PANEL DATA FROM TIME-SERIES OF CROSS-SECTIONS

Angus Deaton

Princeton University

Econometric Research Program
Research Memorandum No. 317

April 1984, Revised January 1985

I am grateful to Badi Baltagi, Bill Barnett, Martin Browning, Margaret Irish, Whitney Newey, and an anonymous referee for help and comments.

Econometric Research Program
Princeton University
207 Dickinson Hall
Princeton, New Jersey 08544

In many countries, there are few or no panel data, but there exists a series of independent cross-sections. For example, in the United Kingdom, there are no panel data on consumers' expenditure or on household labor supply, but there are several large household surveys that are carried out every year. Samples for these surveys are drawn anew each year, so that individual households cannot be traced over time. This paper is concerned with the possibility of tracking "cohorts" through such data. A "cohort" is defined as a group with fixed membership, individuals of which can be identified as they show up in the surveys. The most obvious example is an age cohort, for example, all males born between 1945 and 1950, but there are many other possibilities; consider Korean war veterans, or founder members of the Econometric Society. For large enough cohorts, or large enough samples, successive surveys will generate successive random samples of individuals from each of the cohorts. Summary statistics from these random samples generate a time-series that can be used to infer behavioral relationships for the cohort as a whole just as if panel data were available. Procedures for constructing such cohorts and for estimation using the resulting data are discussed in this paper.

I consider economic relationships that are linear in parameters, though not necessarily in data, and that may or may not contain individual fixed effects. Corresponding to these individual relationships, there will exist averaged versions of the same form for the cohort population, but with unobservable data points. If there are additive individual fixed effects, there will be corresponding additive cohort fixed effects for the cohort population. Furthermore, the sample

cohort means from the surveys are consistent but error-ridden estimates of the unobservable cohort population means. Since the micro data are used to construct the means, they can also be used to construct estimates of the variances and covariances of the sample means. It is therefore possible to use errors-in-variable estimators to estimate consistently the population relationships. Sections 2 and 3 of this paper derive appropriate estimators in the absence and in the presence of individual fixed effects.

Section 1, below, presents some of the models for which the technique is designed with particular emphasis on models of consumption and labor supply. I suggest that the estimation procedures discussed here may shed light on some long-standing puzzles in empirical demand analysis, and that they are likely to be useful for the estimation of life-cycle models in the absence of panel data.

Although the methods discussed here are primarily formulated as a response to the absence of panel data, it is not necessarily the case that they will give inferior results. The attrition problem that effectively curtails the useful length of much panel data is absent here. Because new samples are drawn each year, representativeness is constantly maintained. Indeed, to the extent that long running panels replace respondents that drop out by "look alikes", the resulting data will have many of the features discussed here. Of course, the errors-in-variables nature of the current methodology is absent in genuine panel studies, but I suspect that the difference is more apparent than real. Survey statisticians who collect panel data are in little doubt as to the

magnitude of response error, particularly in the differenced data for individual respondents. And as Ashenfelter (1983) has shown, it is extremely difficult to interpret the diversity of labor supply elasticities obtainable from the Michigan PSID data (certainly the most heavily used panel data set among American economists), without assigning a central role to large and persistent errors of measurement. The technique discussed here has the advantage of recognizing measurement error from the outset and explicitly controlling for it.

1. Model formation

The models discussed in this section are of substantive interest in their own right and form the basis for current research using the methods discussed below. The main purpose in presenting them here, however, is to motivate the methodological discussion in the following sections. Further, the provision of concrete economic examples will justify some of the specific issues of estimation that are dealt with later.

In all of this work, my ultimate aim has been to bring to bear the methodology of panel data on problems of consumer demand analysis. Even in the United States, there is little panel information on the details of consumer expenditure, and a few isolated examples apart, the same is true of the rest of the world. For cross-sectional consumer expenditure surveys however, most countries of the world are better supplied than is the United States. For example, the British Family Expenditure Survey is in continuous operation and surveys some 7000 households annually; questions on income, labor supply, and a very detailed disaggregation of

consumers' expenditure provide a mass of high quality data. There are also many excellent series of household surveys from LDC's; India's National Sample Survey Office has run some twenty nationwide household expenditure surveys since independence, and both Indonesia and Sri Lanka have closely comparable household surveys for two or more separate years. There is therefore a large potential for any method that can "convert" these data sets into panel data.

To illustrate some of the more important issues, consider the Engel curve model.

$$q_{iht} = f_{iht}\,(x_{ht}, \underline{a}_{ht}) \tag{1}$$

for quantity purchased of good (or leisure) i by household h in period t, household total outlay x_{ht} and vector of socio-economic or demographic characteristics $\underline{a}_{ht}$. A convenient functional form is provided by taking the budget share $w_{iht} = p_{it}q_{iht}/x_{ht}$ as dependent variable and writing

$$w_{iht} = \alpha_i + \beta_i \log x_{ht} + \sum^{J} \gamma_{ij} a_{jht} + \varepsilon_{iht} \tag{2}$$

where there are J socio-economic characteristics, α, β, and γ are parameters, and ε_{iht} represents an error term. Equation (2) is typically estimated in one of two derived forms. In the first, using a single cross-section, the t subscript is dropped, and systems of Engel curves are estimated. In the second, the equation is aggregated over h to give, for example,

$$\tilde{w}_{it} = \alpha_i + \beta_i \log \tilde{x}_t + \Sigma \gamma_{ij} \tilde{a}_{jt} + \tilde{\varepsilon}_{it} \tag{3}$$

where $\tilde{w}_{it}$, $\tilde{a}_{jt}$ and $\tilde{\varepsilon}_{it}$ are weighted averages using $x_h/\Sigma x_h$ as weights, and $\tilde{x}_t$ is a representative budget level defined by the aggregation procedure (see Deaton and Muellbauer (1980) for a full discussion of the model). The weighting procedure guarantees that $\tilde{w}_{it}$ is the share of good i in the total of consumers expenditures, and provided the distribution over households of x does not change over time (as measured by Theil's entropy measure of inequality), $\tilde{x}_t$ can be replaced by $\bar{x}_t$. Consequently, apart from the substitution of demographics for prices, neither of which explains very much in aggregate, (3) is a conventional aggregate demand system.

