nesis
2810

Accessing the New Earnings Survey Panel Dataset

Efficient techniques and applications

Felix Ritchie

96

Department of Economics
University of Stirling

Contents

Abstract

The New Earnings Survey Panel Dataset is one of the largest datasets of its kind in the world. Its size and confidentiality restrictions present considerable difficulties for analysis using standard econometric packages.

This thesis presents a number of methods for accessing the information held within the panel relatively efficiently, based upon the use of cross-product matrices and on data compression techniques. These methods allow, for the first time, the panel aspect of the dataset to be used in analysis. The techniques described here are then employed to produce an overview of changes in the UK labour market from 1975 to 1990 and detailed estimates of male and female earnings over a fourteen year period.

These are the first panel estimates on the dataset, and they indicate the importance of allowing the parameters of any labour market model to vary over time. This is significant as panel estimators typically impose structural stability on the coefficients. A comparison of cross-section and panel estimates of earnings functions for males indicate that the allowance for individual heterogeneity also has a notable effect on the estimates produced, implying simple cross-sections may be significantly biased. Some preliminary estimates of the male-female wage gap indicate that variation over time has an important part to play in accounting for the differences in wages, and that "snapshot" studies may not capture dynamic changes in the labour market. Individual differences also play a significant role in the explanation of the wage gap.

List of abbreviations

AV	Attrition variable
CS	Cross-section
DE	UK Department of Employment
DEG	Department of Employment Gazette
FE	Fixed-effects
FES	Family Expenditure Survey
IV	Instrumental variable
LFS	Labour Force Survey
NES	New Earnings Survey
NESPD	New Earnings Survey Panel Datasets
NI	National Insurance
NiNo	National Insurance number
OH	Observation history
QCC	Quasi-complete cohort
SCELI	Social Change and Economic Life Initiative
TVCS	Time-varying cross-section covariance estimator
TVFE	Time-varying fixed-effects covariance estimator
TFFEIV	Time-varying fixed-effects instrumental variable estimator
WC	Wages Council
WES	Women and Employment Survey
WIRS	Workplace Industrial Relations Survey

List of tables page

List of figures

Acknowledgements

This thesis was written as a full-time student and subsequent lecturer at the University of Stirling. I am grateful to my principal and secondary supervisors, David Bell and Brian Main, for advice and support over this period. I would also like to thank Martyn Andrews, Bob Elliott, Bob Hart, Eric Levin, and Peter Sloane for general comments on the thesis and specific comments on chapters, sections and techniques. I am also grateful for the comments received at various seminars from my colleagues at the University of Stirling. Wiji Arulampalam and Phil Murphy examined the thesis in detail, correcting many errors and pointing out a number of inconsistencies and inaccuracies. All remaining errors and omissions are my own.

Elizabeth Roberts and Richard Upward provided many useful comments and suggestions on the software developed here, and were invaluable in the testing process - even if testing to destruction was depressingly easy. Robert Jukes and Paul Lanser at the Department of Employment, who manage the dataset and who oversaw the development of the software, were instrumental in providing both the incentive for the methods developed here and the access to the dataset to obtain results.

Finally, my studies have been carried out under the aegis of the Scottish Doctoral Programme in Economics. The regular seminars at Stirling and Crieff, the extensive links between researchers and institutions, and the network of academics and students have all contributed to this thesis presented here. I would like to thank my friends and colleagues on the programme for four years of advice, encouragement, linguistic advancement and fun.

Felix Ritchie

June 1995
Revised February 1996

Chapter 1

Introduction

This thesis grew out of a number of research projects carried out by researchers at the Universities of Manchester and Stirling into the New Earnings Survey Panel Dataset (NESPD). The NESPD is the largest dataset of its type in Europe but its size presents some difficulties for research. It is too large for practical access by conventional statistical programs (except for generating descriptive statistics), and the data as supplied to the Department of Employment (DE) is not in a convenient form for analysis. The data is also subject to confidentiality restrictions and so cannot be analysed outwith the DE except in some aggregated form.

In 1991 the University of Stirling won a research contract to develop some basic software for the DE which would allow for the creation of data files suitable for the statistical package SPSS and the matrix programming language GAUSS. This software was devised and written by Elizabeth Roberts, a researcher at the University of Stirling. The SPSS files have been used by researchers at several institutions to create cross-tabulations, totals, and other descriptive statistics. In addition, some researchers have run cross-sectional analyses, and some attempt has been made to run non-linear models on a subset of the data.

The approach at Stirling was to use the GAUSS matrix programming language to develop a software package that could perform micro-level regressions on the dataset without the need to reread the data whenever a new set of variables is desired. The key is that the OLS regression is a linear combination of cross-product matrices; therefore, by appropriate construction of such matrices with all the variables of interest included, it is possible to run a large number of regressions with a single pass through the dataset to collect the data. In addition to the computational efficiency of this approach, it has the added benefit of avoiding the DE confidentiality restrictions.

Over time this program has been expanded to incorporate covariance estimation, differencing approaches, and linear instrumental variable specifications. Although these techniques in themselves are not new, the method of calculating them from cross-product matrices is unusual. Moreover, the TVFE/TVCS estimator in particular is unique in being the only software package to present time-varying coefficients for panel and cross-section models as the standard option. While models with time-varying coefficients can be specified in other packages by appropriate use of dummy variables, the default is to force a common set of slope coefficients on the model. The extra inconvenience involved in generating a time-varying-coefficients model, along with the requirement for more degrees of freedom, may explain the almost complete absence of this type of estimator in the literature on panel data. The only significant exception to this is Chamberlain's (1984) minimum-distance estimator, which has received some theoretical attention but little application.

The approach at Stirling University has been to make the time-varying coefficients model (for both panels and cross-sections) the basic specification. The methodological justification for this is provided by Hendry's general-to-specific approach; the empirical justification is that the hypothesis of constant slope coefficients is comprehensively rejected in almost all cases. Fortunately, the NES has sufficient observations to meet the demands on degrees of freedom with no difficulty.

The purpose of this thesis, then, is to present models and estimation methods implemented in the software and to discuss some of the results to emerge from the application of these techniques to the NESPD. This is essentially a practical thesis, and the theoretical content is relatively small. However, the techniques to be discussed and the results obtained have a number of implications for other research in panel data studies and for the analysis of large datasets generally.

Also introduced in this thesis is a data structure called the "observation history", which

presents semi-aggregated data in an extremely compact form. These allow for the quick calculation of an enormous number of descriptive statistics and the simple analysis of transitions between states (for example, from missing to observed or from union to non-union). Perhaps more importantly, they enable the creation of pseudo-panel datasets which may be removed from the DE. Unlike pseudo-panels created from aggregated data, these have the characteristics of true panels and so allow for straightforward analysis of a wide variety of models, including non-linear ones. However, these are a relatively new development and in this thesis are used mainly to generate descriptive statistics.

The structure of this work is as follows. Chapter two presents some of the basic ideas about estimation with panel datasets, paying particular attention to linear models and the choice of fixed- or random-effects specifications. Chapter three describes the NES and the difficulties associated with accessing the NESPD. Chapter four outlines the main solution taken at Stirling to the access problem. Chapter five describes in detail the construction of cross-product matrices for different specifications and estimation methods; chapter six describes the operation of the analytical software, the extension to instrumental variables, and the calculation of summary statistics. Chapter seven describes the construction and potential uses of the observation histories.

In the applied section, chapter eight uses the observation histories and some cohort data to provide insights into the data, the labour market, and the implications for various model specifications. Chapter nine estimates a Mincerian fixed-effects wage equation on the NESPD males and compares it with a cross-section. It also investigates the scope for more restricted specifications than the time-varying coefficients model. Chapter ten estimates a similar equation on female data, and uses this to discuss the gender gap in earnings. Both of these chapters are, I believe, the first estimates on the NESPD to allow for individual heterogeneity, and the evaluation of male earnings in particular has some pertinent comments to make about cross-sectional analyses. Finally, chapter eleven concludes the thesis.

Chapter 2

Exploiting panel datasets

2.1 The uses of panel data

Much econometric modelling is based on 'one-dimensional' data sets: time-series or cross-sectional analysis. Both these methods have their problems. Time-series models excel in their treatment of dynamic effects, but suffer from the multicollinearity of series. Cross-sectional analysis makes use of a wide variety of functional forms, but is necessarily limited in its treatment of dynamic effects.

Pooling data on individuals over time into one dataset allows the econometrician to deal with a range of relationships between units of information within a single coherent structure. Panel data can be seen as cross-sections observed and linked over time, or multiple structural time-series. It is argued that by combining the best of both worlds better estimators result[1]. The use of all information available within the same model makes for inherently more efficient estimators, while the larger amount of data increases the degrees of freedom for hypothesis testing. This latter effect produces increased flexibility in model design by allowing more scope for the use of instrumental variables, simultaneous equation specifications, lagged variables and other techniques needing many degrees of freedom.

More importantly, a panel dataset enables the researcher to discriminate between competing hypotheses indistinguishable under simpler models. Consider estimating the success rates over time of a training program. The hypothesised relationship may be

[1] In the context of this discussion "better" merely refers to some arbitrary criterion such as mean-squared error or efficiency used to evaluate different estimators.

$$y_{it} = f(x_{it}, \beta, \alpha_i) \qquad (2.1)$$

where α_i is some element specific to the individual i ("individual heterogeneity" or an "individual specific effect") which does not vary over time; for example, an element of "motivation". Initial results find that 40% of the class pass each exam, but do the same 40% pass every exam, or does everyone have a 40% chance of passing? In other words, is this unobserved effect significant in determining the outcome? If the term α_i was identified and found to be a significant factor in the probability of passing in any particular period, this would imply that someone passing one exam is more likely to pass or have passed other exams. The hypothesis that the same 40% pass each time appears more likely.

This effect could not have been identified by treating the data as purely cross-sectional (that is, with no connection between observations in different periods). Treating each period equation separately means that the individual-specific effect is not identified and must be subsumed into the constant term. Pooling all the data would appear to show serial correlation in the errors due to the unobserved heterogeneity. However, a panel model could determine the relative importance of the unobserved heterogeneity, and distinguish it from any "true" serial correlation in the results (see Hsiao (1986) p 174).

Panel models provide the opportunity to test and control for a much wider range of measurement errors and unmeasurable effects then the simple example above. By using the panel to its full extent, both intercepts and slope coefficients which vary over time and/or individuals can be estimated from structural or reduced forms. This gives great scope for flexibility in the model without having to identify all the relevant variables: the ability to "group" observations by period or individual is all that is needed in many cases.

Hsiao (1986, pp5-7) provides some examples, reproduced in part in Figure 2.1, where the apparent cross-sectional relationship is belied by the panel estimates. The dotted lines

represent pooled estimates (that is, ignoring any panel structure), while the others represent the "true" structure which could be revealed by the appropriate panel estimators. In Figure 2.1(a), the panel estimates have common positive slope coefficients, as does the pooled estimate. However, as each individual has a different intercept, the pooling estimate is clearly inefficient. In Figure 2.1(b) the effect of ignoring different intercepts means that the pooled estimate no longer even has the correct sign for the slope. Thus, although only intercepts vary over individuals, this suffices for the pooled results to give an entirely erroneous view of the relationship[2]. Moreover, there is no a priori indication of how the pooled slope is biased from true. Identical slope coefficients but different unmeasured effects have led to very different pooled estimates in (a) and (b).

In Figures 2.1(c) and (d) the slope coefficients also vary. Clearly a pooled regression on the individuals in Figure 2.1(c) would indicate little or no relationship between the variables on the x and y axes, while in Figure 2.1(d) the pooled regression appears to produce a nonlinear relationship. A properly specified panel model would be able to determine the true structure of the relationship.

The ability of panel techniques to combine information on individuals and time is their strongest asset. Unfortunately, this ability can also cause significant problems. The rationale for panel models is that interrelationships over time and between individuals are constant and so can be factored out. If the assumed interrelationship is wrong, then the error may affect all elements of the regression. A misspecified cross-section in period t (for example) should not affect estimation of the relationship in t+1 which uses different data; but if an individual specific effect is misspecified, it may corrupt the results from all periods. This is especially

---

[2] For example, some cross-section studies carried out by the author appeared to reveal a positive relationship between the proportion of manufacturing in GDP and energy consumption in developed countries. A simple panel study using the same data showed a negative relationship, a reverse of case (b) above. The implication was that the cross-section results were spurious, arising from significant national differences, and the original estimator was too crude to pick this up.

relevant in non-linear models; see section 2.4.

The most obvious, and important, source of misspecification is selection bias. Panels, by their very nature, are more susceptible than other data sets to missing observations, a problem which increases as the panel grows over time. For example, consider a two-period panel composed of welfare recipients. Those who remain in the panel for the second period may be less "employable", if those who have found jobs leave the panel. Whether this attrition is random or correlated with the dependent variables is crucial for the results of any estimation. At best it reduces efficiency; at worst, it can distort results significantly.

This issue is extremely complex for panel models, and currently unsolved in the general case. There is some current research on this issue (see Ridder (1990) and Ritchie (1994) for a theoretical treatment; Bell and Ritchie (1993b, 1994) for a study of selection bias in the NES); but even simple static models with multinormal spherical errors present formidable computational difficulties. The applied work for this thesis presents a practical but rather ad hoc approach to selection bias.

A second (and much less frequently discussed) source of error peculiar to panels is an overdependence on the ability to account for unmeasured variables. Consider the training example taken over two years, when the underlying "motivation" changes significantly and that the change is reflected across all individuals. Four regression models may be considered:

$$\begin{aligned} &(a)\quad y_{it} = \mu + x_{it}\beta + u_{it} \\ &(b)\quad y_{it} = \mu_t + x_{it}\beta_t + u_{it} \\ &(c)\quad y_{it} = \mu + x_{it}\beta + \alpha_i + u_{it} \\ &(d)\quad y_{it} = \mu_t + x_{it}\beta_t + \alpha_i + u_{it} \end{aligned} \qquad (2.2)$$

for t=1,2, i=1..N. (2.2a) and (2.2b) are cross-sectional models; (2.2c) and (2.2d) are panel models. However, in (2.2a) and (2.2c) the coefficients are assumed to be constant over time, and so only (2.2b) and (2.2d) can identify the structural shift between years one and two.

Clearly (2.2a) is the most restricted model and (2.2d) the least, and the performance of estimators of the models will reflect this; but it is difficult to say whether the flexible cross-section (2.2b) or the poorly-specified panel model (2.2c) will perform better. Chapter nine returns to this in an applied context.

The problem becomes more important when variation in the slope coefficients is allowed. If performance in a training program is improving as recruitment, teaching and testing methods improve, it may be desirable to allow the slope coefficients to vary over time. Separate cross-sections, such as (2.2b), allow for this. So may a panel model, and one such as(2.2d) is at least as efficient as the cross-sections. However, the overwhelming majority of estimators used in applied work are of the form of (2.2c) with a time-varying intercept. If the slope coefficients vary significantly the cross-section may give better results. Relying on the panel attributes of parsimonious models instead of using more general specifications can lead to poor outcomes.

This question of misspecification is not limited to panels, and the particular problems caused by the use of panel models do not raise any significant new issues. A general discussion of the misspecification of panel models is beyond the scope of this thesis, and so is only considered in relation to the particular matter at hand. Unless explicitly stated otherwise, it is assumed that the models are correctly specified.

2.2 Fixed and random effects

The unmeasured effects in a panel model structure may be "fixed" or "random" with respect to the model. This has practical and theoretical consequences, and leads to different estimators, and possibilities for inference. Consider a simple linear panel model:

$$y_{it} = x_{it}\beta + \mu + u_{it} \qquad u_{it} = \alpha_i + \epsilon_{it} \tag{2.3}$$

In this model, an individual-specific term is assumed to capture all the omitted information. μ is the mean intercept for all participants; α_i is the individual variation around μ. This term is sometimes called an "incidental parameter", as the focus of interest is the value of β. The concern about the α_i term arises from its effect on the other variables; identifying its value is often not required.

The individual-specific terms may be deemed "fixed" (that is, parameters of the model). They then amount to a set of coefficients on individual-specific dummy variables which may be estimated by OLS (or any method appropriate to the assumed error structure of ϵ_{it}). All results are then conditional on these parameters. Note that OLS estimates of the dummy coefficients are inconsistent while T remains small (see section 2.3), but this does not affect the consistency of the estimates of β and μ.

However, if α_i is considered "random" (that is, a random component of the variance of the dependent variable), then autocorrelated residuals from OLS estimates will reflect the distribution of both ϵ_{it} and α_i. Note that the error term in (2.3) in vector form becomes,

$$\begin{aligned} u_i &= J_T\alpha_i + \epsilon_i \\ E(u_i) = 0 \qquad E(u_i u_i') &= J_T J_T'\sigma_\alpha^2 + I_T\sigma_\epsilon^2 \end{aligned} \tag{2.4}$$

where J_T is a T-vector of ones. OLS is unbiased and consistent, but inefficient unless σ_α^2 is zero (and assuming zero correlation between x_{it} and α_i). GLS solution methods may be used to identify the two distributions and estimate (2.3) efficiently.

Section 2.3 concentrates on the practical effect of different assumptions; for the moment, consider the theoretical implications. The choice is largely a subjective one, and most texts on panel data consider how this choice may be made. Hsiao(1992) suggests that the key theoretical issues come down to (a) what is the purpose of the study? and (b) what is the

context of the data?

The argument is usually based around sample versus population study. If interest lies in the characteristics of participants in the sample, or if the participation list is exhaustive, then a fixed effects model may be most appropriate. If the aim is to determine population parameters from a sample, then the random-effects specification may be more useful.

For example, in a study of training programmes in different industries over time, a significant industry-specific effect may be the result of institutionally-based practices. The size of the effect may be useful information in itself, enabling predictions for each industry and facilitating inter-industry comparisons. In such a case it appears reasonable to treat these industry-specific factors as fixed, and allowing a different constant term for each industry captures that effect. Interest lies in using the heterogeneity as a predictive and explanatory tool for the behaviour of individual industries in the sample. If all industries of interest are included in the survey, then being unable to predict the size of this effect for other industries - a consequence of the fixed-effect assumption - is irrelevant.

Alternatively, consider the effect of training programmes on individuals over time. It is plausible to assume that each employee responds to the programme in a unique manner which persists over time. However, the interest is less in these individual differences but in the overall effect of the programme. Making general predictions for the programme requires the distribution of these individual effects over the workforce. Accordingly, the trainees in question are assumed to be random drawings from the population of workers with correspondingly random unobserved traits. Then the performance of a new participant on the program can be predicted with more confidence than by extrapolation from the specific (fixed) heterogeneity of current trainees.

As a third example, the applied work in this thesis centres around Mincer-type wage

equations where the worker is the observation unit. While acknowledging that the NES remains a sample drawing, with hundreds of thousands of individuals appearing in the dataset there is some justification for approaching it as a population. The random-effect and fixed-effect specifications can both be justified on the population/sample argument. On the other hand, the unobserved characteristics of each person are of less interest than how these characteristics manifest themselves over the populace as a whole. The aim is to be able to make predictions about the population, not to identify any one particular individual's wage. Thus a random-effects model may be deemed appropriate[3].

With heterogeneity over both individuals and time, a combination of fixed and random effects may be employed. Consider extending (2.2) to include an effect specific to period t, λ_t:

$$y_{it} = x_{it}\beta + \mu + u_{it} \qquad u_{it} = \alpha_i + \lambda_t + \epsilon_{it} \tag{2.5}$$

A common estimator in applied work has one effect fixed and one random; estimators with both effects fixed or random are less common. Often this is for practical reasons: introducing dummy variables for the "large" dimension may be cumbersome and inefficient, whilst random-effects estimation of the "small" dimension tends to be complex and inefficient. As panels tend to be "short and fat" (that is, with T relatively small and N large), a common solution is to estimate random individual effects and to introduce time dummies for the time effects.

To some extent this also reflects research interests. Much panel work is done on micro-data, with population inferences being drawn. There is little interest in the performance of individuals as opposed to the whole, and the number of parameters in a fixed-effects specification may be very large. A better solution is to look for any overall distribution of such individual effects.

---

3 In fact, the estimation method to be discussed uses the fixed-effects methods, for practical reasons outlined in the next section, although knowledge of the individual effects has little practical use.

For time effects the opposite holds: calculating a "distribution" of time intercepts is likely to be fairly meaningless, but by treating intertemporal differences as parameters of the model there is more scope for comparing directly different periods. Fixed effects are particularly useful when slope coefficients are allowed to vary: the evolution of coefficients through time may be very enlightening.

For example, estimation on the full NES panel has little of interest to say about the "motivation" of some hundred thousand individuals; but an observable shift in the intercept over time is of interest. thus a random individual effect and a fixed time effect appears sensible.

All this assumes that the underlying structure of the model is known. If the true structure is unknown, then the feasibility, efficiency and consistency of the specifications must be considered.

The fixed-effects model is distribution free; it is conditioned on the extant values for α_i without the need to describe the source or distribution of this effect. It is robust to alternative specifications of the individual heterogeneity because the distribution function is irrelevant to the estimation method.

Estimation of the random effect requires further assumptions about the error terms. This is not necessarily a significant drawback. Although ML estimation of the specification in (2.3) requires a specific functional form for the panel effects, GLS estimation is feasible and practical with merely a consistent estimate of the covariance matrix, given by fixed-effects estimates.

Much more serious is the necessary assumption of independence of the explanatory variables and the random effect. Mundlak (1978) argued that the fixed effects model is conditional on

the explanatory variables, and that the random-effects model is a misspecification that fails to take account of this conditioning. An appropriate model should use $E(\alpha_i \mid x_i)$. Replacing α_i in (2.3) by a linear approximation

$$E(\alpha_i|x_i) = \sum_t x_{it}'a_t + \omega_i \qquad \omega \sim N(0, \sigma_\omega^2) \tag{2.6}$$

where a_t is a vector of constants to be estimated, allows for correlation between the unmeasurable and explanatory variables. Mundlak then suggested restricting the model to a function of the mean value of the explanatory variables

$$E(\alpha_i|x_i) = \bar{x}_i'a + \omega_i \qquad \omega \sim N(0, \sigma_\omega^2) \tag{2.7}$$

and it can be shown (Hsiao (1986) pp 44-45) that the GLS estimator of β collapses to the fixed-effects estimator; the difference between the "true" and "false" GLS estimates of β is the GLS estimate of a. Therefore, there are not two models and two estimators: the apparent difference is a specification error.

In a more general approach, Chamberlain (1984) notes that, if α_i is correlated with $(x_{i1}..x_{iT})$, y_{it} is potentially a function of all the explanatory variables and any estimator should take account of all lead and lag values of x_{it}. This gives a multivariate regression of all the ys (Tx1) on all the xs (TxKT) with an arbitrary error structure:

$$\begin{aligned} y_{it} &= Z_i'\zeta_t + u_{it} \\ E(u_{it}) &= 0 \qquad E(u_{it}u_{is}) = \sigma_{ts}^2 \end{aligned} \tag{2.8}$$

where $Z_i = [x_{i1} \ldots x_{iT}]$. This is the basis for Chamberlain's "minimum-distance" panel estimator, described in more detail in section 2.6.2.

Mundlak's result depends on the very restrictive assumption of the source of the heterogeneity, while Chamberlain's is much more general, but both raise an important point: the random effects estimator is unbiased and consistent only if the explanatory variables and the heterogeneity are independent of each other. The fixed-effects estimator, giving parameter

estimates conditioned on the explanatory variables, is unaffected. This idea of the fixed effects estimators as conditioning estimators (as opposed to the marginal estimator of the random-effects model) acknowledges the fact that the former is valid for both fixed- and random-effects specifications[4].

Fixed-effects estimators are therefore flexible and robust; however, generally they are inefficient. This is because the fixed-effects approach seeks to isolate individuals or periods, and so restricts itself to smaller samples relative to the number of parameters to be identified; random-effects estimators look for common characteristics and make holistic assessments of the data. For example, as the number of periods shrinks, making efficient use of information across individuals becomes increasingly important; the large number of parameters in fixed-effects estimators becomes a growing burden. As N large and T small is a common structure for panel datasets, then a consistent random-effects estimator is generally more efficient. Taylor (1980) estimated that, if the assumptions of the random-effects model hold, then this construction is more efficient for the cases (T>2, N-K>8) and (T>1, N-K>9).

2.3 Estimation of static linear panel data models

This section is only intended as a brief introduction to some aspects of panel estimation methods, so the mathematics are kept to a minimum and areas covered are selective[5]. This section concentrates on a few basic estimators, with N individuals and T periods, as the qualitative aspects of these carry over in a straightforward manner to more complex specifications. In chapter 5, the fixed-effects specifications implemented in the analysis software will be discussed in more detail.

---

4 Note that, if the assumption of zero correlation between the explanatory variables and the individual heterogeneity is violated, then cross-section estimates, which ignore the heterogeneity completely, will also be biased and/or inconsistent.

5 Full discussion is provided in Hsiao (1986) or Matyas and Sevestre (1992). In the following discussion a balanced panel is assumed; that is, $T_i=T_j=T$ for all i, j. This does not change the results materially.

As focus of the concern here is in illuminating certain aspects of panel estimators (and not developing the econometric methodology), this section begins with the simplest models. The general linear model is

$$y_{it} = x_{it}\beta_{it} + u_{it} \tag{2.9}$$

with i=1..N, t=1..T, and x_{it} being a (1xK) vector of explanatory variables. This allows any parameter to vary over time and individuals, giving (NTxK) parameters in NT equations; the model is unidentified without some restrictions on the parameters. The type of restriction imposed can lead to very different estimators; so that, if one is removing or adding variables before re-estimating, the choice of initial estimator may influence the path taken. For example, a pooled estimator on the data in Figure 2.1(d) may indicate that a quadratic form is needed, whereas an initial panel estimator could indicate that this is unnecessary. Thus, while panel estimates may be fairly robust in many cases, this does not obviate the need for general tests on the specification of the model.

