TABLE 12.3 Estimated Log Wage Equations

	Variables	OLS	GLS/RE	LSDV	HT/IV-GLS	HT/IV-GLS
x_1	Experience	0.0132	0.0133	0.0241	0.0217	
		(0.0011)[a]	(0.0017)	(0.0042)	(0.0031)	
	Bad health	−0.0843	−0.0300	−0.0388	−0.0278	−0.0388
		(0.0412)	(0.0363)	(0.0460)	(0.0307)	(0.0348)
	Unemployed Last Year	−0.0015	−0.0402	−0.0560	−0.0559	
		(0.0267)	(0.0207)	(0.0295)	(0.0246)	
	Time	NR[b]	NR	NR	NR	NR
x_2	Experience					0.0241
						(0.0045)
	Unemployed					−0.0560
						(0.0279)
z_1	Race	−0.0853	−0.0878		−0.0278	−0.0175
		(0.0328)	(0.0518)		(0.0752)	(0.0764)
	Union	0.0450	0.0374		0.1227	0.2240
		(0.0191)	(0.0296)		(0.0473)	(0.2863)
	Schooling	0.0669	0.0676			
		(0.0033)	(0.0052)			
	Constant	NR	NR	NR	NR	NR
z_2	Schooling				0.1246	0.2169
					(0.0434)	(0.0979)
	σ_ε	0.321	0.192	0.160	0.190	0.629
	$\rho = \sqrt{\sigma_u^2/(\sigma_u^2+\sigma_\varepsilon^2)}$		0.632		0.661	0.817
	Spec. Test [3]		20.2		2.24	0.00

[a]Estimated asymptotic standard errors are given in parentheses.
[b]NR indicates that the coefficient estimate was not reported in the study.

11.8.3 ~~11.5.3~~ EFFICIENT ESTIMATION OF DYNAMIC PANEL DATA MODELS—THE ~~ANDERSON/HSIAO AND~~ ARELLANO/BOND ESTIMATORS

KT A leading contemporary application of the methods of this chapter ~~and Chapter 8~~ is the **dynamic panel data model,** which we now write

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it}.$$

Several applications are described in Example ~~9.18~~ 11.21. The basic assumptions of the model are

1. Strict exogeneity: $E[\varepsilon_{it} \mid \mathbf{X}_i, c_i] = 0$,
2. Homoscedasticity: $E[\varepsilon_{it}^2 \mid \mathbf{X}_i, c_i] = \sigma_\varepsilon^2$,
3. Nonautocorrelation: $E[\varepsilon_{it}\varepsilon_{is} \mid \mathbf{X}_i, c_i] = 0$ if $t \neq s$,
4. Uncorrelated observations: $E[\varepsilon_{it}\varepsilon_{js} \mid \mathbf{X}_i, c_i, \mathbf{X}_j, c_j] = 0$ for $i \neq j$ and for all t and s,

where the rows of the $T \times K$ data matrix $\mathbf{X}_i$ are $\mathbf{x}_{it}'$. We will not assume mean independence. The "effects" may be fixed or random, so we allow

$$E[c_i \mid \mathbf{X}_i] = g(\mathbf{X}_i).$$

(See Section ~~9~~11.2.1.) We will also assume a fixed number of periods, T, for convenience. The treatment here (and in the literature) can be modified to accommodate unbalanced panels, but it is a bit inconvenient. (It involves the placement of zeros at various places

11-71

in the data matrices defined below and, of course, changing the terminal indexes in summations from 1 to T.)

The presence of the lagged dependent variable in this model presents a considerable obstacle to estimation. Consider, first, the straightforward application of assumption A.I3 in Section ~~12.2:~~ The compound disturbance in the model is $(c_i + \varepsilon_{it})$. The correlation between $y_{i,t-1}$ and $(c_i + \varepsilon_{i,t})$ is obviously nonzero because $y_{i,t-1} = \mathbf{x}'_{i,t-1}\boldsymbol{\beta} + \delta y_{i,t-2} + c_i + \varepsilon_{i,t-1}$:

8.2

$$\mathrm{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \delta\,\mathrm{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})].$$

If T is large and $-1 < \delta < 1$, then this covariance will be approximately $\sigma_c^2/(1-\delta)$. The large T assumption is not going to be met in most cases. But, because δ will generally be positive, we can expect that this covariance will be at least larger than σ_c^2. The implication is that both (pooled) OLS and GLS in this model will be inconsistent. Unlike the case for the static model ($\delta = 0$), the fixed effects treatment does not solve the problem. Taking group mean differences, we obtain

$$y_{i,t} - \overline{y}_{i.} = (\mathbf{x}_{i,t} - \overline{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \delta(y_{i,t-1} - \overline{y}_{i.}) + (\varepsilon_{i,t} - \overline{\varepsilon}_{i.}). \qquad \text{~~11-60~~ ~~(12-30)~~}$$

As shown in Anderson and Hsiao (1981, 1982),

$$\mathrm{Cov}[(y_{i,t-1} - \overline{y}_{i.}), (\varepsilon_{i,t} - \overline{\varepsilon}_{i.})] \approx \frac{-\sigma_\varepsilon^2}{T^2}\,\frac{(T-1) - T\delta + \delta^T}{(1-\delta)^2}. \qquad \text{~~11-61~~ ~~(12-31)~~}$$

This result is O(1/T), which would generally be no problem if the asymptotics in our model were with respect to increasing T. But, in this panel data model, T is assumed to be fixed and relatively small. For conventional values of T, say 5 to 15, the proportional bias in estimation of δ could be on the order of, say, 15 to 60 percent.

Neither OLS nor GLS are useful as estimators. There are, however, instrumental variables available within the structure of the model. Anderson and Hsiao (1981, 1982) proposed an approach based on first differences rather than differences from group means,

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1}.$$

For the first full observation,

66

$$y_{i3} - y_{i2} = (\mathbf{x}_{i3} - \mathbf{x}_{i2})'\boldsymbol{\beta} + \delta(y_{i2} - y_{i1}) + \varepsilon_{i3} - \varepsilon_{i2}, \qquad \text{11-62 ~~(12-32)~~}$$

the variable y_{i1} (assuming initial point $t = 0$ is where our data generating process begins) satisfies the requirements, because ε_{i1} is predetermined with respect to $(\varepsilon_{i3} - \varepsilon_{i2})$. [That is, if we used only the data from periods 1 to 3 constructed as in (~~12-32~~), then the instrumental variables for $(y_{i2} - y_{i1})$ would be $\mathbf{z}_{i(3)}$ where $\mathbf{z}_{i(3)} = (y_{1,1}, y_{2,1}, \ldots, y_{n,1})$ for the n observations.] For the next observation, 66 11-63

$$y_{i4} - y_{i3} = (\mathbf{x}_{i4} - \mathbf{x}_{i3})'\boldsymbol{\beta} + \delta(y_{i3} - y_{i2}) + \varepsilon_{i4} - \varepsilon_{i3}, \qquad \text{~~11-63~~ ~~(12-33)~~}$$

variables y_{i2} and $(y_{i2} - y_{i1})$ are both available. ~~It then becomes a question whether the twice lagged levels or the twice lagged first differences will be preferable. Arellano (1989) and Kiviet (1995) find evidence that suggests that the asymptotic variance of the estimator is smaller with the levels as instruments than with the differences.~~