The point I want to emphasize is that the values of β_i estimated from cross-sections tend to differ substantially from those estimated using time-series. Such contradictions were first extensively documented by Kuznets (1962), not only for aggregate consumption, but also for the components of consumption. It is widely known that savings ratios rise more in cross-sections than in aggregate time series. It is less well-known that it is generally true that total expenditure elasticities for many commodities and groups of commodities are further dispersed from unity when estimated on cross-sections than when estimated on time series. For example, the food share in England in 1955 was almost identical to its value a century earlier in spite of a manyfold increase in real income and in spite of the repeated confirmation of Engel's law in every household survey during that century, see Deaton (1975), Stone and Rowe (1966) and Clark (1957) for further details. The presence of

such phenomena also poses problems for forecasting demands in those situations where only cross-sectional data are available.

In terms of the foregoing model, goods are necessities if $\beta_i < 0$, luxuries if $\beta_i > 0$ and neither if $\beta_i = 0$. The Kuznets' finding, so stylized, is that β_i is closer to zero in time-series than in cross-sections. Presumably, the problem lies in inadequate statistical control. Expenditure differences between poor and rich consumers are not likely to be replicated by making a poor man rich unless the poor and rich consumers are otherwise identical. Controlling for even a long list of socio-economic characteristics is not satisfactory compared with the opportunity yielded only by panel data to use individuals as their own controls. Recognizing this, write (2) as

$$w_{iht} = \alpha_i + \beta_i \log x_{ht} + \Sigma \gamma_{ij} a_{jht} + \theta_{ih} + \varepsilon_{iht} \qquad (4)$$

for individual fixed effect θ_{ih}. Since, in general, θ_{ih} will be correlated with the other explanatory variables, such an equation can only be consistently estimated from panel data. Consider, however, the case where h is a member of a well-defined cohort group that can be tracked via its (randomly chosen) representatives through successive surveys. Let h belong to a cohort c, and take simple population averages of (4) over all h belonging to c to obtain

$$w^*_{ict} = \alpha_i + \beta_i (\log x_{ct})^* + \Sigma \gamma_{ij} a^*_{jct} + \theta^*_{ic} + \varepsilon^*_{ict} \qquad (5)$$

where asterisks denote population (i.e. cohort population) means. If it were possible to observe the true cohort means, equation (5) would hold for each cohort in each time period rather than for each household in each time period, and could be directly estimated using cohort dummy variables for the cohort fixed effects θ^*_{ic}. This would be feasible since each cohort appears in each time period; it is of course infeasible on the individual model (4) since each individual household appears only once. In practice, the other starred variables can only be proxied by cohort means from the sample; these will contain sampling errors and if used without appropriate correction will generally lead to inconsistent estimates since the model is effectively one of errors in variables with all variables (except dummies) subject to error. However, the sample can be used in the standard way to derive estimates of sampling variances and covariances and these estimates can be used to derive consistent estimators using more or less standard errors in variables procedures, see Sections 2 and 3 below.

Two other features of equation (5) should be noted. First, the total outlay variable is the mean of the logarithms, not the logarithm of the means. There is no need in this context to fudge the issue since the sample can be used just as easily to estimate the mean of a non-linear function as to estimate the non-linear function of the mean. Second, it is usually possible to select cohorts that are more or less broadly defined. Ultimately, the cohort that is all inclusive is the total population and (5) becomes a macroeconomic aggregate time-series model. In consequence, selection of cohort size allows us to move by degrees

from micro to macro data; this is ideal for detecting the roots of a contradiction between micro and macro results.

In the foregoing example, the formation of cohorts can be thought of as an instrumentation procedure that removes the inconsistencies associated with the fixed effects. In my second example, the cohort structure arises naturally out of the formulation of the problem. Consider an individual household choosing consumption and labor supply in an intertemporal setting to maximize the expectation

$$E_t \{\sum_{\tau=t}^{L} \upsilon_\tau(q_\tau)\} \tag{6}$$

subject to an evolving and uncertain budget constraint

$$W_{t+1} = (1 + i_{t+1}) \{W_t + y_t - p_t \cdot q_t\} \tag{7}$$

where, as before, q_τ includes leisure demands, υ_τ is period τ's utility function, W_t is assets at t, y_t is income, and i_{t+1} is the money interest rate from t to t+1.

In Browning, Deaton and Irish (1985), it is shown that the solution to this problem can be straightforwardly characterized in terms of Frisch demand/supply functions

$$q_{it} = f_{it} (r_t, p_t) \tag{8}$$

the vector of which is the gradient with respect to p_{it} of a period t "profit function" $\pi_t(r_t, p_t)$, a convex linearly homogeneous function. The quantity r_t, the period t price of utility, evolves stochastically according to

$$E_t \left\{\frac{1+i_{t+1}}{r_{t+1}}\right\} = \frac{1}{r_t}. \tag{9}$$

Once again, the discussion is more useful given a specific functional form. Browning, Deaton and Irish show that the following is consistent with the theory

$$q_{it} = \alpha_{it} + \beta_i \log p_{it} + \sum_{j \neq i} \theta_{ij}\{p_{jt}/p_{it}\}^{\frac{1}{2}} - \beta_i \log r_t. \tag{10}$$

