Chapter 6. Panel Data

Joan Llull

Quantitative & Statistical Methods II
Master in Economics of Public Policy
Barcelona GSE

I. Introduction

This chapter is aimed to provide an introduction to general panel data techniques. These techniques are applicable both to the treatment effects literature, and, more generally, to more classic econometric problems. For this reason, the notation and examples we will follow are for generic (potentially continuous) regressors, of which a treatment dummy is a particular case.

The term *panel data* is used in econometrics to refer to data sets with repeated observations for a given cross-section of units. Units can be persons, households, firms, countries,..., and observations can be repeated over time or other dimensions (e.g. twins: for each unit —a family– we have two observations —the twins). It is different from repeated cross-sections in that the same (identifiable) units are observed at different points in time. In this chapter, we generally use individuals, denoted by i, to refer to units, and time, denoted by t, to refer to the different observations we have for a unit.

The main advantages of panel data are two. First, they allow us to deal with permanent *unobserved heterogeneity*, i.e. potentially relevant variables that are fixed over time, but unobserved by the econometrician. Second, it allows us to analyze dynamic responses and error components. The latter consists of analyzing separate pieces of the unobserved term (e.g. calculate the variance of the permanent and the transitory parts of the unobservable). The former allow for feedback from past variables into future outcomes.

Over this chapter, we use the following example. Consider the estimation of the demand for cigarettes. Let C_{it} denote the number of cigarettes consumed per day by individual i in year t, let P_{it} denote the price at which this person is exposed, and let Y_{it} denote income level. Holding income level constant might not be enough to capture systematic differences across individuals. Different individuals may have different propensity for being a smoker, which might correlate both with the prices they are exposed to and their income. Thus, the final demand for cigarettes can be expressed as:

$$\ln C_{it} = \beta_0 + \beta_1 \ln P_{it} + \beta_2 \ln Y_{it} + \eta_i + v_{it}, \tag{1}$$

where $\eta_i + v_{it}$ is a random variable that is not observed by the econometrician, in which eta_i is constant for the different observations of a given individual, and v_{it} is i.i.d. across individuals and over time. We will consider different methods depending on the assumptions about η_i.

II. Static Models

We initially consider the static panel data model. Let us first introduce some general notation. Our model is rewritten as:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + (\eta_i + v_{it}), \tag{2}$$

where y_{it} and $\boldsymbol{x}_{it}$ are observed by the econometrician, and η_i and v_{it} are not observed. Sub-indexes are intentionally used both for random variables and for observations to distinguish variables that are individual and time-varying, constant individual-specific. Let $\{y_{it}, \boldsymbol{x}_{it}\}_{i=1,...,N}^{t=1,...,T}$ be our sample. We define $\boldsymbol{y}_i \equiv (y_{i1}, ..., y_{iT})'$, $\boldsymbol{y} \equiv (\boldsymbol{y}_1', ..., \boldsymbol{y}_N')'$, $X_i \equiv (\boldsymbol{x}_{i1}, ..., \boldsymbol{x}_{iT})'$, $X \equiv (X_1', ..., X_N')'$, $\boldsymbol{\eta}_i \equiv \eta_i \iota_T$ where ι_T is a size T vector of ones, $\boldsymbol{\eta} \equiv (\boldsymbol{\eta}_1', ..., \boldsymbol{\eta}_N')'$, $\boldsymbol{v}_i \equiv (v_{i1}, ..., v_{iT})'$, and $\boldsymbol{v} \equiv (\boldsymbol{v}_1', ..., \boldsymbol{v}_N')'$. Therefore, for our sample, we can rewrite the model as:

$$\boldsymbol{y}_i = X_i\boldsymbol{\beta} + (\boldsymbol{\eta}_i + \boldsymbol{v}_i), \tag{3}$$

for $\{\boldsymbol{y}_i, X_i\}_{i=1}^N$, or:

$$\boldsymbol{y} = X\boldsymbol{\beta} + (\boldsymbol{\eta} + \boldsymbol{v}). \tag{4}$$

Both compact notations are very useful in derivations.

For static models, we assume the following:

- Fixed effects: $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq 0$, or random effects: $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$.
- Strict exogeneity: $\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = 0 \ \forall s, t$. This assumption rules out effects of past v_{is} on current $\boldsymbol{x}_{it}$ (e.g. $\boldsymbol{x}_{it}$ cannot include lagged dependent variables).
- Error components: $\mathbb{E}[\eta_i] = \mathbb{E}[v_{it}] = \mathbb{E}[\eta_i v_{it}] = 0$.
- Serially uncorrelated shocks: $\mathbb{E}[v_{it}v_{is}] = 0 \ \forall s \neq t$.
- Homoskedasticity and i.i.d. errors: $\eta_i \sim iid(0, \sigma_\eta^2)$ and $v_{it} \sim iid(0, \sigma_v^2)$, which does not affect any crucial result, but simplifies some derivations.