Based on the preceding paragraph, one might begin to suspect that there is, in fact, rather than a paucity of instruments, a large surplus. In this limited development, we have a choice between differences and levels. Indeed, we could use both and, moreover, in any period after the fourth, not only is y_{i2} available as an instrument, but so also is y_{i1}, and so on. This is the essential observation behind the Arellano, Bover, and Bond (1991, 1995) estimators, which are based on the very large number of candidates for instrumental variables in this panel data model. To begin, with the model in first differences form, for $y_{i3} - y_{i2}$, variable y_{i1} is available. For $y_{i4} - y_{i3}$, y_{i1} and y_{i2} are both available; for $y_{i5} - y_{i4}$, we have y_{i1}, y_{i2}, and y_{i3}, ~~etc~~. Consider, as well, that we have not used the exogenous variables. With strictly exogenous regressors, not only are all lagged values of y_{is} for s previous to $t-1$, but all values of $\mathbf{x}_{it}$ are also available as instruments. For example, for $y_{i4} - y_{i3}$, the candidates are y_{i1}, y_{i2} and $(\mathbf{x}'_{i1}, \mathbf{x}'_{i2}, \ldots, \mathbf{x}'_{iT})$ for all T periods. The number of candidates for instruments is, in fact, potentially huge. [See Ahn and Schmidt (1995) for a very detailed analysis.] If the exogenous variables are only predetermined, rather than strictly exogenous, then only $E[\varepsilon_{it} \mid \mathbf{x}_{i,t}, \mathbf{x}_{i,t-1}, \ldots, \mathbf{x}_{i1}] = 0$, and only vectors $\mathbf{x}_{is}$ from 1 to $t-1$ will be valid instruments in the differenced equation that contains $\varepsilon_{it} - \varepsilon_{i,t-1}$. [See Baltagi and Levin (1986) for an application.] This is hardly a limitation, given that in the end, for a moderate sized model, we may be considering potentially hundreds or thousands of instrumental variables for estimation of what is usually a small handful of parameters.

and so on.

AQ: OK to spell out "etc."?

We now formulate the model in a more familiar form, so we can apply the instrumental variable estimator. In terms of the differenced data, the basic equation is

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + \varepsilon_{it} - \varepsilon_{i,t-1},$$

or

$$\Delta y_{it} = (\Delta \mathbf{x}_{it})'\boldsymbol{\beta} + \delta(\Delta y_{i,t-1}) + \Delta\varepsilon_{it}, \qquad \text{~~(12-34)~~ 11-64 67}$$

where Δ is the first difference operator, $\Delta a_t = a_t - a_{t-1}$ for any time-series variable (or vector) a_t. (It should be noted that a constant term and any time-invariant variables in $\mathbf{x}_{it}$ will fall out of the first differences. We will recover these below after we develop the estimator for $\boldsymbol{\beta}$.) The parameters of the model to be estimated are $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$ and σ_ε^2. For convenience, write the model as

$$\tilde{y}_{it} = \tilde{\mathbf{x}}'_{it}\boldsymbol{\theta} + \tilde{\varepsilon}_{it}$$

We are going to define an instrumental variable estimator along the lines of (~~12-8~~ 8-9) and (~~12-9~~ 8-10). Because our data set is a panel, the counterpart to

$$\mathbf{Z}'\tilde{\mathbf{X}} = \sum_{i=1}^{n} \mathbf{z}_i \tilde{\mathbf{x}}'_i \qquad \text{~~(12-35)~~ 11-65 68}$$

in the cross-section case would seem to be

$$\mathbf{Z}'\tilde{\mathbf{X}} = \sum_{i=1}^{n}\sum_{t=3}^{T} \mathbf{z}_{it}\tilde{\mathbf{x}}'_{it} = \sum_{i=1}^{n} \mathbf{Z}'_i\tilde{\mathbf{X}}'_i \qquad \text{~~(12-36)~~ 11-66 69}$$

$$\tilde{\mathbf{y}}_i = \begin{bmatrix} \Delta y_{i3} \\ \Delta y_{i4} \\ \vdots \\ \Delta y_{iT_i} \end{bmatrix}, \quad \tilde{\mathbf{X}}_i = \begin{bmatrix} \Delta \mathbf{x}'_{i3} & \Delta y_{i2} \\ \Delta \mathbf{x}'_{i4} & \Delta y_{i3} \\ \cdots & \\ \Delta \mathbf{x}'_{iT} & \Delta y_{i,T-1} \end{bmatrix},$$

where there are $(T-2)$ observations (rows) and $K+1$ columns in $\tilde{\mathbf{X}}_i$. There is a complication, however, in that the number of instruments we have defined may vary by period, so the matrix computation in (~~12-36~~) 11-66 appears to sum matrices of different sizes.

Consider an alternative approach. If we used only the first full observations defined in (~~12-34~~) 11-64, then the cross-section version would apply, and the set of instruments $\mathbf{Z}$ in (~~12-35~~) 11-65 with strictly exogenous variables would be the $n \times (1+KT)$ matrix

$$\mathbf{Z}_{(3)} = \begin{bmatrix} y_{1,1}, \mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \cdots \mathbf{x}'_{1,T} \\ y_{2,1}, \mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \cdots \mathbf{x}'_{2,T} \\ \vdots \\ y_{n,1}, \mathbf{x}'_{n,1}, \mathbf{x}'_{n,2}, \cdots \mathbf{x}'_{n,T} \end{bmatrix},$$

and the instrumental variable estimator of (~~12-8~~) 8-9 would be based on

$$\tilde{\mathbf{X}}_{(3)} = \begin{bmatrix} \mathbf{x}'_{1,3} - \mathbf{x}'_{1,2} & y_{1,4} - y_{1,3} \\ \mathbf{x}'_{2,3} - \mathbf{x}'_{2,2} & y_{2,4} - y_{2,3} \\ \vdots & \vdots \\ \mathbf{x}'_{n,3} - \mathbf{x}'_{n,2} & y_{n,4} - y_{n,3} \end{bmatrix} \text{ and } \tilde{\mathbf{y}}_{(3)} = \begin{bmatrix} y_{1,3} - y_{1,2} \\ y_{2,3} - y_{2,2} \\ \vdots \\ y_{n,3} - y_{n,2} \end{bmatrix}.$$

The subscript "(3)" indicates the first observation used for the left-hand side of the equation. Neglecting the other observations, then, we could use these data to form the IV estimator in (~~12-8~~) 8-9, which we label for the moment $\hat{\theta}_{\text{IV}(3)}$. Now, repeat the construction using the next (fourth) observation as the first, and, again, using only a single year of the panel. The data matrices are now

$$\tilde{\mathbf{X}}_{(4)} = \begin{bmatrix} \mathbf{x}'_{1,4} - \mathbf{x}'_{1,3} & y_{1,3} - y_{1,2} \\ \mathbf{x}'_{2,4} - \mathbf{x}'_{2,3} & y_{2,3} - y_{2,2} \\ \vdots & \vdots \\ \mathbf{x}'_{n,4} - \mathbf{x}'_{n,3} & y_{n,3} - y_{n,2} \end{bmatrix}, \tilde{\mathbf{y}}_{(4)} = \begin{bmatrix} y_{1,4} - y_{1,3} \\ y_{2,4} - y_{2,3} \\ \vdots \\ y_{n,4} - y_{n,3} \end{bmatrix}, \text{ and} \qquad 11\text{-}67 \; (\cancel{12\text{-}37})$$

$$\mathbf{Z}_{(4)} = \begin{bmatrix} y_{1,1}, y_{1,2}, \mathbf{x}'_{1,1}, \mathbf{x}'_{1,2}, \cdots \mathbf{x}'_{1,T} \\ y_{2,1}, y_{2,2}, \mathbf{x}'_{2,1}, \mathbf{x}'_{2,2}, \cdots \mathbf{x}'_{2,T} \\ \vdots \\ y_{n,1}, y_{n,2}, \mathbf{x}'_{n,1}, \mathbf{x}'_{n,2}, \cdots \mathbf{x}'_{n,T} \end{bmatrix}$$

AQ: This KT is not in chap. list

and we have a second IV estimator, $\hat{\theta}_{\text{IV}(4)}$, also based on n observations, but, now, $2+KT$ instruments. And so on.