This model is correct both under certainty and under uncertainty. In the former case, (9) holds without the expectation operator so that r_t is simply proportional to a discount factor $\Pi(1 + i_\tau)$ relative to some arbitrary date. Reintroducing the household subscript h, the certainty version of (10) is therefore

$$q_{iht} = \alpha_{iht} + \beta_i \log \tilde{p}_{it} + \sum_{j \neq i} \theta_{ij} \left(\frac{p_{jt}}{p_{it}}\right)^{\frac{1}{2}} - \beta_i \log r_{oh}. \tag{11}$$

where r_{oh} is independent both of time and the commodity under consideration and $\tilde{p}_{it}$ is p_{it} discounted back to the arbitrary date 0. It is therefore an individual fixed effect which is essentially a sufficient statistic for the influence of current and future values of assets,

prices, interest rates, and wages, see MaCurdy (1981). Since r_{oh} is the price of life-time utility to h, it is an increasing function of life-time real wealth given concavity of (6). Consequently r_{oh} will vary with h and thus with cohorts in the cohort version of (11). Moreover, since younger cohorts are on average wealthier than older ones, we should expect the cohort dummy variables to be monotonically related to cohort age. The cohort structure here not only has the advantage of linking micro with macro, but also explicitly recognizes the life-cycle nature of consumption and labor supply. Indeed cohort methods have been widely used in work with life-cycle models, see e.g. Ghez and Becker (1975) and Smith (1977), though these authors work with single cross-sections which lack the panel element introduced here.

Under uncertainty, (9) can be written approximately as

$$\Delta \log r_{t+1} = \log (1+i_{t+1}) + v_{t+1} \tag{12}$$

where $E_t(v_{t+1}) = 0$. Taking differences of (10) and substituting

$$\Delta q_{iht} = \Delta\alpha_{iht} - \beta_i \log (1 + \rho_{it}) + \Sigma\, \theta_{ij}\, \Delta\, (p_{jt}/p_{it}) + v_{th}^{\frac{1}{2}} \tag{13}$$

where ρ_{it} is the real commodity i rate of interest. Note that in this case, even if the shocks to the system, v_t, are stationary, log r_t will be non-stationary, so that differencing is required to obtain consistent estimates. In general, the v_t can only be guaranteed to be mean stationary, and further assumptions will be required for the consistency

of techniques applied to (13). Even so, the differenced version is likely to be a better starting point for estimation than the original version in levels once uncertainty is taken into account. Once again, the aggregation to cohorts provides the repeated observations necessary for differencing, while the micro data provide the estimates of cohort means together with their sampling errors. The differenced versions will have a different measurement error structure than the levels models, and this is discussed below in Sections 2 and 3. Note also that in (13), to the extent that the current prices and interest rates contain relevant new information, the innovation v_{th} will be correlated with the explanatory variables necessitating an estimator than can deal with both errors of measurement and simultaneity.

2. Estimation of models in levels

Before presenting the estimator to be discussed, consider an alternative, and perhaps more obvious approach to the estimation of an equation like (4) of section 1. To unify notation, rewrite this in standard form as

$$y_{ht} = \underline{x}_{ht} \cdot \underline{\beta} + \theta_h + \varepsilon_{ht} \tag{14}$$

where the i subscript is no longer required, y_{ht} is the dependent variable for individual h at t, $\underline{x}_{ht}$ is a vector of explanatory variables, and θ_h is the fixed effect. Aggregate first over those h belonging to cohort c that happen to be observed in the survey taken at t. We then get <u>observed</u> <u>sample</u> cohort means which satisfy the relationship

$$\bar{y}_{ct} = \bar{\underline{x}}_{ct} \cdot \bar{\underline{\beta}} + \bar{\theta}_{ct} + \bar{\varepsilon}_{ct}. \tag{15}$$

Note that $\bar{\theta}_{ct}$ is the average of the fixed effects for those members of c that show up in the survey; unlike the unobserved fixed effect for the cohort population mean, θ_c, say, $\bar{\theta}_{ct}$ is <u>not</u> constant over time. Furthermore, $\bar{\theta}_{ct}$ is unobserved and, in general, is correlated with the $\bar{x}_{ct}$. Hence, although (15) may be useful for "signing" the bias in regressing $\bar{y}_{ct}$ on $\bar{x}_{ct}$, it is not an appropriate basis for consistent estimation any more than is (14), unless the cohort sample sizes are so large that $\bar{\theta}_{ct}$ is a very good approximation for θ_c. In this case, (15) can be estimated by replacing $\bar{\theta}_{ct}$ by dummy variables, one for each cohort.

Consider, instead of (15), the cohort <u>population</u> version of (14). I write this

$$y^*_{ct} = \underline{x}^*_{ct} \cdot \underline{\beta} + \theta_c + \varepsilon_{ct} \tag{16}$$

where y^*_{ct} and $\underline{x}^*_{ct}$ are the unobservable cohort population means, and θ_c is the cohort fixed effect. Since the population belonging to the cohort is assumed to be fixed through time, θ_c is a constant for each c and can be replaced in (16) by cohort dummies. The y^*_{ct} and $\underline{x}^*_{ct}$ cannot be observed, but the cohort sample means $\bar{y}_{ct}$ and $\bar{\underline{x}}_{ct}$ are error-ridden estimators, with variances that can also be estimated from the micro survey data. Equation (16) can then be estimated by errors in variables techniques where <u>all</u> variables, except the dummies, are measured subject to error.

Equation (16) can now be written in convenient standard form

$$y_t^* = \underline{x}_t^* \cdot \underline{\beta} + \varepsilon_t \qquad (17)$$

where the cohorts and surveys have been "stacked" into a single index t, running from 1 to T where T is the product of the number of surveys and the number of cohorts. The cohort dummies θ_c have been absorbed into the $\underline{x}_t^*$'s; there is no loss of generality since the dummies can be thought of as being measured with an error that has zero mean and variance. To fix ideas, take the British Family Expenditure Survey as an illustration. Currently, there are about ten years of data available, with about 7000 observations per year. In Browning, Deaton and Irish's (1985) work on consumption and labor supply, various selection criteria (which are always likely to be present in one form or another) reduce this to between 2500 and 3000 observations, which were formed into sixteen cohorts of five-year age bands. Hence, T=80, but the cohorts, with a maximum size of 300, are not large enough for us to ignore the sampling errors in estimating y_t^* and $\underline{x}_t^*$ by $\bar{y}_t$ and $\bar{\underline{x}}_t$. Since, in this context (and in many others) there is a new survey every year, it is sensible to construct estimators that are consistent as T tends to infinity; with sixteen cohorts $T \to \infty$ sixteen times faster than annual, and four times as fast as quarterly time series data. The cohort size, however, is held fixed as T becomes large.