A simple approach to estimate $\boldsymbol{\beta}$ is to ignore the error structure. This is, define $\boldsymbol{u} \equiv \boldsymbol{\eta} + \boldsymbol{v}$, and estimate $\boldsymbol{\beta}$ by OLS:

$$\hat{\boldsymbol{\beta}}_{OLS} = (X'X)^{-1}X'\boldsymbol{y}. \tag{5}$$

The properties of $\hat{\boldsymbol{\beta}}_{OLS}$ depend on $\mathbb{E}[\boldsymbol{x}_{it}\eta_i]$, as $\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = 0$ ($\boldsymbol{x}_{it}$. If $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$ (random effects), $\hat{\boldsymbol{\beta}}_{OLS}$ is consistent as $N \to \infty$ or $T \to \infty$ or both. However, it is efficient only if $\sigma^2_\eta = 0$, i.e. if $\eta_i = 0\ \forall i$. If $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq 0$ (fixed effects), $\hat{\boldsymbol{\beta}}_{OLS}$ is inconsistent as $N \to \infty$ or $T \to \infty$ or both. Note that cross-section results are also inconsistent, and panel data helps in constructing a consistent alternative.

In our example, this would imply redefining Equation (1) as:

$$\ln C_j = \beta_0 + \beta_1 \ln P_j + \beta_2 \ln Y_j + u_j, \tag{6}$$

where subindex j is used to emphasize that observations it is considered to be independent from each other. OLS estimation of this equation delivers the following result (standard errors in parentheses):

	OLS	
ln Prices (β_1)	-0.083	(0.015)
ln Income (β_2)	-0.032	(0.006)

Note that this estimation may suffer from potential problems. Some individuals may have a higher propensity to smoke than others, everything else equal. This higher propensity could be correlated with observable factors. For example, maybe individuals whose parents used to smoke are more likely to smoke, and that may correlate with income. Likewise, individuals at different ages may have taste for different brands of tobacco that may have different pricing. In both cases, these omitted variables introduce some spurious correlation between regressors and dependent variable that can induce obtaining biased estimates of employment demands.

A. The Fixed Effects Model. Within Groups Estimation

We first assume correlated (fixed) effects: $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq 0$. As already noted, OLS estimation of such model delivers inconsistent estimates. As an alternative, consider the transformation of the model in deviations from individual means, $\tilde{y}_{it} \equiv y_{it} - \bar{y}_i$, where $\bar{y}_i \equiv T^{-1}\sum_{t=1}^{T} y_{it}$:

$$\tilde{y}_{it} = (\boldsymbol{x}_{it} - \bar{\boldsymbol{x}}_i)'\boldsymbol{\beta} + (\eta_i - \bar{\eta}_i) + (v_{it} - \bar{v}_i) = \tilde{\boldsymbol{x}}'_{it}\boldsymbol{\beta} + \tilde{v}_{it}. \tag{7}$$

Note that the transformation eliminates the individual effect, as $\bar{\eta}_i = \eta_i$. As a result, and given our earlier assumptions, $\mathbb{E}[\tilde{\boldsymbol{x}}_{it}\tilde{v}_{it}] = 0$, and OLS on the transformed model delivers consistent estimates. Such estimator is called the *within*

groups estimator, and can be written as:

$$\hat{\boldsymbol{\beta}}_{WG} = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y}. \tag{8}$$

This estimator is consistent regardless of whether $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq 0$ or $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$ if strict exogeneity, $\mathbb{E}[\boldsymbol{x}_{it}v_{is}] = 0\ \forall s,t$, is satisfied. Strict exogeneity plays an important role for consistency when $N \to \infty$ and T is fixed. If this assumption is not satisfied, $\mathbb{E}[\tilde{\boldsymbol{x}}_{it}\tilde{v}_{it}] \neq 0$ even if $\mathbb{E}[\boldsymbol{x}_{it}v_{it}] = 0$. To see it, recall that $\tilde{\boldsymbol{x}}_{it} = \boldsymbol{x}_{it} - T^{-1}(\boldsymbol{x}_{i1} + ... + \boldsymbol{x}_{iT})$, and $\tilde{v}_{it} = v_{it} - T^{-1}(v_{i1} + ... + v_{iT})$. Dynamic panel data models were developed, in part, to relax this assumption.

The consistency of the within groups estimator whether we are in a correlated or uncorrelated effects situation (when the other assumptions of static models are satisfied) is an important advantage of the estimator compared to OLS and GLS (the latter, to be seen below). But this advantage comes at a cost: (i) it is not efficient if $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$ when $N \to \infty$ but T is fixed (it is efficient when both $N, T \to \infty$), and even if $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq 0$, it is still not efficient, because when implementing OLS on the transformed model we do not take into account the autocorrelation introduced from the fact that $\tilde{v}_{it}$ includes $\bar{v}_i$ in all observations for individual i; and (ii) it does not allow to identify coefficients for time-invariant regressors (it also poorly identifies coefficients of almost invariant regressors), and it only individuals for whom regressors vary for identification of the corresponding coefficients.