The simplest parameterisation is the pooled model: in system form,

$$y = X\beta + u \tag{2.10}$$

where y, X and u are stacked to give NTx1, NTxK, and NTx1 matrices. There is assumed to be no significant consistent variation in the coefficients. Separate time-series or cross-section estimates give the same result as the pooled estimator, but the greater combined number of observations in the panel lead to smaller standard errors. Estimates are consistent whether N, T or both tend to infinity. The estimator is efficient under the assumption that the residual errors have no time- or individual-specific element. If the disturbance is non-spherical for other forms of heteroscedasticity or autocorrelation, any of the usual transformation or estimation methods for the particular form of the error terms is appropriate.

2.3.1 The covariance approach

A first extension to (2.10) is to let an intercept vary across individuals[6], as in (2.3) above. For true panel specifications, a number of solution methods become appropriate. This section demonstrates a common approach, using techniques from variance analysis.

If the individual variation α_i is treated as a fixed effect, it becomes a parameter to be estimated. A simple solution is to employ N dummy variables, such that

$$y = X\beta + I_N \otimes J_T A + u \tag{2.11}$$

where $A = [\alpha_1 \ \alpha_2 \ .. \ \alpha_N]'$, an Nx1 vector of the individual fixed effects, I_N is the N-element identity matrix, and J_T is a T-vector of ones. Equation (2.11) is known as the least-squares dummy-variable (LSDV) specification. The N+K parameters in this equation can be estimated by OLS[7]. This solution has a significant drawback. Consider the normal equations:

$$\begin{bmatrix} \beta \\ A \end{bmatrix} = \begin{pmatrix} X'X & X'I_N \otimes J_T \\ I_N \otimes J_T' X & I_N \otimes J_T' J_T \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ I_N \otimes J_T' y \end{pmatrix} \tag{2.12}$$

This requires the potentially formidable task of inverting an (N+K)x(N+K) matrix. A more practical alternative is to take deviations from individual means. For an individual equation, let

$$\tilde{y}_{it} \equiv y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)\beta + u_{it} - \bar{u}_i \equiv \tilde{x}_{it}\beta + \tilde{u}_{it} \tag{2.13}$$

The individual effect, constant over time, drops out of the regression[8]. The system equations become

$$\tilde{y} = \tilde{X}\beta + \tilde{u} \tag{2.14}$$

and the normal equations are now

---

6 Time-specific effects are dealt with in a qualitatively identical manner. The relevant equations are found by swapping T and N and the t-i subscripts.

7 Note that estimates of the individual-specific effect only become consistent for large T not large N. The estimates of the slope coefficients are consistent for N and/or T large.

8 An equivalent route to this result is via the Frisch-waugh theorem for partitioned regressions, where y is replaced by y adjusted for individual intercepts and then the adjusted y regressed on X alone.

$$\hat{\beta}_w = (\tilde{X}'\tilde{X})^{-1}(\tilde{X}'\tilde{y}) \tag{2.15}$$

which only involves a KxK inversion. The panel effects can be found from the individual means:

$$\hat{\alpha}_i = \bar{y}_i - \bar{x}_i\hat{\beta}_w \tag{2.16}$$

(2.15) is the "within" or "covariance" estimator (from the analysis-of-covariance technique used). The estimated slopes in (2.15) are BLUE and consistent if N and/or T is large; however, the estimates of the individual intercepts, while unbiased, are not consistent unless T is large. Expanding (2.16):

$$\begin{aligned} \hat{\alpha}_i &= (\bar{x}_i\beta + \bar{u}_i + \alpha_i) - \bar{x}_i\hat{\beta}_w \\ &= \bar{x}_i(\beta - \hat{\beta}_w) + \alpha_i + \frac{1}{T}\sum_t u_{it} \end{aligned} \tag{2.17}$$

As $E(u_{it}) = 0$, the estimate of α_i is unbiased, but remains inconsistent unless $T \to \infty$ as

$$\underset{N\to\infty}{\hat{\alpha}_i} \to \alpha_i + \frac{1}{T}\sum_t u_{it} \tag{2.18}$$

This is because the number of parameters to estimate increases with N and there remains insufficient variation amongst individuals to uncover heterogeneity whilst T is small.

Now consider random effects, distributed over individuals according to some function. The random individual effects and the residual errors have to be estimated together. Define

$$v_{it} \equiv \alpha_i + u_{it} \qquad \alpha_i \sim N(0, \sigma_\alpha^2) \qquad u_{it} \sim N(0, \sigma_u^2) \tag{2.19}$$

Other covariances are zero. The structure of the covariance matrix for an individual is

$$E(v_i'v_i) = J_T J_T'\sigma_\alpha^2 + I_T\sigma_u^2 \equiv \Omega_i \tag{2.20}$$

This is called the "variance components" or "error components" model, for obvious reasons. OLS is inefficient, as the error terms v_{it}, v_{is} are serially correlated. However, GLS is both feasible and practical. Under the assumptions of (2.19) the covariance matrix for the regression is block-diagonal, as is its inverse:

$$\Omega = diag(\Omega_1, \Omega_2, \ldots \Omega_N)$$
$$\Omega^{-1} = diag(\Omega_1^{-1}, \Omega_2^{-1}, \ldots \Omega_N^{-1}) \tag{2.21}$$

with

$$\Omega_i^{-1} = \frac{1}{\sigma_u^2}\left[\left(I_T - \frac{1}{T}J_T J_T'\right) + \psi\frac{1}{T}J_T J_T'\right] \qquad \psi \equiv \frac{\sigma_u^2}{\sigma_u^2 + T\sigma_\alpha^2} \tag{2.22}$$

The first part in the inverted variance term calculates deviations from the mean of any matrix to which it is applied; it is the same matrix used to generate the within estimate, β_W. The second part calculates the mean of a matrix and multiplies it by a constant which reflects the relative variances of the two error components. The normal equations for the GLS solution are

$$\begin{bmatrix}\hat{\beta} \\ \hat{\mu}\end{bmatrix}_{gls} = \left(\sum_i X_i'\Omega_i^{-1}X_i\right)^{-1}\left(\sum_i X_i'\Omega_i^{-1}y_i\right) \tag{2.23}$$

The covariance matrix, and thus the solution for β, contains two additive terms. The structure in (2.22) indicates one term should reflect "within-group" variation, and the other the differences between group means - "between-group" variation. It can be shown (see, for example, Hsiao (1986) pp34-41) that the GLS solution breaks down into

$$\hat{\beta}_{gls} = \Delta\hat{\beta}_b + (I_K - \Delta)\hat{\beta}_w \tag{2.24}$$

where the "between" estimator

$$\hat{\beta}_b = \left(\sum_i (\bar{X}_i-\bar{X})'(\bar{X}_i-\bar{X})\right)^{-1}\left(\sum_i (\bar{X}_i-\bar{X})'(\bar{y}_i-\bar{y})\right) \tag{2.25}$$

is calculated by taking the deviations of inter-group means from the whole mean of the regression. The GLS estimate is a weighted average of the within and between estimates, with the weights given by

$$\Delta = \left(\sum_i\sum_t (X_{it}-\bar{X}_i)'(X_{it}-\bar{X}_i) + \psi T\sum_i (\bar{X}_i-\bar{X})'(\bar{X}_i-\bar{X})\right)^{-1}\left(\sum_i (\bar{X}_i-\bar{X})'(\bar{X}_i-\bar{X})\right) \tag{2.26}$$

The between estimate, showing the inter-individual variation, is effectively the OLS estimate

on grouped data ignoring all panel aspects. The weighting is provided by the relative importance of the error components and the size of T. If T is one or if σ_u^2 is very large relative to σ_α^2, then the weights are allocated evenly between the two estimators, and the pooled OLS estimate results. In this case either variation among the observations for an individual is small or the number of observations for an individual is small; random variation between individuals is the dominant force. Alternatively, if σ_α^2 or T is large, the within estimator dominates. If σ_α^2 is large, then individuals are sufficiently different to make separate estimation on each individual a sensible strategy. If T is large, between-group variation becomes irrelevant: there are enough observations for each individual to be treated as a separate model. Each α_i can be thought of as drawn once (randomly) and then fixed for N sample drawings which are large enough to be estimated separately.

Feasible estimation of (2.23) requires known or consistently estimated error components. Consistent estimates are given by the residuals from separate covariance and pooled OLS regressions. The error components model can also be estimated by ML. When T is small and N large, this method is wholly consistent; however, if N is small and T large, the estimate of σ_α is inconsistent. When T is large, the model becomes a series of N separate regressions; the ML estimate collapses to the covariance estimator (see Hsiao (1986) for details).

The covariance estimator can be used on the random effects model; obviously, the transformation removes the individual effect whatever its nature. For large T the two estimators coincide, but by ignoring the information on within-group variance rather than using it, the covariance estimator of a random-effects model is less efficient than ML or GLS when T is small. However, the random-effects estimators requires zero correlation between the error-component and the explanatory variables for consistency and unbiasedness. This is not an issue for the fixed-effects estimators which are conditioned on the parameters. Thus covariance estimation of a random-effects model may be consistent when GLS estimation is not.

(2.11) is the simplest of panel models and can be expanded in many ways. Consider adding time- or individual-specific variables; that is, the coefficients are common to all, but the variables are not:

$$y_{it} = x_{it}\beta + z_i\delta + \mu + \alpha_i + u_{it} \qquad (2.27)$$

z_i is a vector of individual data which does not change over time. In this model, a fixed α_i cannot be separated from the coefficient z_i if the covariance transformation is used. Random effects can still be identified. The minimum-distance estimator of Chamberlain (1984) uses invariant individual-specific vectors for its initial estimates of the parameters.

Another simple extension is to have both time- and individual-specific effects:

$$y_{it} = x_{it}\beta + \mu + \alpha_i + \lambda_t + u_{it} \qquad (2.28)$$

This presents no new qualitative aspects. If both effects are fixed, then some restrictions are needed to prevent exact collinearity between the dummies. If both effects are random, then a variance component for time has to be calculated. This requires a third, "between-periods" estimator, which is the equivalent of the between-groups estimator used to find the individual variance.

However, the quantitative effects of incorporating more fixed or random variables are quite different, especially if the panel is unbalanced (that is, individuals have different numbers of observations). Adding time dummies is a straightforward matter and the balance of the panel has little practical effect. However, calculating a new error-component is complex, and if the panel is unbalanced the problem is significantly harder. Generalising (2.22) for unbalanced panels merely requires the substitution of T_i for T, but the three-component equivalent of (2.22) has six additive terms; with unbalanced panels the inverse is exceedingly complex[9].

---

9 See Wansbeek and Kapteyn (1989) for the solution to the unbalanced three-component inverse.

In addition to the theoretical considerations introduced in section 2.2, two additional elements have been introduced. First, the covariance estimator is consistent (if not necessarily efficient) under a wider range of assumptions. Secondly, the covariance estimator is easier to estimate than the random effects one. The practical differences increase if the panel is unbalanced. If the efficiency loss is small, or if collinearity between the individual heterogeneity and the explanatory variables is suspected, then error-components estimators lose their appeal. The implication for the choice of fixed or random effects estimators is clear: the fixed-effects estimators, although inefficient, are both tractable and robust.

Linear panel models more complex than those discussed above add little to the methodological issues raised, and are not considered here. However, three general features of linear models can be identified. Firstly, the dichotomy between fixed and random effects remains strong as models become more complex. Although some solution methods, such as analysis of covariance, may be applicable to both, the results for one or other assumption are general sub-optimal in some way. This is because the choice of random or fixed effects leads to qualitatively different assessments of intergroup or interperiod influences; applying one estimator to both hypotheses implies either irrelevant or relevant but unused information.

Secondly, the general structure of the solutions outlined persists. The fixed-effects specification may be seen as a question of effective use of dummy variables or covariance transformations. The random-effects model has a complex error structure requiring GLS or ML estimation. Choice of fixed or random effect determines the whole approach to estimation.

Thirdly, the more complex the models, the bigger the practical advantage of the fixed-effects approach. The difficulties of estimating fixed-effects models increase arithmetically with the complexity of the models; for random-effects, the relationship is more closely exponential. While the fixed-effects model does involve lots of 'incidental' parameters, generally the model

can be transformed to remove these effects[10]. Although the number of parameters in the random-effects model rises much more slowly, isolating the distribution of these effects becomes increasingly problematic.

2.3.2 The differencing approach

Where there is only one random effect, differencing of the equations is a practical alternative to the covariance approach. In both (2.11) and (2.19), differencing removes the individual-specific effect:

$$\tilde{y}_{it} \equiv y_{it} - y_{it-1} = (x_{it} - x_{it-1})\beta + u_{it} - u_{it-1} \equiv \tilde{x}_{it}\beta + \tilde{u}_{it} \qquad (2.29)$$

with $E(\tilde{u}_{it})=0$. Estimation can then proceed as usual. When T=2, the differencing approach and the covariance estimator coincide.

Differencing has significant advantages over the covariance estimator in dynamic linear models, to be detailed below. It also shares the distributional advantage of the covariance estimator: namely, that no distribution for the individual effect needs to be specified and that any correlation between the individual-specific effect and the other explanatory variables will not lead to inconsistent or biased estimators. However, for static models it is less appealing.

Firstly, the differencing estimator is less efficient than the covariance estimator - compare (5.90) and (5.124) for the expected variance of the regression. This is because, although both estimators involve N restrictions on the model, y_{i1} and y_{iT} each only contribute once to the information set of the differencing estimator, whereas all observations contribute equally to the covariance estimator.

---

10 Often a combination of dummy variables and transformations is used. In the estimators to be described in chapter 5, individual effects are transformed out while dummies are used for time effects.

Secondly, differencing is clearly not appropriate for a random-effects specification. The implication of the differencing approach is that the individual effects are fixed nuisance parameters to be removed before estimation of the main coefficients proceeds. Thus, for a random-effects specification, differencing will be less efficient than GLS estimation for the same reason that the fixed-effect covariance estimator is inefficient.

Thirdly, the estimates of the differencing approach are less amenable to interpretation if the coefficients are allowed to vary over time. Consider

$$y_{it} = x_{it}\beta_t + \alpha_i + \lambda_t + u_{it} \qquad \textbf{(2.30)}$$

Estimation of this equation by differencing leads to the estimator

$$\tilde{y}_{it} = \tilde{x}_{it}\beta_t + x_{it-1}\tilde{\beta}_t + \tilde{\lambda}_t + \tilde{u}_{it} \qquad \textbf{(2.31)}$$

where $\tilde{y}_{it} \equiv y_{it} - y_{it-1}$ and the other variables are similarly defined (see equation (5.129) for details). For the slope coefficients both levels and differences in the coefficients are being estimated, and with a little manipulation all the slope coefficients are identifiable. However, only the change in the time intercept is being estimated, and so only the relative values of the constant (relative to either λ_0 or λ_T) can be identified from the estimates. The levels of the intercepts can of course be recovered from the equations by calculating the predicted means of the regression for one period but the fact that the returned coefficients are changes rather than levels seems to be ignored by most authors[11]. Section 5.4 discusses this issue and a restricted approach which falls between the estimators of (2.29) and (2.30).

2.4 Non-linear models

Allowing for individual heterogeneity has serious disadvantages when a non-linear functional form is specified. Because of the additive nature of the linear equation, estimation of the

[11] For example, the GAUSS program DPD returns "intercepts"; the fact that only T-1 are returned indicates that these are actually changes in the intercept.

main coefficient vector and the panel effects could be separated, for example by covariance or GLS estimation. Under a non-linear specification, this may no longer be possible.

Take a more general form of (2.5):

$$y_{it} = F(x_{it}, \beta, \mu, \alpha_i, \lambda_t) \qquad (2.32)$$

where F(...) is some function non-linear in the parameters. As mentioned earlier, if the panel terms to be fixed coefficients, these will only be estimated consistently if the right "dimension" of the panel is large; that is, a consistent estimate of a time-specific effect requires large N, and a consistent estimate of an individual-specific effect needs large T.

Assuming a typical panel structure with T small and N large, consistent estimates of λ_t are possible but not α_i[12]. In a linear regression with additive terms this does not affect the consistency of the main parameter estimates. However, (2.32) involves maximising the joint probability of all the parameters, and so inconsistent estimates of the incidental parameters will lead to the main parameters being inconsistently estimated (see Hsiao (1986) pp159-161 for an example).

This problem was considered in some detail by Neyman and Scott (1948), who suggested finding alternative functions for the main parameters which are independent of the incidental parameters[13]. Hsiao (1986) illustrates some simple cases, but notes that in general such functions are difficult to find. Perhaps most importantly, there does not seem to be one for the probit model; that is, there is no consistent estimator for small T for a fixed individual-effect probit model.

---

12 This is for the same reason as in the linear case; namely, that there are insufficient observations on each individual, but there are a large number of observations for each period.

13 An alternative under investigation by the author is to use linear approximations to the non-linear functions; for example, replacing a probit or logit form with a linear probability model.

An alternative is to use a random-effects estimator, leading to a multivariate non-linear regression with the likelihood function of the main parameters augmented by the marginal distribution of the random effects. The random-effects specification does at least provide consistent estimators, and simplifying assumptions can be made to reduce some of the complexity of the estimators (see Hsiao (1986) pp164-167). It has the obvious disadvantage of needing a specific distribution for the random effect. There is also the potential for correlation between the random effects and the explanatory variables, which is less easy to resolve than in the linear case: using the linear specifications of Mundlak (1978) or Chamberlain (1984) may impose unwarranted restrictions on a non-linear model.

Thus, fixed-effects estimators are relatively simple but may not provide consistent estimates of any parameters. In contrast, random-effects estimates are consistent, but only as long as assumptions on the distribution and independence of the effects is justified.

2.5 Dynamic panel models

Dynamic panel models can be extremely informative. Apparent dynamic effects could result from heterogeneity, serial correlation, or (in the case of non-linear models) state dependence. Consider finding that a candidate who has passed one exam is more likely to pass another. This could be because candidates differ in their abilities (heterogeneity); because those who pass have acquired (randomly) some extra knowledge which is useful in both exams (serial correlation); or because those who pass avoid resits and so have more time to study for remaining exams ("true" state dependence). If a dynamic model can be constructed, then testing for combinations of these various forms of true and spurious state dependence is a possibility.

2.5.1 Linear models

Dynamic models for panels are more complex than static ones. The reason is the initial condition of the dependent variable. Let

$$y_{it} = x_{it}\beta + y_{it-1}\gamma + \mu + u_{it} \qquad u_{it} = \alpha_i + \lambda_t + \epsilon_{it} \tag{2.33}$$

y_{it} and y_{it-1} depend on α_i, which implies that the initial condition y_{i0} does too - and, potentially, so does the entire pre-sample history. The problem is to remove or control for the individual-specific effect without introducing correlation between the dependent variable and the error term. For example, taking first-differences of (2.33) leads to

$$(y_{it}-y_{it-1}) = (x_{it}-x_{it-1})\beta + (y_{it-1}-y_{it-2})\gamma + (\lambda_t-\lambda_{t-1}) + (\epsilon_{it}-\epsilon_{it-1}) \tag{2.34}$$

where the individual effect has been removed, but the error term is now correlated with the explanatory variables as $E(y_{it-1}\epsilon_{it-1})\neq 0$.

Consistent ML and GLS estimates may be available for a range of assumptions about the initial conditions. However, a relatively simple and popular solution to this problem is to difference the model, as in (2.34), and then use instrumental variables estimation. The panel structure itself provides instruments in the form of lagged or lagged-differenced dependent variables, an instrument set which grows over time as more lagged variables become available (Arellano and Bond (1991)). Unbiased and consistent estimators therefore exist for both the fixed and random effects models.

The dynamic specification and estimators have received some criticism. The use of lagged dependent variables as instruments has been criticised on the grounds of a poor correlation over the long lags necessary to instrument differenced models; adding more and more instruments a la Arellano and Bond will not necessarily improve estimates but makes estimation more complex. More fundamentally, it has been suggested that variable-coefficient models which take account of cross-period correlation a more general way, such as (2.8) or (2.30), offer both flexibility and consistency without the need for instrumentation; for example, Sevestre and Trognon (1992) use Chamberlain's (1984) estimator as an alternative

route to a dynamic specification. Apparently dynamic models may arise from a number of causes including misspecification of an underlying static model (see Raj and Ullah (1980)), and to some extent these arguments can be seen more as a debate over the value of estimating "structural" rather than "reduced form" parameters.

Clearly, the variable-coefficient approach is also open to criticism about the validity of its assumptions. This thesis does not intend to discuss the relative merits of these two approaches. Both IV estimates of (2.34) and more general variable-coefficient estimators are available under the software to be described later. However, most of the applied work on the NES by the author and others has utilised cross-sectional analysis or the variable-coefficients approach.

2.5.2 Non-linear models

Estimation of the linear dynamic model is relatively straightforward. This is not the case for non-linear specifications. Consider the dynamic non-linear form:

$$y_{it} = F(x_{it}, y_{it-1}, \beta, \mu, \alpha_i, \lambda_t) \qquad (2.35)$$

where F(...) is once again some arbitrary non-linear function. In this case panel estimators run into severe difficulties. The problem, as for the linear dynamic model, is the determination of initial conditions, but it has been complicated by the non-separability of the terms in the model as discussed in section 2.4. If the parameter estimates are jointly calculated, then estimates of the initial conditions also need to be jointly calculated. Unlike the estimators in section 2.5.1, the incidental parameters can no longer be divorced from the ones of interest - and therefore estimation of the initial conditions (and possibly pre-sample history) must be included in the maximisation procedure.

On the assumption that the panel effects are fixed, the earlier conclusion that the ML

estimates are inconsistent as long as T is small still holds. Moreover, Monte Carlo tests by Heckman (1981b) suggest that this inconsistency seems much more significant than for the static case discussed above.

The random-effects assumption needs information in y_{i0}, just as for the linear case. Unless y_{i0} is assumed to be independent of α_i, the marginal distribution of α_i needs to be integrated over y_{i0} as well; therefore information on y_{i0} is needed. And if y_{i0} is to be calculated, this may require information on y_{i-1}, y_{i-2}, etcetera if they too depend on α_i. Solutions to this issue have been suggested, but they are all unsatisfactory for one reason or another[14]. At the time of writing there appears to be no general consistent estimator for dynamic non-linear models.

2.6 General estimation techniques

In recent years there has been some interest in more generalised estimation methods. Two of these techniques are briefly considered here.

2.6.1 Generalised method of moments (GMM)

GMM developed in the early eighties as a unifying approach to a wide range of problems[15]. Its name comes from estimators derived by minimising a set of moment conditions in a quadratic form. As such it shares characteristics with 'traditional' methods such as OLS, but the minimisation criterion is specified in such a way that wide range of problems may be treated within the same overall framework. Thus OLS, linear and non-linear IV, GLS,

---

[14] Suggestions including assuming that the initial conditions are truly exogenous and independent of α_i, or assuming that the pre-sample process is in equilibrium. Both these assumptions are very strong and hard to justify in applied work. Heckman (1981b) drops the pretence of consistency altogether and instead offers relatively practical approximating solution.

[15] Hall (1993) and Ogaki (1993) provide comprehensive surveys of GMM techniques and applications.

etcetera can be seen as restricted versions of the same basic estimator. A properly-specified GMM estimator is consistent and efficient.

GMM has found a strong foothold in the estimation of time series models. In panel models its main application has been in the area of linear dynamic models, particularly in the methodology of Arellano and Bond (1991). Their estimation program, DPD, uses the GMM nomenclature in allowing for a range of dynamic specifications and has had a notable impact in some areas of econometrics.

Because GMM is largely a unifying terminology until the appropriate functional forms are specified, GMM as an estimation method per se will not be pursued. The techniques in this thesis can be seen as particular forms of GMM estimators (for example, the linear IV estimator and Sargan's test are expounded using the GMM terminology), but this is a methodological issue. All of the methods used have been chosen for their practicality without reference to an overall framework.

2.6.2 Minimum-distance estimation (MDE)

Chamberlain's (1984) minimum-distance estimator is a general estimator in that it allows for a variety of specifications. While GMM and MDE both emphasise flexibility in the functional forms, MDE aims to provide robust and efficient estimation of unknown specifications rather than an efficient technique for known problems.

The starting point is the recognition that a random individual-specific effect could be correlated with any or all of the explanatory variables, and therefore a proper specification for a random-effects model should be

$$y_{it} = x_{it}\beta + Z_i'\zeta + u_{it} \quad (2.36)$$

where Z_i is a 1xKT vector of all the x variables, as in (2.8). Moreover, the variance of u_{it} should be generously specified given the potential for heteroscedasticity and (particularly) autocorrelation in a panel dataset. Given these requirements, Chamberlain recommends estimating T separate equations of the form

$$y_{it} = Z_i'\pi_t + u_{it} \quad (2.37)$$

Define $\Pi' = [\pi_1' \ldots \pi_T']$. Then, for a general specification of $Var(u_{it})$, the separate estimates of π_t are asymptotically normally distributed with mean Π and variance given by the covariance matrix of these separate estimates. This information is then used in a second stage regression to impose restrictions on the structure of Π in a manner analogous to that of simultaneous equation systems or GMM methods.