We now need to reconcile the $T-2$ estimators of θ that we have constructed, $\hat{\theta}_{\text{IV}(3)}$, $\hat{\theta}_{\text{IV}(4)}, \ldots, \hat{\theta}_{\text{IV}(T)}$. We faced this problem in Section ~~10.3.2~~ 11.5.8 where we examined Chamberlain's formulation of the fixed effects model. The minimum distance estimator suggested there and used in Carey's (1997) study of hospital costs in Example ~~10.6~~ 11.10 provides a means of efficiently "averaging" the multiple estimators of the parameter vector. We will (as promised) return to the MDE in Chapter ~~15~~ 13. For the present, we consider, instead, **Arellano and Bond's** (1991) [and Arellano and Bover's (1995)] **approach** to this problem. We will collect the full set of estimators in a counterpart to (~~10-26~~) 11-53 and (~~10-27~~) 11-54. First, combine the sets of instruments in a single matrix, $\mathbf{Z}$, where for each individual, we obtain the $(T-2) \times L$ matrix $\mathbf{Z}_i$. The definition of the rows of $\mathbf{Z}_i$ depend on whether the regressors are assumed to be strictly exogenous or predetermined. For

11-74

strictly exogenous variables,

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} & 0 & \cdots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T} \end{bmatrix},$$

71

(~~12-38a~~) 11-68a

and $L = \sum_{i=1}^{T-2}(i + TK) = (T-2)(T-1)/2 + (T-2)TK$. For only predetermined variables, the matrix of instrumental variables is

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \cdots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \mathbf{x}'_{i,3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2}, \ldots \mathbf{x}'_{i,T-1} \end{bmatrix},$$

71

(~~12-38b~~) 11-68b

and $L = \Sigma_{i=1}^{T-2}(i(K+1)+K) = [(T-2)(T-1)/2](1+K)+(T-2)K$. This construction does proliferate instruments (moment conditions, as we will see in Chapter ~~15~~ 13). In the application in Example ~~12.8~~ 11.15, we have a small panel with only $T = 7$ periods, and we fit a model with only $K = 4$ regressors in $\mathbf{x}_{it}$, plus the lagged dependent variable. The strict exogeneity assumption produces a $\mathbf{Z}_i$ matrix that is (5×135) for this case. With only the assumption of predetermined $\mathbf{x}_{it}$, $\mathbf{Z}_i$ collapses slightly to (5×95). For purposes of the illustration, we have used only the two previous observations on $\mathbf{x}_{it}$. This further reduces the matrix to

$$\mathbf{Z}_i = \begin{bmatrix} y_{i,1}, \mathbf{x}'_{i,1}, \mathbf{x}'_{i,2} & 0 & \cdots & 0 \\ 0 & y_{i,1}, y_{i,2}, \mathbf{x}_{i,2}, \mathbf{x}'_{i,3} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & y_{i,1}, y_{i,2}, \ldots, y_{i,T-2}, \mathbf{x}'_{i,T-2}, \mathbf{x}'_{i,T-1} \end{bmatrix},$$

71

(~~12-38c~~) 11-68c

which, with $T = 7$ and $K = 4$, will be (5×55). [Baltagi (2005, Chapter 8) presents some alternative configurations of $\mathbf{Z}_i$ that allow for mixtures of strictly exogenous and predetermined variables.]

Now, we can compute the two-stage least squares estimator in (~~12-8~~) 11-~~9~~10 using our definitions of the data matrices $\mathbf{Z}_i$, $\tilde{\mathbf{X}}_i$, and $\tilde{\mathbf{y}}_i$ and (~~12-36~~) 11-66 69. This will be

$$\hat{\boldsymbol{\theta}}_{\text{IV}} = \left[\left(\sum_{i=1}^{n}\tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\tilde{\mathbf{X}}_i\right)\right]^{-1} \times \left[\left(\sum_{i=1}^{n}\tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\tilde{\mathbf{y}}_i\right)\right]. \qquad \text{(~~12-39~~) 11-69}$$

72

The natural estimator of the asymptotic covariance matrix for the estimator would be

73

$$\text{Est. Asy. Var}\left[\hat{\boldsymbol{\theta}}_{\text{IV}}\right] = \hat{\sigma}^2_{\Delta\varepsilon}\left[\left(\sum_{i=1}^{n}\tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1}. \qquad \text{(~~12-40~~) 11-70}$$

11-75

where

$$\hat{\sigma}^2_{\Delta\varepsilon} = \frac{\sum_{i=1}^{n}\sum_{t=3}^{T}[(y_{it} - y_{i,t-1}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\hat{\boldsymbol{\beta}} - \hat{\delta}(y_{i,t-1} - y_{i,t-2})]^2}{n(T-2)}. \qquad 74 \quad 11\text{-}71 \quad \cancel{(12\text{-}41)}$$

However, this variance estimator is likely to understate the true asymptotic variance because the observations are autocorrelated for one period. Because $(y_{it} - y_{i,t-1}) = \tilde{\mathbf{x}}_{it}'\boldsymbol{\theta} + (\varepsilon_{it} - \varepsilon_{i,t-1}) = \tilde{\mathbf{x}}_{it}'\boldsymbol{\theta} + v_{it}$,

$$\text{Cov}[v_{it}, v_{i,t-1}] = \text{Cov}[v_{it}, v_{i,t+1}] = -\sigma_\varepsilon^2.$$

AU: In (11-75) confirm that capital Ω's are OK

Covariances at longer lags or leads are zero. In the differenced model, though the disturbance covariance matrix is not $\sigma_v^2\mathbf{I}$, it does take a particularly simple form.

$$\text{Cov}\begin{pmatrix} \varepsilon_{i,3} - \varepsilon_{i,2} \\ \varepsilon_{i,4} - \varepsilon_{i,3} \\ \varepsilon_{i,5} - \varepsilon_{i,4} \\ \cdots \\ \varepsilon_{i,T} - \varepsilon_{i,T-1} \end{pmatrix} = \sigma_\varepsilon^2 \begin{bmatrix} 2 & -1 & 0 & \ldots & 0 \\ -1 & 2 & -1 & \ldots & 0 \\ 0 & -1 & 2 & \ldots & 0 \\ \ldots & \ldots & -1 & \ldots & -1 \\ 0 & 0 & \ldots & -1 & 2 \end{bmatrix} = \sigma_\varepsilon^2\boldsymbol{\Omega}_i. \qquad 75 \quad 11\text{-}72 \quad \cancel{(12\text{-}42)}$$

74 11-71

The implication is that the estimator in (~~12-41~~) estimates not σ_ε^2 but $2\sigma_\varepsilon^2$. However, simply dividing the estimator by two does not produce the correct asymptotic covariance
73 matrix because the observations themselves are autocorrelated. As such, the matrix in 9.1
11-70 (~~12-40~~) is inappropriate. (We encountered this issue in Theorem ~~8.1~~ and in Sections ~~8.2.3~~, 9.2.3
~~8.4.3~~, and ~~9.3.2~~.) An appropriate correction can be based on the counterpart to the White
9.4.3 estimator that we developed ~~in Chapter 9,~~ in (~~9~~-3). For simplicity, let 11
11.3.2

$$\hat{\mathbf{A}} = \left[\left(\sum_{i=1}^{n}\tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\tilde{\mathbf{X}}_i\right)\right]^{-1}.$$

Then, a robust covariance matrix that accounts for the autocorrelation would be

$$\hat{\mathbf{A}}\left[\left(\sum_{i=1}^{n}\tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{Z}_i'\tilde{\mathbf{X}}_i\right)\right]\hat{\mathbf{A}}. \qquad 76 \quad 11\text{-}73 \quad \cancel{(12\text{-}43)}$$

11-73 11-72

[One could also replace the $\hat{\mathbf{v}}_i\hat{\mathbf{v}}_i'$ in (~~12-43~~) with $\hat{\sigma}_\varepsilon^2\boldsymbol{\Omega}_i$ in (~~12-42~~) because this is the known expectation.]