In our example, within groups estimates deliver the following results:

	OLS		WG	
ln Prices (β_1)	-0.083	(0.015)	-0.292	(0.023)
ln Income (β_2)	-0.032	(0.006)	0.107	(0.019)

The within groups estimator can also be obtained by including a set of N individual dummy variables in the regression:

$$y_{it} = \boldsymbol{x}_{it}'\boldsymbol{\beta} + \eta_1 D_{1i} + ... + \eta_N D_{Ni} + v_{it}, \tag{9}$$

where $D_{hi} = \mathbb{1}\{h = i\}$ (e.g. D_{1i} takes a value of 1 for the observations on individual 1 and 0 for all other observations). OLS estimation of this model gives estimates that are numerically equivalent to within groups. For this reason, the within groups estimator is also known as *least squares dummy variables* estimator.

An alternative transformation of the model that also eliminates the individual

effects is first-differencing:

$$\Delta y_{it} = \Delta \boldsymbol{x}_{it}' \boldsymbol{\beta} + \Delta v_{it}, \tag{10}$$

where $\Delta y_{it} = y_{it} - y_{it-1}$. The fact that this transformation eliminates individual effects is evident from that they are time-invariant ($\Delta \eta_i = \eta_i - \eta_i = 0$). Therefore, OLS estimation on the differenced model is consistent. An advantage of *first-differenced least squares* is that it only requires $\mathbb{E}[\Delta \boldsymbol{x}_{it} \Delta v_{it}] = 0$ for consistency, which is implied by but weaker than strict exogeneity. However, under the classical assumptions, within groups is more efficient. First-differenced least squares, nonetheless, is more efficient if v_{it} is a random walk.

In our example, Least Squares Dummy Variables (LSDV) and and first-differenced least squares (FDLS) are presented in the following table:

	OLS	WG	LSDV	FDLS
ln Prices (β_1)	-0.083	-0.292	-0.292	-0.413
	(0.015)	(0.023)	(0.023)	(0.035)
ln Income (β_2)	-0.032	0.107	0.107	0.178
	(0.006)	(0.019)	(0.019)	(0.055)
Individual 1 ($\beta_0 + \eta_1$)			2.804	
			(0.288)	
Individual 2 ($\beta_0 + \eta_2$)			3.455	
			(0.398)	
Individual 3 ($\beta_0 + \eta_3$)			2.891	
			(0.416)	
Individual 4 ($\beta_0 + \eta_4$)			2.908	
			(0.384)	
Individual 5 ($\beta_0 + \eta_5$)			3.490	
			(0.433)	
Individual 6 ($\beta_0 + \eta_6$)			2.092	
			(0.325)	
Individual 7 ($\beta_0 + \eta_7$)			1.769	
			(0.393)	
...			...	

B. The Random Effects Model. Error Components

Now we turn into the assumption of uncorrelated or random effects: $\mathbb{E}[\boldsymbol{x}_{it} \eta_i] = 0$. In this case, OLS is consistent, but not efficient. The inefficiency is provided by the serial correlation introduced by η_i:

$$\mathbb{E}[u_{it} u_{is}] = \mathbb{E}[(\eta_i + v_{it})(\eta_i + v_{is})] = \mathbb{E}[\eta_i^2] = \sigma_\eta^2. \tag{11}$$

Likewise, the variance of the unobservables is:

$$\mathbb{E}[u_{it}^2] = \mathbb{E}[\eta_i^2] + \mathbb{E}[v_{it}^2] = \sigma_\eta^2 + \sigma_v^2. \tag{12}$$

Therefore, the variance-covariance matrix of the unobservables is a block-diagonal matrix formed by elements:

$$\mathbb{E}[\boldsymbol{u}_i \boldsymbol{u}_i'] = \begin{pmatrix} \sigma_\eta^2 + \sigma_v^2 & \sigma_\eta^2 & \dots & \sigma_\eta^2 \\ \sigma_\eta^2 & \sigma_\eta^2 + \sigma_v^2 & \dots & \sigma_\eta^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\eta^2 & \sigma_\eta^2 & \dots & \sigma_\eta^2 + \sigma_v^2 \end{pmatrix} = \Omega_i, \tag{13}$$

so that:

$$\mathbb{E}[\boldsymbol{u}\boldsymbol{u}'] = \begin{pmatrix} \Omega_1 & 0 & \dots & 0 \\ 0 & \Omega_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \Omega_N \end{pmatrix} = \Omega. \tag{14}$$

Under the classical assumptions and random effects, a GLS estimator that incorporates this covariance structure, as noted by Balestra and Nerlove (1966), is consistent and efficient:

$$\hat{\boldsymbol{\beta}}_{GLS} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\boldsymbol{y}. \tag{15}$$

This estimator is unfeasible, as σ_η^2 and σ_v^2 are unknown.

Consistency crucially relies on $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$. To see it, it is convenient to rewrite $\hat{\boldsymbol{\beta}}_{GLS}$ as the OLS estimation on the *theta-differenced* model:

$$y_{it}^* = \boldsymbol{x}_{it}^{*\prime}\boldsymbol{\beta} + u_{it}^*, \quad \text{with} \quad y_{it}^* \equiv y_{it} - (1-\theta)\bar{y}_i \quad \text{and} \quad \theta^2 = \frac{\sigma_v^2}{\sigma_v^2 + T\sigma_\eta^2}. \tag{16}$$

Note that this transformation does not eliminate the individual effect in general, so if $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] \neq 0$, GLS is inconsistent. This transformation illustrates two interesting extreme cases: (i) if $\sigma_\eta^2 = 0$, the estimator boils down to OLS, and, hence, OLS is efficient; (ii) if $T \to \infty$, then $\theta \to 0$, $y_{it}^* \to \tilde{y}_{it} = y_{it} - \bar{y}_{it}$, and the estimator reduces to within groups. Therefore, within groups is efficient if $T \to \infty$.