The error ε_t in (17) is assumed to be normal, independent over t, and homoskedastic; if the cohorts are very different in size, this will require that each observation be weighted by the square root of the cohort size. I shall assume this has been done as necessary. The model is completed by adding to (17) the assumed measurement structure. The

cohort means, $\bar{y}_t$ and $\bar{\underline{x}}_t$ are observed; dropping the overbars -- from now on these are the basic data -- I assume

$$\begin{pmatrix} \bar{y}_t \\ \bar{\underline{x}}_t \end{pmatrix} \sim N\left(\begin{matrix} y_t^* \\ \underline{x}_t^* \end{matrix} ; \begin{matrix} \sigma_{00} & \underline{\sigma}' \\ \underline{\sigma} & \Sigma \end{matrix} \right) \tag{18}$$

Given the sampling structure, the normality does not seem to be an implausible assumption. However, the error variances σ_{ij} will in general have to be estimated by their sample counterparts s_{ij} based on the micro survey data. Note that, in estimating the σ_{ij}'s, all T observations can be pooled, so that, if there are n_c observations in each cohort, the sampling variance of s_{ij} diminishes in proportion to $(Tn_c)^{-1}$ and that of $\bar{y}_t$ and $\bar{\underline{x}}_t$ as $(n_c)^{-1}$. The former is (a) smaller, and (b) tends to zero as $T \to \infty$ instead of remaining fixed, so that it may be reasonable to assume that the σ's are known in carrying out the estimation. Nevertheless, I shall derive formulae for both cases, when σ_{00}, $\underline{\sigma}$ and Σ are known, and when they are estimated by s_{00}, $\underline{s}$ and S.

The model is now in the form in which I can apply the results of Fuller (1975, 1981); indeed, in the rest of this section, I essentially repeat Fuller's (1975) formulae in the current context and notation. The interested reader is referred to that paper for further details.

Assume that means have been removed from all data and let the sample moments and cross-product matrices of X and y be M_{xx}, $\underline{m}_{xy}$ and m_{yy} in an obvious notation. Write σ_ε^2 for the variance of ε_t and Ω for the moment matrix of the unobservable x_t^*'s. Hence

$$E(M_{xx}) = \Omega + \Sigma = \Sigma_{xx}, \text{ say} \tag{19}$$

$$E(\underline{m}_{xy}) = \Omega\underline{\beta} + \underline{\sigma} = \underline{\sigma}_{xy}, \text{ say} \tag{20}$$

$$E(m_{yy}) = \underline{\beta}'\Omega\underline{\beta} + \sigma_{00} + \sigma_{\varepsilon}^2 = \sigma_{yy}, \text{ say} \tag{21}$$

The estimator $\tilde{\underline{\beta}}$ is then clearly consistent as $T \to \infty$, where

$$\tilde{\underline{\beta}} = (M_{xx} - \Sigma)^{-1}(\underline{m}_{xy} - \underline{\sigma}) \tag{22}$$

$$= (X'X - T\,\Sigma)^{-1}(X'\underline{y} - T\,\underline{\sigma}) \tag{23}$$

and, provided $(X'X-T\Sigma)$ is positive definite, will be a MLE under the normality assumptions on $\underline{x}_t^*$ that are sometimes made in errors in variables models. I shall not assume normality of $\underline{x}_t^*$ here. Note that if Σ and $\underline{\sigma}$ are replaced by estimators S and $\underline{s}$, $\tilde{\underline{\beta}}$ is replaced by

$$\underline{\beta}^* = (M_{xx} - S)^{-1}(\underline{m}_{xy} - \underline{s}) = (X'X - TS)^{-1}(X'y - T\underline{s}) \tag{24}$$

The formula for the variances are derived first for the case where Σ, $\underline{\sigma}$ and σ_{00} are known, i.e. for $\tilde{\underline{\beta}}$. Expanding (22) yields

$$\tilde{\underline{\beta}} - \underline{\beta} = \Omega^{-1}(\underline{m}_{xy} - \underline{\sigma}_{xy}) - \Omega^{-1}(M_{xx} - \Sigma_{xx})\Omega^{-1}(\underline{\sigma}_{xy} - \underline{\sigma}) + 0_p(T^{-1}) \tag{25}$$

But, from (20), $\Omega^{-1}(\underline{\sigma}_{xy}-\underline{\sigma})=\underline{\beta}$, so that

$$\tilde{\underline{\beta}} - \underline{\beta} = \Omega^{-1}[(m_{xy} - M_{xx}\underline{\beta}) - (\underline{\sigma}_{xy} - \Sigma_{xx}\underline{\beta})] + 0_p(T^{-1}) \tag{26}$$

The variance of $\tilde{\underline{\beta}}$ thus depends asymptotically on the variance covariance matrix of $\underline{m}_{xy} - M_{xx}\underline{\beta}$. But

$$\{\underline{m}_{xy} - M_{xx}\underline{\beta}\}_i = \frac{1}{T} \sum_t x_{ti} (y_t - \sum_k x_{tk}\beta_k) \tag{27}$$

$$= \frac{1}{T} \sum_t (x^*_{ti} + u_{ti})(\varepsilon_t + u_{t0} - \sum_k u_{tk}\beta_k) \tag{28}$$

where $u_{t0} = y_{t0} - y^*_{t0}$ and $u_{ti} = x_{ti} - x^*_{ti}$.