The slope coefficients in (2.36) may be allowed to vary over time, which makes the second-round estimates more complex but otherwise has little effect on the technique. This approach can also be applied to the non-linear counterpart of (2.36), with appropriate adjustments. However, the linear specification of the panel effect does imply a restriction on non-linear models, which may or may not be warranted.

MDE is very flexible as it places no restrictions on the error terms; it therefore has relatively low information requirements and is robust to the specification of u_{it}. However, the price paid is that MDE is only efficient within a class of estimators imposing no restrictions on the error term. Any prior knowledge about the true distribution of the error term implies a more efficient estimator exists.

A second difficulty with MDE is the large number of parameters to be investigated as T grows. This is a particular problem if N is relatively small or K large. Clearly, MDE is also unlikely to be attractive for random time-effects which would involve a matrix inversion of

order NK. However, the variable-coefficient fixed-effects estimator to be described in Chapter 4 shares many features with MDE (including the data requirement), and it seems likely that MDE is a practical option[16].

2.7 Summary

As discussed in section 2.1, the panel structure allows for a much richer range of models and estimators. However, it is apparent that the usefulness of the more complex panel models is restricted by the feasibility of the estimator. This is especially true of non-linear models. The rest of this thesis is concerned with static, linear models, as estimators for these have been implemented in the analysis program to be outlined in chapters 5 and 6. The rationale for this decision is examined in chapter 4.

16 The possibility of implementing this estimator in the NES analysis software is currently being investigated.

Chapter 3

The New Earnings Survey and the NES Panel Dataset

3.1 The history of the NES

The New Earnings Survey is a product of the UK Department of Employment (DE). It was originally a one-off survey of the workplace conducted in 1968, which became a yearly survey from 1970. In the early years the information was collected by the Department of Health and Social Security from the National Insurance (NI) deduction card records. In 1975 the survey was expanded, the sample selection method was changed and the responsibility for data collection was transferred to the Inland Revenue and their tax records. As there was only a 25% overlap between the participants in the 1974 and 1975 surveys the NES has been treated as a new survey from 1975.

The survey is composed of all those in employment (apart from the self-employed and some other exemptions) whose NI numbers end with a particular two-digit code. It has been held on computer since its inception. Each participant is identified by a search of the Inland Revenue's PAYE records. The identification code has been the same since 1975, and therefore individuals who remain in work will make a number of appearances in the survey. This fact was largely ignored during much of the life of the NES, its only useful contribution being a loose check on the integrity of data being entered; but in the early 1980s it was realised that the DE had the makings of a remarkable panel dataset in its computer banks. A decision was made to restructure many years of isolated survey data into a single integrated dataset which could be used for longitudinal studies. This was duly completed by the end of the decade.

The published NES remained unaffected by these changes, but researchers wishing to have more detailed information on hours and earnings could contact the DE with particular queries.

This was very much on an ad hoc basis.

With the completion of the first panel dataset (covering the years 1975-1990), it was decided that public access to the data was to be encouraged in order to make the best use of this new resource. This access to this data is limited for a number of reasons to be outlined in the next chapter, but it has generally taken one of two routes. One method is to use SPSS: a copy of the dataset is held in SPSS format and the DE will run SPSS programs on the data, returning the results of regressions or cross-tabulations to the researcher. The alternative route was taken by the research team at Stirling University, in association with a number of other researchers. A second copy of the dataset is in the format of the GAUSS matrix language[1]. Programs written in GAUSS extract aggregate data which can be removed from the DE and analysed at leisure. The core of this dissertation focuses on two aspects of this, the cross-product matrix and the observation history; see chapters five and seven.

3.2 Survey contents

The survey data is split into "fixed" and "variable" sections. The variable section holds information which is recorded every year: general information about hours, earnings and the employer. These questions are the same very year, and an individual will have one record for each year of the survey in which he appears[2]. These repeated questions concern:

- components of pay (weekly; basic, gross, overtime, shift, bonuses; affected by absence)
- hours of work (actual and normal)
- age and sex (which is surprisingly variable, hopefully due to coding errors...)
- occupation (in KOS, SIC and manual/non-manual breakdowns)

---

1 The raw NES data is held in ASCII format. The fact that copies are held in the SPSS and GAUSS formats reflects the uses made of the data, not any limitation of software or hardware.

2 The physical split of the data into fixed and variable files does not correspond exactly to the breakdown used here.

- time in the job (twelve months or more)
- location and business of employer
- coverage by Wages Board or collective agreement
- full-time or part-time
- sector (private, government, public corporation)

The repeated information thus contains a wealth of information on the workplace, allowing for some detailed analysis of work trends. The information available in the dataset is much richer than the published figures suggest; for example, location of employer is coded down to town/district level, although only county/regional figures are given in the published Survey.

In addition to this, occasional questions are asked. These are the "fixed" fields, as they are only requested once or are seldom repeated. There are a variety of reasons for asking these extra questions. Some are requested by researchers, some by the DE, and some by other bodies; for example, several extra questions were asked in 1979 on behalf of the Statistical Office of the European Community.

Information collected has included data on tenure, collective agreement, holiday entitlement, training, company size, and so on. This information is of limited use in panel studies (time-invariant variables are indistinguishable from individual heterogeneity in differencing or means-deviations estimators), but they have been used successfully in a number of cross-section studies; for example, Coleman (1994) used the occasional questions on tenure for an analysis of the relationship between tenure, sector and wages.

An obvious omission from this list of variables is any personal information about the employee other than age and sex. The NES contains no data on ethnicity, education, family background, et cetera. This is one of the major flaws in the NES, and particularly relevant to cross-sectional studies. To some extent, a panel analysis allowing for individual

heterogeneity will reduce the impact of these influences, as they are generally time-invariant: individuals tend to complete their general education before commencing a career (see Elliott (1991); Dolton and Kidd(1994); Vella (1994), for example).

This is not a very satisfactory solution, as the panel analysis is being used to control for an "unmeasurable" effect which is clearly measurable in a more general sense (even if some of the measures used, such as ordered dummies for levels of education, are of debatable worth). An effective valuation of education could lead to much better cross-sectional analysis and more informative panel studies if the education variable changes over time. This is even more true of family background. Some authors have argued that, for example, earnings and participation of women are affected by the number and age of any children (Dolton and Makepeace (1987); Elias (1988); Elias and Main (1982); Joshi and Newell (1985)). Both of these may vary over time, and a panel analysis will have little more success than cross-sections in controlling for this unmeasured effect.

Some of the information on employers is also relatively limited: details of company size, establishment size, number of employees et cetera is only available for particular years. This information cannot be subsumed into individual heterogeneity unless the individual always works for companies of a similar type, although it could be argued that, once occupation, industry, region, sector and so on are taken into account, the remaining inter-company differences should be small.

3.3 Survey coverage and missing data

By using a two-digit NI code for identification, the DE hoped to achieve a one percent sample of the labour force in employment. In fact, the overall participation rate is around 70% of this level (using Labour Force/General Household Survey estimates of the workforce). The survey forms are sent to the individual's last recorded employer, who is obliged to

complete the form under the 1947 Statistics of Trade Act. Return rates are typically 95% or better, but not all of these returns contain usable information.

The missing data is due to a number of causes. Firstly, rather less than the full number of forms is sent out. Some employers are exempt, principally the armed forces or the self-employed. In theory, those earning insufficient amounts to pay NI or tax should have no records and so be left out of the survey; but because tax records are held over, even if no tax is paid in a particular year, rather more are included than might be expected[3]. There is also the possibility of the employer having changed address or name. Adams and Owen (1989, Table 1) report that, over the period 1975-1986, only around 90% of the desired number of forms are sent out.

When the forms are returned, about 80% are usable. Of the missing individuals, a number have moved "out of scope", into occupational pensions or unpaid work. However, by far the largest number of unusable responses is due to employers replying that the employee in question no longer worked for them (Adams and Owen (1989)). This could be due to a refusal of employers to co-operate, but it is more likely to be due to unemployment or a change of jobs. These last two arise because the form is sent to the last recorded employer, and there is a lag in updating records when a change of status occurs.

This missing data is a source of some concern. About 240,000 men and 190,000 women have appeared in the NES over the period 1975-1990[4]. Almost all of these have some missing observations, and around one-fifth have only one observation. Some 99,000 men and 56,000 women joined in 1975. Of these, only 9,200 men and 3,300 women have a complete set of observations up to 1990.

---

3 Studies by Bob Hart and Elizabeth Roberts on the micro-data do indicate that a substantial number of people earning below the NI limit are included in the NES.

4 All the analysis in this thesis is based on the 1975-1990 dataset, as later versions are not yet available.

If the data are missing randomly (that is, observability does not vary systematically with the variables of interest in a study), the net effect of this will be to reduce the precision of estimates but not to invalidate them. Recent studies do not indicate that random attrition is the case. Discussion on the quality of the data in the NES is very sparse, with barely six papers in twenty years[5]. Bell and Ritchie (1993b, 1994) claim that non-observation is correlated with almost every variable in the dataset to some degree, although these findings are largely descriptive and so the magnitude of this correlation is difficult to assess.

The question of missing data is an enormous issue, but is largely ignored in the literature. This thesis follows the trend and avoids the issue too[6]. This is due to the difficulty of constructing realistic consistent dynamic non-linear models, as discussed in chapter two. However, chapters nine and ten on applied earnings analysis do attempt to allow for the missing data problem. As Heckman-type corrections for panels are restrictive in their assumptions and inappropriate for the NES, an ad hoc approach using proxy variables is taken.

3.4 Validation and measurement errors

An issue related to that of missing data is measurement error. Much of the NES data is in the form of categorical variables, and it might be assumed that this data is reasonably accurate. However, the measurement of hours and earnings needs to be re-examined.

The DE does not carry out separate validation checks on the NESPD. Instead, the only works so far to carry out a comprehensive comparison of the NES and an independent data source (the FES and other household surveys) are the papers by Atkinson, Micklewright and Stern

5 Atkinson, Micklewright and Stern (1981, 1982); Micklewright and Trinder (1981); Adams and Owen (1989); Bell and Ritchie (1993b, 1994). Adams and Owen is an in-house report by the DE. Bell and Ritchie (1993b, Appendix) survey all four.

6 An investigation into the problem of attrition in panels is being carried out by the author.

(1981, 1982)[7]. The most important finding is that hours for non-manual workers are consistently lower in the NES than those reported by the household surveys. The reason is probably due to the fact that the household surveys ask employees what their "normal hours" are; in the NES the employer is asked. The NES response is significantly more concentrated around a standard working week of 38-40 hours, while the household surveys report higher normal hours on average. The suggestion is not that employers under-report their employees' hours, but that they may not have any clear idea about the normal hours for non-manual workers.

The importance of this for estimation depends on the model studied. Models of labour supply based on individual utility may be biased if an employee's perception of working time differs significantly from the employer's impression as recorded in the NES. Similarly, studies of labour hoarding and employment on the intensive margin may be distorted by an apparently arbitrary hours measure. Moreover, the bias imparted by the use of a standard measure of hours may vary over observed characteristics: for example, it is more likely to be significant for workers paid on weekly or hourly rates than for salaried employees.

Atkinson *et al*'s comparison of earnings in the NES and the FES does not produce any clear results, and it may be expected that employers' reporting of earnings to the tax office is reasonably accurate. However, the potential for error in the hours worked will lead to error in the hourly wage rates reported by the NES, as these are calculated as reported earnings over reported hours. This increased measurement error in hourly earnings has already been noted for one major US dataset, the Panel Study of Income Dynamics (Bound, Brown, Duncan and Rodgers (1994), using a follow-up Validation Survey).

---

[7] The review of Atkinson *et al* gives some guidelines as to the accuracy of the NESPD but has some drawbacks. Firstly, the authors did not have access to the NES micro-data and so had to rely on published aggregates for NES numbers. Secondly, the various datasets did not record the same information and so only a limited analysis of the categorical variableswas possible. Most importantly, their analysis concentrated on the years 1971-1977 and seemed to indicate a change in the NES after the 1975 change in administration, and so extrapolating these results to the post-1975 NESPD may not be justified.

It is a well-known result (for example, Johnston (1984)) that measurement error in the dependent variable will reduce the efficiency of an estimator; and that measurement error in explanatory variables lead to biased and inconsistent OLS estimates of the true coefficients. The effect of measurement error may be insignificant (and indeed is often assumed to be). However, in panel models these errors become more important because of the nature of the estimators used. All the linear estimators in chapter two (including the differencing estimators) involve converting the data to deviations from some level. If the observed variables have little variation, then these transformations may increase the measurement errors relative to the observed variables. In other words, the signal-to-noise ratio decreases, and bias and/or inefficiency may increase (see Biorn(1992); Bound *et al* (1994); Hsiao(1986)).

This has recieved most attention in the area of union status, where it can be shown that measurement error leads to an underestimate of the union effect (Freeman (1984)). This has led some authors to argue that cross-sectional estimates may be better than longitudinal studies: while the covariance estimator is in general more efficient than the cross-sectional estimator, in the presence of significant measurement error this may not be the case (see Freeman (1984); Card (1994); Booth (1995, p176) for example).

The estimates to be presented in chapters nine and ten do not include hours or earnings as explanatory variables, and so are free of error from this source. The dependent variable of log hourly earnings is liable to be subject to some measurement error, but earnings do vary over time, and so the efficiency loss due to measurement error and the covariance transformation should be small. The problem of measurement error in union status is potentially serious; however, because the relevant variable is whether an individual is affected by any collective agreements (not whether the individual is a union member), and because this information is provided by the employer, it may be justifiable to assume that the classification error is small enough to be ignored.

Chapter 4

The econometric problem

Since the start of the decade, researchers have been granted general access to the individual data underlying the NES. This gives rise to two difficulties. Firstly, the size of the dataset is a very significant drawback to using it. Secondly, this public access is limited by various confidentiality restrictions. This chapter discusses these problems and outlines the solution path taken.

4.1 The size thing

Econometric packages calculate the data needed for each particular regression whenever a regression is run. Data is held as a set of vectors to manipulated in appropriate ways. This has several advantages. It clearly allows much more flexibility in the type of regression allowed than if the data was held in an aggregate form, such as a cross-product matrix. It also makes the creation of new variables simple, including predictions for instruments. It allows for the analysis of the residuals and the creation of estimated covariance matrices. Finally, non-linear models and solution methods are feasible.

Given the speed and capacity of modern computers, holding raw data and manipulating it at will is a sensible way to approach most datasets. However, the NES panel dataset is very large. The raw ASCII files for the first sixteen years are some 600Mb in total, and conversion to binary format does not reduce this as most of the variables are qualitative and so take small values. Reading the data can take some hours.

This causes problems for standard statistical analysis. First of all, some packages are not able to cope with such a large dataset, or only with extreme hardware requirements. However, even if the admissible size of the raw data is unlimited, running regressions becomes an

extremely tedious and time consuming process. Misspecification of an equation is heavily penalised, if the regression has to be run more than once; and so is "exploratory" analysis, where the researcher would hope to try a variety of specifications.

This limits severely the advantage of traditional econometric packages in analysing the data. While SPSS has been used by a number of researchers to create cross-tabulations, running regressions is still a slow business. As the matrices used in a standard analysis are created anew for every regression, each estimate is likely to involve a complete run through the data. If the program automatically creates diagnostic statistics involving residuals, then a second run will have to take place.

This is the only practical solution for non-linear estimators, those requiring numerical optimisation, or those which need several steps. However, for one-stage, linear estimators (in other words, OLS), this is hugely inefficient. This is because linear combinations of variables can be stored very compactly in aggregate form and manipulated to produce further linear combinations. The implications of this are considered in section 4.3.

4.2 The confidentiality thing

While access to the data is complete, the amount of information which can be taken "outside" the DE is not. The NES contains information on individuals, their work histories, and their wages, and is subject to, among others, the 1947 Statistics of Trade and 1986 Data Protection Acts. The legal position of the DE is that no information can be removed from its premises which would allow the wages of an individual to be identified. This applies to both computer media and printed materials[1].

[1] In the mid-1980s the DE constructed a small dataset by aggregating individuals into cells of three and allowed general access. However, although this got round the confidentiality restrictions and led to some analysis, it did not prove as popular as the DE hoped and was dropped in favour of the panel dataset.

This means that the NESPD must remain on the DE's computers only. The practical upshot is that standard regressions (or almost any statistical analysis using standard packages) have to be run at the DE's head office by DE staff or visiting researchers. This leads to time being wasted, either by DE staff being recalled from other duties to service the dataset, or by researchers needing to travel to London. It also limits the scope for exploratory analysis by introducing a further delay between requesting information and receiving results.

Finally, no information on residuals, for example, or other disaggregated data can be removed. Therefore, diagnostic tests and second-stage regressions have to be run at the DE too. Although researchers can be given summary statistics from running the regression, any further analysis has to return to the DE for processing.

4.3 Outline of the solution methods

The main solution taken at the University of Stirling is, as hinted at earlier, to create linear combinations of the variables and then to analyse them using programs devised for this purpose. This method is extremely efficient because OLS is a simple multiplication of two summation terms. If the sums of several variables are calculated at the same time, then multiple regressions can be run simply by multiplying the appropriate sums together.

Consider the simple OLS regression:

$$y = X\beta + u \tag{4.1}$$

with the normal equations

$$\hat{\beta} = (X'X)^{-1}X'y \tag{4.2}$$

and the calculation of errors which needs

$$TSS = y'y \qquad ESS = \hat{\beta}'X'y \qquad RSS = TSS - ESS \tag{4.3}$$

Just three matrices, X'X, X'y, and y'y, are sufficient to estimate β and calculate R^2, σ^2, and the standard error of the coefficients. Define a vector W≡[X y]. Then the moment matrix W'W contains all the information needed for the OLS estimation of (4.1):

$$W'W = \begin{bmatrix} X'X & X'y \\ y'X & y'y \end{bmatrix} \tag{4.4}$$

Suppose X and y are NxK and Nx1 matrices respectively, so that W is a Nx(K+1) matrix. W'W could be created as the moment of the W matrix; however, W'W could also be constructed without having to create the W matrix, for W'W is merely a sum of the moments of each individual row w_i of W:

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix} = \begin{bmatrix} x_1 & y_1 \\ x_2 & y_2 \\ & \vdots \\ x_N & y_N \end{bmatrix} \Rightarrow W'W = \sum_i^N w_i'w_i = \begin{bmatrix} \sum x_i'x_i & \sum x_i'y_i \\ \sum y_i'x_i & \sum y_i'y_i \end{bmatrix} \tag{4.5}$$

Therefore W'W can be created without any need to store anything bigger than the (K+1)x(K+1) moment matrix.

Thus, the fact that N is enormous for the NES is no longer a restriction. The size of the cross-product matrix to be calculated depends only on the number of variables, not the number of observations. Moreover, the moment matrix is flexible in its definition of variables. The choice of X and y as explanatory and dependent variables, respectively, is for notational convenience. As far as the moment matrix is concerned, there is no difference between the two.

Suppose W is an Nx4 matrix:

$$W = \begin{bmatrix} 1 & x_1 & y_1 & z_1 \\ 1 & x_2 & y_2 & z_2 \\ \vdots & & \vdots & \\ 1 & x_N & y_N & z_N \end{bmatrix} \tag{4.6}$$

Then W'W is a 4x4 matrix

$$W'W = \begin{bmatrix} N & \sum x_i & \sum y_i & \sum z_i \\ \sum x_i & \sum x_i x_i & \sum x_i y_i & \sum x_i z_i \\ \sum y_i & \sum y_i x_i & \sum y_i y_i & \sum y_i z_i \\ \sum z_i & \sum z_i x_i & \sum z_i y_i & \sum z_i z_i \end{bmatrix} \tag{4.7}$$

This gives a range of possible regressions to run. Any of x, y, or z could be used as the dependent variable with any or all of the others as explanatory variables, including constants if so wished. For example, to regress x on y only requires

$$\hat{\beta} = \left(\sum y_i y_i\right)^{-1} \sum y_i x_i \tag{4.8}$$

while the regression of z on x, y, and a constant gives the estimate

$$\hat{\beta} = \begin{bmatrix} N & \sum x_i & \sum y_i \\ \sum x_i & \sum x_i x_i & \sum x_i y_i \\ \sum y_i & \sum y_i x_i & \sum y_i y_i \end{bmatrix}^{-1} \begin{bmatrix} \sum z_i \\ \sum x_i z_i \\ \sum y_i z_i \end{bmatrix} \tag{4.9}$$

Even the simple 4x4 matrix in (4.7) allows for twenty-one possible regressions (excluding using the constant as a dependent variable). Moreover, adding further variables to this matrix involves relatively little extra space. In this compact form, an extra variable increases the size of the matrix by 2K+1 elements, where K is the previous number of variables. In contrast, adding a new variable to be stored in raw form requires the space to store an extra N observations, and in the NES N could be almost three million.

The efficiency of this approach is due to the fact that each moment matrix contains the information for a number of regressions. If a moment matrix was created for the purposes of running one regression only, then this method has no significant advantage over other

solutions in producing the estimates. However, when several regressions are to be run then the time saving of this method is large. The time taken to create moment matrices containing more variables is small relative to the time saved when regressions are run. Moment matrices calculated for the University of Stirling typically include between sixty and one hundred variables for each year, and experience has shown that there is relatively little difference in the time taken to create these matrices[2].

This is especially important if two-stage methods are being considered. Suppose predicted values are to be included in the regression as instruments. Which predictions are to be used? In this case, the solution is to include all the various predicted values in the moment matrix, as the time and space cost is usually small. Then all the predicted values are available for use at estimation time.

A second advantage of this solution relates to the confidentiality issue. Although these cross-product matrices can be used to run OLS regressions as if they contained disaggregated data, because the data is actually aggregated the matrices can be taken out of the DE and used by researchers at their own institutions[3]. This gives external researchers the chance to do their own analysis, rather than having to inform the DE of their instructions. It also relieves the pressure on the DE's staff resources.

Finally, the cross-product matrix contains useful information other than that needed to run regressions. For example, the W'W matrix in (4.7) could be used to discover, amongst other things, the number of observations N, the mean of all of the variables, the variance and covariances of the variables, and hence the correlation between variables.

---

2 The time to create the fixed-effects estimators of sections 5.2 and 5.3 is much greater, but this is because the moment matrix is of size TK, rather than K, and T is the main factor affecting speed.

3 Some confidentiality checks have to be made in case some individuals can still be identified, but the effect of this is negligible when the full dataset is used.

In this chapter a simple OLS example has illustrated the main method of analysis used in this regression. In fact, a number of estimators can be estimated from properly constructed moment matrices, and the next chapter describes those currently available. Analysis of the moment matrix requires some particular software, described in chapter six. However, some standard statistical packages (for example, STATA) can work with moment matrices instead of raw data, and so the software supplied by Stirling University is not a necessity for analysis.

One significant disadvantage of this method is that only a limited amount of diagnostic testing can be carried out. This is because the residuals are not available for analysis. For example, to construct White standard errors, requires the product of squared residuals and explanatory variables; Breusch-Pagan test statistics for heteroscedasticity involves the recreation of an X'X matrix with the residuals calculated and added back in. Even carrying out relatively straightforward tests such as these becomes a major invesment in time and energy. To some extent the X'X approach mitigates this by the ability to collect the information for many diagnostic tests in one go (for example, having run several regression, the appropriate matrices for all of them can be calculated in one pass of the dataset), but this is clearly less satisfactory than having the results to hand immediately. This is primarily a consequence of the confidentiality restrictions placed on the data, but, ironically, the very quantity of data available makes diagnostic testing rather less appealing. The results presented in chapters nine and ten are thus presented without detailed specification tests other than those available from the X'X matrix[4].

Two other forms of analysis have been applied to the NES by the researcher team at Stirling. One is the construction of cohort matrices, containing information on cohorts of people of particular ages and in the panel at particular times. The second is the construction of observation histories, to be described in further detail later in chapter seven. Some other

4 Analysis of regression residuals is being carried out separately.

researchers have also made use of cross-tabulations (cross-product matrices in all but name), which satisfy the confidentiality requirement and so can be removed from the DE. All these are also subject to confidentiality checks.

The cross-product route is not appropriate for non-linear models or solution methods, and although there is some non-linear analysis currently being carried out, this is necessarily limited to a sample of the NES[5]. Some interesting results have been achieved from linear approximations to the non-linear estimators, and further work on this is under way. However, it seems that non-linear analysis exploiting the full potential of the NES remains impractical for the time being.

4.4 Varying coefficients and fixed-effects

The estimators used for the software package all allow coefficients to vary over time. They allow for fixed time effects, and the "fixed effects" and differencing estimators allow for fixed individual effects. The time-varying coefficients are "fixed" in a similar way to the fixed individual effect, in that the estimation of fixed (T sets of K) parameters rather than the characteristics of a distribution is involved.

The choice of a fixed-effects specification for individual heterogeneity is twofold. From a practical perspective, the two stages necessary for GLS estimation of a random-effects specification constitute a serious drawback. The covariance method is a single-stage effort and so is far more practical given the nature of the data. This does not stop random-effects models being estimated; however, the onus of calculating the first-stage component estimates, recreating the cross-product matrix and re-estimating is shifted onto the user.

5 Source: DE

There are also theoretical considerations for using fixed-effects, the most prominent being the potential for collinearity between the individual effects and explanatory variables. Given the self-selecting nature of many labour force decisions, the assumption that unchanging personal characteristics are not correlated with the choice of occupation, industry, and so on seems unlikely. For example, Vella(1994) argues strongly that a significant factor in job choice is due to something called "attitude". As was discussed in chapter two, these potential correlations can lead to the random-effects estimator being biased or inconsistent while the fixed-effects estimator remains valid. Although the fixed-effects estimator is relatively inefficient for random-effects models, the large number of individuals in the NES make the trade of efficiency for robustness appealing.