A feasible GLS estimator is obtained by recovering consistent estimates of σ_η^2 and σ_v^2. A consistent estimator of σ_v^2 is provided by the within groups residuals:

$$\hat{\tilde{v}}_{it} \equiv \tilde{y}_{it} - \tilde{\boldsymbol{x}}_{it}'\hat{\boldsymbol{\beta}}_{WG} \quad \Rightarrow \quad \hat{\sigma}_v^2 = \frac{\hat{\tilde{\boldsymbol{v}}}'\hat{\tilde{\boldsymbol{v}}}}{N(T-1) - K}. \tag{17}$$

To obtain an estimate of σ_η^2, we need to recover the residuals of a *between groups* estimation, which is the OLS estimation on the cross-section of individual aver-

ages:

$$\bar{y}_i = \bar{\boldsymbol{x}}_i'\boldsymbol{\beta} + \bar{\eta}_i + \bar{v}_i, \quad i = 1, ..., N. \tag{18}$$

In particular, σ_η^2 is recovered from the residuals of that regression, using the estimated $\hat{\sigma}_v^2$:

$$\hat{\bar{u}}_i \equiv \bar{y}_i - \bar{\boldsymbol{x}}_i'\hat{\boldsymbol{\beta}}_{BG} \quad \Rightarrow \quad \hat{\sigma}_{\bar{u}}^2 = \widehat{\left(\sigma_\eta^2 + \frac{1}{T}\sigma_v^2\right)} = \frac{\hat{\bar{\boldsymbol{u}}}'\hat{\bar{\boldsymbol{u}}}}{N-K} \quad \Rightarrow \quad \hat{\sigma}_\eta^2 = \hat{\sigma}_{\bar{u}}^2 - \frac{1}{T}\hat{\sigma}_v^2. \tag{19}$$

In our example, Feasible GLS results are in the following table:

	OLS	WG	FGLS
ln Prices (β_1)	-0.083	-0.292	-0.122
	(0.015)	(0.023)	(0.014)
ln Income (β_2)	-0.032	0.107	-0.012
	(0.006)	(0.019)	(0.004)

C. Testing for Correlated Individual Effects

Given the efficiency advantage of feasible GLS compared to within groups when the random effects assumption is satisfied, but its inconsistency otherwise, it is useful to test whether we are in a random or fixed effects situation. As $\hat{\boldsymbol{\beta}}_{WG}$ is consistent under the two situations, but $\hat{\boldsymbol{\beta}}_{FGLS}$ is only consistent in the random effects case, we can test wether they are similar.

This comparison is done by the Hausman test (Hausman, 1978):

$$h \equiv \hat{\boldsymbol{q}}'[avar(\hat{\boldsymbol{q}})]^{-1}\hat{\boldsymbol{q}} \overset{a}{\sim} \chi^2(K) \tag{20}$$

under the null hypothesis of $\mathbb{E}[\boldsymbol{x}_{it}\eta_i] = 0$, where:

$$\hat{\boldsymbol{q}} = \hat{\boldsymbol{\beta}}_{WG} - \hat{\boldsymbol{\beta}}_{FGLS}, \quad \text{and} \quad avar(\hat{\boldsymbol{q}}) = avar\left(\hat{\boldsymbol{\beta}}_{WG}\right) - avar\left(\hat{\boldsymbol{\beta}}_{FGLS}\right). \tag{21}$$

The test requires the classical assumptions to be satisfied, in order to ensure that feasible GLS is more efficient than within groups.

In our example:

	Statistic	P-value
Hausman test	24.661	0.000

III. Dynamic Models

We now include feedback from past shocks into future outcomes, relaxing strict exogeneity. In our cigarette consumption model, one may want to introduce past consumption of cigarettes to capture the effect of tobacco addiction (the more you smoke one period, the more you want to smoke next period). We focus on first order autoregressive models, but the results can be generalized to other forms of persistence, and to the inclusion of regressors.

A. Autoregressive Models with Individual Effects

We consider the following model:

$$y_{it} = \alpha y_{it-1} + \eta_i + v_{it} \quad |\alpha| < 1. \tag{22}$$

We assume that we observe y_{i0}, and, hence, our sample includes observations for $i = 1, ..., N$, and $t = 1, ..., T$. We keep the error components and serially uncorrelated shocks assumptions (and potentially the homoskedasticity/iid), but we relax strict exogeneity. Instead, we assume a much weaker condition, namely predetermined initial condition: $\mathbb{E}[y_{i0}v_{it}] = 0$ for $t = 1, ..., T$. Note that we are in a fixed effects situation as $\mathbb{E}[y_{it-1}\eta_i] \neq 0$ by construction. Likewise, $\mathbb{E}[y_{it-1}v_{it-1}] \neq 0$.

