{-1}$ is provided by $ns^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$ where s^2 is the usual residual variance estimate, so that according to the Cramér Theorem, the usual t statistics for the significance of individual regressors are asymptotically $\mathrm{N}(0,1)$ on the null hypothesis. Similarly, the 'F tests' of linear restrictions in §2.4.3 can be performed in the usual manner, although the degrees of freedom corrections are optional and the tests are often interpreted as chi-squared tests; that is to say, the test statistic in (2.4.12) multiplied by r, the number of restrictions under test, is treated as asymptotically $\chi^2(r)$ under the null hypothesis. In practice, as argued in §3.5.5, the $F(n-k,r)$ distribution may provide at least as good an approximation to the true unknown distribution of the F statistic, with finite n, as does $\chi^2(r)/r$.

7.1.4 A VAR Application

As an example that extends the analysis of Chapter 6 in the crucial way, suppose the equation of interest is embedded in a structural VAR of the form (4.2.9). To fix ideas, consider the simple system

$$y_t = \beta x_t + \gamma y_{t-1} + u_t \tag{7.1.16a}$$

$$x_t = \lambda x_{t-1} + v_t \tag{7.1.16b}$$

in which u_t at least satisfies Assumptions 7.1.1 and 7.1.2, and the disturbances are contemporaneously uncorrelated, so that the first equation is a legitimate regression model with $\mathcal{I}_t = \sigma(x_{t-j}, y_{t-j-1},\ j > 0)$. The key condition to be checked, subject to these assumptions, is (7.1.3). The reduced or VAR form of the system is

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \gamma & \beta\lambda \\ 0 & \lambda \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_t + \beta v_t \\ v_t \end{bmatrix} \tag{7.1.17}$$

or in vector form,

$$\boldsymbol{z}_t = \boldsymbol{\Lambda}\boldsymbol{z}_{t-1} + \boldsymbol{\varepsilon}_t. \tag{7.1.18}$$

The characteristic equation of the system is

$$|\boldsymbol{\Lambda} - \mu \boldsymbol{I}| = \mu^2 - (\gamma + \lambda)\mu + \lambda\gamma = 0 \tag{7.1.19}$$

which has solutions of γ and λ so the stability conditions, as is evident by inspection, are $|\gamma| < 1$ and $|\lambda| < 1$.

Subject to stability, the sample moments $n^{-1}\sum_{t-2}^{n} x_t^2$, $n^{-1}\sum_{t-2}^{n} y_{t-1}^2$ and $n^{-1}\sum_{t-2}^{n} x_t y_{t-1}$ must be shown to converge in probability and the nature of the limits determined. The first of these problems has already been dealt with in Chapter 6, and one possible approach is to undertake a vector generalization of the analysis in §6.3.1. However, this approach is based on Assumption 6.3.1, which as noted there, is quite restrictive in requiring the model to be 'correct' in essential respects. In particular, the m.d. property was required of the shocks, which in this context would have to be extended to all the equations of the system, not just the equation of interest. It is clearly better to be able to assume that the specification

is incomplete, such that some elements of $\boldsymbol{\varepsilon}_t$ (v_t, in the example) are mixing processes, but not necessarily serially uncorrelated. Therefore we adopt the approach of checking the conditions of Theorem 6.4.4. This also provides a good excuse to explore the use of this theory, whose application is potentially wider than just to linear models.

The next part of the analysis applies to VARs of any dimension (assume p) and also any order, via the companion form. To check the near-epoch dependence condition expand the VAR as in (4.3.6) and use the approach of (6.4.7). Assume the eigenvalues of $\boldsymbol{\Lambda}$ are distinct, so that the decomposition $\boldsymbol{\Lambda} = \boldsymbol{QMQ}^{-1}$ holds as in (4.3.2), where $\boldsymbol{M}$ is diagonal. Defining the vectors $\boldsymbol{z}_t^* = \boldsymbol{Q}^{-1}\boldsymbol{z}_t$ and $\boldsymbol{\varepsilon}_t^* = \boldsymbol{Q}^{-1}\boldsymbol{\varepsilon}_t$, note that $\boldsymbol{z}_t^*$ has the representation as a vector of univariate AR processes with the form