11-69 72

It will be useful to digress briefly and examine the estimator in (~~12-39~~). The computations are less formidable than it might appear. Note that the rows of $\mathbf{Z}_i$ in (~~12-38~~a,b,c) are orthogonal. It follows that the matrix

11-68 71

$$\mathbf{F} = \sum_{i=1}^{n}\mathbf{Z}_i'\mathbf{Z}_i$$

73 11-69

in (~~12-39~~) is block-diagonal with $T-2$ blocks. The specific blocks in $\mathbf{F}$ are

$$\mathbf{F}_t = \sum_{i=1}^{n}\mathbf{z}_{it}\mathbf{z}_{it}' = \mathbf{Z}_{(t)}'\mathbf{Z}_{(t)},$$

11-76

71 11-68 for $t = 3, \ldots, T$. Because the number of instruments is different in each period—see (~~12-38~~)—these blocks are of different sizes, say, $(L_t \times L_t)$. The same construction shows that the matrix $\sum_{i=1}^{n} \tilde{\mathbf{X}}_i'\mathbf{Z}_i$ is actually a partitioned matrix of the form

$$\sum_{i=1}^{n} \tilde{\mathbf{X}}_i'\mathbf{Z}_i = \left[\tilde{\mathbf{X}}_{(3)}'\mathbf{Z}_{(3)} \quad \tilde{\mathbf{X}}_{(4)}'\mathbf{Z}_{(4)} \quad \ldots \quad \tilde{\mathbf{X}}_{(T)}'\mathbf{Z}_{(T)}\right],$$

where, again, the matrices are of different sizes; there are $T-2$ rows in each but the number of columns differs. It follows that the inverse matrix, $(\sum_{i=1}^{n} \mathbf{Z}_i'\mathbf{Z}_i)^{-1}$, is also block-diagonal, and that the matrix quadratic form in (~~12-39~~) can be written 72 11-69

$$\left(\sum_{i=1}^{n} \tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n} \tilde{\mathbf{Z}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{Z}_i'\tilde{\mathbf{X}}_i\right) = \sum_{t=3}^{T} \left(\tilde{\mathbf{X}}_{(t)}'\mathbf{Z}_{(t)}\right)\left(\mathbf{Z}_{(t)}'\mathbf{Z}_{(t)}\right)^{-1}\left(\mathbf{Z}_{(t)}'\tilde{\mathbf{X}}_{(t)}\right)$$

$$= \sum_{t=3}^{T} \left(\hat{\tilde{\mathbf{X}}}_{(t)}'\hat{\tilde{\mathbf{X}}}_{(t)}\right)$$

$$= \sum_{t=3}^{T} \mathbf{W}_{(t)},$$

[see (~~12-8~~) and the preceding result]. Continuing in this fashion, we find

8-9

$$\left(\sum_{i=1}^{n} \tilde{\mathbf{X}}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{n} \tilde{\mathbf{Z}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{n} \mathbf{Z}_i'\tilde{\mathbf{y}}_i\right) = \sum_{t=3}^{T} \hat{\tilde{\mathbf{X}}}_{(t)}'\mathbf{y}_{(t)}.$$

From (~~12-9~~), we can see that

8-10

$$\hat{\tilde{\mathbf{X}}}_{(t)}'\mathbf{y}_{(t)} = \left(\hat{\tilde{\mathbf{X}}}_{(t)}'\hat{\tilde{\mathbf{X}}}_{(t)}\right)\hat{\boldsymbol{\theta}}_{\mathrm{IV}(t)}$$

$$= \mathbf{W}_{(t)}\hat{\boldsymbol{\theta}}_{\mathrm{IV}(t)}.$$

73 11-69

Combining the terms constructed thus far, we find that the estimator in (~~12-39~~) can be written in the form

$$\hat{\boldsymbol{\theta}}_{\mathrm{IV}} = \left(\sum_{t=3}^{T} \mathbf{W}_{(t)}\right)^{-1}\left(\sum_{t=3}^{T} \mathbf{W}_{(t)}\hat{\boldsymbol{\theta}}_{\mathrm{IV}(t)}\right)$$

$$= \sum_{t=3}^{T} \mathbf{R}_{(t)}\hat{\boldsymbol{\theta}}_{\mathrm{IV}(t)},$$

where

$$\mathbf{R}_{(t)} = \left(\sum_{t=3}^{T} \mathbf{W}_{(t)}\right)^{-1}\mathbf{W}_{(t)} \text{ and } \sum_{t=3}^{T} \mathbf{R}_{(t)} = \mathbf{I}.$$

In words, we find that, as might be expected, the Arellano and Bond estimator of the parameter vector is a matrix weighted average of the $T-2$ period specific two-stage least squares estimators, where the instruments used in each period may differ. Because the estimator is an average of estimators, a question arises, is it an efficient average—are the weights chosen to produce an efficient estimator? ~~This is precisely the question that~~

~~arose, with a similar estimation problem, in Section 10.3.2, and Example 10.6 where we considered pooling a set of estimators for a generalized regression model.~~ Perhaps not surprisingly, the answer for this $\hat{\boldsymbol{\theta}}$ is no; there is a more efficient set of weights that can be constructed for this model. We will assemble them when we examine the generalized method of moments estimator in Chapter ~~15.~~ 13 11-64

There remains a loose end in the preceding. After (~~12-34~~), it was noted that this treatment discards a constant term and any time invariant variables that appear in the model. The Hausman and Taylor (1981) approach developed in the preceding section suggests a means by which the model could be completed to accommodate this possibility. Expand the basic formulation to include the time-invariant effects, as

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \varepsilon_{it}, \qquad (\cancel{12\text{-}44})$$

AU: Is "IV" OK ← italic?

where $\mathbf{f}_i$ is the set of time-invariant variables and $\boldsymbol{\gamma}$ is the parameter vector yet to be estimated. This model is consistent with the entire preceding development, as the component $\alpha + \mathbf{f}_i'\boldsymbol{\gamma}$ would have fallen out of the differenced equation along with c_i at the first step at (~~12-30~~). Having developed a consistent estimator for $\boldsymbol{\theta} = (\boldsymbol{\beta}', \delta)'$, we now turn to estimation of $(\alpha, \boldsymbol{\gamma}')'$. The residuals from the IV regression (~~12-39~~), 11-69 72

13 11-60

$$w_{it} = \mathbf{x}_{it}'\hat{\boldsymbol{\beta}}_{IV} - \hat{\delta}_{IV} y_{i,t-1}$$

are pointwise consistent estimators of

$$\omega_{it} = \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \varepsilon_{it}.$$

Thus, the group means of the residuals can form the basis of a second-step regression;

$$\overline{w}_i = \alpha + \mathbf{f}_i'\boldsymbol{\gamma} + c_i + \overline{\varepsilon}_i + \eta_i \qquad (\cancel{12\text{-}45}) \text{ 11-75 } 76 \text{ 74}$$

where $\eta_i = (\overline{w}_i. - \overline{\omega}_i.)$ is the estimation error that converges to zero as $\hat{\boldsymbol{\theta}}$ converges to $\boldsymbol{\theta}$. The implication would seem to be that we can now linearly regress these group mean residuals on a constant and the time invariant variables $\mathbf{f}_i$ to estimate α and $\boldsymbol{\gamma}$. The flaw in the strategy, however, is that the initial assumptions of the model do not state that c_i is uncorrelated with the other variables in the model, including the implicit time invariant terms, $\mathbf{f}_i$. Therefore, least squares is not a usable estimator here unless the random effects model is assumed, which we specifically sought to avoid at the outset. As in Hausman and Taylor's treatment, there is a workable strategy if it can be assumed that there are some variables in the model, including possibly some among the $\mathbf{f}_i$ as well as others among $\mathbf{x}_{it}$ that are uncorrelated with c_i and ε_{it}. These are the $\mathbf{z}_1$ and $\mathbf{x}_1$ in the Hausman and Taylor estimator (see Step 2 in the development of the preceding section). Assuming that these variables are available—this is an identification assumption that must be added to the model—then we do have a usable instrumental variable estimator, using as instruments the constant term (1), any variables in $\mathbf{f}_i$ that are uncorrelated with the latent effects or the disturbances (call this $\mathbf{f}_{i1}$), and the group means of any variables in $\mathbf{x}_{it}$ that are also exogenous. There must be enough of these to provide a sufficiently large set of instruments to fit all the parameters in (~~12-45~~). 11-76 74 76

This is, once again, the same identification we saw in step 2 of the Hausman and Taylor estimator, K_1, the number of exogenous variables in $\mathbf{x}_{it}$ must be at least as large as L_2, which is the number of endogenous variables in $\mathbf{f}_i$. With all this in place, we then have the instrumental variable estimator in which the dependent variable is $\overline{w}_i.$, the right-hand-side variables are $(1, \mathbf{f}_i)$, and the instrumental variables are $(1, \mathbf{f}_{i1}, \overline{\mathbf{x}}_{i1}.)$.