Treating the x^*_{ti} as fixed but unknown constants, and using the standard properties of the normal distribution yields

$$TV\ \{\underline{m}_{xy} - M_{xx}\underline{\beta}\} = \Sigma_{xx}(\sigma^2_\varepsilon + \sigma_{00} + \underline{\beta}'\Sigma\underline{\beta} - 2\underline{\sigma}'\beta) + \tag{29}$$

$$+ (\underline{\sigma} - \Sigma\underline{\beta})(\underline{\sigma} - \Sigma\underline{\beta})'$$

Hence the asymptotic variance covariance matrix of $\tilde{\beta}$ is given by

$$TV(\tilde{\underline{\beta}}) = \Omega^{-1}[\Sigma_{xx}\omega^2 + (\underline{\sigma} - \Sigma\underline{\beta})(\underline{\sigma} - \Sigma\underline{\beta})']\Omega^{-1} \tag{30}$$

where

$$\omega^2 = \sigma^2_\varepsilon + \sigma_{00} + \underline{\beta}'\Sigma\underline{\beta} - 2\underline{\sigma}'\underline{\beta} \tag{31}$$

An estimate of (30) is straightforwardly derived from the observable moment matrices. From (19) to (21)

$$\tilde{\Omega} = M_{xx} - \Sigma \tag{32}$$

$$\tilde{\sigma}^2_\varepsilon = m_{yy} - \sigma_{00} - \tilde{\underline{\beta}}'\tilde{\Omega}\tilde{\underline{\beta}} \ . \tag{33}$$

Substitution in (31) yields an estimate of ω^2, i.e.

$$\tilde{\omega}^2 = \frac{1}{T}(\underline{y} - X\tilde{\underline{\beta}})'(\underline{y} - X\tilde{\underline{\beta}}) = \frac{1}{T}\underline{e}'\underline{e}, \text{ where} \tag{34}$$

$$\underline{e} = \underline{y} - X\tilde{\underline{\beta}} \tag{35}$$

$$\text{But } \Sigma\tilde{\underline{\beta}} - \underline{\sigma} = (\Sigma\tilde{\underline{\beta}} + \Omega\tilde{\underline{\beta}}) - (\underline{\sigma} + \Omega\tilde{\underline{\beta}}) \tag{36}$$

$$= M_{xx}\tilde{\underline{\beta}} - \underline{\sigma}_{xy} = \frac{1}{T}X'\underline{e} \tag{37}$$

Hence, the estimated variance-covariance matrix is given by

$$T\tilde{V}(\tilde{\underline{\beta}}) = \tilde{\Omega}^{-1}[T^{-1}M_{xx}\,\underline{e}'\underline{e} \ + \ T^{-2}X'\underline{e}\,\underline{e}'X]\tilde{\Omega}^{-1} \tag{38}$$

which is straightforwardly evaluated in practice.

The derivation of the variance-covariance matrix of $\underline{\beta}^*$, the estimator using the <u>estimated</u> error variances, requires only minor modification of the above. I assume that the estimates s_{ij} of σ_{ij}, i, j = 0, 1,...K, the dimension of $\underline{x}$, are based on νT degrees of freedom. If

all cohorts are pooled in estimating the s_{ij}'s, $\nu=n_c$, the number of observations per cohort, but clearly other schemes are possible. Some estimate of the variances and covariances of the s_{ij} is also required; to focus attention I shall use that derived from sampling under normality. Hence, I assume that s_{ij} is consistent for σ_{ij}, and that asymptotically,

$$(\nu T)\ E\{(s_{ij} - \sigma_{ij})(s_{k\ell} - \sigma_{k\ell})\} = \sigma_{ik}\sigma_{j\ell} + \sigma_{i\ell}\sigma_{jk} \tag{39}$$

The derivation proceeds as before except that the expansion (25) has an additional term corresponding to the stochastic variation in S and $\underline{s}$. Hence, (26) becomes

$$\underline{\beta}^* - \underline{\beta} = \Omega^{-1}[(\underline{m}_{xy} - M_{xx}\underline{\beta}) - (\underline{\sigma}_{xy} - \Sigma_{xx}\underline{\beta})]$$

$$- \Omega^{-1}[(\underline{s} - S\underline{\beta}) - (\underline{\sigma} - \Sigma\underline{\beta})] + 0_p(T^{-1}) \tag{40}$$

By the properties of sampling under normality, the first and second terms are independent, so that, asymptotically

$$TV(\underline{\beta}^*) = \Omega^{-1}[\Sigma_{xx}\omega^2 + (\underline{\sigma} - \Sigma\underline{\beta})(\underline{\sigma} - \Sigma\underline{\beta})']\Omega^{-1}$$

$$+ \nu^{-1}\Omega^{-1}V(\underline{s} - S\underline{\beta})\Omega^{-1} \tag{41}$$

Elementary manipulation yields

$$V(\underline{s} - S\underline{\beta}) = \Sigma(\sigma_{00} - 2\underline{\sigma}'\underline{\beta} + \underline{\beta}'\Sigma\underline{\beta}) + (\underline{\sigma} - \Sigma\underline{\beta})(\underline{\sigma} - \Sigma\underline{\beta})' \tag{42}$$

Note that if ν is large, (41) reduces to (30), the case of known variances; this latter is likely to be a formula that would normally be adequate in practice.