$$z_{jt}^* = \mu_j z_{j,t-1}^* + \varepsilon_{jt}^* = \sum_{k=0}^{m} \mu_j^k \varepsilon_{j,t-k}^* + \mu_j^{m+1} z_{j,t-m-1}^* \tag{7.1.20}$$

for $j = 1, \dots, p$.[2] Letting $\mathcal{F}_s^t = \sigma(\boldsymbol{\varepsilon}_s, \dots, \boldsymbol{\varepsilon}_t)$ for $t \geq s$, and applying (6.4.7),

$$\|z_{jt}^* - E(z_{jt}^* | \mathcal{F}_{t-m}^{t+m})\|_2 \leq 2|\mu_j|^{m+1} \|z_{j,t-m-1}^*\|_2. \tag{7.1.21}$$

In a stable system with $|\mu_j| < 1$ for each j the elements of $\boldsymbol{z}_t^*$ are accordingly L_2-NED on $\boldsymbol{\varepsilon}_t$ of size $-\infty$. Finally consider the vector $\boldsymbol{z}_t = \boldsymbol{Q}\boldsymbol{z}_t^*$. Minkowski's inequality gives

$$\|z_{it} - E(z_{it} | \mathcal{F}_{t-m}^{t+m})\|_2 \leq \sum_{j=1}^{p} |q_{ij}| \|z_{jt}^* - E(z_{jt}^* | \mathcal{F}_{t-m}^{t+m})\|_2 \tag{7.1.22}$$

which shows that these elements share the L_2-NED property.

The weak law of large numbers is to be proved for the squares and cross products of the variables, but these are L_1-NED of the same size as the variables themselves by Theorem 6.4.3(i). It only remains to check the uniform moment condition imposed by Theorem 6.4.4, and in view of the Cauchy–Schwarz inequality it is enough to check this for the squared terms. First note that, for $1 \leq i \leq m$,

$$\|z_{it}^2\|_{1+\delta}^{1/2} = \|z_{it}\|_{2+2\delta} \leq \sum_{j=1}^{p} |q_{ij}| \|z_{jt}^*\|_{2+2\delta} \tag{7.1.23}$$

and substituting from (7.1.20) with $m = \infty$, the Minkowski inequality gives

$$\|z_{jt}^*\|_{2+2\delta} \leq \sum_{k=0}^{\infty} |\mu_j|^k \|\varepsilon_{j,t-k}^*\|_{2+2\delta} \leq \sum_{k=0}^{\infty} |\mu_j|^k \sum_{i=1}^{p} |q^{ji}| \|\varepsilon_{i,t-k}\|_{2+2\delta} \tag{7.1.24}$$

where the q^{ji} denote the elements of $\boldsymbol{Q}^{-1}$. Since the $\{|\mu_j|^k, k = 0, 1, 2, \dots\}$ form a summable sequence when $|\mu_j| < 1$, putting together (7.1.23) and (7.1.24) leads to

[2] Although in the example the eigenvalues are real they can be complex in general, and then the z_{jt}^* are complex-valued. The following analysis applies unchanged, noting that the L_2-norm of a complex r.v. x is $(E|x|^2)^{1/2}$ where $|x|$ is the modulus of x.

the conclusion that, if $\|\varepsilon_{j,t-k}\|_{2+2\delta} < \infty$ uniformly in t, the conditions of Theorem 6.4.4 are satisfied. Note that for this result, the only restriction needed on the distribution of the shocks is that they are mixing, and possess the moments of the required order.

Having established the existence of the limit $\boldsymbol{M}_{XX}$, the next step (in principle at least) should be to determine its form, and check that it is positive definite. While the following analysis is perfectly general, it is easiest to revert to the simple example to see how it works since the calculations are potentially burdensome. Also for the sake of simplicity let u_t and v_t satisfy Assumptions 7.1.1 and 7.1.2, with variances σ_u^2 and σ_v^2. The idea is to form suitable equations in the second moments by operating directly on (7.1.16). Inspection easily shows that the following relations hold:

$$E(y_t^2) = \beta^2 E(x_t^2) + \gamma^2 E(y_{t-1}^2) + 2\beta\gamma E(x_t y_{t-1}) + \sigma_u^2 \tag{7.1.25a}$$