75 11-73 There is yet another direction that we might extend this estimation method. In (~~12-43~~), we have implicitly allowed a more general covariance matrix to govern the generation of the disturbances ε_{it} and computed a robust covariance matrix for the simple IV estimator. We could take this a step further and look for a more efficient estimator. As a library of recent studies have shown, panel data sets are rich in information that allows the analyst to specify highly general models and to exploit the implied relationships among the variables to construct much more efficient generalized method of moments (GMM) estimators. [See, in particular, Arellano and Bover (1995) and Blundell and Bond (1998).] We will return to this development in Chapter ~~15~~. has 13

11-15 **Example ~~12.8~~ Dynamic Labor Supply Equation**

In Example ~~12.3~~, we used instrumental variables fit a labor supply equation, 8.5

$$Wks_{it} = \gamma_1 + \gamma_2 \ln Wage_{it} + \gamma_3 Ed_i + \gamma_4 Union_{it} + \gamma_5 Fem_i + u_{it}.$$

To illustrate the computations of this section, we will extend this model as follows:

$$Wks_{it} = \beta_1 \ln Wage_{it} + \beta_2 Union_{it} + \beta_3 Occ_{it} + \beta_4 Exp_{it} + \delta\, Wks_{i,t-1} + \alpha + \gamma_1 Ed_i + \gamma_2 Fem_i + c_i + \varepsilon_{it}.$$

(We have rearranged the variables and parameter names to conform to the notation in this section.) We note, in theoretical terms, as suggested in the earlier example, it may not be appropriate to treat $\ln Wage_{it}$ as uncorrelated with ε_{it} or c_i. However, we will be analyzing the model in first differences. It may well be appropriate to treat changes in wages as exogenous. That would depend on the theoretical underpinnings of the model. We will treat the variable as predetermined here, and proceed. There are two time-invariant variables in the model, Fem_i, which is clearly exogenous, and Ed_i, which might be endogenous. The identification requirement for estimation of $(\alpha, \gamma_1, \gamma_2)$ is met by the presence of three exogenous variables, $Union_{it}$, Occ_{it}, and Exp_{it} ($K_1 = 3$ and $L_2 = 1$).

The differenced equation analyzed at the first step is

$$\Delta Wks_{it} = \beta_1 \Delta \ln Wage_{it} + \beta_2 \Delta Union_{it} + \beta_3 \Delta Occ_{it} + \beta_4 \Delta Exp_{it} + \delta \Delta Wks_{i,t-1} + \varepsilon_{it}.$$

one 72 11-69

FN 26 76 11-73 TB 11.12 We estimated the parameters and the asymptotic covariance matrix according to (~~12-39~~) and (~~12-43~~). For specification of the instrumental variables, we used the ~~two~~ one previous observation~~s~~ on $\mathbf{x}_{it}$, as shown in the text.[~~18~~ 26] Table ~~12.4~~ 11.12 presents the computations with several other inconsistent estimators.

The various estimates are quite far apart. In the absence of the common effects (and autocorrelation of the disturbances), all five estimators shown would be consistent. Given the very wide disparities, one might suspect that common effects are an important feature of the data. The second standard errors given with the IV estimates are based on the uncorrected matrix in (~~12-40~~ 11-70) with $\hat{\sigma}^2_{\Delta\varepsilon}$ in (~~12-41~~ 11-71) divided by two. We found the estimator to be quite volatile, as can be seen in the table. The estimator is also very sensitive to the choice of instruments that comprise $\mathbf{Z}_i$. Using (~~12-38a~~ 11-68a) instead of (~~12-38b~~ 11-68b) produces wild swings in the estimates and, in fact, produces implausible results. One possible explanation in this particular example is that the instrumental variables we are using are dummy variables that have relatively little variation over time.

~~18~~ 26 This estimator and the GMM estimators in Chapter ~~15~~ 13 are built into some contemporary computer programs, including *NLOGIT* and *Stata*. Many researchers use Gauss programs that are distributed by M. Arellano, http://www.cemfi.es/%7Earellano/#dpd, or program the calculations themselves using *MatLab* or *R*. We have programmed the matrix computations directly for this application using the matrix package in *NLOGIT*. URL

11-79

11.~~13~~ 12

TABLE ~~12.4~~ Estimated Dynamic Panel Data Model Using Arellano and Bond's Estimator

	OLS, Full Eqn.		*OLS, Differenced*		*IV, Differenced*		*Random Effects*		*Fixed Effects*	
Variable	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***	***Estimate***	***Std. Err.***
ln *Wage*	0.2966	0.2052	−0.1100	0.4565	−1.1402	0.2639 0.8768	0.2281	0.2405	0.5886	0.4790
Union	−1.2945	0.1713	1.1640	0.4222	2.7089	0.3684 0.8676	−1.4104	0.2199	0.1444	0.4369
Occ	0.4163	0.2005	0.8142	0.3924	2.2808	1.3105 0.7220	0.5191	0.2484	1.0064	0.4030
Exp	−0.0295	0.00728	−0.0742	0.0975	−0.0208	0.1126 0.1104	−0.0353	0.01021	−0.1683	0.05954
Wks_{t-1}	0.3804	0.01477	−0.3527	0.01609	0.1304	0.04760 0.02131	0.2100	0.01511	0.0148	0.01705
Constant	28.918	1.4490			−0.4110	0.3364	37.461	1.6778		
Ed	−0.0690	0.03703			0.0321	0.02587	−0.0657	0.04988		
Fem	−0.8607	0.2544			−0.0122	0.1554	−1.1463	0.3513		
Sample	$t = 2 - 7$ $n = 595$		$t = 3 - 7$ $n = 595$		$t = 3 - 7$; $n = 595$ Means used $t = 7$		$t = 2 - 7$ $n = 595$		$t = 2 - 7$ $n = 595$	

11-80

~~The constant terms and coefficients on the lagged dependent variable are country specific. The remaining coefficients are treated as random, normally distributed, with means β_k and unrestricted variances. They are modeled as uncorrelated, so in (9-49), Γ is diagonal. [In the form of the model in Section 9.8.2, the diagonal matrix would be Λ in (9-55).] The model was estimated using a modification of the Hildreth–Houck–Swamy method described in Section 9.8.1.~~

11.8.4 ~~9.8.6~~ NONSTATIONARY DATA AND PANEL DATA MODELS

Some of the discussion thus far (and to follow) focuses on "small T" statistical results. Panels are taken to contain a fixed and small T observations on a large n individual units. Recent research using cross-country data sets such as the Penn World Tables URL (http://pwt.econ.upenn.edu/php_site/pwt_index.php), which now include data on nearly 200 countries for well over 50 years, have begun to analyze panels with T sufficiently large that the time-series properties of the data become an important consideration. In particular, the recognition and accommodation of nonstationarity that is now a standard part of single time-series analyses (as in Chapter ~~22~~ 23) are now seen to be appropriate for large scale cross-country studies, such as income growth studies based on the Penn World Tables, cross-country studies of health care expenditure, and analyses of purchasing power parity.

AU: OK to spell out "etc."?