Even if $\mathbb{E}[y_{it-1}v_{it}] = 0$, OLS is biased. In particular:

$$\operatorname*{plim}_{N\to\infty} \hat{\alpha}_{OLS} > \alpha, \tag{23}$$

because $\mathbb{E}[y_{it-1}\eta_i] = \sigma_\eta^2 \left(\frac{1-\alpha^{t-1}}{1-\alpha}\right) + \alpha^{t-1}\,\mathbb{E}[y_{i0}\eta_i] > 0$. Likewise, within groups is biased because $\mathbb{E}[\tilde{y}_{it-1}\tilde{v}_{it}] \neq 0$. In particular:

$$\operatorname*{plim}_{N\to\infty} \hat{\alpha}_{WG} < \alpha, \tag{24}$$

because $\mathbb{E}[\tilde{y}_{it-1}\tilde{v}_{it}] = -A\sigma_v^2 < 0$, with $\left(A = \frac{(1-\alpha)\left(1+T\left(1-\alpha^{t-1}-\alpha^{T-1-t}\right)\right)+\alpha T\left(1-\alpha^{T-1}\right)}{T^2(1-\alpha)^2}\right)$. Note that the within groups bias vanishes as $T \to \infty$, but, in practice, the bias is not small even with $T = 15$. Given the sign of the biases, OLS and within groups give interesting bounds, and estimators that give $\hat{\alpha} >> \alpha_{OLS}$ or $\hat{\alpha} << \hat{\alpha}_{WG}$ should be seen with suspicion.

A seminal approach to correct these biases was proposed by Anderson and Hsiao (1981, 1982). Consider the model in first differences:

$$\Delta y_{it} = \alpha \Delta y_{it-1} + \Delta v_{it}. \tag{25}$$

OLS in first differences is inconsistent, because $\mathbb{E}[\Delta y_{it-1}\Delta v_{it}] = -\sigma_v^2 < 0$. However, if the serially uncorrelated shocks assumption holds, $\mathbb{E}[y_{it-1}v_{it}] = 0$ and y_{it-2} or Δy_{it-2} are valid instruments for Δy_{it-1}. In particular, they satisfy relevance, as $\mathbb{E}[\Delta y_{it-2}\Delta y_{it-1}] \neq 0$ and $\mathbb{E}[y_{it-2}\Delta y_{it-1}] \neq 0$, and orthogonality, as $\mathbb{E}[\Delta y_{it-2}\Delta v_{it}] = \mathbb{E}[y_{it-2}\Delta v_{it}] = 0$. Therefore, the Anderson-Hisao estimator is:

$$\hat{\alpha}_{AH} = \left(\widehat{\Delta \boldsymbol{y}'_{-1}}\widehat{\Delta \boldsymbol{y}_{-1}}\right)^{-1}\widehat{\Delta \boldsymbol{y}'_{-1}}\Delta \boldsymbol{y}, \tag{26}$$

where

$$\widehat{\Delta \boldsymbol{y}_{-1}} = Z\left(Z'Z\right)^{-1}Z'\Delta \boldsymbol{y}_{-1}, \tag{27}$$

and Z is either $\boldsymbol{y}_{-2}$ or $\Delta \boldsymbol{y}_{-2}$. A minimum of three periods ($T = 2$ plus y_{i0}) is needed to implement it. This approach is only efficient if $T = 2$ (otherwise, additional instruments are available, as discussed below).

In our example, we redefine the model to be an AR(1) process (for now without regressors):

$$n_{it} = \alpha n_{it-1} + \eta_i + v_{it} \tag{28}$$

The Anderson-Hsiao results (together with OLS and WG) are:

	OLS	WG	Anderson-Hsiao
Lagged cons. ($\ln C_{it-1}$)	0.982	0.884	1.395
	(0.003)	(0.061)	(0.090)

B. A small digression: quick review of Generalized Method of Moments (GMM)

We are interested in a parameter vector $\boldsymbol{\beta}$ that is defined by the set of moments (or orthogonality conditions) given by:

$$\mathbb{E}[\psi(\boldsymbol{x};\boldsymbol{\beta})] = \mathbf{0}, \tag{29}$$

where $\boldsymbol{x}$ is a (vector) random variable, $\boldsymbol{\beta}$ is the parameter vector, and $\psi(\ )$ is a vector function such that $\dim(\psi) \geq \dim(\boldsymbol{\beta})$. Therefore, the problem specifies $\dim(\psi)$ moment conditions.

For example, consider a regression model. The parameter vector, $\boldsymbol{\beta}$, is such that:

$$\mathbb{E}[\boldsymbol{z}u] = \mathbf{0}, \tag{30}$$

where:

$$u \equiv y - f(\boldsymbol{x};\boldsymbol{\beta}), \quad \text{and} \quad \boldsymbol{z} \equiv g(\boldsymbol{x}), \tag{31}$$

with $\dim(\boldsymbol{z}) \geq \dim(\boldsymbol{\beta})$.