$$E(x_t^2) = \lambda^2 E(x_{t-1}^2) + \sigma_v^2 \tag{7.1.25b}$$

$$E(x_t y_t) = \beta E(x_t^2) + \gamma E(x_t y_{t-1}) \tag{7.1.25c}$$

$$E(x_t y_{t-1}) = \lambda E(x_{t-1} y_{t-1}). \tag{7.1.25d}$$

Equations (7.1.25a) and (7.1.25b) are got by squaring both sides of (7.1.16a) and (7.1.16b), and (7.1.25c) and (7.1.25d) by multiplying them through by x_t and y_{t-1} respectively. Since the system is stable and the shocks are white noise the variables are wide-sense stationary, so $E(y_t^2) = E(y_{t-1}^2)$, $E(x_t^2) = E(x_{t-1}^2)$ and $E(x_t y_t) = E(x_{t-1} y_{t-1})$. These substitutions leave four equations in four unknowns, of which three are of interest. It is easy to verify that the solutions are

$$E(x_t^2) = \frac{\sigma_v^2}{1-\lambda^2} \tag{7.1.26a}$$

$$E(x_t y_{t-1}) = \frac{\lambda\beta\sigma_v^2}{(1-\gamma\lambda)(1-\lambda^2)} \tag{7.1.26b}$$

$$E(y_{t-1}^2) = \frac{\sigma_u^2}{1-\gamma^2} + \frac{\beta^2(1+\gamma\lambda)\sigma_v^2}{(1-\gamma^2)(1-\lambda^2)(1-\gamma\lambda)}. \tag{7.1.26c}$$

Subject to the stability conditions these solutions exist and, as can be verified, form a positive definite matrix.

7.1.5 Asymptotic Efficiency

The efficiency analysis in §3.6 extends more or less unchanged to the present case, once the CAN property is established, with a few additional considerations. We considered a class of linear estimators of the form

$$\boldsymbol{\beta}_L = \sum_{t=1}^{n} \boldsymbol{l}_{nt} y_t \tag{7.1.27}$$

and then a subclass $\boldsymbol{\beta}_W$ for which

$$\boldsymbol{l}_{nt} = \boldsymbol{P}_n \boldsymbol{w}_t. \tag{7.1.28}$$

To apply Theorems 3.6.1 and 3.6.2 to Model C, replace Assumptions 3.6.1 and 3.6.2 with $\boldsymbol{w}_t \in \mathcal{I}_t$. Subject to the usual regularity conditions, the results

$$\operatorname{plim} \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{w}_t u_t = \mathbf{0} \tag{7.1.29}$$

$$\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{w}_t u_t \xrightarrow{\text{D}} \text{N}(\mathbf{0}, \sigma^2 \boldsymbol{M}_{WW}) \tag{7.1.30}$$

follow from Assumptions 7.1.1 and 7.1.2, and the analysis can proceed as before.

There is virtually no loss of generality in confining attention to the class of instrumental variables estimators, defined by (7.1.28). Argue as follows. Any member of the class $\boldsymbol{\beta}_L$ can be represented using the form (7.1.28) if $\boldsymbol{w}_t$ is allowed to bear an array subscript, as $\boldsymbol{w}_{nt}$. One may simply write $\boldsymbol{l}_{nt} = \boldsymbol{P}_n \boldsymbol{w}_{nt}$ and if necessary choose $\boldsymbol{P}_n = n^{-1} \boldsymbol{I}_k$. Since any multiplicative scaling factors can be incorporated in $\boldsymbol{P}_n$, the main distinctive role for the array subscript would be to allow $\boldsymbol{w}_{nt}$ to depend in some way on the whole sample of data. Results (7.1.29) and (7.1.30) hold under the assumptions because $\boldsymbol{w}_t \in \mathcal{I}_t$, but do *not* hold in general if $\boldsymbol{w}_t \in \mathcal{I}_s$ for $s > t$. In particular, $\boldsymbol{w}_t$ is only allowed to depend on data up to and including date t. The only way such a vector could not be unambiguously labelled with the single subscript t would be due to some form of non-multiplicative transformation of the data depending on sample size. If such unusual possibilities are neglected, (7.1.28) represents the maximum generality compatible with consistency.

7.2 Extensions of the Basic Model

7.2.1 Dummy Regressors

We have a good idea of what is required to enforce condition (7.1.3) when the regressors are stochastic. The conditions of a weak law of large numbers must be satisfied, which, if the data are generated by a VAR, involves stable roots and suitable moment and/or dependence restrictions on the innovations. However, the case of dummy (deterministic) regressors, as discussed in §4.1.2, also needs consideration, and some care is needed in treating these when the sample size is being extended to infinity.