The analysis of long panels, such as in the growth and convergence literature, typically involves dynamic models, such as

$$y_{it} = \alpha_i + \gamma_i y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta}_i + \varepsilon_{it}. \qquad (\text{11-77}) \; \cancel{(11\text{-}75)} \; \cancel{(9\text{-}66)}$$

In single time-series analysis involving low-frequency macroeconomic flow data such as income, consumption, investment, the current account deficit, etc., it has long been recognized that estimated regression relations can be distorted by nonstationarity in the data. What appear to be persistent and strong regression relationships can be entirely spurious and due to underlying characteristics of the time-series processes rather than actual connections among the variables. Hypothesis tests about long-run effects will be considerably distorted by unit roots in the data. It has become evident that the same influences, with the same deletarious effects, will be found in long panel data sets. The panel data application is further complicated by the possible heterogeneity of the parameters. The coefficients of interest in many cross-country studies are the lagged effects, such as γ_i in ~~(9-66)~~ 11-77 ~~11-75~~, and it is precisely here that the received results on nonstationary data have revealed the problems of estimation and inference. Valid tests for unit roots in panel data have been proposed in many studies. Three that are frequently cited are Levin and Lin (1992), Im, Pesaran, and Shin (2003) and Maddala and Wu (1999).

and so on

There have been numerous empirical applications of time series methods for nonstationary data in panel data settings, including Frankel and Rose's (1996) and Pedroni's (2001) studies of purchasing power parity, Fleissig and Strauss (1997) on real wage stationarity, Culver and Papell (1997) on inflation, Wu (2000) on the current account balance, McCoskey and Selden (1998) on health care expenditure, Sala-i-Martin (1996) on growth and convergence, McCoskey and Kao (1999) on urbanization and production, and Coakely et al. (1996) on savings and investment.[32] 34

~~34~~ ~~32~~ ~~A more complete~~ An extensive enumeration appears in Baltagi (2005, Chapter 12).

Move into body of text

11-81

11-~~75~~ 77

A subtle problem arises in obtaining results useful for characterizing the properties of estimators of the model in ~~(9-66)~~. The asymptotic results based on large n and large T are not necessarily obtainable simultaneously, and great care is needed in deriving the asymptotic behavior of useful statistics. Phillips and Moon (1999, 2000) are standard references on the subject.

We will return to the topic of nonstationary data in Chapter ~~22~~ 23. This is an emerging literature, most of which is well beyond the level of this text. We will rely on the several detailed received surveys, such as Bannerjee (1999), Smith (2000), and Baltagi and Kao (2000) to fill in the details.

9.9 CONSISTENT ESTIMATION OF DYNAMIC PANEL DATA MODELS

As prelude to the further developments of Chapters 12 and 13, we return to a homogeneous dynamic panel data model,

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + c_i + \varepsilon_{it}, \tag{9-67}$$

where c_i is, as in the preceding sections of this chapter, individual unmeasured heterogeneity, that may or may not be correlated with $\mathbf{x}_{it}$. We consider methods of estimation for this model when T is fixed and relatively small, and n may be large and increasing.

Pooled OLS is obviously inconsistent. Rewrite (9-67) as

$$y_{it} = \gamma y_{i,t-1} + \mathbf{x}_{it}'\boldsymbol{\beta} + w_{it}. \tag{9-68}$$

The disturbance in this pooled regression may be correlated with $\mathbf{x}_{it}$, but either way, it is surely correlated with $y_{i,t-1}$. By substitution,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] = \sigma_c^2 + \gamma\text{Cov}[y_{i,t-2}, (c_i + \varepsilon_{it})],$$

and so on. By repeated substitution, it can be seen that for $|\gamma| < 1$ and moderately large T,

$$\text{Cov}[y_{i,t-1}, (c_i + \varepsilon_{it})] \approx \sigma_c^2/(1-\gamma). \tag{9-69}$$

[It is useful to obtain this result from a different direction. If the stochastic process that is generating (y_{it}, c_i) is *stationary*, then $\text{Cov}[y_{i,t-1}, c_i] = \text{Cov}[y_{i,t-2}, c_i]$, from which we would obtain (9-69) directly. The assumption $|\gamma| < 1$ would be required for stationarity. We will return to this subject in Chapters 20 and 21.] Consequently, OLS and GLS are inconsistent. The fixed effects approach does not solve the problem either. Taking deviations from individual means, we have

$$y_{it} - \bar{y}_{i.} = (\mathbf{x}_{it} - \bar{\mathbf{x}}_{i.})'\boldsymbol{\beta} + \gamma(y_{i,t-1} - \bar{y}_{i.}) + (\varepsilon_{it} - \bar{\varepsilon}_{i.}).$$

Anderson and Hsiao (1981, 1982) show that

$$\begin{aligned} Cov[(y_{it} - \bar{y}_{i.}), (\varepsilon_{it} - \bar{\varepsilon}_{i.})] &\approx \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2}\left[\frac{(T-1) - T\gamma + \gamma^T}{T}\right] \\ &= \frac{-\sigma_\varepsilon^2}{T(1-\gamma)^2}\left[(1-\gamma) - \frac{1-\gamma^T}{T}\right]. \end{aligned} \tag{9-70}$$

11-82

~~so the estimator of the asymptotic covariance matrix for m is~~

$$\text{Est. Asy. Var}[\mathbf{m}] = \mathbf{G}\mathbf{V}_b^*\mathbf{G}'.$$

~~One might be tempted to treat λ_i as a constant, in which case only the first term in the quadratic form would appear and the computation would amount simply to multiplying the asymptotic standard errors for $\mathbf{b}_p$ by λ_i. This approximation would leave the asymptotic t ratios unchanged, whereas making the full correction will change the entire covariance matrix. The approximation will generally lead to an understatement of the correct standard errors.~~

11.9 NONLINEAR REGRESSION WITH PANEL DATA ~~11.7 PANEL DATA APPLICATIONS~~

The extension of the panel data models ~~of Chapter 9~~ to the nonlinear regression case is, perhaps surprisingly, not at all straightforward. Thus far, to accommodate the nonlinear model, we have generally applied familiar results to the linearized regression. This approach will carry forward to the case of clustered data. (See Section ~~9.3.3~~ 11.3.3) Unfortunately, this will not work with the standard panel data methods. The nonlinear regression will be the first of numerous panel data applications that we will consider in which the widsom of the linear regression model cannot be extended to the more general framework.

11.9.1 A ROBUST COVARIANCE MATRIX FOR NONLINEAR LEAST SQUARES

The counterpart to (~~9~~11-3) or (~~9~~11-4) would simply replace $\mathbf{X}_i$ with $\hat{\mathbf{X}}_i^0$ where the rows are the pseudoregressors for cluster i as defined in (~~11-9~~ 7-12) and "^" indicates that it is computed using the nonlinear least squares estimates of the parameters.

Example ~~11.10~~ 11.16 Health Care Utilization

The recent literature in health economics includes many studies of health care utilization. A common measure of the dependent variable of interest is a count of the number of encounters with the health care system, either through visits to a physician or to a hospital. These counts of occurrences are usually studied with the Poisson regression model described in ~~Example 11.9~~ Section 19.2. The nonlinear regression model is

$$E[y_i \mid \mathbf{x}_i] = \exp(\mathbf{x}_i'\beta).$$

A recent study in this genre is "Incentive Effects in the Demand for Health Care: A Bivariate Panel Count Data Estimation" by Riphahn, Wambach, and Million (2003). The authors were interested in counts of physician visits and hospital visits. In this application, they were particularly interested in the impact that the presence of private insurance had on the utilization counts of interest, ~~i.e.~~ that is, whether the data contain evidence of moral hazard.

The raw data are published on the *Journal of Applied Econometrics* data archive website. URL The URL for the data file is http://qed.econ.queensu.ca/jae/2003-v18.4/riphahn-wambach-million/. The variables in the data file are listed in Appendix Table ~~F11.1~~ F7.1. The sample is an unbalanced panel of 7,293 households, the German Socioeconomic Panel data set. The number of observations varies from one to seven (1,525; 1,079; 825; 926; 1,311; 1,000; 887) with a total number of observations of 27,326. We will use these data in several examples here and later in the book.