Equations (23), (24), (38) and (41) are the basic results of this section. I conclude with four issues of practical importance. First, the error variances σ_{00}, $\underline{\sigma}$ and Σ will generally vary from survey to survey and cohort to cohort. Write σ_{00}^t, $\underline{\sigma}^t$ and Σ^t for the values at observation t, so that (19) and (20) become

$$E(M_{xx}) = \Omega + \bar{\Sigma} \tag{19'}$$

$$E(\underline{m}_{xy}) = \Omega\underline{\beta} + \bar{\underline{\sigma}} \tag{20'}$$

where $\bar{\Sigma}$. $\bar{\underline{\sigma}}$ are the mean values over the t observations. The analysis then goes through with σ_{00}, $\underline{\sigma}$ and Σ (or s_{00}, $\underline{s}$ and S in the case of estimated variances) replaced by their means. Given the nature of the variation with t, an appropriate variance-covariance matrix for the $\bar{S}_{ij}$'s can be derived and substituted for $V(\underline{s} - S\underline{\beta})$ in (41). Second, it is necessary to allow for the presence of some $\underline{x}$ variables that are measured without errors. For example, relationships like (5) of Section 1 contain cohort dummies that are clearly error free. Other variables may not be drawn from the surveys but from other data sources; macroeconomic variables that are the same for all cohorts but vary with time (prices) are the obvious examples, and, exceptionally, there may be other relevant data on the cohorts themselves. One way to proceed is to introduce additional error free variables to the right-hand side of (14) and to track them through the analysis. This turns out to be equivalent to the

simpler (and intuitive) procedure of setting the appropriate elements of $\underline{\sigma}$ and rows and columns of Σ to zero; the formulae for $\tilde{\beta}$, $\underline{\beta}^*$ and their asymptotic variances then remain unchanged. Third, for the reasons discussed in Section 1, it may be necessary to recognize contemporaneous correlations between some of the $\underline{x}_t^*$'s and ε_t. Instrumental variables will typically be available; in the example of Section 1, in the form of lagged cohort wages or prices. On the assumption that the instrument vector $\underline{w}_t$ is constructed from a survey prior to that used for $\underline{x}_t$ so that their errors of measurement are uncorrelated, the appropriate instrumental variable estimator is

$$\tilde{\underline{\beta}}_{IV} = [W'X(W'W - T\Sigma_w)^{-1}X'W]^{-1}[W'X(W'W - T\Sigma_w)^{-1}W'\underline{y}] \qquad (43)$$

with a variance matrix calculable by the methods given above. If measurement errors in W and X are correlated, the obvious additional corrections can be made. Fourth, and finally, note that there will typically be some flexibility in constructing cohorts. If cohorts are constructed by age, the age bands can be taken broad or narrow (e.g., a five year window versus a one year window), and other qualifying characteristics can be left unspecified or tightly defined. Clearly, the construction of cohorts with members that are distinct from one another and internally homogeneous will minimize the errors in variable problem and enhance precision. Beyond that, it is possible to use trial cohorts to estimate $\tilde{\underline{\beta}}$ and its variance, and to use these consistent estimates to gauge the consequences of combining or separating cohorts.

3. Estimation of Models in Differences

In this section, I develop the estimators appropriate for the case where the model, like the second model of Section 1, requires differencing prior to estimation. The previous results do not go through directly because the measurement errors induced by the sampling scheme now have a MA(1) structure relative to the unobservable true first differences.

I now write the model in the form

$$\Delta y_t^* = \Delta \underline{x}_t^* \cdot \underline{\beta} + \varepsilon_t \tag{44}$$

for the true unobservable first differences. Corresponding to (18), the measurement structure is

$$\Delta y_t = \Delta y_t^* + v_{t0} \tag{45}$$

$$\Delta \underline{x}_t = \Delta \underline{x}_t^* + \underline{v}_t \tag{46}$$

and

$$v_{t0} = u_{t0} - u_{t-10} \tag{47}$$

$$v_{ti} = u_{ti} - u_{t-1i} \tag{48}$$

with u_{t0} and u_{ti} the original measurement errors on the y_t and x_{ti} variables respectively. The relationship between Δy_t and $\Delta \underline{x}_t$ is therefore given by

$$\Delta y_t = \underline{\beta} \cdot \Delta \underline{x}_t + (\varepsilon_t + v_{t0} - \underline{\beta} \cdot \underline{v}_t). \tag{49}$$

In passing note a tempting but ineffective possible route to estimation. Consider $Nx_t = x_t + x_{t-1}$, the moving average, as a possible instrument for Δx_t, the first-difference, and for simplicity, assume x_t is scalar. Since $Nx_t = Nx_t^* + u_t + u_{t-1}$, and since $u_t + u_{t-1}$ is independent of $v_t = u_t - u_{t-1}$, Nx_t is orthogonal to the compound error in (49). However, in large samples Nx_t is also orthogonal to the regressor Δx_t; $E(Nx_t.\Delta x_t) = x_t^{*2} - x_{t-1}^{*2}$, so that $E(T^{-1}\Sigma Nx_t.\Delta x_t) = T^{-1}(x_T^{*2} - x_0^{*2})$ with a limit as $T \to \infty$ of zero. In consequence, instrumental variables estimation of this type will not work. It is therefore necessary to follow through a scheme similar to that of Section 2.

To ease notation, write $n_t \equiv \Delta y_t$ and $\underline{z}_t \equiv \Delta \underline{x}_t$ so that, corresponding to (19) to (21) the moment matrices are now

$$E(M_{zz}) = W + 2\Sigma = \Sigma_{zz}, \text{ say} \tag{50}$$

$$E(\underline{m}_{zn}) = W\underline{\beta} + 2\underline{\sigma} = \underline{\sigma}_{zn}, \text{ say} \tag{51}$$

$$E(\underline{m}_{nn}) = \underline{\beta}'W\underline{\beta} + \sigma_\varepsilon^2 + 2\sigma_{00} = \sigma_{nn}, \text{ say} \tag{52}$$

where W is the sample moment matrix of the $\Delta\underline{x}_t^* \equiv \underline{z}_t^*$ variables. The doubled role of measurement error comes from the moving average errors in (47) and (48). The first-difference estimator, $\tilde{\underline{\beta}}_\Delta$ is immediately given as

$$\tilde{\underline{\beta}}_\Delta = (M_{zz} - 2\Sigma)^{-1} (\underline{m}_{zn} - 2\underline{\sigma}) \tag{53}$$

or equivalently

$$\tilde{\underline{\beta}}_\Delta = (Z'Z - 2T\Sigma)^{-1} (Z'\underline{n} - 2T\underline{\sigma}). \tag{54}$$