Write $\boldsymbol{x}_t = (\boldsymbol{s}_t', \boldsymbol{d}_t')'$, where $\boldsymbol{s}_t$ denotes stochastic variables, and $\boldsymbol{d}_t$ nonstochastic variables. A sequence of real numbers $\{a_t,\ t = 1, 2, 3, \dots\}$ is said to be *Cesàro-summable* if it has the property $|\bar{a}| < \infty$, where

$$\bar{a} = \lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} a_t. \tag{7.2.1}$$

The constant $\bar{a}$ is called the Cesàro sum of the sequence. Clearly, to satisfy (7.1.3) the elements of the matrix $\boldsymbol{d}_t \boldsymbol{d}_t'$ must be Cesàro-summable. If this holds, and the elements of $\boldsymbol{s}_t$ obey the WLLN, there should typically be no difficulty in showing that elements of the product matrix $\boldsymbol{d}_t \boldsymbol{s}_t'$ also obey the WLLN. The intercept

(equals 1 in every period) satisfies the Cesàro-summability condition trivially, and the seasonal dummies, taking the value 1 every fourth (or twelfth) period and zero otherwise, have Cesàro sums of 1/4 or 1/12, as the case may be. Any variable that cycles through a fixed pattern of values in this way should satisfy Cesàro-summability, and the Cesàro sum is identical to the integral with respect to the probability measure that assigns the relative frequency of each value in the sequence.[3]

Intervention dummies are variables designed to isolate a particular period of the sample (a war, strike, policy reform, or similar) being set equal to 1 in the period in question, 0 otherwise. They ought probably to be regarded as stochastic since, unlike the case of seasonal patterns, it would not make sense to argue that these phenomena always fall in the same periods, when considering the ensemble distribution of the time series. The motivation for 'dummying' such effects is not that they are non-random, but that the assumptions about the disturbance distribution would otherwise be violated. Any sequence that is different from zero only for a finite set of time periods, as $n \to \infty$, has a Cesàro sum of zero. The corresponding column of $\boldsymbol{M}_{XX}$ is likewise zero. In the analyses of either §3.5 or §7.1 such cases must obviously be excluded. The only kind of intervention admissible in an asymptotic inference framework is one where a fixed fraction of the observations are dummied as n increases.

Strictly speaking, 'unique' interventions, which will not recur as the sample is extended, ought to be modelled stochastically, incorporating them into the error distribution. By definition, the influence of these observations is negligible in the limit. If in practice they are unduly influential in the sample available, they ought to be dropped or modified, and dummying can be thought of as an ad hoc device for achieving this. Note that the dummying of a single observation has exactly the same effect on the estimates as dropping this observation from the sample. The CAN property cannot be claimed for the coefficients of such dummies, and tests of significance have no validity on asymptotic criteria.[4]

The other type of dummy variable often introduced is the polynomial trend dummy, having the form $d_t = t^r$ for $r \geq 1$. Such series are not Cesàro-summable, so once again the conventional analysis breaks down. However, they can be dealt with by applying the modifications described in the next section.

7.2.2 Global Nonstationarity

The analyses of §3.5 and §7.1 built in a simplifying but unnecessary feature, that the convergence of all the estimates is at the rate $\sqrt{n}$. While it is possible to define data series that do not yield this result, the CAN property can in some cases still be obtained with a suitable normalization. Define a diagonal $k \times k$ matrix $\boldsymbol{K}_n$ whose diagonal elements are positive functions of n for $i = 1, \ldots, k$. This allows a different scaling to be applied to each regressor, although setting $\boldsymbol{K}_n = n\boldsymbol{I}_k$ in

[3] See Gallant (1977), or Burguete, Gallant and Souza (1982) for further details.