We have a sample of N observations $\{\boldsymbol{x}_i\}_{i=1}^N$. The estimation is based on the sample analog of (29):

$$\boldsymbol{b}_N(\boldsymbol{\theta}) = \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \psi(\boldsymbol{x}_i; \boldsymbol{\beta}). \tag{32}$$

The GMM estimator is given by the value of $\boldsymbol{\beta}$ that minimizes the quadratic distance of $b_N(\boldsymbol{\beta})$ from zero:

$$\hat{\boldsymbol{\beta}}_{GMM} = \arg\min_{\beta\in\Theta} \boldsymbol{b}_N(\boldsymbol{\beta})' W_N \boldsymbol{b}_N(\boldsymbol{\beta}), \tag{33}$$

where W_N is a squared semi-positive definite weighting matrix that satisfies the rank condition $\mathrm{rank}(W_N) \geq \dim(\boldsymbol{\beta})$. Note that, if the problem is just-identified, this is when $\dim(\psi) = \dim(\boldsymbol{\beta})$, the weighting matrix becomes irrelevant, and the GMM estimator satisfies:

$$\boldsymbol{b}_N(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \tag{34}$$

Building on the previous example, consider the linear regression model, i.e. $f(\boldsymbol{x}; \boldsymbol{\beta}) = \boldsymbol{x}'\boldsymbol{\beta}$. Then:

$$\boldsymbol{b}_N(\boldsymbol{\beta}) = N^{-1}\sum_{\mathrm{i}=1}^{\mathrm{N}} \boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\beta}) = N^{-1}Z'(\boldsymbol{y} - X\beta), \tag{35}$$

and $\hat{\boldsymbol{\beta}}_{GMM}$ satisfies:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{GMM} &= \arg\min_{\beta} N^{-2}(\boldsymbol{y} - X\beta)' Z W_N Z' (\boldsymbol{y} - X\beta) \\ &= (X'ZW_NZ'X)^{-1}X'ZW_NZ'\boldsymbol{y}, \end{aligned} \tag{36}$$

which is a familiar expression, as it equals 2SLS if $W_N = (Z'Z)^{-1}$.

Under some general conditions, $\hat{\boldsymbol{\beta}}_{GMM}$ is a consistent estimator of $\boldsymbol{\beta}$. Additionally, it is asymptotically normal, with the following variance:

$$avar(\hat{\boldsymbol{\beta}}_{GMM}) = (D'WD)^{-1}D'WS_0WD(D'WD)^{-1}, \tag{37}$$

where $D \equiv \mathrm{plim}_{N\to\infty} \frac{\partial \boldsymbol{b}_N(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}$, $W \equiv \mathrm{plim}_{N\to\infty} W_N$, and $S_0 \equiv \frac{1}{N}\sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{E}[\boldsymbol{z}_i u_i u_i' \boldsymbol{z}_i']$.

Even though any semi-positive definite weighting matrix that satisfies the rank condition provides a consistent estimate of $\boldsymbol{\beta}$, not all of them form an efficient estimator. Efficiency is achieved with any W_N that implies $W_0 = \kappa S_0^{-1}$, for a positive κ. This includes $W_N = S_0^{-1}$ (unfeasible), but also $W_N = \hat{S}_N^{-1}$, where $\hat{S}_N$ is any consistent estimator of S_0. In practice, the Optimal GMM estimator is implemented in two steps:

1) Obtain $\hat{\boldsymbol{\beta}}_{GMM}(W_N^0)$ for an initial guess W_N^0.

2) Re-estimate using $W_{opt} = (\sum_{\mathrm{i}=1}^{\mathrm{N}} \psi(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}_{GMM}(W_N^0))\psi(\boldsymbol{x}_i; \hat{\boldsymbol{\beta}}_{GMM}(W_N^0))')^{-1}$ as the new weighting matrix.

C. Difference GMM Estimation

In a very influential paper, Arellano and Bond (1991), proposed a GMM estimation that uses all available exogenous variation in the estimation. In particular, they use the fact that for a given period t, not only y_{it-2} satisfy relevance and orthogonality conditions, but also $y_{it-3}, ..., y_{i0}$ do so. Therefore, the Arellano-Bond is defined by the following $(T-1)T/2$ orthogonality conditions:

$$\mathbb{E}[\Delta v_{i2} y_{i0}] = 0, \quad \mathbb{E}\left[\Delta v_{i3} \begin{pmatrix} y_{i0} \\ y_{i1} \end{pmatrix}\right] = \mathbf{0}, \quad \ldots, \quad \mathbb{E}\left[\Delta v_{iT} \begin{pmatrix} y_{i0} \\ y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT-2} \end{pmatrix}\right] = \mathbf{0}. \tag{38}$$

It is convenient to write these moment conditions as $\mathbb{E}[Z_i' \Delta \boldsymbol{v}_i] = 0$, where:

$$Z_i = \begin{pmatrix} y_{i0} & 0 & 0 & 0 & \ldots & 0 & 0 & \ldots & 0 \\ 0 & y_{i0} & y_{i1} & 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \ldots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & y_{i0} & y_{i1} & \ldots & y_{iT-2} \end{pmatrix} \text{ and } \Delta \boldsymbol{v}_i = \begin{pmatrix} \Delta v_{i2} \\ \Delta v_{i3} \\ \vdots \\ \Delta v_{iT} \end{pmatrix}. \tag{39}$$