The following model uses a simple specification for the count of number of visits to the physican in the observation year,

$$\mathbf{x}_{it} = (1, age_{it}, educ_{it}, income_{it}, kids_{it})$$

11-83

~~11.17~~

11.13 **TABLE ~~11.8~~** Nonlinear Least Squares Estimates of a Utilization Equation

Begin NLSQ iterations. Linearized regression.
Iteration = 1; Sum of squares = 1014865.00; Gradient = 156281.794
Iteration = 2; Sum of squares = 8995221.17; Gradient = 8131951.67
Iteration = 3; Sum of squares = 1757006.18; Gradient = 897066.012
Iteration = 4; Sum of squares = 930876.806; Gradient = 73036.2457
Iteration = 5; Sum of squares = 860068.332; Gradient = 2430.80472
Iteration = 6; Sum of squares = 857614.333; Gradient = 12.8270683
Iteration = 7; Sum of squares = 857600.927; Gradient = 0.411851239E-01
Iteration = 8; Sum of squares = 857600.883; Gradient = 0.190628165E-03
Iteration = 9; Sum of squares = 857600.883; Gradient = 0.904650588E-06
Iteration = 10; Sum of squares = 857600.883; Gradient = 0.430441193E-08
Iteration = 11; Sum of squares = 857600.883; Gradient = 0.204875467E-10
Convergence achieved

Variable	*Estimate*	*Standard Error*	*Robust Standard Error*
Constant	0.9801	0.08927	~~1.01613~~ 0.12522
Age	0.01873	0.001053	~~0.01107~~ 0.00142
Education	−0.03613	0.005732	~~0.06881~~ 0.00780
Income	−0.5911	0.07173	~~0.7182~~ 0.09702
Kids	−0.1692	0.02642	~~0.2735~~ 0.03330

TB 11.13

11.13

Table ~~11.8~~ details the nonlinear least squares iterations and the results. The convergence criterion for the iterations is $\mathbf{e}^{0\prime}\mathbf{X}^0(\mathbf{X}^{0\prime}\mathbf{X}^0)\mathbf{X}^{0\prime}\mathbf{e}^0 < 10^{-10}$. Although this requires 11 iterations, the function actually reaches the minimum in seven. The estimates of the asymptotic standard errors are computed using the conventional method, $s^2(\hat{\mathbf{X}}^{0\prime}\hat{\mathbf{X}}^0)^{-1}$ and then by the cluster correction in (8-4). The corrected standard errors are considerably larger, as might be expected given that these are a panel data set.

11

11.~~7~~9.2 FIXED EFFECTS

The nonlinear panel data regression model would appear

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + \varepsilon_{it}, t = 1, \ldots, T_i, i = 1, \ldots, n.$$

Consider a model with latent heterogeneity, c_i. An ambiguity immediately emerges; how should heterogeneity enter the model. Building on the linear model, an additive term might seem natural, as in

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}) + c_i + \varepsilon_{it}, t = 1, \ldots, T_i, i = 1, \ldots, n. \quad \text{(~~11-32~~) 11-~~76~~ 78}$$

But we can see in the previous application that this is likely to be inappropriate. The loglinear model of the previous section is constrained to ensure that $E[y_{it} \mid \mathbf{x}_{it}]$ is positive. But an additive random term c_i as in (~~11-32~~) 11-~~76~~78 could subvert this; unless the range of c_i is restricted, the conditional mean could be negative. The most common application of nonlinear models is the **index function model,**

KT

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + c_i) + \varepsilon_{it}.$$

AU: KT "index function model" is not in chap list

11-84

This is the natural extension of the linear model, but only in the appearance of the conditional mean. Neither the fixed effects nor the random effects model can be estimated as they were in the linear case ~~in Chapter 9~~.

Consider the fixed effects model first. We would write this as

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) + \varepsilon_{it},$$

where the parameters to be estimated are $\boldsymbol{\beta}$ and $\alpha_i, i = 1, \ldots, n$. Transforming the data to deviations from group means does not remove the fixed effects from the model. For example,

$$y_{it} - \bar{y}_{i.} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) - \frac{1}{T_i}\sum_{s=1}^{T_i} h(\mathbf{x}_{is}'\boldsymbol{\beta} + \alpha_i), \qquad \text{79 } \text{11-77 } \cancel{(11\text{-}33)}$$

which does not simplify things at all. Transforming the regressors to deviations is likewise pointless. To estimate the parameters, it is necessary to minimize the sum of squares with respect to all $n + K$ parameters simultaneously. Because the number of dummy variable coefficients can be huge—the preceding example is based on a data set with 7,293 groups—this can be a difficult or impractical computation. A method of maximizing a function (such as the negative of the sum of squares) that contains an unlimited number of dummy variable coefficients is shown in Chapter ~~23~~. As we will examine later in the book, the difficulty with nonlinear models that contain large numbers of dummy variable coefficients is not necessarily the practical one of computing the estimates. That is generally a solvable problem. The difficulty with such models is an intriguing phenomenon known as the **incidental parameters problem.** In most (not all, as we shall find) nonlinear panel data models that contain n dummy variable coefficients, such as the one in (~~11-33~~ 11-77), as a consequence of the fact that the number of parameters increases with the number of individuals in the sample, the estimator of $\boldsymbol{\beta}$ is biased and inconsistent, to a degree that is $O(1/T)$. Because T is only 7 or less in our application, this would seem to be a case in point.

17

KT

AU: KT "incidental parameters problem" is not in chap. list.

79

11-77

Example 11.17 Exponential Model with Fixed Effects

The exponential model of the preceding example is actually one of a small handful of known special cases in which it is possible to "condition" out the dummy variables. Consider the sum of squared residuals,

$$S_n = \frac{1}{2}\sum_{i=1}^{n}\sum_{t=1}^{T_i}[y_{it} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i)]^2.$$

The first order condition for minimizing S_n with respect to α_i is

$$\frac{\partial S_n}{\partial \alpha_i} = \sum_{t=1}^{T_i} -[y_{it} - \exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i)]\exp(\mathbf{x}_{it}'\boldsymbol{\beta} + \alpha_i) = 0. \qquad \text{80 } \text{11-78 } \cancel{(11\text{-}34)}$$

Let $\gamma_i = \exp(\alpha_i)$. Then, an equivalent necessary condition would be

$$\frac{\partial S_n}{\partial \gamma_i} = \sum_{t=1}^{T_i} -[y_{it} - \gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})][\gamma_i \exp(\mathbf{x}_{it}'\boldsymbol{\beta})] = 0,$$

11-85

or

$$\gamma_i \sum_{t=1}^{T_i} [y_{it} \exp(\mathbf{x}_{it}'\boldsymbol{\beta})] = \gamma_i^2 \sum_{t=1}^{T_i} [\exp(\mathbf{x}_{it}'\boldsymbol{\beta})]^2.$$

Obviously, if we can solve the equation for γ_i, we can obtain $\alpha_i = \ln \gamma_i$. The preceding equation can, indeed, be solved for γ_i, at least conditionally. At the minimum of the sum of squares, it will be true that

$$\hat{\gamma}_i = \frac{\sum_{t=1}^{T_i} y_{it} \exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})}{\sum_{t=1}^{T_i} [\exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})]^2}. \qquad \text{11-79 } \text{81 } (\cancel{11\text{-}35})$$

11-~~78~~ 80 11-79 81

We can now insert (~~11-35~~) into (~~11-34~~) to eliminate α_i. (This is a counterpart to taking deviations from means in the linear case. As noted, this is possible only for a very few special models—this happens to be one of them. The process is also known as "concentrating out" the parameters γ_i. Note that at the solution, $\hat{\gamma}_i$, is obtained as the slope in a regression without a constant term of y_{it} on $\hat{z}_{it} = \exp(\mathbf{x}_{it}'\hat{\boldsymbol{\beta}})$ using T_i observations.) The result in (~~11-35~~) 81 11-79 must hold at the solution. Thus, (~~11-35~~) inserted in (11-34) restricts the search for $\boldsymbol{\beta}$ to those values that satisfy the restrictions in (~~11-35~~). The resulting sum of squares function is now a function only of the data and $\boldsymbol{\beta}$, and can be minimized with respect to this vector of K parameters. With the estimate of $\boldsymbol{\beta}$ in hand, α_i can be estimated using the log of the result in (~~11-35~~) (which is positive by construction).