[4] See §7.2.2, and also §7.6.5 for an example.

what follows will reproduce the previous analysis. Instead of (7.1.3), assume that

$$\begin{aligned}\operatorname{plim} \boldsymbol{K}_n^{-1/2} \sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t' \boldsymbol{K}_n^{-1/2} &= \lim_{n\to\infty} \boldsymbol{K}_n^{-1/2} \sum_{t=1}^{n} E(\boldsymbol{x}_t \boldsymbol{x}_t') \boldsymbol{K}_n^{-1/2} \\ &= \boldsymbol{M}_{XX} < \infty \end{aligned} \tag{7.2.2}$$

(positive definite). Since $\operatorname{Var}(\boldsymbol{\lambda}'\boldsymbol{x}_t u_t) = O(\boldsymbol{\lambda}'\boldsymbol{K}_n\boldsymbol{\lambda}/n)$, (7.1.12) must be replaced by

$$\sqrt{n}\frac{\max_{1\le t\le n} \|\boldsymbol{\lambda}'\boldsymbol{x}_t u_t\|_{2+\delta}}{(\boldsymbol{\lambda}'\boldsymbol{K}_n\boldsymbol{\lambda})^{1/2}} \le B < \infty \quad \delta > 0,\ n \ge 1,\ \forall \text{ fixed } \boldsymbol{\lambda} \text{ with } \boldsymbol{\lambda}'\boldsymbol{\lambda} = 1. \tag{7.2.3}$$

Theorems 6.2.3 and 3.3.3 then allow us to say that

$$\boldsymbol{K}_n^{-1/2} \sum_{t=1}^{n} \boldsymbol{x}_t u_t \xrightarrow{\mathrm{D}} \mathrm{N}(0, \sigma^2 \boldsymbol{M}_{XX}). \tag{7.2.4}$$

Instead of (7.1.15) the conclusion becomes

$$\boldsymbol{K}_n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N}(0, \sigma^2 \boldsymbol{M}_{XX}^{-1}). \tag{7.2.5}$$

Global stationarity describes the situation in which, if the sample is split into contiguous parts of fixed relative size, for example, $t = 1, \dots, [n/2]$ and $t = [n/2]+1, \dots, n$, each subsample shows the same average behaviour in the limit.[5] This is weaker than stationarity because it does not rule out local heterogeneity such as seasonal or cyclical patterns. Stationarity is not necessary for (7.1.3), but global stationarity is necessary.

A case of nonstationarity often considered is so-called *trend stationarity*,[6] meaning that the time series can be thought of as the weighted sum of a linear trend term and a stationary process. If y_t is of this form, the regression model that explains it must include another trend-stationary variable in the regressors. The simplest case is where this regressor is just the trend dummy. In other words, suppose that $x_{1t} = t$, the other regressors satisfying the usual assumptions. Cesàro-summability does not hold, as noted, and (7.1.3) fails. However, since $n^{-3}\sum_{t=1}^{n} t^2 \to 1/3$,[7] setting $(\boldsymbol{K}_n)_{11} = n^3$ will satisfy (7.2.2). With $\lambda_1 \neq 0$, $\|\boldsymbol{\lambda}'\boldsymbol{x}_t u_t\|_{2+\delta} = O(t)$, and hence its maximum is of $O(n)$ which matches the order of magnitude of $(\boldsymbol{\lambda}'\boldsymbol{K}_n\boldsymbol{\lambda}/n)^{1/2}$, and (7.2.3) is satisfied. The coefficient of a linear trend is asymptotically Gaussian, but the rate of convergence is $n^{3/2}$, not $\sqrt{n}$. This is a case of 'super-consistency'.

---

[5] This is a somewhat simplified description. See Davidson (1994a) §13.4 for further discussion of this question.

[6] This terminology is accepted but somewhat misleading. 'Trend-nonstationarity' would be a more descriptive usage here.

[7] Two useful formulae are $\sum_{t=1}^{n} t = n(n+1)/2$, and $\sum_{t=1}^{n} t^2 = n(n+1)(2n+1)/6$. Gauss is said to have derived the first as a schoolboy, perhaps the reader can! A useful hint is to consider the squares of the chessboard – how many squares fall on or below the diagonal?