The sample analogue is:

$$\boldsymbol{b}_N(\alpha) = \frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} Z_i' \Delta \boldsymbol{v}_i(\alpha), \tag{40}$$

and the Arellano-Bond estimator is:

$$\begin{aligned} \hat{\alpha}_{GMM} &= \arg\min_{\alpha} \left(\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \Delta \boldsymbol{v}_i'(\alpha) Z_i \right) W_N \left(\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} Z_i' \Delta \boldsymbol{v}_i(\alpha) \right) \\ &= (\Delta \boldsymbol{y}_{-1}' Z W_N Z' \Delta \boldsymbol{y}_{-1})^{-1} \Delta \boldsymbol{y}_{-1}' Z W_N Z' \Delta \boldsymbol{y}. \end{aligned} \tag{41}$$

In order to obtain efficient estimates, the optimal weighting matrix should be used. The unfeasible optimal weighting matrix is:

$$W_N = \left(\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \mathbb{E}[Z_i' \Delta \boldsymbol{v}_i \Delta \boldsymbol{v}_i' Z_i] \right)^{-1}, \tag{42}$$

and a feasible and asymptotically equivalent alternative is obtained in two-steps as:

$$W_N = \left(\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} [Z_i' \widehat{\Delta \boldsymbol{v}_i(\hat{\alpha})} \widehat{\Delta \boldsymbol{v}_i'(\hat{\alpha})} Z_i] \right)^{-1}. \tag{43}$$

Windmeijer (2005) proposes a finite sample correction of the variance that takes into account that α is estimated.

A common one-step weighting matrix (often used also as a first-step when the optimal two-step is calculated) uses:

$$\mathbb{E}[\Delta \boldsymbol{v}_i \Delta \boldsymbol{v}_i'] = \sigma_v^2 \begin{pmatrix} 2 & -1 & 0 & 0 & \dots & 0 \\ -1 & 2 & -1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 0 & 0 & 0 & 0 & \dots & 2 \end{pmatrix}. \tag{44}$$

Note that the weighting matrix can be multiplied by any positive scale factor without affecting the results, so there is no need to know σ_v^2.

There are two main shortcomings with this approach. The first one relates to weak instruments. If $\alpha \to 1$, relevance of the instruments decreases. As long as $\alpha < 1$, the instruments are still valid, but have very poor small sample properties. Monte Carlo evidence shows that with $\alpha > 0.8$, the estimator behaves poorly unless really huge samples are available. There are alternatives in the literature (like System-GMM) that mitigate this problem. The second one relates to overfitting. If T is relatively large compared to N, there might be "too many" instruments. In that case, we might want to restrict the number of instruments to be used. In general, it is good practice to check the robustness of the results to different combinations of instruments.

The GMM results in our AR(1) example (including one-step, two-step, and two-step with small sample correction) are:

	Coefficient	Standard Error
Least Squares (OLS)	0.982	(0.003)
Within Groups (WG)	0.884	(0.061)
Anderson-Hsiao	1.395	(0.090)
One-step GMM	1.023	(0.104)
Two-step GMM	0.994	(0.040)
Two-step GMM small sample	0.994	(0.121)

The extension of this approach to models that include regressors is straightforward. Consider the following model:

$$y_{it} = \alpha y_{it-1} + \boldsymbol{x}_{it}' \boldsymbol{\beta} + \eta_i + v_{it} \quad |\alpha| < 1. \tag{45}$$

We maintain the previous assumptions: error components, serially uncorrelated shocks, and predetermined initial conditions. Therefore, the moment conditions in Equation (38) are still valid. Different assumptions regarding $\boldsymbol{x}_{it}$ will add to them different additional orthogonality conditions. Specifically, $\boldsymbol{x}_{it}$ can be correlated or uncorrelated with $\boldsymbol{\eta_i}$, and $\boldsymbol{x}_{it}$ can be endogenous, predetermined, or strictly exogenous with respect to v_{it}. In either of this cases, different instruments will be used in the new orthogonality conditions. For instance, if assumptions are analogous to those for y_{it-1}, we may use $\boldsymbol{x}_{it-1}$ and previous lags as instruments for $\boldsymbol{x}_{it}$. In that case, the matrix of instruments would be expanded as follows:

$$Z_i = \begin{pmatrix} y_{i0} & \boldsymbol{x}'_{i0} & \boldsymbol{x}'_{i1} & \dots & 0 & \dots & 0 & \mathbf{0}' & \dots & \mathbf{0}' \\ \vdots & \vdots & \vdots & \ddots & \vdots & \dots & \vdots & \vdots & \vdots & \vdots \\ 0 & \mathbf{0}' & \mathbf{0}' & \dots & y_{i0} & \dots & y_{iT-2} & \boldsymbol{x}'_{i0} & \dots & \boldsymbol{x}'_{iT-1} \end{pmatrix}. \tag{46}$$