The choice of varying coefficients likewise has a practical and theoretical component. From a practical point of view, the time penalty for estimating cross-sectional models with varying coefficients as opposed to pooled CS models is relatively small. For the fixed-effects estimators the penalty is more significant; but the extra data requirements are relatively minor and the programming is straightforward. Therefore the practical cost of allowing for varying-coefficient models is not prohibitively high, and given the compelling theoretical arguments for allowing coefficients to vary over time, the trade-off seems reasonable.

Whilst many models have been developed which allow coefficients to vary over individuals (Hsiao (1992) surveys these), relatively few authors have considered coefficients which vary over time. Theoretical work and applied work with time-varying coefficients is almost non-existent; see Bell and Ritchie (1996) for a discussion. However, the assumed stability of the structure of labour market is open to doubt. Surveys of the UK labour market (for example, Robinson (1994)) lead to the overwhelming conclusion that the employment in the UK has changed considerably in recent years. To assert that these changes have not been reflected at a micro-level, without testing the alternative hypothesis of varying coefficients, seems an unjustifiable presumption.

Moreover, there are many reasons for parameters to vary over time even if the "true" model is time-invariant. Obvious examples are misspecified models, omitted variables, incorrect lag structures, aggregation bias, and so on; see Diechartz(1988) or Raj and Ullah(1981, ch 1) for a discussion. As a specific example, consider estimating a linear model where all parameters are time-invariant but different individuals are observed each year and there is a significant element of individual heterogeneity. Separate cross-sections will appear to show that the intercept varies over time as it is reflecting the average intercept for different individuals in different periods.

Given the lack of information about changes in the parameters of the market, imposing a particular structural form on changing coefficients (for example, evolutionary coefficients or random-effects about a stable mean) is unlikely to improve on the constant-coefficient assumption except by luck. Instead, the approach taken here is to allow for a different set of slope parameters for each year. This allows any changes in the labour market to occur in a relatively unrestricted way, by letting the responses of the population vary without reference to any specific structure. The default models therefore condition on time-varying slope coefficients in the same manner as the fixed-effects estimator conditions on individual heterogeneity. Just as a fixed-effects estimator is less efficient than a random-effects estimator, this slope conditioning is less efficient than a properly-specified model which takes account of the structure of coefficients over time (for example, models with systematic evolution of the coefficients). However, the lack of evidence on trends in the parameters of the labour market would seem to justify this approach for the moment[6].

6 Chapters eight and nine consider this issue of structural change in more detail, and at this point it may be noted that F-tests on varying-coefficients model comprehensively reject the hypothesis of parametric stability.

Chapter 5

Linear fixed-effects models: matrix creation

In this chapter two models, four estimators and the data matrices needed for them are defined[1]. The first model is the cross-section: although the estimator for this incorporates some panel aspects, it is little different from a dummy variable cross-section. The second (panel) model allows for individual heterogeneity. The three estimators for this model treat this as a fixed effect to be removed. The first panel estimator uses the covariance transformation. An alternative approach to the heterogeneity problem is time-differencing of the data. Section 5.3 and 5.4 consider differencing in balanced and unbalanced panels[2]. These are the differencing estimators.

For both the panel and cross-section model, three further specifications are considered:

unrestricted	Slopes and intercepts vary over time
pooled	Slopes and intercepts are constant over time
restricted	Slopes are constant, but intercepts may vary over time.

5.1 Cross-sections: the simple panel model

This model has no individual heterogeneity but allows for slopes and intercepts to vary over time. Models are estimated by a covariance transformation.

---

[1] This chapter involves a large amount of matrix algebra which is straightforward but extensive. Although the different models and estimators are defined by similar equations, the basic equations are described in some detail here as they are directly implemented in the software and so form part of the validation of the programs. A shorter version of this chapter is available as a discussion paper.

[2] The issue of panel balance does not materially affect the cross-section or fixed effects, although it simplifies the data requirements for the latter. As this can be done post extraction, we ignore the issue here and outline the adjustments in the next chapter. However, the balance of the dataset will determine whether the matrices for a differencing model have to be formed during extraction or whether they can be created post-extraction.

5.1.1 The unrestricted case

The "unrestricted" regression is

$$y_{it} = x_{it}\beta_t + \lambda_t + u_{it}$$
$$E(u_{it}) = 0 \qquad E(u_{it}u_{is}) = \sigma_{ts} \tag{5.1}$$

where x_{it} is a $1xK_x$ row vector, β_t is a $K_x x1$ column vector, and the other terms are all scalars. Stacked over all individuals for time t,

$$y_t = X_t\beta_t + J_t\lambda_t + u_t \tag{5.2}$$

where J_t is an N_t vector of ones and X_t is the $N_t xK_x$ matrix of the x_{it} stacked. Define

$$Q_t \equiv I_t - \frac{1}{N_t}J_tJ_t' \tag{5.3}$$

where I_t is the $N_t xN_t$ identity matrix. Note that

$$Q_t = Q_t' = Q_tQ_t' \qquad Q_tJ_t = 0 \tag{5.4}$$

Then premultiplying (5.2) by Q_t will remove the time effects:

$$\begin{aligned} Q_ty_i &= Q_tX_t\beta_t + Q_tJ_t\lambda_t + Q_tu_t \\ &= Q_tX_t\beta_t + Q_tu_t \end{aligned} \tag{5.5}$$

For the system of equations over all t and n, the equivalent of (5.5) is

$$PY = PZ\zeta + PU \tag{5.6}$$

where

$$Y \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \quad Z \equiv \begin{bmatrix} X_1 & 0 & & \\ 0 & X_2 & & \\ & & \ddots & \\ & & & X_T \end{bmatrix} \quad U \equiv \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_T \end{bmatrix} \quad \zeta \equiv \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} \tag{5.7}$$

and P is the system equivalent of Q_t, namely

$$P \equiv \begin{bmatrix} Q_1 & 0 & & \\ 0 & Q_2 & & \\ & & \ddots & \\ & & & Q_T \end{bmatrix} \tag{5.8}$$

P is also symmetric and idempotent. The OLS solution to this will be to minimise

$$U'PU = Y'PY + \zeta'Z'PZ\zeta - 2\zeta'Z'PY \tag{5.9}$$

which gives the normal equations

$$\hat{\zeta} = (Z'PZ)^{-1}Z'PY \tag{5.10}$$

The constituents of (5.10) are block-diagonal:

$$Z'PZ = \begin{bmatrix} S_1 & 0 & 0 \\ 0 & S_2 & 0 \\ & & \ddots \\ 0 & 0 & S_T \end{bmatrix} \quad Z'PY = \begin{bmatrix} R_1 & 0 & 0 \\ 0 & R_2 & 0 \\ & & \ddots \\ 0 & 0 & R_T \end{bmatrix} \tag{5.11}$$

where

$$S_t = X_t'Q_tX_t \qquad R_t = X_t'Q_ty_t \tag{5.12}$$

and therefore (5.10) is equivalent to T separate estimations of

$$\hat{\beta}_t = S_t^{-1}R_t = (X_t'Q_tX_t)^{-1}X_t'Q_ty_t \tag{5.13}$$

Note that

$$\begin{aligned} X_t'Q_tX_t &= X_t'X_t - \frac{1}{N_t}X_t'J_tJ_t'X_t \\ &= X_t'X_t - N_t\bar{X}_t'\bar{X}_t \end{aligned} \tag{5.14}$$

where

$$\bar{X}_t = \frac{1}{N_t}J_t'X_t = \frac{1}{N_t}\sum_i x_{it} \tag{5.15}$$

that is, the mean of x_{it} over all individuals for period t. Similarly,

$$X_t' Q_t y_t = X_t' y_t - N_t \bar{X}_t' \bar{y}_t \tag{5.16}$$

and

$$\bar{y}_t = \frac{1}{N_t} J_t' y_t = \frac{1}{N_t} \sum_i y_{it} \tag{5.17}$$

Create a matrix $v_i'v_i$ by summing over cross-products for individuals for a period t:

$$v_t = [X_t \;\; J_t \;\; y_t]$$

$$v_t' v_t = \begin{bmatrix} \sum_i x_{it}' x_{it} & \sum_i x_{it} & \sum_i x_{it}' y_{it} \\ \sum_i x_{it}' & N_t & \sum_i y_{it} \\ \sum_i y_{it} x_{it} & \sum_i y_{it} & \sum_i y_{it}^2 \end{bmatrix} \tag{5.18}$$

Clearly $X_t'X_t$ is the top-left corner of $v_t'v_t$ and X_ty_t the top right, but the matrix also contains all the other information necessary to calculate the terms in (5.14) and (5.16).

The OLS solution requires the summation of $v_t'v_t$ over i for each separate t. This is the format created by the extraction software.

Finally, note that the value of the time effects can be calculated from the mean for each period:

$$\begin{aligned} \frac{1}{N_t}\sum_i y_{it} &= \frac{1}{N_t}\sum_i x_{it}\beta_t + \frac{1}{N_t}\sum_i \lambda_t + \frac{1}{N_t}\sum_i u_{it} \\ \bar{y}_t &= \bar{x}_t\beta_t + \lambda_t + \bar{u}_t \\ \lambda_t &= \bar{y}_t - \bar{x}_t\beta_t - \bar{u}_t \\ \hat{\lambda}_t &= \bar{y}_t - \bar{x}_t\hat{\beta}_t \end{aligned} \tag{5.19}$$

This information is readily obtained from the cross-product matrix.

5.1.2 The pooled case

The "pooled" model constrains both slopes and intercepts to be constant over all periods:

$$y_{it} = x_{it}\beta + \lambda + u_{it} \tag{5.20}$$

Stacking over N and T:

$$Y = X\beta + J_{NT}\lambda + U \tag{5.21}$$

where $X = [\, X_1' \;\; X_2' \;\; \ldots \; X_T' \,]'$ and $NT = \sum N_t$. Premultiplying by the NT x NT matrix Q_{NT} (as defined above) gives

$$\begin{aligned} Q_{NT}Y &= Q_{NT}X\beta + Q_{NT}J_{NT}\lambda + Q_{NT}U \\ &= Q_{NT}X\beta + Q_{NT}U \end{aligned} \tag{5.22}$$

The time effect has been removed. The transformation matrix takes means over the whole regression, as all observations are treated alike. The OLS solution is

$$\hat{\beta} = (X'Q_{NT}X)^{-1}X'Q_{NT}Y \tag{5.23}$$

This differs from the unrestricted version in that the regressor matrices are no longer block diagonal and the summation is taken over the whole regression. This time the regressors split into

$$\begin{aligned} X'Q_{NT}X &= X'X - \frac{1}{NT}XJ_{NT}J_{NT}'X \\ &= X'X - NT\bar{X}'\bar{X} \\ &= \sum_t \sum_i x_{it}'x_{it} - \frac{1}{NT}\sum_t \sum_i x_{it}' \sum_t \sum_i x_{it} \end{aligned} \tag{5.24}$$

In other words, the mean is this time taken from the whole set of observations. Summing the raw cross-product matrix gives

$$\sum_t v_t'v_t = \begin{bmatrix} \sum_t \sum_i x_{it}'x_{it} & \sum_t \sum_i x_{it} & \sum_t \sum_i x_{it}'y_{it} \\ \sum_t \sum_i x_{it}' & NT & \sum_t \sum_i y_{it} \\ \sum_t \sum_i y_{it}x_{it} & \sum_t \sum_i y_{it} & \sum_t \sum_i y_{it}^2 \end{bmatrix} \tag{5.25}$$

and so the cross-product matrix once more provides all the information necessary to calculate the estimator. In this case the intercept is found from

$$\frac{1}{NT}\sum_t\sum_i y_{it} = \frac{1}{NT}\sum_t\sum_i x_{it}\beta_t + \frac{1}{NT}\sum_t\sum_i \lambda_t + \frac{1}{NT}\sum_t\sum_i u_{it} \qquad (5.26)$$

$$\hat{\lambda} = \bar{y} - \bar{x}\hat{\beta}$$

with the means taken over all variables.

5.1.3 The restricted case

In the time-specific intercept case with constant slopes (the "within" estimator) the model is

$$y_{it} = x_{it}\beta + \lambda_t + u_{it} \qquad (5.27)$$

To remove the time effect, stack over all individuals and premultiply by Q_t as for the unrestricted estimator

$$\begin{aligned} Q_t y_i &= Q_t X_t \beta + Q_t J_t \lambda_t + Q_t u_t \\ &= Q_t X_t \beta + Q_t u_t \end{aligned} \qquad (5.28)$$

For the system of TxN equations, the appropriate transformation matrix is P, above:

$$PY = PX\beta + PU \qquad (5.29)$$

but note that $X\beta$ is as defined for the pooled estimator, instead of the $Z\zeta$ in the unrestricted estimator. Again, the normal equations are

$$\hat{\beta} = (X'PX)^{-1}X'PY \qquad (5.30)$$

Unlike the unrestricted estimator, these terms are no longer block-diagonal; however,

$$X'PX = \sum_t X_t'Q_tX_t \qquad X'PY = \sum_t X_t'Q_ty_t \qquad (5.31)$$

and from (5.14) and (5.16) it may be observed that

$$\begin{aligned} \sum_t X_t'Q_tX_t &= \sum_t X_t'X_t - \sum_t N_t\bar{X}_t'\bar{X}_t \\ \sum_t X_t'Q_ty_t &= \sum_t X_t'y_t - \sum_t N_t\bar{X}_t'\bar{y}_t \end{aligned} \qquad (5.32)$$

Therefore, the elements of the regression in this case are the sum of those in the unrestricted

case after the latter has been adjusted to take deviations from time means (in the pooled case, the relevant matrices were summed before taking deviations). Thus the within estimator is also achievable from the $v_t'v_t$ cross-product matrix.

For the within case, the estimates of λ_t are given by

$$\frac{1}{N_t}\sum_i y_{it} = \frac{1}{N_t}\sum_i x_{it}\beta + \frac{1}{N_t}\sum_i \lambda_t + \frac{1}{N_t}\sum_i u_{it} \qquad (5.33)$$

$$\hat{\lambda}_t = \bar{y}_t - \bar{x}_t\hat{\beta}$$

that is, by taking means for each period.

This method of taking deviations from time means is a one-way analysis-of-covariance approach. Clearly the coefficients could also be estimated by using time dummies, so why bother taking deviations to remove these dummies? The main reason is that it simplifies testing the different specifications, as the tests are carried out on the same number of coefficients in all three estimators. A second reason is that the analysis-of-covariance method merely tests for constancy over time; using the standard F-tests in the time-dummy estimator tests for both constancy and a common level of the intercept in all three models[3].

5.1.4 Variances and testing in the simple model

Consider variances for the unrestricted estimator first. Defining e_{it} as the residual error and E its NTx1 system equivalent, then

$$e_{it} = y_{it} - x_{it}\hat{\beta}_t - \hat{\lambda}_t \qquad (5.34)$$

or

[3] A levels cross-section with time dummies included is available in the software, although this option does not calculate the specification tests. The structure of the input matrix is block-diagonal as before, and the basic data requirement is still the matrix $v_i'v_i$.

$$PE = PY - PZ\hat{\zeta} \tag{5.35}$$

using the same notation of (5.6). The residual sum of squares is given by

$$\begin{aligned} E'PE &= Y'PY - 2Y'PZ\hat{\zeta} + \hat{\zeta}'Z'PZ\hat{\zeta} \\ &= Y'PY - (2Y'PZ - Y'PZ(Z'PZ)^{-1}Z'PZ)\hat{\zeta} \\ &= Y'PY - (2Y'PZ - Y'PZ)\hat{\zeta} \\ &= Y'PY - Y'PZ\hat{\zeta} \end{aligned} \tag{5.36}$$

or RSS=TSS-ESS. Substituting the estimated coefficients again,

$$\begin{aligned} E'PE &= Y'PY - Y'PZ(Z'PZ)^{-1}Z'PY \\ &= Y'P(I_{NT} - PZ(Z'PZ)^{-1}Z')PY \\ &= (U'P + Z'P)(I_{NT} - PZ(Z'PZ)^{-1}Z')(PZ + PU) \end{aligned} \tag{5.37}$$

where NT is $\Sigma_t N_t$. As the middle PZ terms drop out,

$$\begin{aligned} E'PE &= U'P(I_{NT} - PZ(Z'PZ)^{-1}Z')PU \\ &= U'PU - U'PZ(Z'PZ)^{-1}Z'PU \end{aligned} \tag{5.38}$$

E'PE is a scalar, and so the solution to (5.38) is the trace of E'PE. Taking expected values,

$$\begin{aligned} \mathcal{E}(E'PE) &= \mathcal{E}(tr[U'PU - U'PZ(Z'PZ)^{-1}Z'PU]) \\ &= \mathcal{E}(tr[PUU' - PZ(Z'PZ)^{-1}Z'PUU']) \\ &= tr[P\mathcal{E}(UU') - PZ(Z'PZ)^{-1}Z'P\mathcal{E}(UU')] \end{aligned} \tag{5.39}$$

On the assumption that $\mathcal{E}(UU')=\sigma_u^2 I_{NT}$,

$$\begin{aligned} \mathcal{E}(E'PE) &= \sigma_u^2 tr[P - PZ(Z'PZ)^{-1}Z'P] \\ &= \sigma_u^2(trP - tr((Z'PZ)^{-1}Z'PZ)) \\ &= \sigma_u^2(trP - tr(I_K)) \\ &= \sigma_u^2(\sum_t N_t(1 - \frac{1}{N_t}) - K) \\ &= \sigma_u^2(\sum_t (N_t - 1) - K) \end{aligned} \tag{5.40}$$

Therefore

$$\hat{\sigma}_u^2 = \frac{E'PE}{\sum_t N_t - T - K} \tag{5.41}$$

A similar result holds for the pooled and within estimators. The main differences are the value of "K" and the trace of the first matrix. If K_X is the number of variables in x_{it}, then

$$\hat{\sigma}_u^2 = \frac{E'PE_u}{\sum_t N_t - T - TK_x}$$
$$\hat{\sigma}_p^2 = \frac{E'PE_p}{\sum_t N_t - 1 - K_x} \qquad (5.42)$$
$$\hat{\sigma}_r^2 = \frac{E'PE_r}{\sum_t N_t - T - K_x}$$

where the u, p, and r subscripts refer to the unrestricted, pooled and restricted estimators (that is, $E'PE_u$ is the sum of squared errors for the unrestricted estimator, for example). On these error assumptions, F-statistics for testing hypotheses of the latter two specifications are

$$F_{\substack{up \\ u\ vs.\ p}} = \frac{(E'PE_p - E'PE_u)/((T-1)(K_x+1))}{E'PE_u/(\sum_T N_t - T - TK_x)}$$
$$F_{\substack{ur \\ u\ vs.\ r}} = \frac{(E'PE_r - E'PE_u)/(K_x(T-1))}{E'PE_u/(\sum_T N_t - T - TK_x)} \qquad (5.43)$$
$$F_{\substack{rp \\ r\ vs.\ p}} = \frac{(E'PE_p - E'PE_r)/(T-1)}{E'PE_r/(\sum_t N_t - T - K_x)}$$

Large values imply a rejection of the more restricted hypothesis.

One refinement is to note that, as ζ is being estimated for the unrestricted model over T separate regressions, it is relatively easy to calculate separate estimates of the variance for each period. In this case, the time-heteroscedastic errors for period t are (from equation (5.41)):

$$\hat{\sigma}_{ut}^2 = \frac{e_t' Q_t e_t}{N_t - 1 - K} \qquad (5.44)$$

These are the errors reported by the regression program. However, the F-tests in (5.43) make the assumption that the variance is homoscedastic; the potential benefit of time-heteroscedastic errors appears small relative to the additional complexity of the corrected F-statistic.

5.2 Fixed-effects: allowing for individual heterogeneity

The models in this section are the more usual "panel" models in that they allow for individual heterogeneity. This is treated as a fixed effect and removed by taking deviations from individual means. Time dummies are left in the regression. This is because the transformation matrix which removes individual dummies cannot remove time dummies, and vice versa. It is possible to construct a matrix which removes both effects, but the structure of the resulting matrices are too complicated for our purposes. In any case, it will be demonstrated that it is not necessary to remove the time dummies to construct estimators for the stability of coefficients over time[4]. These estimators are the ones used to produce the results in chapters nine and ten.

5.2.1 The unrestricted case

Let the fundamental equation be

$$y_{it} = x_{it}\beta_t + \alpha_i + \lambda_t + u_{it} \qquad (5.45)$$
$$E(u_{it}) = 0 \qquad E(u_{it}u_{is}) = \sigma_{ts}$$

or, stacked for individual i,

$$y_i = w_i\beta + J_i\alpha_i + L_i\lambda + u_i \qquad (5.46)$$

where J_i is a T_i vector of ones, λ is a vector of the time effects, L is a T_ixT matrix formed by removing the rows from an identity matrix where an observation is missing, and

[4] Perfect collinearity between the time dummies and the individual dummies means that one time dummy should be dropped to remove the linear dependency. However, this can easily be done post extraction, and it does not change the qualitative results of this chapter at all. Thus, although the X'X of this section is actually singular, this is ignored solely to simplify the exposition. The next chapter discusses appropriate corrections.

$$w_i \equiv \begin{bmatrix} x_{i1} & 0 & & \\ 0 & x_{i2} & & \\ & & \ddots & \\ & & & x_{iT} \end{bmatrix} \quad \beta \equiv \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_T \end{bmatrix} \tag{5.47}$$

with rows of w_i similarly removed where observations are missing. Define

$$Q_i \equiv I_i - \frac{1}{T_i} J_i J_i' \tag{5.48}$$

where I_i is the $T_i x T_i$ identity matrix. Note that

$$Q_i = Q_i' = Q_i Q_i' \qquad Q_i J_i = 0 \tag{5.49}$$

Assume the individual effects α_i are fixed (which enables single-stage regression). To remove them from the equation, premultiply by Q_i:

$$\begin{aligned} Q_i y_i &= Q_i w_i \beta + Q_i J_i \alpha_i + Q_i L_i \lambda + Q_i u_i \\ &= Q_i w_i \beta + Q_i L_i \lambda + Q_i u_i \\ &= Q_i z_i \zeta + Q_i u_i \end{aligned} \tag{5.50}$$

with $z_i=[w_i\ L_i]$ and $\zeta=[\beta'\ \lambda']'$. The OLS solution to this will be to minimise

$$\sum_i^N u_i' Q_i u_i = \sum y_i' Q_i y_i + \sum \zeta' z_i' Q_i z_i \zeta - 2 \sum y_i' Q_i z_i \zeta \tag{5.51}$$

The normal equations for this are

$$\hat{\zeta} = \left(\sum_i^N z_i' Q_i z_i \right)^{-1} \sum_i^N z_i' Q_i y_i \tag{5.52}$$

The breakdown of the first element here is:

$$\begin{aligned} z_i' Q_i z_i &= z_i' (I_i - \frac{1}{T_i} J_i J_i') z_i \\ &= z_i' z_i - \frac{1}{T_i} z_i' J_i J_i' z_i \\ &= z_i' z_i - T_i \bar{z}_i' \bar{z}_i \end{aligned} \tag{5.53}$$

Similarly,

$$z_i' Q y_i = z_i' y_i - T \bar{z}_i' \bar{y}_i \tag{5.54}$$

where

$$\bar{z}_i \equiv \frac{1}{T_i} J_i' z_i \qquad \bar{y}_i \equiv \frac{1}{T_i} \sum_t y_{it} \tag{5.55}$$

are the mean values of the elements of z_i and y_i. Because of the block nature of w_i and L_i, this is equivalent to creating a vector of $1/T_i$ times all variables:

$$\begin{aligned} \bar{z}_i &= \begin{bmatrix} \frac{1}{T_i} & \frac{1}{T_i} & \cdots & \frac{1}{T_i} \end{bmatrix} \begin{bmatrix} x_{i1} & 0 & & 1 & 0 \\ 0 & x_{i2} & & 0 & 1 \\ & & \ddots & & \ddots \\ & & x_{iT} & & 1 \end{bmatrix} \\ &= \frac{1}{T_i} [x_{i1}\ x_{i2}\ \ldots x_{iT}\ \ 1\ 1\ \ldots\ 1] \end{aligned} \tag{5.56}$$

It can be seen that

$$z_i' z_i = \begin{bmatrix} x_{i1}' x_{i1} & 0 & & x_{i1}' & 0 \\ 0 & x_{i2}' x_{i2} & & 0 & x_{i2}' \\ & & \ddots & & \ddots \\ & & x_{iT}' x_{iT} & & x_{iT}' \\ x_{i1} & 0 & & 1 & 0 \\ 0 & x_{i2} & & 0 & 1 \\ & & \ddots & & \ddots \\ & & x_{iT} & & 1 \end{bmatrix} \tag{5.57}$$

Therefore, if defining

$$v_i \equiv [x_{i1}\ x_{i2}\ \ldots\ x_{iT}\ \ 1\ 1\ \ldots\ 1] \tag{5.58}$$

to give the cross-product matrix

$$v_i'v_i = \begin{bmatrix} x_{i1}'x_{i1} & x_{i1}'x_{i2} & & & x_{i1}' & x_{i2}' & \\ x_{i2}'x_{i1} & x_{i2}'x_{i2} & & & x_{i1}' & x_{i2}' & \\ & & \ddots & & & & \ddots \\ & & & x_{iT}'x_{iT} & & & x_{iT}' \\ x_{i1} & x_{i2} & & & 1 & 1 & \\ x_{i1} & x_{i2} & & & 1 & 1 & \\ & & \ddots & & & & \ddots \\ & & & x_{iT} & & & 1 \end{bmatrix} \quad \textbf{(5.59)}$$

then it is clear that $T_i\bar{z}_i'\bar{z}_i = T_i^{-1}v_i'v_i$ and $z_i'z_i$ is the block diagonal of $v_i'v_i$. A similar story holds for $z_i'Q_iy_i$. As y_i is a T_i x 1 column vector,

$$v_i'y_i' = \begin{bmatrix} x_{i1}'y_{i1} & x_{i1}'y_{i2} & & \\ x_{i2}'y_{i1} & x_{i2}'y_{i2} & & \\ & & \ddots & \\ & & & x_{iT}'y_{iT} \\ y_{i1} & y_{i2} & & \\ y_{i1} & y_{i2} & & \\ & & \ddots & \\ & & & y_{iT} \end{bmatrix} \quad \textbf{(5.60)}$$

Then $T_i\bar{z}_i'\bar{y}_i$ is the horizontal summation of $T_i^{-1}v_i'y_i'$ and $z_i'y_i$ are the diagonal terms $x_{i1}'y_{i1}..x_{iT}'y_{iT}$ and $y_{i1}..y_{iT}$ of $v_i'y_i'$.