This is the situation that obtains when there is just one trending variable among the regressors. However, an interesting situation arises if two or more regressors are trend-stationary. Suppose for illustration that $\boldsymbol{x}_t = \boldsymbol{v}_t + \boldsymbol{\gamma} t$ where $n^{-1}\sum_{t=1}^{n} \boldsymbol{v}_t \boldsymbol{v}_t' \xrightarrow{\text{pr}} \boldsymbol{M}_{vv}$ and $\boldsymbol{\gamma}$ is a vector of constants.[8] In this case, $\boldsymbol{K}_n = n^3 \boldsymbol{I}_k$ would appear to be the appropriate scaling matrix. However,

$$\frac{1}{n^3}\sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t' \xrightarrow{\text{pr}} \tfrac{1}{3}\boldsymbol{\gamma}\boldsymbol{\gamma}' \tag{7.2.6}$$

which is a singular matrix, having rank 1. The least squares estimator is therefore inconsistent, being undefined in the limit. This can be interpreted as a case of extreme limiting multicollinearity where the deterministic components of $\boldsymbol{x}_t$, which cannot be distinguished, eventually dominate the stochastic components.

Strangely enough, this problem is resolved rather easily by adding the trend dummy to the regression.[9] This does not, as might be thought, make the problem worse. Writing the model as

$$y_t = \boldsymbol{\beta}'\boldsymbol{x}_t + u_t = \boldsymbol{\beta}'\boldsymbol{v}_t + \boldsymbol{\beta}'\boldsymbol{\gamma} t + u_t \tag{7.2.7}$$

it can be seen that the effects of $\boldsymbol{x}_t$ on y_t can be separated into the stationary components and a single trend component with coefficient $\boldsymbol{\beta}'\boldsymbol{\gamma}$. Switching to full-sample notation, let $\boldsymbol{d} = (1, 2, \ldots, n)'$ denote the dummy vector and consider running the regression

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \theta \boldsymbol{d} + \boldsymbol{u} \tag{7.2.8}$$

where the true value of θ is actually 0. The partitioned algebra of §1.2.5 reveals the properties of this regression. Writing $\boldsymbol{X} = \boldsymbol{V} + \boldsymbol{d}\boldsymbol{\gamma}'$, observe that $\boldsymbol{M}_d \boldsymbol{X} = \boldsymbol{V}$ where $\boldsymbol{M}_d = \boldsymbol{I}_n - \boldsymbol{d}(\boldsymbol{d}'\boldsymbol{d})^{-1}\boldsymbol{d}'$. Therefore, if $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{M}_d\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{M}_d\boldsymbol{y}$ note that

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \Big(\frac{\boldsymbol{V}'\boldsymbol{V}}{n}\Big)^{-1} \frac{\boldsymbol{V}'\boldsymbol{u}}{\sqrt{n}} \xrightarrow{\text{D}} \text{N}(\boldsymbol{0},\, \sigma^2 \boldsymbol{M}_{VV}^{-1}). \tag{7.2.9}$$

On the other hand, the coefficient of the dummy is

$$\hat{\theta} = (\boldsymbol{d}'\boldsymbol{d})^{-1}\boldsymbol{d}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = (\boldsymbol{d}'\boldsymbol{d})^{-1}\boldsymbol{d}'\boldsymbol{M}_V\boldsymbol{u} - \boldsymbol{\gamma}'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = O_p(n^{-1/2}) \tag{7.2.10}$$

where $\boldsymbol{M}_V = \boldsymbol{I}_n - \boldsymbol{V}(\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'$. Therefore, the usual scale factor of $\sqrt{n}$ is appropriate to this case. The penalty of having collinear trends explaining the dependent variable, relative to a single trend, is that the 'super-consistency' represented by the $n^{3/2}$ rate of convergence is lost.

There are other varieties of global nonstationarity, that exhibit 'super-consistency' and can in principle be analysed in this way, but caution is needed since the limiting distributions are generally not Gaussian. These are the cases where

[8] The constant term is omitted for simplicity here.

[9] It is assumed here that the trend dummy is not one of the variables already included. If it is, the following analysis applies with minor modifications, allowing the dummy to have a nonzero true coefficient.

$\boldsymbol{x}_t$ contains integrated stochastic processes, having the form $\sum_{s=1}^{t} z_s$ where z_t is a stationary process and not the difference of a stationary process. Then there is no choice of $\boldsymbol{K}_n$ that will satisfy (7.2.2) with $\boldsymbol{M}_{XX}$ a non-stochastic matrix. Part IV of the book is devoted to the cases of this type.