Expansion, as in Section 2, yields

$$\tilde{\underline{\beta}}_{\Delta} - \underline{\beta} = (\Sigma_{zz} - 2\Sigma)^{-1} \{(m_{zn} - M_{zz}\underline{\beta}) - (\sigma_{zn} - \Sigma_{zz}\underline{\beta})\} + 0_p(T^{-1}) \tag{55}$$

so that if C is the variance-covariance matrix of $\underline{m}_{zn} - M_{zz}\underline{\beta}$, the asymptotic variance of $\tilde{\underline{\beta}}_{\Delta}$ is given by

$$TV(\tilde{\underline{\beta}}_{\Delta}) = (\Sigma_{zz} - 2\Sigma)^{-1} C(\Sigma_{zz} - 2\Sigma)^{-1}. \tag{56}$$

Now $\quad (\underline{m}_{zn} - M_{zz}\underline{\beta})_i = T^{-1} \Sigma z_{ti}(n_t - \underline{\beta}.\underline{z}_t)$, so that

$$(\underline{m}_{zn} - M_{zz}\underline{\beta})_i = \frac{1}{T} \Sigma\ (z^*_{ti} + v_{ti})(\varepsilon_t + v_{t0} - \sum_k \beta_k v_{tk}). \tag{57}$$

The variance of this is tedious to calculate, particularly given the MA structure (47) and (48). In the appendix it is shown that

$$\begin{aligned} C = (W + 2\Sigma)(\sigma^2_{\varepsilon} + 2\sigma^2_A) - \sigma^2_A\ (W^+ + W^- - \Sigma) \\ + 14\ (\underline{\sigma} - \Sigma\beta)(\underline{\sigma} - \Sigma\beta)' \end{aligned} \tag{58}$$

where

$$\sigma^2_A = \sigma^2_{00} - 2\underline{\sigma}.\underline{\beta} + \underline{\beta}'\Sigma\underline{\beta} \tag{59}$$

and

$$W^+ = \frac{1}{T} \Sigma \underline{z}^*_t \underline{z}^{*\prime}_{t-1}. \tag{60}$$

$$W^- = \frac{1}{T} \Sigma \underline{z}^*_{t-1} \underline{z}^{*\prime}_t. \tag{61}$$

The presence of W^+ and W^- reflects the autocorrelation in the measurement error. Comparing with (29), and taking the case where $W^+ = W^- = 0$, the measurement errors now play a much larger role in determining the variance. Put differently, given the same amount of variance in the true unobservables under levels and differences, estimation precision will be lower in the latter case. This result, which is not surprising, can be enhanced or modified by positive or negative autocorrelation in the true $\underline{z}_t^*$ series.

The asymptotic variance covariance matrix of $\tilde{\underline{\beta}}_\Delta$ is obtained by substituting (58) into (56). In practice, an estimation formula can be obtained by noting that, from (52)

$$\tilde{\sigma}_\varepsilon^2 = m_{nn} - 2\sigma_{00} - \tilde{\underline{\beta}}_\Delta' \tilde{W} \tilde{\underline{\beta}}_\Delta \tag{62}$$

Hence, from (59)

$$\tilde{\sigma}_\varepsilon^2 + 2\sigma_A^2 = m_{nn} - 2\tilde{\underline{\beta}}_\Delta' (2\underline{\sigma} + \tilde{W}\tilde{\underline{\beta}}_\Delta) + \tilde{\underline{\beta}}_\Delta'(\tilde{W} + 2\Sigma)\tilde{\underline{\beta}}_\Delta \tag{63}$$

$$= m_{nn} - 2\tilde{\underline{\beta}}_\Delta' \underline{m}_{zn} + \tilde{\underline{\beta}}_\Delta' M_{zz} \tilde{\underline{\beta}}_\Delta \tag{64}$$

$$= T^{-1} \underline{e}' \underline{e} \text{ , where} \tag{65}$$

$$\underline{e} = \underline{n} - Z.\tilde{\underline{\beta}}_\Delta = \Delta \underline{y} - \Delta X.\tilde{\underline{\beta}}_\Delta \tag{66}$$

Similarly

$$2\Sigma\tilde{\underline{\beta}}_\Delta - 2\underline{\sigma} = 2\Sigma\tilde{\underline{\beta}}_\Delta + \tilde{W}\tilde{\underline{\beta}}_\Delta - \tilde{W}\tilde{\underline{\beta}}_\Delta - 2\underline{\sigma} = M_{zz}\tilde{\underline{\beta}}_\Delta - \underline{m}_{zn} = T^{-1} Z' \underline{e} \tag{67}$$

Finally, therefore, writing M^{+}_{zz} for $T^{-1}\ \Sigma \underline{z}_t \underline{z}_{t-1}$ and M^{-}_{zz} similarly, the estimated variance matrix of $\tilde{\underline{\beta}}_{\Delta}$ is given by

$$T\hat{V}\ (\tilde{\underline{\beta}}_{\Delta}) = \tilde{W}^{-1}[M_{zz}\frac{1}{T}\ \underline{e}'\underline{e} - \tilde{\sigma}^2_A(M^{+}_{zz} + M^{-}_{zz}) + \frac{7}{2T^2}\ Z'\underline{ee}'Z]\tilde{W}^{-1} \quad (68)$$

$$\tilde{\sigma}^2_A = \sigma^2_{00} - 2\underline{\sigma}.\tilde{\underline{\beta}}_{\Delta} + \tilde{\underline{\beta}}'_{\Delta}\Sigma\tilde{\underline{\beta}}_{\Delta}. \quad (69)$$

The modification of these formulae for the case where σ_{00}, $\underline{\sigma}$ and Σ are replaced by estimates is straightforward and is left to the reader.