Unlike the simple case, the time effects λ are now directly estimated, rather than being extracted from the time means as in Section 5.1. This makes no real difference to the outcome. The issue of testing for levels and constancy of the intercepts does not arise as all three estimates are based around deviations from the mean of the whole regression; therefore the test for a constant intercept in all periods amounts to testing for zero intercepts in all periods.

5.2.2 The pooled case

For the pooled model, the hypothesis is that β and λ are constant over time:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \tag{5.61}$$

or, stacked for individual i,

$$y_i = X_i\beta + J_i\alpha_i + u_i \tag{5.62}$$

where

$$X_i \equiv \left[x_{i1}' \quad x_{i2}' \quad \ldots \quad x_{iT}' \right]' \tag{5.63}$$

and x_{it} contains the constant term[5]. Using Q_i as above and still assuming the individual effects α_i are fixed, the latter are removed by premultiplying by Q_i:

$$\begin{aligned} Q_i y_i &= Q_i X_i\beta + Q_i J_i\alpha_i + Q_i u_i \\ &= Q_i X_i\beta + Q_i u_i \end{aligned} \tag{5.64}$$

The OLS minimisation problem is

$$\sum_i^N u_i' Q_i u_i = \sum y_i' Q_i y_i + \sum \beta' X_i' Q_i X_i \beta - 2\sum y_i' Q_i X_i \beta \tag{5.65}$$

giving

$$\hat{\beta} = \left(\sum_i^N X_i' Q_i X_i \right)^{-1} \sum_i^N X_i' Q_i y_i \tag{5.66}$$

Again, this breaks down into:

---

5 Including the constant term within the x_{it} at this stage is merely a simplification and has no bearing on the results.

$$\begin{aligned} X_i'Q_iX_i &= X_i'X_i - \frac{1}{T_i}X_i'J_iJ_i'X_i \\ &= X_i'X_i - T_i\bar{X}_i'\bar{X}_i \\ X_i'Q_iy_i &= X_i'y_i - T_i\bar{X}_i'\bar{y}_i \end{aligned} \tag{5.67}$$

with

$$\bar{X}_i \equiv \frac{1}{T_i}J_i'X_i \qquad \bar{y}_i \equiv \frac{1}{T_i}\sum_t y_{it} \tag{5.68}$$

the mean values of the elements of X_i and y_i. Note that the mean of X_i is different from the mean of z_i because of the block nature of the latter. In this case

$$X_i'X_i = \sum_t x_{it}'x_{it} \tag{5.69}$$

and, summing over i,

$$\sum_i X_i'X_i = \sum_i \sum_t x_{it}'x_{it} = \sum_t \sum_i x_{it}'x_{it} \tag{5.70}$$

In addition

$$\begin{aligned} \bar{X}_i &= \begin{bmatrix} \frac{1}{T_i} & \frac{1}{T_i} & \cdots & \frac{1}{T_i} \end{bmatrix} \begin{bmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{iT} \end{bmatrix} \\ &= \frac{1}{T_i}\sum_t x_{it} \end{aligned} \tag{5.71}$$

Summing over i gives

$$\sum_i \frac{1}{T_i}\sum_t x_{it}'\sum_t x_{it} = \sum_i \frac{1}{T_i}\sum_t \sum_s x_{it}'x_{is} = \sum_t \sum_s \sum_i \frac{1}{T_i}x_{it}'x_{is} \tag{5.72}$$

All the summations over time are from 1 to T. Although T_i can vary from individual to individual, missing values are set to zero and so the mean calculations are correct whether the summation is over T or T_i.

Using the definition of v_i from (5.58), it is apparent that $X_i'X_i$ is the sum of the diagonal

blocks of $v_i'v_i$ and $T_i\bar{X}_i'\bar{X}_i$ is the sum of the all of the blocks of the means matrix. $X_i'Q_iy_i$ has a similar structure:

$$\begin{aligned}\sum_i X_i'y_i &= \sum_i \sum_t x_{it}'y_{it} = \sum_t \sum_i x_{it}'y_{it} \\ \sum_i \frac{1}{T_i}\sum_t x_{it}'\sum_t y_{it} &= \sum_i \frac{1}{T_i}\sum_t \sum_s x_{it}'y_{is} = \sum_t \sum_s \sum_i \frac{1}{T_i}x_{it}'y_{is}\end{aligned} \tag{5.73}$$

5.2.3 The restricted case

Now consider β constant and λ varying over time:

$$y_{it} = x_{it}\beta + \alpha_i + \lambda_t + u_{it} \tag{5.74}$$

or, stacked,

$$y_i = X_i\beta + J_i\alpha_i + L_i\lambda + u_i \tag{5.75}$$

where all the terms are as defined above. As for Section 5.2.1, the constant term is separated out from x_{it}. This is the commonest form of linear panel model found in applied work, even though it involves significant (and often not tested) restrictions on the basic model (5.45). Premultiplying by Q_i to remove the individual effects:

$$\begin{aligned}Q_iy_i &= Q_iX_i\beta + Q_iJ_i\alpha_i + Q_iL_i\lambda + Q_iu_i \\ &= Q_iX_i\beta + Q_iL_i\lambda + Q_iu_i \\ &= Q_iC_i\chi + Q_iu_i\end{aligned} \tag{5.76}$$

with $C_i=[X_i\ L_i]$ and $\chi=[\beta'\ \lambda']'$. The OLS solution is

$$\hat{\chi} = \left(\sum_i^N C_i'Q_iC_i\right)^{-1}\sum_i^N C_i'Q_iy_i \tag{5.77}$$

Breaking this down,

$$\begin{aligned}C_i'Q_iC_i &= C_i'C_i - \frac{1}{T_i}C_i'J_iJ_i'C_i \\ &= C_i'C_i - T_i\bar{C}_i'\bar{C}_i \\ C_i'Q_iy_i &= C_i'y_i - T_i\bar{C}_i'\bar{y}_i\end{aligned} \tag{5.78}$$

with

$$\bar{C}_i \equiv \frac{1}{T_i} J_i' C_i \qquad \bar{y}_i \equiv \frac{1}{T_i} \sum_t y_{it} \tag{5.79}$$

for the mean values. The mean of y_i is the same in all these alternative hypotheses, but again, the mean of C_i has some slightly different elements:

$$\begin{aligned} \bar{C}_i &= \left[\frac{1}{T_i} \; \frac{1}{T_i} \; \cdots \; \frac{1}{T_i} \right] \begin{bmatrix} x_{i1} & 1 & 0 & \\ x_{i2} & 0 & 1 & \\ \vdots & & & \ddots \\ x_{iT} & & & 1 \end{bmatrix} \\ &= \frac{1}{T_i} [\sum_t x_{it} \;\; 1 \; 1 \; \ldots \; 1] \end{aligned} \tag{5.80}$$

Therefore

$$C_i' C_i = \begin{bmatrix} \sum_t x_{it}' x_{it} & 0 & 0 & & 0 \\ & & & & \\ 0 & & 1 & 0 & \\ 0 & & 0 & 1 & \\ & & & & \ddots \\ 0 & & & & 1 \end{bmatrix} \tag{5.81}$$

and

$$T_i \bar{C}_i' \bar{C}_i = \frac{1}{T_i} \begin{bmatrix} \sum_t x_{it}' \sum_t x_{it} & \sum_t x_{it}' & \cdots & \sum_t x_{it}' \\ \sum_t x_{it} & 1 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ \sum_t x_{it} & 1 & \cdots & 1 \end{bmatrix} \tag{5.82}$$

As for the pooled estimator, $\Sigma x_{it}' \Sigma x_{it}$ can be calculated post extraction. For the simple means of x_{it} over t, if column c of x_{it} contains the constant term then

$$\frac{1}{T} \sum_t x_{it}' x_{itc} = \frac{1}{T} \sum_t x_{it}' 1 = \frac{1}{T_i} \sum_t x_{it} \tag{5.83}$$

From the definition of v_i it is apparent that $C_i' C_i$ is the block diagonal of $v_i' v_i$ and $T_i \bar{C}_i' \bar{C}_i$ is

T_i^{-1} times one of the constant intersections of $v_i'v_i$. However, in both cases the non-constant sections (the x_{it} bits) must be summed. Again, $C_i'Q_iy_i$ has a similar structure, with

$$C_i'y_i = \begin{bmatrix} \sum_t x_{it}'y_{it} \\ y_{i1} \\ y_{i2} \\ \vdots \\ y_{iT} \end{bmatrix} \qquad T_i\bar{C}_i'\bar{y}_i = \frac{1}{T_i}\begin{bmatrix} \sum_t x_{it}'\sum_t y_{it} \\ \sum_t y_{it} \\ \sum_t y_{it} \\ \vdots \\ \sum_t y_{it} \end{bmatrix} \tag{5.84}$$

Clearly all that is needed to for the three estimators are the two matrices $\Sigma v_i'v_i$ and $\Sigma T_i^{-1}v_i'v_i$. As these can be created piecewise (that is, for each individual i, $v_i'v_i$ is found and then summed into a totalling cross-product matrix), and as this matrix can be created without reference to the particular variables used in a particular regression (that is, $\Sigma v_i'v_i$ contains all variables of interest of which a subset are used in any particular regression), this presents no real difficulties to the extraction or analysis software.[6]

It is clear that the data matrices for the pooled and restricted versions are constructed merely by summing over time the relevant elements from the unrestricted matrix. This is what might be expected. In the simple cross-section estimators, taking deviations for each period led to different matrices being needed for different hypotheses about variation over time, as the mean used depended upon the model in question. In the fixed effects estimators, the transformation is being made with respect to deviations from individual means. In that context, different assumptions about time-varying slopes and intercepts merely amounts to a rearrangement of the variables; the same mean (over all periods for one individual) is used. Had assumptions been made about coefficients varying over individuals, then testing the different models would have involved calculating different means - as the cross-section

[6] The fact that the constant terms are sometimes included in x_{it} and sometimes represented as separate elements is merely for notational convenience and makes little difference to the analysis. However, for analytical purposes it is easier if the constant terms are always included in x_{it} rather than being grouped separately, and so this is the structure the extraction software uses.

estimators necessitated.

Note that there are two significant disadvantages to the fixed-effects formulation used here. Firstly, each cross-product means matrix $\Sigma T_i^{-1} v_i' v_i$ is dependent upon a particular value of T: as the divisor in the means matrix is T_i, and as $T_i \leq T$, a different value of T alters the range of acceptable T_is. Only in the case of the balanced matrix can the means for several years be calculated from one $\Sigma v_i' v_i$ matrix, because in this case the constant divisor T can be taken outside all the summations over i. In the general case of an unbalanced panel, it is no longer possible to run regressions on multiple combinations of years using the same cross-product means matrix. Separate means matrices need to be created for each selection of years.

Secondly, creation of these matrices involves much computer time and memory - significantly more than for the simple panel estimator. However, note that $\Sigma v_i' v_i$ for a given T contains $\Sigma v_i' v_i$ for any smaller T. As the two matrices can be created independently, it is sensible to extract $\Sigma v_i' v_i$ in one run and then the means matrix $\Sigma T_i^{-1} v_i' v_i$ separately - possibly for several different values of T. This is a relatively efficient solution to the extraction problem, and is implemented in the extraction software.

In fact, the data requirement is less onerous than stated above. Consider the two matrices requested for this section and the previous one. It is clear that each of the diagonal blocks in $\Sigma v_i' v_i$ (5.59) corresponds to one of the $v_t' v_t$ matrices in (5.18). There is no necessity to calculate the whole matrix (5.59); the time-effects cross-products matrices will do equally well. The analysis software takes account of this. However, as the practical difference between creating $\Sigma v_i' v_i$ and $\Sigma T_i^{-1} v_i' v_i$ is negligible, the extraction software writer allows for the creation of full $\Sigma v_i' v_i$ matrices. Such a matrix can contain useful information on interperiod correlations; in addition, it allows for time-differencing estimators to be

constructed, as will be shown in Section 5.3[7].

5.2.4 Variances and testing in the covariance estimators

Consider variances for the general model first. The sum of squared residuals is

$$\begin{aligned} \sum e_i'Q_ie_i &= \sum\left[y_i'Q_iy_i - 2y_i'Q_iz_i\hat{\zeta} + \hat{\zeta}z_i'Q_iz_i\hat{\zeta}\right] \\ &= \sum y_i'Q_iy_i - \sum y_i'Q_iz_i\left(\sum z_i'Q_iz_i\right)^{-1}\sum z_i'Q_iy_i \\ RSS &= TSS + ESS \end{aligned} \tag{5.85}$$

where e_i is the residual error. Define the symmetric, idempotent matrix P:

$$P \equiv \begin{bmatrix} Q_1 & 0 & 0 \\ 0 & Q_2 & 0 \\ & & \ddots \\ 0 & 0 & Q_N \end{bmatrix} \tag{5.86}$$

Then

$$\sum_i^N e_i'Q_ie_i = E'PE \quad \sum_i^N y_i'Q_iy_i = Y'PY \quad \sum_i^N z_i'Q_iz_i = Z'PZ \quad \sum_i^N z_i'Q_iy_i = Z'PY \tag{5.87}$$

where E=[e_1' e_2' ... e_N']', Y=[y_1' y_2' ... y_N']', and Z=[z_1' z_2' ... z_N']'. Define U=[u_1' u_2' ... u_N']' and I_{NT} as the identity matrix with ΣT_i rows. Substituting (5.87) in (5.85) gives

$$\begin{aligned} E'PE &= Y'PY - Y'PZ(Z'PZ)^{-1}Z'PY \\ &= (U'P + Z'P)(I_{NT} - PZ(Z'PZ)^{-1}Z')(PZ + PU) \\ &= U'PU - U'PZ(Z'PZ)^{-1}Z'PU \end{aligned} \tag{5.88}$$

As for section 5.1, E'PE is a scalar with its solution equal to its trace. Taking expected values,

$$\begin{aligned} \mathcal{E}(E'PE) &= \mathcal{E}(tr[U'PU - U'PZ(Z'PZ)^{-1}Z'PU]) \\ &= tr[P\,\mathcal{E}(UU') - PZ(Z'PZ)^{-1}Z'P\,\mathcal{E}(UU')] \end{aligned} \tag{5.89}$$

Maintaining the earlier assumption that $\mathcal{E}(UU')=\sigma_u^2 I_{NT}$ leads to

---

[7] The full $\Sigma v_i'v_i$ matrix should also allow for a "minimum-distance" model to be generated, which allows for a more general error structure (Hsiao (1986) ch.3; Chamberlain (1984)). This is currently being investigated.

$$\begin{aligned} \mathcal{E}(E'PE) &= \sigma_u^2 tr[P - PZ(Z'PZ)^{-1}Z'P] \\ &= \sigma_u^2(tr P - tr(I_K)) \\ &= \sigma_u^2(\sum_i T_i - N - K) \end{aligned} \tag{5.90}$$

Therefore

$$\hat{\sigma}_u^2 = \frac{E'PE}{\sum_i T_i - N - K} \tag{5.91}$$

This result holds for all three estimators, as the structure of P is identical in all three. The only difference between them is the value of K. If K_x is the number of variables in x_{it} (excluding any constant term), then

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{E'PE_u}{\sum_i T_i - N - T(K_x+1)} \\ \hat{\sigma}_p^2 &= \frac{E'PE_p}{\sum_i T_i - N - (K_x+1)} \\ \hat{\sigma}_r^2 &= \frac{E'PE_r}{\sum_i T_i - N - (K_x+T)} \end{aligned} \tag{5.92}$$

where the u, p, and r subscripts refer to the unrestricted, pooled and restricted estimators.

On the error assumptions, F-statistics for testing hypotheses of unrestricted, pooled and time-dummy are

$$\begin{aligned} F_{\substack{up \\ u\ vs.\ p}} &= \frac{(E'PE_p - E'PE_u)/((T-1)(K_x+1))}{E'PE_u/(\sum_i T_i - N - T(K_x+1))} \\ F_{\substack{ur \\ u\ vs.\ r}} &= \frac{(E'PE_r - E'PE_u)/(K_x(T-1))}{E'PE_u/(\sum_i T_i - N - T(K_x+1))} \\ F_{\substack{pt \\ r\ vs.\ p}} &= \frac{(E'PE_p - E'PE_r)/(T-1)}{E'PE_r/(\sum_i T_i - N - (K_x+T))} \end{aligned} \tag{5.93}$$

One difficulty is the calculation of N and ΣT_i. However, from the bottom-right hand corner

of $T_i^{-1}v_i'v_i$ we have the $T_i x T_i$ block

$$\frac{1}{T_i}\begin{bmatrix} 1 & 1 & & 1 \\ 1 & 1 & & 1 \\ & \vdots & & \\ 1 & 1 & & 1 \end{bmatrix} \tag{5.94}$$

This will actually be stored as a TxT block with zeroes in the appropriate places, but this is qualitatively the same result. Summing the diagonal gives

$$\frac{1}{T_i}\sum_{t=1}^{T_i} 1 = \frac{T_i}{T_i} = 1 \tag{5.95}$$

Therefore, summing the constant diagonal for the whole matrix gives

$$\sum_{i=1}^{N}\left(\frac{1}{T_i}\sum_{t=1}^{T_i} 1\right) = \sum_{i=1}^{N} 1 = N \tag{5.96}$$

Meanwhile, for the total number of observations, let d_{it} be a marker, 1 if individual i was observed in period t and 0 otherwise. Then

$$\sum_{i=1}^{N} T_i = \sum_{i=1}^{N}\sum_{t=1}^{T_i} 1 = \sum_{i=1}^{N}\sum_{t=1}^{T} d_{it} = \sum_{t=1}^{T}\sum_{i=1}^{N} d_{it} = \sum_{t=1}^{T} N_t \tag{5.97}$$

Clearly ΣT_i is then the sum of the diagonal of the units section of $\Sigma v_i'v_i$, as this gives the number of observations in each period.

Observe that N comes from the means-matrix and thus is time dependent. This is correct: if an individual only has observations outside the range of years used for a particular regression, then he may not be included in the "N" for that regression. the total number of observations is just a straightforward tally of the number of observations in each period, and so the total number of times an individual has been observed is irrelevant.

5.3 Differencing estimators (complete observations sets)

An alternative approach to individual heterogeneity is to take time differences. This also removes the individual effect; however, for unbalanced panels the arithmetic is more complicated. This section considers the case where all individuals have the same number of observations.

5.3.1 The unrestricted case

The unrestricted case is the same as for the heterogenous model of Section 5.2:

$$y_i = w_i\beta + J_i\alpha_i + L_i\lambda + u_i \tag{5.98}$$

where the matrices are as defined for equation (5.46). To remove heterogeneity by time-differencing, the transformation matrix is

$$\underset{T_i-1\times T_i}{Q_i} \equiv \begin{bmatrix} -1 & 1 & 0 & & \\ 0 & -1 & 1 & & \\ & & & \ddots & \\ & & & 0 & -1 & 1 \end{bmatrix} \tag{5.99}$$

We still have $Q_iJ_i=0$; however, Q_i is no longer idempotent. Instead,

$$\underset{T_i\times T_i}{Q_i'Q_i} = \begin{bmatrix} 1 & -1 & 0 & & & \\ -1 & 2 & -1 & & & \\ & & & \ddots & & \\ & & & -1 & 2 & -1 \\ & & & 0 & -1 & 1 \end{bmatrix} = I_i + S_i \tag{5.100}$$

where

$$S_i \equiv \begin{bmatrix} 0 & -1 & 0 & & & \\ -1 & 1 & -1 & & & \\ & & & \ddots & & \\ & & & -1 & 1 & -1 \\ & & & 0 & -1 & 0 \end{bmatrix} \tag{5.101}$$

The normal equations for the coefficients are (from (5.52)):

$$\hat{\zeta} = \left(\sum_i^N z_i' Q_i' Q_i z_i\right)^{-1} \sum_i^N z_i' Q_i' Q_i y_i \tag{5.102}$$

Note that one time dummy will have to be deleted because the matrix is not of full rank. Section 5.4 discusses the issue of identification of time dummies in more detail. Breaking down the components of (5.102),

$$z_i' Q_i' Q_i z_i = z_i' z_i + z_i' S_i z_i \qquad z_i' Q_i' Q_i y_i = z_i' y_i + z_i' S_i y_i \tag{5.103}$$

Define $\bar{x}_{it}=(-x_{it})$; that is, the negative of x_{it}. Then, from the definition of S_i and z_{it},

$$z_i' S_i z_i = \begin{bmatrix} 0 & \bar{x}_{i1}' x_{i2} & 0 & & 0 & \bar{x}_{i1}' & 0 \\ \bar{x}_{i2}' x_{i1} & x_{i2}' x_{i2} & \bar{x}_{i2}' x_{i3} & & \bar{x}_{i2}' & x_{i2}' & \bar{x}_{i2}' \\ 0 & \bar{x}_{i3}' x_{i2} & x_{i3}' x_{i3} & & 0 & \bar{x}_{i3}' & x_{i3}' \\ & & & \ddots & & & \ddots \\ & & & 0 & & & 0 \\ 0 & \bar{x}_{i1} & 0 & & 0 & -1 & 0 \\ \bar{x}_{i2} & x_{i2} & \bar{x}_{i2} & & -1 & 1 & -1 \\ 0 & \bar{x}_{i3} & x_{i3} & & 0 & -1 & 1 \\ & & & \ddots & & & \ddots \\ & & & 0 & & & 0 \end{bmatrix} \tag{5.104}$$

It can be seen that $z_i'S_iz_i$ is just the Hadamard product of $z_i'z_i$ and the (K+1)x(K+1) equivalent of S_i:

$$z_i' S_i z_i = z_i' z_i \odot H_i \tag{5.105}$$

where J is a K-vector of ones and

$$H_i \equiv \begin{bmatrix} 0_{KK} & -JJ' & 0_{KK} & 0_{KK} & & 0_{KI} & -J & 0_{KI} & 0_{KI} \\ -JJ' & JJ' & -JJ' & 0_{KK} & & -J & J & -J & 0_{KI} \\ 0_{KK} & -JJ' & JJ' & -JJ' & & 0_{KI} & -J & J & -J \\ & \ddots & & & & \ddots & & & \\ & -J' & J' & -J' & 0_{1K} & & -1 & 1 & -1 & 0 \\ & 0_{1K} & -J' & J' & -J' & & 0 & -1 & 1 & -1 \\ & 0_{1K} & 0_{1K} & -J' & 0_{1K} & & 0 & 0 & -1 & 0 \end{bmatrix} \quad \textbf{(5.106)}$$

Therefore, all the information to calculate this estimator is in the matrix $v_i'v_i$. As the matrix H_K is a constant matrix depending only on the years of the analysis and not on any individual variables, then $\Sigma_i v_i'v_i$ can be built up and used for regressions on multiple years.

Finally, calculating $z_i'Q_i'Q_iy_i$ requires

$$z_i'S_iy_i = \begin{bmatrix} \bar{x}_{i1}'y_{i2} \\ \bar{x}_{i2}'y_{i1} + x_{i2}'y_{i2} + \bar{x}_{i2}'y_{i3} \\ \vdots \\ \bar{x}_{iT-1}'y_{iT-2} + x_{iT-1}'y_{iT-1} + \bar{x}_{iT-1}'y_{iT} \\ \bar{x}_{iT}'y_{iT-1} \\ -y_{i2} \\ -y_{i1} + y_{i2} - y_{i3} \\ \vdots \\ -y_{iT-2} + y_{iT-1} - y_{iT} \\ -y_{iT-1} \end{bmatrix} \quad \textbf{(5.107)}$$

which has a similar structure to $z_i'S_iz_i$ and can also be calculated from $v_i'v_i$.

Restrictions can be placed on the specification in a manner analogous to the earlier sections.