Other cases where the asymptotic analysis does not apply, for wholly different reasons, are the intervention dummies discussed in §7.2.1, which are characterized as equal to zero except for a finite number of time periods. For these it is appropriate to put the corresponding elements of $\boldsymbol{K}_n$ to $n^0 = 1$. Under this normalization, $\boldsymbol{M}_{XX}$ is finite and nonsingular in the limit. The failure of asymptotic normality of the coefficients is evident on setting the other elements of $\boldsymbol{\lambda}$ (relating to the regular regressors) to 0, for then the denominator in (7.2.3) is $O(n^{-1})$, and the numerator is $O(1)$. However, as can be shown by considering the relevant partition of $\hat{\boldsymbol{\beta}}$, the CAN property continues to hold for the coefficients of the regular regressors, for which one can set $(\boldsymbol{K}_n)_{ii} = n^r$ for $r \geq 1$.

7.3 Consequences of Misspecification

Assumptions 7.1.1 and 7.1.2 define a model that is correctly specified in the strong sense, when $\mathcal{I}_t$ is the maximal set of valid conditioning variables. However, it is reasonable to ask the following question: if all the asymptotic distribution results go through when the conditioning set is defined to be smaller than $\mathcal{I}_t$, how does it matter if these conditions are violated? In other words, are some misspecifications innocuous?

7.3.1 Misspecification in Mean

Suppose first that $(y_t, \boldsymbol{x}_t')'$ is an i.i.d. vector. In this case, for reasons related to the discussion in §1.4.2, all our results hold in the case $\mathcal{I}_t = \sigma(\boldsymbol{x}_t)$. The interpretation of the estimated parameter vector will of course depend on the choice of $\boldsymbol{x}_t$, and a 'relevant' variable, in this context, might be defined as either one whose coefficient is a parameter of interest, or one to whose inclusion the values of parameters of interest are sensitive, in the sense that $\boldsymbol{\beta}$ in (1.4.9) is different from $\boldsymbol{\gamma}$ in (1.4.12). However, so long as $\boldsymbol{\beta}$ is defined in terms of the equation

$$E(y_t|\boldsymbol{x}_t) = \boldsymbol{\beta}'\boldsymbol{x}_t \tag{7.3.1}$$

the CAN property of OLS estimates will go through for *any* choice of $\boldsymbol{x}_t$. In fact, even linearity of $E(y_t|\boldsymbol{x}_t)$ can be dispensed with, as indicated in §1.4.1.

In the time series context, things are different because of serial dependence. If there is serial correlation of the disturbance terms, equation (7.1.9) no longer holds and while the CAN property may still hold,[10] the asymptotic variance formula in (7.1.14) is no longer valid, since

$$\operatorname{Var}\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{x}_t u_t\right) \neq \frac{1}{n}\sum_{t=1}^{n}\operatorname{Var}(\boldsymbol{x}_t u_t). \tag{7.3.2}$$

[10] For a justification of this statement see §10.5.

To validate the asymptotic distribution results, the inclusion of u_{t-j} in the conditioning set for all $j > 0$ is therefore *not* optional. Given equation (7.1.1), this means conditioning on the lags of all included variables, at least. Strictly speaking there is no need to condition on any other variables, but this is the case only because the included variables can act as proxies for the excluded ones.

Consider an example. Let there exist a conditioning variable, z_t say, which explains y_t, but is independent of $\boldsymbol{x}_t$ so that $\boldsymbol{\beta}$ is invariant to its inclusion in the model. In other words,

$$E(y_t|\boldsymbol{x}_t, z_t) = \boldsymbol{\beta}'\boldsymbol{x}_t + \gamma z_t \tag{7.3.3}$$

holds, where $\gamma \neq 0$, but $E(z_t|\boldsymbol{x}_t) = 0$ and hence $\boldsymbol{\beta}$ is the same in (7.3.1) and in (7.3.3). In the case of i.i.d. data, OLS applied to the smaller model certainly yields CAN estimates of $\boldsymbol{\beta}$, although since the residual variance is reduced by including z_t, this would be desirable on the grounds of asymptotic efficiency. In the time series case, however, if the series $e_t = y_t - \boldsymbol{\beta}'\boldsymbol{x}_t - \gamma z_t$ is a martingale difference then $u_t = y_t - \boldsymbol{\beta}'\boldsymbol{x}_t$ cannot also be, in general. This would only be true if z_t itself were an m.d., which would be the exception to the rule for economic time series. If z_t is not in the data set, and the autocorrelation in u_t can be approximated by a finite-order AR form, its omission could be alleviated by including lags of u_t as explanatory variables in the model.[11] Clearly, omission of z_t is not innocuous in the same way as it would be in a random sampling framework. This is a major distinction between sampling model C and the other cases discussed in §1.3.