In our example, we rewrite the log-cigarette consumption equations as follows:

$$\ln C_{it} = \alpha \ln C_{it-1} + \beta_1 \ln P_{it} + \beta_2 \ln Y_{it} + \eta_i + v_{it}. \tag{47}$$

Results are:

	OLS	WG	GMM
Lagged dep (α)	0.947	0.528	0.495
	(0.011)	(0.064)	(0.127)
ln Prices (β_1)	0.010	-0.501	-0.607
	(0.006)	(0.098)	(0.143)
ln Income (β_2)	0.049	0.369	0.338
	(0.011)	(0.044)	(0.051)

D. System GMM Estimation

The System GMM estimator, proposed by Arellano and Bover (1995) uses the assumption $\mathbb{E}[y_{i0}|\eta_i] = \frac{\eta_i}{1-\alpha}$, which provides additional moment conditions. In particular, this implies that $\mathbb{E}[\Delta y_{it}\eta_i] = 0$ for any t, or, equivalently:

$$\mathbb{E}[\Delta y_{iT-s} u_{iT}] = 0, \quad u_{iT} \equiv \eta_i + v_{iT}, \tag{48}$$

for $s = 1, ..., T-1$. Therefore, we rewrite the moment conditions as $\mathbb{E}[(Z^*)' u_i^*] = 0$, with:

$$Z_i^* = \begin{pmatrix} Z_i & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0}' & \Delta y_{i1} & \dots & \Delta y_{iT-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{u}_i^* = \begin{pmatrix} \Delta \boldsymbol{v}_i \\ \eta_i + v_{iT} \end{pmatrix}, \tag{49}$$

and the System GMM estimator is:

$$\hat{\alpha}_{Sys-GMM} = \left(X^{*\prime} Z^* W_N Z^{*\prime} X^*\right)^{-1} X^* Z^* W_N Z^{*\prime} \boldsymbol{y}^*, \tag{50}$$

where:

$$X_i^* = \begin{pmatrix} \Delta \boldsymbol{y}_{-1i} \\ y_{iT-1} \end{pmatrix} \quad \text{and} \quad \boldsymbol{y}_i^* = \begin{pmatrix} \Delta \boldsymbol{y}_i \\ y_{iT} \end{pmatrix}. \tag{51}$$

This estimator is more efficient that differenced GMM, as it uses additional moment conditions. It also reduces the small sample bias, especially when $\alpha \to 1$.

Adding System-GMM results to the previous AR(1) results:

	Coefficient	Standard Error
Least Squares (OLS)	0.982	(0.003)
Within Groups (WG)	0.884	(0.061)
Anderson-Hsiao	1.395	(0.090)
One-step GMM	1.023	(0.104)
Two-step GMM	0.994	(0.040)
Two-step GMM small sample	0.994	(0.121)
One-step System-GMM	0.926	(0.023)
Two-step System-GMM small	0.911	(0.032)

E. Specification Tests

There are several relevant aspects for the validity of the estimation that can be tested formally. The null hypothesis that the orthogonality conditions are satisfied (i.e. moments are equal to zero) can be tested in general, as the estimation problem is typically overidentified (if $T > 2$). The standard Sargan/Hansen overidentifying restrictions test is applicable (Sargan, 1958; Hansen, 1982). The test statistic is:

$$S = N \left(\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} \hat{\hat{\boldsymbol{u}}}_i' Z_i \left(\frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} Z_i' \hat{\boldsymbol{u}}_i \hat{\boldsymbol{u}}_i' Z_i \right)^{-1} \frac{1}{N} \sum_{\mathrm{i}=1}^{\mathrm{N}} Z_i' \hat{\hat{\boldsymbol{u}}}_i \right), \tag{52}$$

where $\hat{\boldsymbol{u}}$ are predicted residuals from the first stage and $\hat{\hat{\boldsymbol{u}}}$ are those predicted from the second stage. Under the null, $S \overset{a}{\sim} \chi^2(L-K)$.

In some cases, we might be more confident on some moment conditions than others. If the problem is overidentified, we can test whether the results are stable to the inclusion/exclusion of the orthogonality conditions associated with such stronger assumptions: if these hold, efficiency is increased by using them, but if not, the estimator is inconsistent. This suggests again a Hausman test for the differences in coefficients, like in the random vs fixed effects assumptions. We

can also do an equivalent test from the moments perspective, testing whether the extra orthogonality conditions evaluated at the estimated parameters are close enough to zero (incremental Sargan test).

Finally, Arellano and Bond (1991) proposed a direct test for serial correlation of shocks, whose absence is crucial for the validity of instruments. In particular, the null hypothesis of the test is the absence of second order autocorrelation in the first-differenced residuals. Specifically:

$$m_2 = \frac{\widehat{\Delta \boldsymbol{v}_{-2}}' \widehat{\Delta \boldsymbol{v}_*}}{se} \overset{a}{\sim} \mathcal{N}(0,1), \tag{53}$$

where $\Delta \boldsymbol{v}_{-2}$ is the second lagged residual in differences, and $\Delta \boldsymbol{v}_*$ is the part of the vector of contemporaneous first differences for the periods that overlap with the second lagged vector. Values close to zero do not allow rejection the hypothesis of no serial correlation.