5.3.2 The pooled case

Let β and λ be constant over time:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \tag{5.108}$$

x_{it} contains the constant term. Stacking over t and using Q_i as above and still assuming the individual effects α_i are fixed, the latter is removed by premultiplying by Q_i:

$$\begin{aligned} Q_i y_i &= Q_i X_i \beta + Q_i J_i \alpha_i + Q_i u_i \\ &= Q_i X_i \beta + Q_i u_i \end{aligned} \tag{5.109}$$

The OLS normal equations are

$$\hat{\beta} = \left(\sum_i^N X_i' Q_i' Q_i X_i \right)^{-1} \sum_i^N X_i' Q_i' Q_i y_i \tag{5.110}$$

Again, this breaks down into:

$$X_i' Q_i' Q_i X_i = X_i' X_i + X_i' S_i X_i \qquad X_i' Q_i' Q_i y_i = X_i' y_i + X_i' S_i y_i \tag{5.111}$$

But in this case

$$\begin{aligned} X_i' S_i X_i &= \bar{x}_{i2}' x_{i1} + (\bar{x}_{i1}' + x_{i2}' + \bar{x}_{i3}') x_{i2} + (\bar{x}_{i2}' + x_{i3}' + \bar{x}_{i4}') x_{i3} + \ldots \\ &= \sum_{t=2}^{T_i - 1} \left(x_{it}' x_{it} - x_{it}' x_{it-1} - x_{it}' x_{it+1} \right) - x_{i2}' x_{i1} - x_{iT-1}' x_{iT} \end{aligned} \tag{5.112}$$

where $\bar{x}_{it}=(-x_{it})$, as before. $X_i'S_iy_i$ has a similar structure:

$$\begin{aligned} X_i' S_i y_i &= \bar{x}_{i2}' y_{i1} + (\bar{x}_{i1}' + x_{i2}' + \bar{x}_{i3}') y_{i2} + (\bar{x}_{i2}' + x_{i3}' + \bar{x}_{i4}') y_{i3} + \ldots \\ &= \sum_{t=2}^{T_i - 1} \left(x_{it}' y_{it} - x_{it}' y_{it-1} - x_{it}' y_{it+1} \right) - x_{i2}' y_{i1} - x_{iT-1}' y_{iT} \end{aligned} \tag{5.113}$$

Once more the result is a constant matrix multiple of the constructed $\Sigma v_i'v_i$; in fact, the result is merely the summation over t of the x'x section of the unrestricted case. Thus the pooled case can be feasibly constructed from a correctly formed cross-product.

5.3.3 The restricted case

Finally, consider β constant and λ varying over time:

$$y_{it} = x_{it}\beta + \alpha_i + \lambda_t + u_{it} \tag{5.114}$$

or, stacked,

$$y_i = X_i\beta + J_i\alpha_i + L_i\lambda + u_i \tag{5.115}$$

where all the terms are as defined above. Again, the constant term is made explicit. Premultiplying by Q_i removes the individual effects:

$$Q_i y_i = Q_i X_i\beta + Q_i L_i\lambda + Q_i u_i \tag{5.116}$$

with $C_i=[X_i\ L_i]$ and $\chi=[\beta'\ \lambda']'$, as in Part II. The OLS solution is

$$\hat{\chi} = \left(\sum_i^N C_i'Q_i'Q_iC_i\right)^{-1}\sum_i^N C_i'Q_i'Q_iy_i \tag{5.117}$$

Breaking this down,

$$C_i'Q_i'Q_iC_i = C_i'C_i + C_i'S_iC_i \qquad C_i'Q_i'Q_iy_i = C_i'y_i + C_i'S_iy_i \tag{5.118}$$

where, as would be expected from previous cases

$$C_i'C_i = \begin{bmatrix} X_i'X_i & 0 & 0 & 0 \\ 0 & 1 & 0 & \\ 0 & 0 & 1 & \\ & & & \ddots \\ 0 & & & 1 \end{bmatrix} \qquad C_i'S_iC_i = \begin{bmatrix} X_i'S_iX_i & & X_i'S_i & \\ & 0 & -1 & \\ S_iX_i & -1 & 1 & \\ & & & \ddots \\ & & & 0 \end{bmatrix} \tag{5.119}$$

with $X_i'S_iX_i$ as defined above and

$$X_i'S_i = \begin{bmatrix} \bar{x}_{i2}' & (\bar{x}_{i1}'+x_{i2}'+\bar{x}_{i3}') & (\bar{x}_{i2}'+x_{i3}'+\bar{x}_{i4}') & \cdots \end{bmatrix} \tag{5.120}$$

Again, this presents no especial difficulties in the construction of the OLS estimator from a cross-product matrix. The information requirements are less than for the estimators in section

5.2, as all that is needed to estimate all models is the matrix $\Sigma v_i'v_i$, which can be created piecewise[8].

Thus, although the time-differencing estimators require more computer power than the simple models of Section 5.1, they are less of a burden than the covariance estimators. In addition, the cross-product matrix used for the time-differencing estimator ($\Sigma v_i'v_i$,) is not dependent on the actual number of years used in regression.

On the other hand, the time-differencing estimator is a less efficient solution than the covariance estimator for two reasons. Firstly, it does not take full account of all observations, as the end observations only contribute once to the estimate; secondly, the deviations estimators use the cross-correlation between all the explanatory variables to determine the coefficients, whereas the time differencing approach merely uses correlations between two periods.

Finally, the differencing estimator described in this section suffers from the problem of missing data. The above solution is only appropriate where all individuals have a full set of observations for the period of interest. A balanced sample is very much the exception in NES extractions, and so section 5.4 considers the issue of time differencing from an unbalanced panel.

5.3.4. Variances and testing in the time-differencing estimator

Again, consider variances for the general model first. As the system is still block-diagonal, the system equivalent of $Q_i'Q_i$ is P:

---

[8] Again, the representation of the constant terms sometimes as being included in x_{it} and sometimes as separate elements is merely for notational convenience.

$$P \equiv \begin{bmatrix} Q_1'Q_1 & 0 & 0 \\ 0 & Q_2'Q_2 & 0 \\ & & \ddots \\ 0 & 0 & Q_N'Q_N \end{bmatrix} \tag{5.121}$$

Using the logic and notation of the earlier estimators

$$\begin{aligned} E'PE &= Y'PY - Y'PZ(Z'PZ)^{-1}Z'PY \\ &= (U'P + Z'P)(I_{NT} - PZ(Z'PZ)^{-1}Z')(PZ + PU) \\ &= U'PU - U'PZ(Z'PZ)^{-1}Z'PU \end{aligned} \tag{5.122}$$

The solution to this scalar is the trace of E'PE. Taking expected values,

$$\begin{aligned} \mathcal{E}(E'PE) &= \mathcal{E}(tr[U'PU - U'PZ(Z'PZ)^{-1}Z'PU]) \\ &= tr[P\mathcal{E}(UU') - PZ(Z'PZ)^{-1}Z'P\mathcal{E}(UU')] \end{aligned} \tag{5.123}$$

and assuming that $\mathcal{E}(UU')=\sigma_u^2 I_{NT}$

$$\begin{aligned} \mathcal{E}(E'PE) &= \sigma_u^2 tr[P - PZ(Z'PZ)^{-1}Z'P] \\ &= \sigma_u^2(trP - tr(I_K)) \\ &= \sigma_u^2(tr(I_{NT}) + \sum_i tr(S_i) - tr(I_K)) \\ &= \sigma_u^2(NT + \sum_i (T-2) - K) \\ &= \sigma_u^2(2N(T-1) - K) \end{aligned} \tag{5.124}$$

Thus

$$\hat{\sigma}_u^2 = \frac{E'PE}{2N(T-1) - K} \tag{5.125}$$

Again, the results is common to all three estimators, with only the value of K changing. Taking K_x as the number of variables in x_{it} (excluding the constant), the relevant adjustments are

$$\begin{aligned} \hat{\sigma}_u^2 &= \frac{E'PE_u}{2N(T-1) - T(K_x+1)} \\ \hat{\sigma}_p^2 &= \frac{E'PE_p}{2N(T-1) - (K_x+1)} \\ \hat{\sigma}_r^2 &= \frac{E'PE_r}{2N(T-1) - (K_x+T)} \end{aligned} \tag{5.126}$$

where the u, p, and r subscripts refer to the unrestricted, pooled and restricted estimates. Note that the inefficiency of the differencing approach is to some extent reflected in the higher denominators in the estimation of the standard error.

The appropriate F-tests for these estimators are

$$
\begin{aligned}
F_{\substack{up \\ u\ vs.\ p}} &= \frac{(E'PE_p - E'PE_u)/((T-1)(K_x+1))}{E'PE_u/(2N(T-1) - T(K_x+1))} \\
F_{\substack{ur \\ u\ vs.\ r}} &= \frac{(E'PE_r - E'PE_u)/(K_x(T-1))}{E'PE_u/(2N(T-1) - T(K_x+1))} \\
F_{\substack{rp \\ r\ vs.\ p}} &= \frac{(E'PE_p - E'PE_r)/(T-1)}{E'PE_r/(2N(T-1) - (K_x+T))}
\end{aligned}
\qquad (5.127)
$$

The value of N can be easily extracted from the cross-product matrix. T and K are known.

5.4 Time differencing with missing observations

If the panels are unbalanced, then the approach of the previous section will not work. This is because the matrix H_i is different for each individual, and so the post facto creation of the necessary data matrices from $\sum v_i'v_i$ is not possible. Observations may only be used where both T and T-1 are observed, and these cannot be identified from any of the group matrices.

In this case, matrices must be constructed initially in first differences. Define

$$
\begin{aligned}
\hat{x}_{it} &= x_{it} - x_{it-1} \qquad x_{it},\ x_{it-1}\ \textit{observed} \\
\hat{x}_{it} &= 0 \qquad\qquad\qquad\qquad \textit{otherwise}
\end{aligned}
\qquad (5.128)
$$

and similar terms for $\hat{y}_{it}$ and $\hat{u}_{it}$. Clearly, as the individual-specific term α_{it} is constant ($\alpha_{it}=\alpha_i$ for all t), it drops out of the final equation. With no individual heterogeneity in the equation for $\hat{y}_{it}$, the approach of Section 5.1 can be used, and so all the results of that section apply here. The fact that the "means transformation" (the matrix Q_t) is being applied to differenced data does not have any implications for the estimation of the slope coefficients.

However, there may be some confusion over the equation to be estimated. Consider a straight differencing of the stacked equation (5.2):

$$\begin{aligned} y_t - y_{t-1} &= X_t\beta_t - X_{t-1}\beta_{t-1} + J_t\lambda_t - J_{t-1}\lambda_{t-1} + u_t - u_{t-1} \\ &= X_t\beta_t - X_{t-1}\beta_{t-1} + J_t(\lambda_t - \lambda_{t-1}) + u_t - u_{t-1} \\ &= (X_t - X_{t-1})\beta_t + X_{t-1}(\beta_t - \beta_{t-1}) + J_t(\lambda_t - \lambda_{t-1}) + u_t - u_{t-1} \end{aligned} \qquad \textbf{(5.129)}$$

where the requirement that an individual must be observed in T and T-1 to be included ensures that $J_t=J_{t-1}$. To estimate this equation requires that both X_t and X_{t-1} (or equivalently, that X_t-X_{t-1} and X_{t-1}) be included explicitly as regressors. This requires a degree of carefulness in ensuring that the correct variables are included. However, the form of (5.129) is identical to (5.2), with dependent and explanatory variables and a unit vector. Premultiplication by the matrix Q_t defined in (5.3) will still remove the time dummies and so, once the matrices have been constructed correctly, all the mathematics of the first section can be applied. The presence of the unit vector in (5.129) means that N_t can be recovered from the cross-product matrix.

A simplifying assumption which may be justifiable on the grounds of practicality may be that $\beta_t=\beta_{t-1}$ for any t considered in isolation[9]. The form of (5.129) with this assumption is

$$\begin{aligned} y_t - y_{t-1} &= (X_t - X_{t-1})\beta_t + J_t(\lambda_t - \lambda_{t-1}) + u_t - u_{t-1} \\ \hat{y}_t &= \hat{X}_t\beta_t + J_t\hat{\lambda}_t + \hat{u}_t \end{aligned} \qquad \textbf{(5.130)}$$

This time-differencing approach implies a restriction on the slope coefficients that is not present in the earlier estimators. A necessary assumption for the differencing approach of (5.130) is that the slopes do not change significantly from one year to the next, although over the whole period the slopes may shift. The calculated βs now represent the best-fitting slope for any <u>two</u> consecutive periods; that is, instead of the ζ of equation (5.7) being estimated slopes for 1975, 1976, 1977... and so on, it now represents a separate slope coefficient for

[9] The NES extraction software makes this assumption by default; that is, the automatically generated procedures produce data appropriate for (5.130). However, these are easily edited to allow for the estimation of (5.129). Obviously, it makes no difference to the analysis program whether variables are included to estimate (5.130) or (5.129).

two years at a time (1975/6, 1976/7, 1977/8...).

This has implications for the interpretation of these coefficients. If the slopes do change over time, then the differencing estimator will show less variation than the basic time-effects estimator, simply because the differencing approach estimates average slopes. A larger degree of autocorrelation may also be expected, with the slopes evolving over time. Models of evolving coefficients have been developed by some authors, but they are outwith the scope of this work. The F-tests described will only check for parameter constancy, not systematic change.

We should remark that, although the practice described above involves a limitation on the values of the coefficients, it remains a more general specification than is often found in differencing models. In most applied analysis, all coefficients save the intercept are kept constant over time; here only constancy over any two consecutive periods is assumed.

For the time dummies, taking time differences means that the change in λ over any two periods is now being estimated. This holds for both (5.129) and (5.130), although few authors seem to recognise this point. As this is a reparameterisation and not a restriction (such as that implied by moving from (5.129) to (5.130)), this is unaffected by any evolution of the intercept, in the sense that the estimation of any change is unbiased. However, in contrast to the slope coefficients, we no longer have an absolute measure of the level of the intercept, and so estimates of the intercept are no longer directly comparable with the results from estimates in levels[10].

10 The levels of the intercepts can be recovered from the means of the regression.

5.4.1. Variances and testing in the differenced unbalanced panels

Taking differences before calculating the matrix means that the expected error term is no longer the same as in section 5.1.4. In the differenced estimator, the error term is (u_t-u_{t-1}). Retaining the assumption that $\mathscr{E}(UU')=I_{NT}\sigma^2$, the correct variance for this term is therefore

$$\mathscr{E}\left[(u_t-u_{t-1})(u_t-u_{t-1})'\right] = \mathscr{E}(u_t'u_{t-1}) + \mathscr{E}(u_tu_{t-1}') = 2I_t\sigma^2 \qquad (5.131)$$

In this case, then the expected value of the error term from (5.40) becomes

$$\mathscr{E}(E'PE) = 2\sigma_u^2(\sum_t (N_t-1) - K) \qquad (5.132)$$

and so the estimated error terms for the three models are now

$$\begin{aligned}\hat{\sigma}_u^2 &= \frac{E'PE_u}{2(\sum_t N_t - T - TK_x)}\\ \hat{\sigma}_p^2 &= \frac{E'PE_p}{2(\sum_t N_t - 1 - K_x)}\\ \hat{\sigma}_r^2 &= \frac{E'PE_r}{2(\sum_t N_t - T - K_x)}\end{aligned} \qquad (5.133)$$

The F-statistics are unchanged as the extra 2s cancel out.

As in the previous sections, note that estimating ζ for the unrestricted model over T separate regressions enables the calculation of time heteroscedastic errors. Allowing for heteroscedasticity in (5.133) gives

$$\mathscr{E}\left[(u_t-u_{t-1})(u_t-u_{t-1})'\right] = I_t(\sigma_t^2 + \sigma_{t-1}^2) \equiv I_t\sigma_{ut}^2 \qquad (5.134)$$

and therefore

$$\hat{\sigma}_{ut}^2 = \frac{e_t'Q_te_t}{N_t - 1 - K} \qquad (5.135)$$

Note that this estimated variance is a compound term, and so there is no need to divide the

RSS by two. These are the reported errors, but again, the F-tests for model restrictions are based on homoscedastic errors for simplicity.

Chapter 6

Linear estimation and testing

The previous chapter outlined various estimators that could be extracted from a properly constituted X'X matrix. These estimators have been implemented in a GAUSS program called XPReg.GP, and this chapter describes briefly the development and working of the program, the extension for instrumental variables, and the hypothesis tests implemented.

6.1 XPReg.GP: Estimation basics

6.1.1 Development of the software

The regression program was first implemented in Autumn 1991, and has passed through various stages reflecting the uses made of it. It was designed originally as a simple (and necessary) tool to obtain statistics and estimates from cross-products, and although in the subsequent development of the program the coding, capabilities and features have changed enormously, the basic principles have changed relatively little.

The five stages of the program have been, roughly,

Version 1	Autumn 1991	Simple OLS cross-section
Version 2	Spring 1992	Time-specific intercepts; analysis of covariance; residual variance analysis
Version 3	Autumn 1992	Fixed-effects (balanced panels only)
Version 4	Spring 1993	Instrumental variables
Version 5	Spring 1994	Proper fixed-effects estimator; time-differencing; instrumental variable and joint significance tests; complete internal rewrite

The original fixed-effects estimator was a relatively simple extension to deal with individual heterogeneity in the manner of section 5.2. However, a flaw in the mathematics meant that the program only dealt with heterogeneity properly for balanced panels. For unbalanced panels the "fixed effects" estimator merely applied a meaningless transformation to the data; this was corrected in Version 5.

The original extraction software was completed and used before the first version of the regression program (the early X'X matrices being used as cross-tabulations), and has remained essentially the same although the speed and efficiency of the programs have improved[1]. In line with Version 5 of the program the extraction software was completely rewritten, the intention being to integrate more fully the complete process from collection of data to analysis of results. Although extraction and analysis are separate tasks, the choice of models available is obviously dependent upon the type of cross-product matrix created, and the type of matrix created depends on the model and estimator.

In early 1994, the University of Stirling agreed with the Department of Employment to provide extraction software enabling general access to the NES in the form of cross-product matrices. Requests to the DE would list the data to be collected, interpreting software would produce extraction software, and the extraction software would produce an X'X matrix to be returned to the researcher. The researcher could then analyse the data using some provided software or his own tools. Under the initial specification the software was designed to produce input for the simple instrumental-variables version of XPReg.GP (Version 4) and a basic working suite of programs was developed. However, the opportunity was taken to reconsider completely the type and nature of potential models and estimators, with a view to implementing those that were both feasible and desirable in the context of an X'X dataset. The result was Version 2 of the extraction software and Version 5 of the analysis program.

---

[1] The original extraction routine was written by Elizabeth Roberts at the University of Stirling.

The extraction software is documented elsewhere, in the NES user instructions to be issued by the DE. This software is the intellectual property of the DE. Stirling University retains control over the regression program code and distribution. Some more basic analysis programs have also been developed and, with a restricted version of XPReg.GP, given to the DE for distribution to users.

It should be noted that, while the extraction software is to some extent specific to the NES, the analytical software is independent of the source of the data. A properly constructed cross-product matrix and some locational information is the sole data requirement.

6.1.2 Collinearity amongst the time dummies

Estimation proceeds using the arithmetic and notation of the previous chapter. The set of valid estimators depend upon the matrices created. Clearly the balanced time differenced estimator can only be run on a balanced full-size dataset as outlined in section 5.3. However, this same dataset could also be used for the fixed-effects model (if T_i=T for all individuals, there is no need for a separate means matrix) and the cross-section models, which treat matrices separately. The actual combinations of matrices, models and estimators are described in a user guide to the software[2].

When the full fixed effects models are being estimated there is a problem of multicollinearity between the time dummies and the individual dummies. Differencing or taking deviations will remove the individual dummies, but will not restore the X matrix to full column rank. This can easily be seen if we consider the deviations transformation on a group of time dummies for T=4:

2 Initial draft available from the DE or the author.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} \\ -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} \\ -\frac{1}{T} & -\frac{1}{T} & 1-\frac{1}{T} & -\frac{1}{T} \\ -\frac{1}{T} & -\frac{1}{T} & -\frac{1}{T} & 1-\frac{1}{T} \end{bmatrix} \tag{6.1}$$

The rank of the transformed matrix is T-1 and not T, and so this matrix is not invertible. However, suppose there are only three observations for one individual. It may thought that this leads to

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1-\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & 1-\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} & 1-\frac{1}{3} \end{bmatrix} \tag{6.2}$$

which has full rank, but this is not the case. In the previous chapter the first matrix in (6.2) was depicted with no zero columns or rows to simplify exposition; in other words the data was packed and so the correct version of (6.2) is

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1-\frac{1}{3} & -\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & 1-\frac{1}{3} & -\frac{1}{3} \\ -\frac{1}{3} & -\frac{1}{3} & 1-\frac{1}{3} \end{bmatrix} \tag{6.3}$$

where the transformed matrix has rank T-1 again. More correctly, note that the proper form of Q_i in (5.48) is always a TxT matrix, but with zeros on the appropriate rows and columns:

$$Q_i = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} - \frac{1}{3}\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix} \tag{6.4}$$

If these zeros did not appear in these places, then the transformation matrix would not sweep out the heterogeneity: spurious values of $-\alpha_i$ would appear in previously blank lines. This then leads to the transformation:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \Rightarrow \begin{bmatrix} 1-\frac{1}{3} & -\frac{1}{3} & 0 & -\frac{1}{3} \\ -\frac{1}{3} & 1-\frac{1}{3} & 0 & -\frac{1}{3} \\ 0 & 0 & 0 & 0 \\ -\frac{1}{3} & -\frac{1}{3} & 0 & 1-\frac{1}{3} \end{bmatrix} \tag{6.5}$$

which again is of rank 2. Dropping one time dummy (one column) will leave this particular matrix still with rank 2. However, this is one individual's record; when the matrix in (6.5) is stacked with the records for other individuals (whose patterns of observations differ) then the overall matrix will be of rank three. Therefore when the moment is taken of every individual's records to produce a 3x3 matrix, it will have full rank and so be invertible.

This makes no qualitative difference to the algebra, and so it was ignored in the previous chapter. As far as estimations goes, the program will automatically drop the first time dummy in a fixed-effects or balanced full-size differencing estimator to make the matrix invertible. In the case of the pooled models, this amounts to dropping the constant completely[3].

The value of this missing constant term can be recovered from the means. In all cases,

---

3 The use of categorical variables also leads to collinearity problems. Selection of these other dummies to be dropped is up to the user.

$$\lambda_1 = \bar{y} - \bar{x}\hat{\beta} \qquad (6.6)$$

where the means are taken over the whole regression. This mean will also incorporate any dummy variables dropped; in other words, it is the expected value of the dependent variable for a "representative individual" - including the mean of the fixed-effects.

The standard errors given in section 5.2 and 5.3 have to be emended for these adjustments. In both cases, one is taken from the denominator of the estimated standard error, reflecting the fall in the number of variables used. Corrected standard errors are given in Appendix A6.

6.1.3 Collinearity between time dummies and incremental variables

In a recent paper, Bell and Ritchie (1996) have shown that allowing coefficients to vary over time has a hitherto unreported side-effect. When variables which increment or decrement periodically over time (such as age, tenure, age of youngest child, et cetera) are included in a regression which has time-varying coefficients, the coefficients on the incrementing variables are poorly identified because of collinearity with the time dummies and any other incrementing variables. The reason is that the addition of an incrementing variable amounts to the inclusion of a person-specific numerator variable and either a trend (in the case of an incrementing cardinal variable) or a secondary set of time dummies (in the case of qualitative variables). The particular effect depends on the model being estimated; for example, a time-invariant model with a cardinal variable loses all useful information except the average numerator value; a model with time-varying parameter retains the trend information but is distorted by cohort effects in an unbalanced panel. For categorical variables, the result is een less clear-cut.

This effect is specific to estimators where the coefficients are estimated jointly over time; thus the cross-section estimator of section 5.1 and the unbalanced differenced estimator of section

5.4 are unaffected because of their block-diagonal nature. However, the unrestricted models of section 5.2 and 5.3 potentially have this problem, and so the interpretation of some coefficients requires some care. Chapter nine discusses a specific example.

6.1.4 Single observations

The fixed effects covariance estimator takes deviations from individual means, and so clearly individuals with only one observation play no significant part in the estimation of the coefficients (although they will affect the calculation of λ_1 in (6.6)). The extraction software creates the main and mean matrices separately, for the reasons of practicality and flexibility discussed at the end of section 5.2.3, and it cannot take account of single-observation cases.

The effect of including single-observation cases and then excluding them is relatively minor, and does not affect the calculation of the coefficients. It does mean that the calculation of the estimated variance in, for example, (5.91) will have values for ΣT_i and N different to those arising from an estimator in which the single-observations are initially excluded. However, the number of single observations in any year is only around 3-5% of the total observed, so, while N may have 20% of single observations (and so be roughly 20% "too big") over the full sixteen years of the survey, ΣT_i is only around 4% "too big". As ΣT_i is easily the dominant term in the calculations for all but very short study periods, it seems likely that the estimated variance is slightly underestimated in the fixed-effects covariance estimator.

The other area where having single observations upsets the results is in the displayed means, which include single observations in the calculations as they represents the mean values of each variable for a particular period. However, they play no part in the fixed effect calculations.

The single observation issue does not affect the cross-section studies, as these are only

concerned with observations within a period and not the correlation between observations over time. The differencing calculations likewise are unaffected: for the balanced panel a single observation period is not feasible, and for the unbalanced panel the extraction software rejects single observations.

6.2 Hypothesis Testing

One of the more serious limitations of the regression program is the area of hypothesis testing. Many of the more informative tests are based on an analysis of the residual errors (serial correlation, heteroscedastic-consistent errors, et cetera). The relevant statistics would have to be calculated by sending a program to the DE offices, and so such statistics are not provided by the program. This is not a very satisfactory solution, but at present there is no alternative.

The analytical features generated automatically by the program are limited to what is available under the X'X format: essentially anything involving the total, estimated and residual sums of squares and other linear combinations of the variables. These are all available from the cross-product matrix by some method or other, and so some useful tests and statistics may be produced.