7.3.2 Misspecification in Variance

The assumption $E(u_t^2|\mathcal{I}_t) = \sigma^2$ is likewise not necessary for the CAN property, but its failure also results in the variance formula being incorrect, in this case because of the failure of equation (7.1.8). This failure is the condition known as conditional heteroscedasticity, although since $\mathcal{I}_t$ can be thought of as including deterministic variables the case of ordinary heteroscedasticity, where $E(u_t^2)$ is a deterministic function of t, is subsumed under the definition. In this case the condition in (7.1.10) must be replaced by the assumption

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{Var}(\boldsymbol{x}_t u_t) \to \boldsymbol{A} \text{ as } n \to \infty \tag{7.3.4}$$

where $\boldsymbol{A}$ is some finite, positive definite matrix. Maintaining the other assumptions as before, so that the vectors $\boldsymbol{x}_t u_t$ form a v.m.d. and are serially uncorrelated,

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} \boldsymbol{x}_t u_t \xrightarrow{\mathrm{D}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{A}) \tag{7.3.5}$$

and hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N}(\boldsymbol{0},\ \boldsymbol{M}_{XX}^{-1}\boldsymbol{A}\boldsymbol{M}_{XX}^{-1}) \tag{7.3.6}$$

[11] See §10.2.3 on the implementation of this technique.

which is, however, different from (7.1.15). This misspecification will therefore result in biased inferences if it is ignored.

It can be overcome by simply using a consistent estimate of the variance, which is generally available. Applying the kind of argument leading to Theorem 6.5.4, it can be assumed that $\boldsymbol{A}$ is consistently estimated by

$$\hat{\boldsymbol{A}} = \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t' \hat{u}_t^2 \tag{7.3.7}$$

where $\hat{u}_t$ denotes the OLS residual. Formula (7.3.6) can therefore be used to generate asymptotically valid standard errors for $\hat{\boldsymbol{\beta}}$, defined as the square roots of the diagonal elements of the matrix $n\hat{\boldsymbol{M}}_{XX}^{-1}\hat{\boldsymbol{A}}\hat{\boldsymbol{M}}_{XX}^{-1}$, where $\hat{\boldsymbol{M}}_{XX} = n^{-1}\sum_{t=1}^{n} \boldsymbol{x}_t \boldsymbol{x}_t'$. Originally proposed by Eicker (1963) and independently by White (1980a), these are usually known as the *White standard errors*.

7.4 The Model Selection Problem

The biggest problem we face in econometrics is the uncertainty about the correct specification of our models. Because of the non-experimental nature of economics, we are never sure how the observed data were generated. The test of any hypothesis in economics always turns out to depend on additional assumptions necessary to specify a reasonably parsimonious model, which may or may not be justified. Specification testing, and diagnostic testing, are synonymous terms used to describe procedures for determining whether a model is contradicted when confronted by a data set. A model that is not rejected in this way may be said to be *data coherent*.[12]

7.4.1 Spurious Regression

Let us first of all dispose of a notorious fallacy. Many textbooks describe what is called the 'test of significance of the regression'. Assuming $\boldsymbol{x}_{kt} = 1$ (the intercept term), this is the joint test of the significance of $\boldsymbol{x}_{1t}, \ldots, \boldsymbol{x}_{k-1,t}$ by the technique of §2.4.5. It is easy to verify that this test statistic takes the form

$$F = \frac{R^2}{1-R^2}.\frac{n-k}{k-1} \tag{7.4.1}$$

where R^2 is defined in (1.2.12). The test formalizes the notion that a large R^2 implies a 'significant' explanation of y_t by $\boldsymbol{x}_t$, and under the assumptions listed in Chapter 2, the statistic is distributed as $F(k-1, n-k)$ on $H_0 : \beta_1 = \cdots = \beta_{k-1} = 0$.

However, it is clear that this test is valid only if y_t is randomly sampled. When the data in question are time series it is generally misleading, in the sense that

[12] The term data coherent was introduced by Hendry and Richard (1982) to describe a property effectively equivalent to this one. Note that Hendry (1995) has defined the term 'data congruent' to refer to a somewhat more elaborate concept of correct specification.