Appendix

Derivation of the variance for the differenced case

Starting from equation (57), define

$$\theta_i = (\underline{m}_{zn} - M_{zz}\underline{\beta})_i \quad (A1)$$

$$\xi_t = \varepsilon_t + v_{t0} - \sum_k \beta_k v_{tk} \quad (A2)$$

so that

$$\theta_i = T^{-1}\ \Sigma(z^{*}_{ti} + v_{ti})\xi_t \quad (A3)$$

To obtain the variance-covariance matrix of θ_i, start from

$$E(\theta_i\theta_j) = E\{T^{-2}\ \sum_{ts}\sum\ (z^{*}_{ti} + v_{ti})\xi_t(z^{*}_{sj} + v_{sj})\xi_s\} \quad (A4)$$

$$= E\{T^{-2}\sum_{ts}\sum(z^*_{ti}z^*_{sj}\xi_t\xi_s + v_{ti}\xi_t v_{sj}\xi_s)\} \tag{A5}$$

since ξ_t and v_t are jointly normal with zero third moments. By the MA(1) structure of the v's

$$E(\theta_i\theta_j) = T^{-1}\{E(z^*_{ti}z^*_{tj}\xi^2_t) + E(z^*_{ti}z^*_{t+1j}\xi_t\xi_{t+1}) +$$

$$+ E(z^*_{t+1i}z^*_{tj}\ \xi_{t+1}\xi_t)+E(v_{ti}\xi_t v_{tj}\xi_t)+2E(v_{ti}\xi_t v_{t+1j}\xi_{t+1})\} \tag{A6}$$

Now $$E(\xi^2_t) = \sigma^2_\varepsilon + 2\sigma^2_{00} + 2\underline{\beta}'\Sigma\underline{\beta} - 4\underline{\beta}'\underline{\sigma} = \sigma^2_\varepsilon + 2\sigma^2_A\ ,\ \text{say} \tag{A7}$$

$$E(\xi_t\xi_{t+1}) = -\ \sigma_{00} + 2\underline{\beta}'\underline{\sigma} - \underline{\beta}'\Sigma\underline{\beta} = -\ \sigma^2_A \tag{A8}$$

Hence, evaluating (A6) term by term using, where necessary, the standard formulae for fourth movements of normals, gives

$$E(z^*_{ti}z^*_{tj}\xi^2_t) = w_{ij}(\sigma^2_\varepsilon + 2\sigma^2_A) \tag{A9}$$

$$E(z^*_{ti}z^*_{t+1j}\xi_t\xi_{t+1}) = -\ w^+_{ij}\sigma^2_A, \tag{A10}$$

with a similar expression for its transpose.

$$E(v_{ti}v_{tj}\xi^2_t)=2\sigma_{ij}(\sigma^2_\varepsilon+2\sigma^2_A)+2\{2\sigma_i-2\sum_k \beta_k\sigma_{ik}\}\{2\sigma_j-2\sum_k\beta_k\sigma_{jk}\}$$

$$= 2\sigma_{ij}(\sigma^2_\varepsilon + 2\sigma^2_A) + 8\ (\underline{\sigma} - \Sigma\underline{\beta})_i(\underline{\sigma} - \Sigma\underline{\beta})_j \tag{A11}$$

$$E(v_{ti}\xi_t v_{t+1j}\xi_{t+1}) = \sigma_{ij}\ \sigma_A^2 + 5(\underline{\sigma} - \Sigma\underline{\beta})_i(\underline{\sigma} - \Sigma\underline{\beta})_j \tag{A12}$$

Hence, collecting terms and subtracting $E(\theta_i)\ E(\theta_j) = 4(\underline{\sigma} - \Sigma\underline{\beta})_i$ $(\underline{\sigma} - \Sigma\underline{\beta})_j$ yields

$$V(\underline{\theta}) = (W + 2\Sigma)(\sigma_\varepsilon^2 + 2\sigma_A^2) - \sigma_A^2\ (W^+ + W^- - \Sigma) +$$

$$+ 14(\underline{\sigma} - \Sigma\underline{\beta})(\underline{\sigma} - \Sigma\underline{\beta}) \tag{A13}$$

which is equation (58) of the main text.

References:

Ashenfelter, O. (1983), "Macroeconomic analyses of labor supply and microeconomic analyses of labor supply", presented to Carnegie-Rochester Conference, Bal Harbor, Florida, November 1983.

Browning, M., A. Deaton and M. Irish (1985), "A profitable approach to labor supply and commodity demands over the life-cycle", Econometrica, forthcoming.

Clark, C. (1957), The conditions of economic progress, London. Macmillan.

Deaton, A. (1975), "The structure of demand 1920-1970", in The Fontana Economic History of Europe, C.M. Cipolla (ed.), Collins-Fontana.

Deaton, A. and J. Muellbauer (1980), "An almost ideal demand system", American Economic Review, vol. 70, pp 312-26.

Fuller, W.A. (1975), "Regression analysis for sample survey"; Sankhya: The Indian Journal of Statistics, Series C, Vol. 37, pp. 117-32.

Fuller, W.A. (1981), Measurement error models, Department of Statistics, Iowa State University, unpublished.

Ghez, G.R. and G.S. Becker (1975) The allocation of time and goods over the life-cycle, New York. NBER.

Kuznets, S. (1962), "Quantitative aspects of the economic growth of nations: VII. The share and structure of consumption", Economic Development and Cultural Change, vol 10, pp 1-92.

MaCurdy, T.E. (1981), "An empirical model of labor supply in a life-cycle setting", Journal of Political Economy, vol. 89, pp. 1059-85.

Smith, J. (1977), "Family labor supply over the life-cycle", Explorations in Research, vol. 4, pp. 205-76.

Stone, R. and D.A. Rowe (1966) The Measurement of consumers' expenditures and behavior in the United Kingdom 1920-1938, vol II, Cambridge. Cambridge University Press.