Given the TSS, ESS and RSS, then R^2 and R^2 adjusted for degrees of freedom may be calculated. The estimates of the variance lead to the t-statistics via the variance of β:

$$Var(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1} \tag{6.7}$$

and F-tests for the general significance of the regression are available from

$$F = \frac{ESS/dof1}{TSS/dof2} \tag{6.8}$$

where dof1 and dof2 are the appropriate degrees of freedom. As noted in chapter five, F-tests

for choosing between the unrestricted, pooled and restricted models can also be calculated. Because the degrees of freedom are more complex for panel models (especially unbalanced ones), these are given in full in Appendix A6.

Equation (6.8) is a restricted form of a more general hypothesis-testing framework whereby sets of hypotheses may be tested jointly. The program can automatically calculate one set. Variables may be defined as parts of "groups"; usually, each of the different dummy variable groupings is generally treated as a set of related variables. The program then tests for the joint significance of all the groups which have two or more members; for example, the program will report an F-ratio for whether the occupation dummies as a whole contribute anything significant to the estimates, as well as the usual t-ratios for each individual occupational dummy.

These F tests are all based on the assumption of normality in the error term. There are as yet no tests in the program for this assumption, as most tests are based on an analysis of the residuals, and these are unavailable to the regression program.

6.3 Estimated variance analysis

The program provides a breakdown of the variance by variable grouping, following a suggestion of Blackburn (1990). First, note that

$$TSS = ESS + RSS = \hat{\beta}'X'y + RSS = \hat{\beta}'X'X\hat{\beta} + RSS \qquad (6.9)$$

Assume that the X variables can be organised into M groups of variables. The number of elements in each group may vary; for example, a set of occupation dummies may count as one group, whereas a wage variable may be thought of as a one-element group. Then

$$X \equiv [X_1 \;\; X_2 \;\ldots\; X_M] \qquad \beta \equiv [\beta_1' \;\; \beta_2' \;\ldots\; \beta_M']' \qquad (6.10)$$

with

$$X'X = \begin{bmatrix} X_1'X_1 & X_1'X_2 & & \\ X_2'X_1 & X_2'X_2 & & \\ & & \ddots & \\ & & & X_M'X_M \end{bmatrix} \tag{6.11}$$

Using these definitions gives

$$ESS = \hat{\beta}'X'X\hat{\beta} = \sum_{m=1}^{M} \hat{\beta}_m'X_m'X_m\hat{\beta}_m + 2\sum_{m=1}^{M}\sum_{n=m+1}^{M} \hat{\beta}_m'X_m'X_n\hat{\beta}_n \tag{6.12}$$

In other words, the explained sum of squares can be broken down into two components: firstly, the contribution of each group to the total explained variance; and secondly, the explained covariance between groups, which may be positive, zero, or negative.

This information is useful as it gives an indication of how the groups interact with one another; more importantly, it weights the results by the estimated coefficients. Thus it can be shown not whether two variables interact (which could be found simply from the covariance matrix of X), but whether that interaction is important to the relationship being studied.

This is perhaps most useful when regressions are run with time-varying coefficients. If, for example, the contribution of age to the variance in wages declines over time, this could be attributable to a decline in the variance of ages (the population is more homogeneous and so age has less chance of explaining wage differentials); a decline in the coefficient values (reflecting a decline in the return to age); or changes in both, not necessarily in the same direction.

One simple way to test this is by studying how the age variance changes over time. However, this does not take account of any scale effects. An alternative suggested by Blackburn (1990) is to apply the coefficients for one "base" year to the covariance matrices for each year in

turn. The result is effectively an index of the relevance of a variable group in the regression[4].

If a large part of the variance of y is explained by the own-variance terms, then this suggests that the various influences on the dependent variable are largely independent of one another. this can be seen by noting that if the variables are independent then as the number of observations becomes large

$$\hat{\beta}_m' X_m' X_m \hat{\beta}_m \Rightarrow +\infty \qquad \hat{\beta}_m' X_m' X_n \hat{\beta}_n \Rightarrow \hat{\beta}_m' \bar{X}_m' \bar{X}_n \hat{\beta}_n \tag{6.13}$$

where the covariances converge to the separate means of each group of variables. If the variables are not independent then

$$\hat{\beta}_m' X_m' X_m \hat{\beta}_m \Rightarrow +\infty \qquad \hat{\beta}_m' X_m' X_n \hat{\beta}_n \Rightarrow \pm\infty \tag{6.14}$$

However, in most of the estimators the analysis is done on deviations from either time or individual means. Thus the means in (6.13) will be the means of the transformed variables. Clearly the mean value of these transformed variables will be zero; moreover, the variables will converge to the sum of their mean values if the number of observations becomes large and the variables are not independent. Thus (6.13) and (6.14) become

$$\hat{\beta}_m' X_m' X_m \hat{\beta}_m \Rightarrow +\infty \qquad \hat{\beta}_m' X_m' X_n \hat{\beta}_n \Rightarrow 0 \tag{6.15}$$

when the variable groups are independent, and

$$\hat{\beta}_m' X_m' X_m \hat{\beta}_m \Rightarrow +\infty \qquad \hat{\beta}_m' X_m' X_n \hat{\beta}_n \Rightarrow \pm\infty \tag{6.16}$$

when they are not. For the balanced time-differenced estimator of section 5.3, the sum of all the transformed X variables is equal to the sum of the first observations for each individual, and so (6.13) and (6.14) will reflect the averages of these first observations.

Note that using (6.15) and (6.16) as indicators of independence will depend to some extent on the scaling of the variables, particularly when categorical and continuous variables are

4 The usual index number problem arises. With no preferences for one particular base over another, the simplest one to implement was chosen.

mixed.

6.4 Instrumental Variables

6.4.1 Instrumental variable regression

One simple extension to the estimators outlined in the previous chapter is to allow for the use of instrumental variables. The linear generalised instrumental variables estimator (GIVE) can be derived in a number of ways; the GMM interpretation is given below (Hall(1993); see Johnston (1984) for a 'traditional' derivation). Let Z be a matrix of instruments uncorrelated with the error vector u. Then

$$\begin{aligned} plim\ \frac{1}{N}(X'(X\beta - y)) &= plim\ \frac{1}{N}(X'u) \neq 0 \\ plim\ \frac{1}{N}(Z'(X\beta - y)) &= plim\ \frac{1}{N}(Z'u) = 0 \end{aligned} \tag{6.17}$$

by assumption. Defining a quadratic form for the sample condition

$$S \equiv (\frac{1}{n}Z'u)'W(\frac{1}{n}Z'u) = (\frac{1}{n}Z'(y - X\beta))'W(\frac{1}{n}Z'(y - X\beta)) \tag{6.18}$$

where W is a weighting matrix not dependent upon β which converges in probability to a positive definite matrix. Differentiating to find the value of β which minimises this expression,

$$\begin{aligned} \frac{\partial S}{\partial \hat{\beta}} = 0 &= -\frac{2}{n^2}X'ZWZ'y + \frac{2}{n^2}X'ZWZ'X\hat{\beta} \\ \hat{\beta} &= (X'ZWZ'X)^{-1}X'ZWZ'y \end{aligned} \tag{6.19}$$

The optimal choice of weighting matrix is $W=n(Z'Z)^{-1}$, and so the linear GIVE is

$$\hat{\beta} = (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'y \tag{6.20}$$

Where Z and X have the same rank (that is, Z'X is square), (6.20) collapses to

$$\hat{\beta}_{iv} = (Z'X)^{-1}Z'y \qquad (6.21)$$

which is the same form as the OLS estimator in (4.2) apart from the substitution of X' by Z'. All the matrix arithmetic of the previous chapters therefore still holds, with the obvious proviso that the rows and columns selected from the $\Sigma v_i'v_i$ matrix will differ if $Z \neq X$[5]. This holds for the means matrix calculations too.

Where Z and X have the same rank, the calculations for standard errors are

$$\hat{\sigma}^2 = \frac{(y - X\hat{\beta}_{iv})'(y - X\hat{\beta}_{iv})}{dof} = \frac{y'y - 2\hat{\beta}_{iv}'X'y + \hat{\beta}_{iv}'X'X\hat{\beta}_{iv}}{dof} \qquad (6.22)$$

$$Var(\hat{\beta}_{iv}) = \hat{\sigma}^2(Z'X)^{-1}(Z'Z)(X'Z)^{-1}$$

where dof is the appropriate number of degrees of freedom for the estimators (Johnston (1984), p366). These are the same as for the OLS estimator, as the transformation matrices, the number of observations and restrictions and the number of periods remain the same in the two estimators.

The instrumental variables estimator therefore requires five cross-product matrices (X'X, X'y, Z'Z, Z'X, Z'y) in contrast to the two needed for OLS (X'X, X'y). However, these can all be created from the raw cross-product matrix as long as the instruments already exist in the matrix. This is a significant disadvantage in that it greatly limits the options for two-stage solutions. For example, to run a two stage least squares regression would involve creating a dataset; running the first stage regression; writing a new extraction program using the estimated coefficients; creating a new dataset; and running the second stage regression. This may be tedious and threatens significant time penalties for a poor choice of regressors for the

5 Other minor changes from the programming point of view are that the matrix is not symmetric, and it is no longer necessarily positive definite.

first round coefficients[6].

It was noted in section 2.3 that dynamic models could be consistently estimated by the use of lagged dependent variables as instruments. The extraction software does allow for lagged variables (including in the differenced format). Therefore dynamic models are feasible, although as the software currently stands this would involve the loss of a number of observations, and the estimator is unlikely to be very efficient. In the longer term a more flexible and efficient approach to dynamic models would be desirable and there are no theoretical difficulties to replicating, for example, the simpler Arellano and Bond (1991) estimators (that is, those with spherical errors).

If the number of instruments exceeds the number of regressors, then no new conceptual or practical difficulties arise. All the data in (6.20) is available from the cross-product matrix and the standard errors of the coefficients in (6.22) also need to be amended:

$$Var(\hat{\beta}_{iv}) = \hat{\sigma}^2[(X'Z(Z'Z)^{-1}Z'X)]^{-1} \qquad (6.23)$$

It is clear that the non-square z'x matrix does not require any additional information: all the data needed is somewhere in the cross-product matrix. This is a consequence of the linear nature of the IV estimator used here.

Note that even if more instruments than variables are included, only the minimum number of instruments play a significant part in the regression; in other words, the effective Z'X matrix is square. This does not mean that choosing minimal instruments will necessarily be an efficient IV solution, but that choosing more instruments than the minimum effective set will increase the variance of the estimates (Bowden and Turkington (1984) pp29-36).

---

6 This does not weaken the claim that the cross-product matrix is an effective tool for linear regression; in fact, the case is more compelling for IV regressions, as the cross-product could contain numerous first-round estimates for a relatively small increase in matrix size.

6.4.2 Testing the IV specification

Two tests on the instruments are easily implemented using the cross-product matrix.

The first test statistic is a Hausman test for regressor-disturbance independence. This relies on the potential inconsistency of the OLS estimator compared to the supposed consistency of the IV estimator to provide a testable distance measure.

The test is between two hypotheses:

$$\begin{aligned} H_0: \quad & plim\ \frac{1}{N}X'u = 0 \qquad plim\ \frac{1}{N}Z'u = 0 \\ H_1: \quad & plim\ \frac{1}{N}X'u \neq 0 \qquad plim\ \frac{1}{N}Z'u = 0 \end{aligned} \tag{6.24}$$

Under the null hypothesis, OLS estimates of the coefficients are consistent and efficient, whereas the IV estimates are consistent but inefficient. However, under the alternative hypothesis, OLS is inconsistent. Define

$$\hat{q} \equiv \hat{\beta}_{iv} - \hat{\beta}_{ols} \tag{6.25}$$

Then the Hausman test is whether $\hat{q}$ is significantly different from 0; that is, whether

$$n\hat{q}'\Omega\hat{q} = 0 \tag{6.26}$$

where Ω is a weighting matrix. The obvious choice for this weighting matrix is the inverse covariance of $\hat{q}$, and it can be shown (Hausman (1978); Bowden and Turkington (1984)), that under a fairly general set of assumptions, an asymptotic test statistic for (6.20) is

$$\hat{q}'\hat{Var}(\hat{q})^{-1}\hat{q} \sim \chi^2(K) \tag{6.27}$$

where K is the number of variables, n the number of observations, and $\hat{Var}(\hat{q})$ is a consistent estimate of the variance of $\hat{q}$,

$$Var(\hat{q}) = Var(\hat{\beta}_{iv}) - Var(\hat{\beta}_{ols}) \tag{6.28}$$

Under the null hypothesis, $Var(\hat{q})$ should be large (as the IV estimate of β is inefficient) and

$\hat{q}$ small, and so a large value for (6.23) indicates rejection of H_0[7].

It should be noted that this test is predicated on the assumption that the IV estimate is consistent even if H_0 is rejected. Without this assumption, this merely amounts to a test for the relative independence of Z compared to X. In other words, the Hausman test compares the relative performance of two estimators, both potentially erroneous. Thus this test does require a degree of confidence about the consistency of the IV estimator[8].

The second test is a general one for the validity of the instruments used, the Sargan test[9]. The test statistic is simply

$$q = \frac{1}{n\hat{\sigma}^2}(Z'e)'W(Z'e) = \frac{(y-X\hat{\beta})'Z(Z'Z)^{-1}Z'(y-X\hat{\beta})}{\hat{\sigma}^2} \sim \chi^2(p-k) \qquad (6.29)$$

where p and k are the number of columns in Z and X respectively. The basic idea behind the test is that, if the instruments are uncorrelated with the error terms, then e'Z(Z'Z)$^{-1}$Z'e should converge to n independent squared errors, and so q should be small. The adjustment for degrees of freedom reflects the fact that, of the p columns in Z, k will be constrained by the action of setting $\partial S/\partial\beta=0$ in (6.19).

If p=k, then clearly the Sargan test is not appropriate. The weighting matrix is irrelevant, and q in (6.29) collapses to $\sigma^2/\sigma^2 = 1$. The reason is that the coefficient vector uses all the information in Z by construction, whereas in the overidentified case only the most effective columns of Z are significant (Bowden and Turkington (1984), p29).

---

7 The "n" in (6.22) and (6.23) relates to the number of observations used to calculate the coefficient estimates, for the Hausman test is based on N repeated observations on a parameter set. For our purposes, n is N_t for the cross-section estimators, and $\Sigma_i T_i$ for the fixed effects estimators.

8 See Bowden and Turkington (1984, pp52-55) for a discussion of what the Hausman test actually measures.

9 The Sargan test as presented here can be seen as a particular form of Hansen's test of "overidentifying restrictions" in the GMM methodology (see Hall (1993) for the derivation of the general test statistic).

Appendix A6 Degrees of Freedom in the Linear Model

The general solution for degrees of freedom are

$$dof = n-k-r \tag{A6.1}$$

where n is the total number of observations, k is the number of variables, and r is the number of other restrictions: for example, taking deviations from time means (as in the cross-section case) means an additional T restrictions in the unrestricted model as the sum of the variables for each of the T periods must sum to zero. The appropriate calculations for these results are

$$\hat{\sigma}^2 = \frac{RSS}{n-k-r} \tag{A6.2}$$

$$\frac{RSS}{\sigma^2} \sim \chi^2(n-k-r) \tag{A6.3}$$

$$R^2 = 1 - \frac{RSS/n}{TSS/n} = \frac{ESS}{TSS}$$
$$\bar{R}^2 = 1 - \frac{RSS/(n-k-r)}{TSS/(n-r)} \tag{A6.4}$$

A general test of q linear restrictions of the form $R\beta=r$ has a χ^2 form:

$$(R\beta - r)'[\sigma^2 R(X'X)^{-1}R']^{-1}(R\beta - r) \sim \chi^2(q) \tag{A6.5}$$

which, combined with (A6.3) to remove the unknown σ, gives the F-test

$$\frac{(R\beta - r)'[R(X'X)^{-1}R']^{-1}(R\beta - r)/q}{RSS/(n-k-r)} \sim F(q, n-k-r) \tag{A6.6}$$

For a test of q variables being zero, this test collapses to

$$\frac{(RSS_r - RSS_u)/q}{RSS_u/(n_u-k_u-r_u)} \sim F(q, n-k-r) \tag{A6.7}$$

where the r subscript refers to the RSS from some restricted regression and u the unrestricted or "base" regression. The value of q is slightly complicated if the number of other restrictions

on the regression changes; for example, in the cross-sectional regressions the number of restrictions "r" may be T or 1, although in the fixed effects regressions the number of restrictions is always N as deviations are taken around the individual mean. This will change the calculations for the estimated variance used to derive (A6.7). The correct value for "q" in (A6.7) is

$$q=(n_u-k_u-r_u) - (n_r-k_r-r_r) \qquad \textbf{(A6.8)}$$

For the joint significance test ($\beta \neq 0$), q will clearly be k: all the model results show deviations from some mean value, and so the appropriate test from (A6.6) is for all remaining variables being zero.

Let T be the number of periods under study, N be the number of individuals observed (not observations), T_i and N_t the observations for individual i and per period t, respectively, and K the number of variables in the X matrix, excluding the constant term. One time dummy is dropped in the fixed-effects and balanced differenced models. Then the degrees of freedom for the reported statistics are given in Table A6.1, overleaf.

Table A6.1 Degrees of freedom

		Unrestricted	Pooled	Restricted
σ^2	Cross-section	ΣN_t - T - TK	ΣN_t - 1 - K	ΣN_t - T - K
	Fixed-effects	ΣT_i - N - TK - (T-1)	ΣT_i - N - K	ΣT_i - N - K - (T-1)
	Balanced differenced	$2\Sigma T_i$ - 2N - TK - (T-1)	$2\Sigma T_i$ - 2N - K	$2\Sigma T_i$ - 2N - K - (T-1)
	Unbalanced differenced	$2(\Sigma N_t$ - T - TK)	$2(\Sigma N_t$ - 1 - K)	$2(\Sigma N_t$ - T - K)
R^2	Cross-section	ΣN_t - T - TK ΣN_t - T	ΣN_t - 1 - K ΣN_t - 1	ΣN_t - T - K ΣN_t - T
	Fixed-effects	ΣT_i - N - TK - (T-1) ΣT_i - N	ΣT_i - N - K ΣT_i - N	ΣT_i - N - K - (T-1) ΣT_i - N
	Balanced differenced	$2\Sigma T_i$ - 2N - TK - (T-1) $2\Sigma T_i$ - 2N	$2\Sigma T_i$ - 2N - K $2\Sigma T_i$ - 2N	$2\Sigma T_i$ - 2N - K - (T-1) $2\Sigma T_i$ - 2N
	Unbalanced differenced	$2(\Sigma N_t$ - T - TK) $2(\Sigma N_t$ - T)	$2(\Sigma N_t$ - 1 - K) $2(\Sigma N_t$ - T)	$2(\Sigma N_t$ - T - K) $2(\Sigma N_t$ - T)
Joint F	Cross-section	TK ΣN_t - T - TK	K ΣN_t - 1 - K	K ΣN_t - T - K
	Fixed-effects	TK + T - 1 ΣT_i - N - TK - (T-1)	K ΣT_i - N - K	K + T - 1 ΣT_i - N - K - (T-1)
	Balanced differenced	TK + T - 1 $2\Sigma T_i$ - 2N - TK - (T-1)	K $2\Sigma T_i$ - 2N - K	K + T - 1 $2\Sigma T_i$ - 2N - K - (T-1)
	Unbalanced differenced	TK $2(\Sigma N_t$ - T - TK)	K $2(\Sigma N_t$ - 1 - K)	K $2(\Sigma N_t$ - T - K)
Specification Tests	CS Pooled v	(T-1)(K+1) ΣN_t - T - TK		
	CS Restricted v	K(T-1) ΣN_t - T - TK	T-1 ΣN_t - T - TK	
	FE Pooled v	(T-1)(K+1) ΣT_i - N - TK - (T-1)		
	FE Restricted v	K(T-1) ΣT_i - N - TK - (T-1)	T-1 ΣT_i - N - K - (T-1)	
	BD Pooled v	(T-1)(K+1) $2\Sigma T_i$ - 2N - TK - (T-1)		
	BD Restricted v	K(T-1) $2\Sigma T_i$ - 2N - TK - (T-1)	T - 1 $2\Sigma T_i$ -2 N - K - (T-1)	
	UD Pooled v	(T-1)(K+1) ΣN_t - T - TK		
	UD Restricted v	K(T-1) ΣN_t - T - TK	T-1 ΣN_t - T - TK	

Chapter 7

Observation histories

The observation histories (OHs) arose from a desire for a practical way of obtaining empirical hazard functions from the NESPD. The hazard function is an indication of the likelihood of an individual dropping out of the data set at a particular time, given that the individual is still in the dataset at that point. Empirical hazard functions, reflecting the observed rate of attrition, can be constructed from

$$\frac{\text{(number in period T) - (number in period T+1)}}{\text{(number in period T)}}$$

The obvious way to create these figures for each period is simply to read the whole dataset and sum the relevant numbers for each period. However, a far more efficient and informative aggregation of the data is possible by reversing the type of information collected: instead of saving information on observations for individuals, the characteristics of individuals with particular "observation histories" or patterns of observation in the dataset are recorded[1].

7.1 Creating the observation histories

The key to the OH analysis is to note that a pattern of observation in the dataset constitutes a binary number. Individuals are identified in the NES by their national insurance numbers (NINos), represented internally as a six digit number. A missing observation has a missing NINo. Create an observation flag, representing a missing NINo by a zero and any valid number by a one. Then the record of observations translates into a sixteen-bit number (a "flag vector").

For example, consider the record of an individual with the NINo "000100" over the sixteen

1 The methods for storage and access of information described in this chapter are variants on a programming technique called "key transformation" or "scatter storage". A discussion can be found in most computer science texts dealing with system-level programming or data manipulation techniques; see, for example, Page and Wilson (1983) pp173-183.

years of the NES[2]. This is represented as Table 7.1.

This individual was observed first in 1978 and on seven subsequent occasions, the last being in 1988. The longest continual period of observation was four years, and the person had three periods of consecutive observation.

Year	Nino	Other data	Flag
1975	(missing)		0
1976	(missing)		0
1977	(missing)		0
1978	000100	.....	1
1979	000100	.....	1
1980	000100	.....	1
1981	000100	.....	1
1982	(missing)		0
1983	000100	.....	1
1984	000100	.....	1
1985	000100	.....	1
1986	(missing)		0
1987	(missing)		0
1988	000100	.....	1
1989	(missing)		0
1990	(missing)		0

Table 7.1 Creating the flag vector

This could have been collected to give totals for the dataset for each possible start year and end year, but there is no feasible way to store, for example, the information on run length so that is accessible.

However, consider the creation of a "hashing vector" of constants in powers of two:

$$h \equiv [1 \;\; 2 \;\; 4 \;\; 8 \;\; 16 \;\; 32 \;\; 64 \;\; 128 \;\; 256 \;\; 512 \;\; 1024 \;\; 2048 \;\; 4096 \;\; 8192 \;\; 16384 \;\; 32768]'$$

Take the inner product of this vector and f, the flag vector from table 7.1:

$$h' \cdot f = (8+16+32+64+256+512+1024+8192) = 10104$$

This is a unique reference, only generated by a particular pattern in the flag vector. No other combination of missing/observed flags will give this number when multiplied by the hash vector. Now consider a 65535x1 column vector, initially set to zero. Adding one to the 10104-th element in this vector records that this individual has the above pattern. If this cell

2 This is an example pattern and does not reflect the actual characteristics of NESPD individual "000100".

now contains, for example, ten, then nine other people have also been found with that same pattern of observations.

Running through the entire database, the end result will be a vector containing the numbers of people with each of the possible 65535 patterns of observation[3]. Because the observation pattern itself provides a unique index into the dataset, there is no need to separately record which patterns were used to generate which totals. Similarly, for any cell in the vector, the row number of that cell allows the identification of the observation pattern experienced by the people counted by that cell as the transformation between pattern and index is a one-to-one mapping.

For example, if the cell in row number 10104 contained "ten", then ten people have same original observation pattern. That pattern can be recovered by repeated modulo division of the row index by the elements of the hashing vector, starting with the largest number[4].

Taking the number 10104, the result of this calculation is given in Table 7.2; it can be seen that the pattern of ones and zeros obtained, once inverted, replicates the original selection pattern in Table 7.1.

Index	Divisor	Result	Remainder → index
10104	32768	0	10104
10104	16384	0	10104
10104	8192	1	1912
1912	4096	0	1912
1912	2048	0	1912
1912	1024	1	888
888	512	1	376
376	256	1	120
120	128	0	120
120	64	1	56
56	32	1	24
24	16	1	8
8	8	1	0
0	4	0	0

Table 7.2 Recovering the observation pattern

Just as the total number of people can be stored, so can other information: for example, age, wage, numbers in the private sector, and so on. However, these are now

---

3 The number of patterns is 2^{16}-1 because there is obviously no possibility of an individual being in the dataset but having no observations at all. For observation histories looking at only a subset of the period of the NES, this is a feasible alternative and account must be taken of this.

4 Modulo division returns the remainder from an integer division; that is

$$x\%y \equiv \text{remainder}(x/y)$$

where "%" represents modulo division and "/" integer division.

aggregated for all people with a particular pattern. These may be stored for each year or for the whole period, as long as they are stored on the correct row of the output matrix so that the pattern can be reconstituted. For example, one extraction run might store age in 1975 and then wages for each period.

This is an extremely efficient way of storing the pattern based information, but it has three drawbacks. Firstly, these observation matrices tend to be large objects, which double in size with every increase in the period under review[5]. Secondly, there is less scope to make individual inferences. Data is only disaggregated down to the level of the OH, and so only total or average figures are available for groups of people with particular observation patterns. Thirdly, as a result of this last point, the variables stored are less informative, particularly with regard to interactive and qualitative variables.