81 11-79 81 11-79 81 11-79

81 11-79

The preceding example presents a mixed picture for the fixed effects model. In nonlinear cases, two problems emerge that were not present earlier, the practical one of actually computing the dummy variable parameters and the theoretical incidental parameters problem that we have yet to investigate, but which promises to be a significant shortcoming of the fixed effects model. We also note, we have focused on a particular form of the model, the "single index" function, in which the conditional mean is a nonlinear function of a linear function. In more general cases, it may be unclear how the unobserved heterogeneity should enter the regression function.

11.~~7~~9.3 RANDOM EFFECTS

The random effects nonlinear model also presents complications both for specification and for estimation. We might begin with a general model

$$y_{it} = h(\mathbf{x}_{it}, \boldsymbol{\beta}, u_i) + \varepsilon_{it}. \qquad \text{11-80 } \text{82 } (\cancel{11\text{-}36})$$

The "random effects" assumption would be, as usual, mean independence,

$$E[u_i \mid \mathbf{X}_i] = 0.$$

Unlike the linear model, the nonlinear regression cannot be consistently estimated by (nonlinear) least squares. In practical terms, we can see why in (~~11-5~~)–(~~11-7~~). In the linearized regression, the conditional mean at the expansion point $\boldsymbol{\beta}^0$ [see (~~11-5~~)] as well as the pseudoregressors are both functions of the unobserved u_i. This is true in the general case (~~11-36~~) as well as the simpler case of a single index model, 7-28 7-30 7-28

81 11-79

$$y_{it} = h(\mathbf{x}_{it}'\boldsymbol{\beta} + u_i) + \varepsilon_{it}. \qquad \text{83 } (11\text{-}81)$$

Thus, it is not possible to compute the iterations for nonlinear least squares. As in the fixed effects case, neither deviations from group means nor first differences solves the problem. Ignoring the problem—that is, simply computing the nonlinear least squares estimator without accounting for heterogeneity—does not produce a consistent estimator, for the same reasons. In general, the benign effect of latent heterogeneity (random

11-86

effects) that we observe in the linear model only carries over to a very few nonlinear models and, unfortunately, this is not one of them.

An approach that can be used, albeit at the cost of an additional assumption, is the simulation based estimator in Section 9.8.2. If we assume that u_i is normally distributed with mean zero and variance σ_u^2, then an analog to (9-57) for least squares would be

$$S_n^S = \frac{1}{2}\sum_{i=1}^{n}\frac{1}{R}\sum_{r=1}^{R}\sum_{t=1}^{T_i}\left[y_{it} - h(\mathbf{x}_{it}, \boldsymbol{\beta}, \sigma_u v_{ir})\right]^2 \qquad \text{11-81} \ (\cancel{\text{11-37}})$$

The approach from this point would be the same as in Section 9.8.2. [If it is further assumed that ε_{it} is normally distributed, then after incorporating σ_ε^2 in the criterion function, (~~11-37~~) would actually be the extension of (9-57) to a nonlinear regression function. The random parameter vector there is specialized here to a nonrandom constant term.]

11-81

11.8 SUMMARY AND CONCLUSIONS

In this chapter, we extended the regression model to a form that allows nonlinearity in the parameters in the regression function. The results for interpretation, estimation, and hypothesis testing are quite similar to those for the linear model. The two crucial differences between the two models are, first, the more involved estimation procedures needed for the nonlinear model and, second, the ambiguity of the interpretation of the coefficients in the nonlinear model (because the derivatives of the regression are often nonconstant, in contrast to those in the linear model). Finally, we added an additional level of generality to the model. Two-step nonlinear least squares is suggested as a method of allowing a model to be fit while including functions of previously estimated parameters.

Key Terms and Concepts

- Box–Cox transformation
- Delta method
- GMM estimator
- Identification condition
- Incidental parameters problem
- Index function model
- Indirect utility function
- Iteration
- Jacobian
- Linearized regression model
- Lagrange multiplier test
- Logit model
- Loglinear model
- Nonlinear regression model
- Normalization
- Nonlinear least squares
- Orthogonality condition
- Overidentifying restrictions
- Pseudoregressors
- Roy's identity
- Semiparametric
- Starting values
- Two-step estimation
- Wald test

Exercises

1. Describe how to obtain nonlinear least squares estimates of the parameters of the model $y = \alpha x^\beta + \varepsilon$.
2. Verify the following differential equation, which applies to the Box–Cox transformation:

$$\frac{d^i x^{(\lambda)}}{d\lambda^i} = \left(\frac{1}{\lambda}\right)\left[x^\lambda(\ln x)^i - \frac{i d^{i-1}x^{(\lambda)}}{d\lambda^{i-1}}\right]. \qquad \textbf{(11-38)}$$

The problem of computing partial effects in a random effects model such as (11-81) 83 is that when $E[y_{it}|\mathbf{x}_{it},u_i]$ is given by (11-81), (83)

$$\frac{\partial E[y_{it}\,|\,\mathbf{x}_{it}'\boldsymbol{\beta}+u_i]}{\partial \mathbf{x}_{it}} = [h'(\mathbf{x}_{it}'\boldsymbol{\beta}+u_i)]\boldsymbol{\beta}$$

is a function of the unobservable u_i. Two ways to proceed from here are the fixed effects approach of the previous section and a random effects approach. The fixed effects approach is feasible, but may be hindered by the incidental parameters problem noted earlier. A random effects approach might be preferable, but comes at the price of assuming that $\mathbf{x}_{it}$ and u_i are uncorrelated, which may be unreasonable. Papke and Wooldridge (2008) examined several cases, and proposed the Mundlak approach of projecting u_i on the group means of $\mathbf{x}_{it}$. The working specification of the model is then

$$E^*[y_{it}|\mathbf{x}_{it}, \bar{\mathbf{x}}_i, v_i] = h(\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha+\bar{\mathbf{x}}_i'\boldsymbol{\theta}+v_i).$$

This leaves the practical problem of how to compute the estimates of the parameters and how to compute the partial effects. Papke and Wooldridge (2008) suggest a useful result if it can be assumed that v_i is normally distributed with mean zero and variance σ_v^2. In that case,

$$E[y_{it}\,|\,\mathbf{x}_{it},\bar{\mathbf{x}}] = E_{v_i}E[y_{it}\,|\,\mathbf{x}_{it},\bar{\mathbf{x}},v_i] = h\left(\frac{\mathbf{x}_{it}'\boldsymbol{\beta}+\alpha+\bar{\mathbf{x}}_i'\boldsymbol{\theta}}{\sqrt{1+\sigma_v^2}}\right) = h\left(\mathbf{x}_{it}'\boldsymbol{\beta}_v+\alpha_v+\bar{\mathbf{x}}_i'\boldsymbol{\theta}_v\right).$$

The implication is that nonlinear least squares regression will estimate the scaled coefficients, after which the average partial effect can be estimated for a particular value of the covariates, $\mathbf{x}_0$, with

$$\hat{\Delta}(\mathbf{x}_0) = \frac{1}{n}\sum_{i=1}^{n} h'\left(\mathbf{x}_0'\hat{\boldsymbol{\beta}}_v+\hat{\alpha}_v+\bar{\mathbf{x}}_i'\hat{\boldsymbol{\theta}}_v\right)\hat{\boldsymbol{\beta}}_v.$$

They applied the technique to a case of test pass rates, which are a fraction bounded by zero and one. Loudermilk (2007) is another application with an extension to a dynamic model.