However, these drawbacks are relatively minor, and the OHs do allow a very different analysis to the cross-product matrices. Although not completely disaggregated, the information available is far more detailed than is possible with any aggregate statistics. Moreover, the techniques described here are easily extended to cope with differing data requirements; for example, allowing for multiple "destinations" (part-time work, full-time work, unemployment). An incidental benefit, but one which is important for the NESPD, is that the data, being aggregated, are not subject to the confidentiality restrictions and so may be removed from the DE and analysed at the researcher's leisure.

Section 7.3 describes potential applications for the OHs. Chapter eight is based on analyses of OHs, and a large number of the statistics dotted throughout the thesis have been generated from them.

---

5 The "table lookup" algorithm described here is the fastest key transformation technique, and also the simplest. More complex methods using less storage space are well-documented, but these rely on a sparse storage vector for their effectiveness and are therefore not often appropriate for the NES. The analytical routines are also greatly complicated.

7.2 Analysing the observation histories

Taking the compacted information and listing the variables associated with each pattern is of limited interest; more useful would be to specify and analyse subsets of the patterns. For example, it could be desirable to select only those who appear in the dataset in 1975 and have at least two consecutive observations.

As for the cross-product matrices, this analysis requires software designed for the purpose. Unlike XPReg, this software is, to some extent, specific to the particular type of OH created, although the same programs may be used for several OHs. However, the basic principle is essentially the same. Similar compression techniques to those outlined above may be used to store useful information about the patterns: number of observations, first observation, longest period of continuous observation, and so on. By reversing the compression operation, as detailed above, vectors of characteristics of the observation patterns are created which may be used in logical operations to select patterns with particular characteristics. The key transformation technique described above thus provides an effective way of both storing data and analysing it[6].

7.3 Using the observation histories

7.3.1 Descriptive analyses

The most obvious use of the OHs is in descriptive statistics of the dataset, such as those presented in chapter eight. As information is identified for every possible pattern, it is relatively easy to form simple statistics such as frequencies of observations and absences, lengths of continuous observation, and so on. When more data then just numbers of

6 The mechanics of extracting information in simple cases is described elsewhere.

individuals for each pattern is recorded, the scope for this analysis increases. Suppose the numbers of individuals moving between towns was noted for each observation pattern. Then it is a simple matter to calculate the frequency of moves for people with a particular OH. Aggregated dataset statistics are straightforward to calculate for years or combinations of years.

A useful feature is the ability to follow cohorts within the dataset. For example, a cohort could be constructed of all those observed the whole time from 1975 to 1990 and a second of all those observed in 1975 and 1990 but with at least one missing observation in the meantime. This could then provide a basis for the comparative analysis of wage growth and the effect of absence.

7.3.2 Analysis of transitions

The above example used the OHs to map an individual's pattern of observation. However, the principle is clearly extensible to the analysis of other changes of state. Two examples illustrate this use.

The first extension is to allow for multiple destinations rather than a simple observed/not observed split. The NESPD has been aligned with the Department of Social Security's Juvos dataset, which records periods of unemployment[7]. Observations or missing values may then be classified as part-time work, full-time work, known unemployment or just missing. This allows transition probabilities (and associated changes in average wages) to be calculated for all possible combinations of the four states. The potential for state dependence in the labour market has been pointed out by many authors and is central to a number of theories of segmented or two-tier labour markets.

---

7 Unfortunately, the Juvos data proved to be initially unreliable, particularly in the first few years. Although the first release now appears to have been sorted out, the data was not available for this thesis.

A second extension is to split observations into categories; for example, covered or not covered by collective agreement. This would allow transitions between union and non-union status to be modelled much more accurately than is possible with summary statistics. Transitions between observed states have received much less attention than moves between full-time and part-time work or employment and unemployment, for example. However, some authors (for example, Card (1994) on union membership) have attempted to analyse and allow for changes in state rather than just using current status. The OH approach is ideally suited to this.

Both of the examples given move away from the binary observed/not observed decision of section 7.1. However, the only significant change is that the hashing vector needs a higher base number: base four in the first example, base three in the second. The OH principle remains the same however many possibilities are analysed.

7.3.3 Estimation: the pseudo-panel dataset

As the OHs amount to a grouped dataset they can be used to specify and tests models in the same way as any other dataset. The necessary corrections for efficient linear estimation on grouped data are straightforward, although for non-linear models the aggregation can cause problems if the number of individuals in the groups is small - which it often is.

However, from an econometric viewpoint, an extremely appealing feature of the OHs is that the same individuals appear (or are absent) in each year for a given pattern. This means that there is a constant panel effect for each pattern, and so panel estimation techniques may be employed as if the OHs constituted a "true" panel dataset.

Consider the usual relationship for individual i in period t allowing for individual heterogeneity:

$$y_{it} = x_{it}\beta + \alpha_i + u_{it} \tag{7.1}$$

Collecting information on all individuals, using some known characteristic grouping variable, means that the model for a pattern p in period t is

$$\bar{y}_{pt} = \bar{x}_{pt}\beta + \bar{\alpha}_{pt} + \bar{u}_{pt} \tag{7.2}$$

where

$$\bar{y}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i \in p} y_{it} \quad \bar{x}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i \in p} x_{it} \quad \bar{\alpha}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i \in p} \alpha_i \quad \bar{u}_{pt} \equiv \frac{1}{N_{pt}}\sum_{i \in p} u_{it} \tag{7.3}$$

There has been some interest in recent years in using repeated cross-sections to construct "pseudo-panels" embodying the relationship in (7.2); Verbeek (1992) surveys these. The attraction of these models is that they enable some panel analysis techniques to be applied to non-panel datasets. However, unlike true panel datasets, the pattern-specific effect α_{pt} is not constant over time. A pattern or group will be formed of different individuals in different years and the number of individuals in a group may vary over time. If

$$plim \quad \frac{1}{N_p}\bar{x}_{pt}\bar{\alpha}_{pt} = 0 \tag{7.4}$$

then OLS estimation of (7.2) is consistent and unbiased; but if (7.2) does not hold then OLS is inconsistent.

The option of true panel models, to avoid the correlation by treating the pattern effect as a fixed parameter, runs up against the identifiability problem as there are PT α-terms to be identified in PT equations. One solution is to consider (7.2) as an errors-in-variables problem, for which appropriate instrumental variable techniques have been developed (see Verbeek (1992); Bowden and Turkington (1984))[8]. However, this reduces the appeal of these pseudo-panels over cross-section specifications.

8 An alternative is to assume that N_{pt} is large and the α term is relatively small and so ignorable.

In the case of the OHs, this problem does not arise. Each pattern represents a particular OH; only individuals with the same OH over the whole period will be marked as belonging to that pattern; and the same individuals appear every year. Therefore, for any one OH,

$$\bar{\alpha}_{pt} = \bar{\alpha}_p \tag{7.5}$$

and so (7.2) becomes

$$\bar{y}_{pt} = \bar{x}_{pt}\beta + \bar{\alpha}_p + \bar{u}_{pt} \tag{7.6}$$

Thus the pattern effect can be validly treated as a fixed parameter which is identifiable; either the covariance estimator or using dummy variables will allow consistent estimation of β. Alternatively, it can be treated as a random effect, with the appropriate estimation method being used.

7.3.4 Hazard and survivor functions

As mentioned, the desire to construct hazard and survival functions was a motivating factor for this work. Consider using the OHs to group individuals. The most obvious choice of "group" is to consider each yearly cohort, but these functions could also be calculated for only those sub-groups have no missing observations; or for those with only start and end observations; or for the whole dataset rebased to a common start period; and so on.

Let N_{gt} represent the number of individuals still deemed to be in group g in year t, a number which is easily retrieved from the OHs. Then empirical hazard and survivor functions for the group g are determined from

$$\textit{hazard rate} \equiv \frac{N_{gt+1} - N_{gt}}{N_{gt}} \qquad \textit{survival rate} \equiv \frac{N_{gt}}{N_{g1}} \tag{7.7}$$

Alternatively, one could create functions which reflect the probability of going missing in a particular year rather than leaving the dataset for good, or the likelihood of returning after a missing observation, and so on. Clearly, a wide variety of probabilistic measures are

available from manipulation of the OHs.

7.4 Summary

The above sections illustrate ways in which the OHs can be constructed and used. Together with the cross-product matrices, the OHs provide an efficient way of extracting large amounts of information from the data in an easily digestible form. Clearly, all the information required could also be calculated by reading the dataset and storing the desired totals. However, by using the OHs, one pass through the dataset provides the potential to construct many more statistics. For example, it may be that, following an analysis of the effect of absence on the wages of the 1975 cohort, it becomes desirable to separate the effects of the timing and length of absence. This information is already available in the OHs without the need to return to the NESPD. The principles of the OH can easily be extended to encompass efficient storage of other information on discrete states.

One final advantage of using the OHs is that, being semi-aggregated data, they do not suffer from the confidentiality restrictions on the NESPD or the practical difficulties caused by its size. They are therefore appealing as a compact panel data set which can be removed from the DE premises and analysed using standard econometric packages under the direct control of the researcher.

This chapter ends the description of the data collection methods used to analyse the NESPD at the University of Stirling. There are a number of other ways of analysing the NESPD, of which the most obvious is in the construction of simple transition matrices (which are to some extent already embodied in the cross-product matrices) and labour market cohorts. All of these are used to some extent in the following chapters, but the bulk of the analysis is performed using the cross-product and the OHs.

Chapter 8

Sex, age, and transitions - the cohort effect

8.1 Introduction

This chapter discusses some characteristics of the participants in the NESPD. Although summary statistics for the NES have been presented in a number of places, the new feature is that the basic unit of analysis is the cohort. Two definitions of cohort are used. The more usual type of cohort, of a group of individuals of the same age followed over their working life, is referred to as the *employee cohort.* In this chapter the *data cohort* is also used, where individuals are grouped according to their first appearance in the dataset. The advantage of the data cohort, constructed from the observation histories, is that patterns of observation can be tracked in detail to illuminate simple counts of numbers employed.

In this chapter the changing nature of the dataset and, by implication, the labour market is examined. As the labour market is followed over time, the individuals who make up the labour market (and population drawings such as the NES) will also change. Even on the simplest assumption that these changes are due simply to ageing (old workers leave, young workers join), the job characteristics for cohorts will still differ from those of the market as a whole. Suppose that wages rise with age. Then the growth in average market wages will understate the total wage growth experienced by individual workers as older, higher-earnings workers are replaced by younger, low-earnings workers.

The situation where the composition of the panel changes on a regular basis with selection and deselection being based on random criteria is called a rotating panel. If the distribution of characteristics of the constituent individuals remains constant as these constituents change then the rotating panel presents few statistical difficulties. An adequate description of the labour market can be achieved by studying one cohort over time or by making simple allowances for

the time of entry into the labour market. "Snapshot" descriptions of the labour market will accurately reflect trends in the labour market. Finally, models of employment, earnings, et cetera may be justifiably simplified by the assumption of time-invariant coefficients.

The assumption of a constant distribution of characteristics does not stand up to a casual analysis of the UK labour market over the period in question. For example, the UK has experienced a shift from manufacturing into services and from full-time to part-time work; union membership has fallen, as has employment in the public sector.

Moreover, the probability of appearing in the panel may not be random. If the experience of unemployment lowers the probability of being employed (Phelps (1972)) then fluctuations in demand can cause persistent differences in the probability of being employed between cohorts. As another example, the full-time and part-time participation rates for women have been rising and the full-time employment of men has fallen. Such fluctuations in labour supply may affect both wages and unemployment rates, and could discourage or encourage potential employees in the long term; for example, Dolton and Mavromaras (1994) argued that the choice of a teaching or non-teaching career for two graduate cohorts was significantly influenced by prevailing labour market conditions.

If individuals joining the labour market have different characteristics from their predecessors, two problems arise. Firstly, snapshot descriptions of the labour market may no longer represent dynamic effects accurately; secondly, econometric analyses of the labour market which do not allow for this changing structure are likely to be inefficient and may give misleading or inaccurate results. The aim of this chapter is describe the characteristics of the labour market to enable the validity of simplifying assumptions to be assessed.

The data presented here is used to show the composition of the NESPD, and to draw some inferences on the UK labour market in general. There are two important caveats to this.

Firstly, the data on part-timers is suspect, given the NES's presumed ignorance of those earning below the National Insurance (NI) limit. Secondly, while appearance in the NES implies employment during the survey week, and therefore participation in the labour market, the opposite is not true. Participation in the labour market does not necessitate employment; employment does not necessitate appearance in the Survey. The results of this chapter are mainly compared against Robinson(1994), a recent survey of UK labour market changes, to provide an alternative base for comparison[1].

Because of the enormous amount of information on cohorts available from the micro-data, the results presented here are necessarily selective. Before discussing these results, a brief description of the cohorts is useful.

8.1.1 Employment cohorts

Employment cohorts were constructed for individuals born in every year from 1910 to 1975, so covering those who were 65 in 1975 and those who were 15 in 1990[2]. For each year from 1975 to 1990, numbers observed, both in total and in each of the dummy variable categories listed in Table A9.3, were stored for those working part-time or full-time in that year. In addition, gross weekly wages were stored for each observation. Thus, for example, it is possible to tell how many individuals born in 1928 were working full-time in London in 1976 and what their average weekly wage was. The data is split into males and females.

8.1.2 Data cohorts: the observation histories

---

[1] Comparative results are based on alternative datasets. Robinson (1994) primarily uses the population census and the Labour Force Survey (LFS); Elias and Main (1982) uses the National Training Survey (NTS); Main (1988a, 1988b) the Women and Employment Survey (WES); Elias (1988) the LFS and WES. Coverage and survey methods for these sources of information differ from the NES and so provide effective comparators.

[2] Those born outside these dates were treated as being born in 1975 or 1910.

The data cohorts are constructed from OHs for the period 1977-1990. OHs were calculated separately for males and females, and were further broken down into full-time workers, part-time workers, and those who had experience of both full- and part-time work[3]. Thus this last group, the "cross-over" workers, is likely to share the characteristics of both full-time and part-time workers. However, it is not possible to identify from these OHs the precise split between full-time and part-time working.

For males, the part-time and cross-over OHs are almost empty. Those males who have only full-time employment experience (in the NESPD) account for around 95% of the male sample, and so for the most part the other two male groups are dropped.

One useful side-effect of the OHs is the ability to identify and eliminate individuals with single observations. These tend to have different characteristics from those observed for longer periods, and given the potential for erroneous data entry (which is not corrected retrospectively by the DE) a working assumption is that a significant proportion of these are due to errors[4]. Single observations can account for anything from 5% to 40% of entrants in any one year, but only comprise around 5% of the total sample at any one time. Single observations are therefore generally excluded from the data cohorts.

For each OH, information is available on the age when joining the NESPD, the number of years spent in the private sector or covered by collective agreement or Wages Board, and wages for each year of observation. This information is disaggregated down to the level of observation patterns.

8.2 Numbers in the dataset

[3] The individuals in the third category could not be broken down further without drastically shortening the survey period, because of the number of potential part-time/full-time work patterns.

[4] This assumption is probably unrealistic for the later years of the survey as we may expect more people to return to the NES after 1990. However, the bulk of entry to the survey occurs in the first few years.

Figure 8.1 shows numbers observed in full-time and part-time employment using the employment cohorts, and figure 8.2 the proportion of total observed. Note that, although the NES is intended to be a 1% sample of those in employment, the numbers in figure 8.1 only show around 0.75% of employees as measured by the LFS, for example. This is partly because the NES does not include the self-employed, the armed services, occupational pensioners and so on. In theory, it should also exclude those earning below the NI limit[5]. Finally, the NES does have difficulties in tracing those who have changed jobs around the survey period (Adams and Owen (1989)). Thus it is not entirely surprising that the NES falls short of its target numbers; however, it still remains a significant survey of the labour market and far larger (in terms of participants) than alternative surveys[6].

Part-time employment is rising for both sexes; however, full-time employment is falling for males but rising for females. This is in line with other studies of the labour market (see Johnes and Taylor (1989); Robinson (1994); DE Gazette (DEG) and LFS). Over the survey period, male full-time employees have fallen from 65% of those employed to 55%, the growth in total employment coming from females. Note that, in contrast to LFS and Census data (Robinson (1994)), these figures do not show part-time male employment increasing its share of total employment; this may be due to the NES's difficulties tracking part-timers.

Table 8.1 depicts the numbers of new entrants to the dataset and those making their last appearance. The rise in the numbers leaving at the end of the period is due to the closure of the survey period. The number of males joining and leaving has always exceeded the numbers of full-time females joining and leaving; however, for females the number joining generally exceeds the number leaving whilst for males the reverse is the case. Thus the changing gender composition of the full-time labour force appears to be due to both a high rate of attrition

---

5 In fact, a notable number do get through the NI contributions barrier, possibly due to inertia in the Inland Revenue records system.

6 See Adams and Owen (1989) for details of the missing 25% of the target sample. Bell and Ritchie (1994) consider the implications for analysis of this missing data.

among males in full-time employment and a net inflow of females into all types of jobs. It is worthwhile noting that, since 1981, more females than males appear to be joining the labour force.

Table 8.1 Numbers joining and leaving the NESPD

		1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
Males	joining	20581	11078	9337	6632	4816	4770	4958	4158	5068	5110	6552	4675
full-time	leaving	7598	8317	9374	7718	7058	6661	6611	6296	7055	6920	9809	12571
Females	joining	6808	4879	4572	3653	2900	3132	3309	3166	3657	4076	5213	3839
full-time	leaving	3553	3872	4281	3407	3222	3105	3144	3005	3227	3198	5240	6138
Females	joining	3730	2842	2485	1792	1448	1453	1645	1386	1852	2073	3053	2091
part-time	leaving	2378	2463	2654	2126	2061	1828	1893	1608	1793	1794	3217	2957
Females	joining	4592	3006	2348	1947	1565	1604	1343	1337	1383	1606	768	407
cross-over	leaving	210	457	650	818	956	1073	1131	1278	1402	1616	2305	3954

8.3 Probabilities of observation

Figure 8.3 shows the probability for an individual observed in period t of not being observed in period t+1, calculated from the OHs. The initial leap in the probability of absence is due to an initialisation effect - moving the start date back to 1975 raises the chance of dropping out in 1977 and 1978 (Bell and Ritchie (1994)). As may be expected from empirical patterns of participation for men and women (Elias and Main (1982), Main (1988a), Elias (1988)), males have the lowest probability of becoming absent in the next year. Females employed part-time have highest chance, which is unsurprising given the NI earnings limit and the transient nature and high turnover rate for part-time jobs.

The downward trend in the disappearance rate may reflect a number of factors: improved administration of the NES, or the filtering out of those with poor employment records to give a dataset composed of "stayers". However, the sharp fall in the disappearance rate from 1979 to 1981 is unexpected, given the changes in the labour market. Over this period the

unemployment inflow rate increased by 50% while outflows stayed roughly constant (Layard and Nickell (1987) pp132-333) and unemployment rose sharply for both males and females (Robinson (1994)). If the NES represents a truly random sample of the population, a rise in the disappearance rate rather than a fall is the more likely outcome.

In the absence of any information from the DE about improved data collection, the implication is that those in the dataset cut down on changes in jobs. The two biggest causes of missing observations in the NES are employees moving "out of scope" (eg self-employment, armed forces, occupational pension) or the last recorded employer having no record of the employee's current situation (Adams and Owen (1989)). Missing observations due to these causes will all fall if employees change jobs less frequently.

Figure 8.4 depicts the proportion in each year who have held their current post for over twelve months[7]. It is clear that there is a strong cyclical element, in that job changes fall markedly in the early 1980s when unemployment is rising and employment falling, and then rise from the mid-1980s as the economy picks up. Note, though, that although total employment was rising from 1983, unemployment rates did not start to fall until 1986/7 (Johnes and Taylor (1989); Robinson (1994)). Thus the turnover variable appears to be tracking employment rather than unemployment trends[8].

These results can be interpreted in the context of a number of hypotheses. For example, a job-search theorist could argue that falling employment levels reduce the expected alternative wage by lowering the probability of finding acceptable job offers, which leads to a reduction

---

[7] Both those who join a new company and those who change jobs whilst remaining within the same company are deemed by the DE to have "changed jobs". Movements within companies are unlikely to affect the probability of observation.

[8] Turnover according to the LFS also appears to show this pattern. The DE Gazette (December 1989, table 1.6 pS11) records that, apart from a sharp fall in 1975, leaving rates over the period in question were fairly constant apart from a slight downward trend; however, engagement rates followed changes in employment.

in search activity and fewer changes in jobs (see Theodossiou (1992), for example). Alternatively, several models of the two-tier or segmented labour markets postulate an "inner" (insider, career, or primary) part of the labour force which has significant ability to protect its employment position against the "outer" (outsider, non-career, secondary) workers. The apparent increase in average tenure is consistent with this latter group (who, by definition, have poor employment records and prospects) bearing the brunt of employment shocks. Finally, a simple last-in-first-out seniority model will also lead to the result shown in figure 8.4.

Figure 8.5 extends this analysis by another year, giving the observation rates for the four possible combinations of observation patterns in the two years following an observation. Figure 8.5(a) shows that males are most likely to be observed for three consecutive periods, as a proportion of those observed in the first period. For all groups this proportion is rising. Figures 8.5(b) and (c) show that the female part-timers are most likely to have two consecutive observations and then to miss an observation, whereas they are least likely to miss an observation and then return the following year. The cross-over female group is most likely to return to the NES after an absence - consistent with a view that this group is at least partly due to women taking breaks from full-time work to raise families and returning to part-time work[9]. Finally, figure 8.5(d) shows that (ignoring single observations) part-timers have much the highest probability of not reappearing in the dataset in the two years following an observation.

Figure 8.5 complements figure 8.3; increasing observation frequency in the former reflects falling disappearance rates in the latter. The near-constant probability of not reappearing in the two years following an observation (figure 8.5(d)) suggests that the improvement in

[9] Although only a one-year gap is allowed for here, and the median time out taken for raising families varies from 3.7 to 9.7 years (Dex and Puttick (1988), p136) depending on the birth cohort. An alternative explanation would be that a large part of employment for this group is in temporary work and so is often missed by the Survey, which could be supported by Dex and Puttick's assertion that increasingly women are taking on short-term jobs between births.

observation in figure 8.3 is due to a fall in the numbers of intermittent missing observations. Again, this may indicate fewer changes in employment, as moves between jobs should manifest themselves as one missing observation at most[10].

It is straightforward to reproduce figure 8.5 for each data cohort but this enormous amount of information is unlikely to be enlightening. Instead figure 8.6 presents, for each group, the numbers observed in each year following the first appearance. Thus, for example, male full-timers have an 80% chance of being observed one year after their first appearance which falls to 50% in the fourteenth year[11]. These are similar to survivor functions except that individuals absent one year are allowed to return in a subsequent year. All four groups follow similar patterns, and the "hazard rate" for both males and females can be shown to be very similar and almost constant for most data cohorts (Bell and Ritchie 1994).

Interestingly, the highest rate of continuing participation is achieved by the female cross-over group. It is also relatively flat, suggesting that the distribution of gaps in the employment record of this group is fairly even. If this group does reflect women moving into and out of part-time employment as family circumstances change (Main and Elias (1986)), then a high reappearance rate is to be expected. This raises the question of why the "survivor function" of female part-timers does not flatten out as well. One possibility is that a significant proportion of those who only work part-time are nearing the end of their working lives and leaving the labour force at a constant rate; age profiles presented in the next section provide some support for this view.

One result not apparent from figure 8.6 or from employment cohorts is that, for all but the

---

10 Micklewright and Trinder (1981) comment on the problems caused for the NES by changing jobs near the end of the tax year. As data is not collected or corrected retrospectively, a missing observation is permanently recorded.

11 These figures are averaged over each cohort in each group, as the disappearance rates are similar for all the data cohorts within a group. Ritchie (1995) presents the disaggregated result.

longest periods, females full-timers are more likely than men to have few or many periods in the dataset. Moreover, it appears that, from the end of the 1970s, females joining the dataset are more likely to have a complete, or almost complete, set of observations in the dataset[12]. Equating appearance in the dataset with employment, this result is unexpected. Elias and Main (1982) and Main (1988a) characterise the employment patterns of men and women, using the NTS and WES. These studies indicate that men undertake paid work for more of their life and change jobs less frequently than women (relative to years of participation), which should be reflected in men appearing in the dataset for long periods if the NES accurately represented work patterns.

In this case the NES and alternative studies of the labour market appear to diverge. However, the NES results are conditioned on the fact that these women are only working full-time; those moving between full-time and part-time employment (the cross-over group) have fewer observations than full-time males (Ritchie (1995) table 10.2). Those women with some or all part-time working comprise two-thirds of the female observations, and so the overall effect is that the NES shows women with fewer observations. However, the suggestion that women only working full-time may have better employment records than their male counterparts is an interesting qualification to this result.

Robinson (1994) notes that female employment was hit harder than male employment in the early 1980s recession, but mainly in the manufacturing industries; in services, which weathered the recession relatively well, the impact on female employment was small. The net effect was that the loss in female employment was in low paid, possibly part-time employment. Similarly, several authors (for example Layard and Nickell (1987); Robinson (1994); Sloane and Theodossiou (1994)) have noted the increase in demand for female-

12 Ritchie (1995). Bell and Ritchie (1993b). using a similar dataset grouped over part-time and full-time workers and including missing observations, argued that women were more likely to have long continuous periods of observation in the dataset and to return to the dataset after a long period of absence.

intensive professional services. The evidence from the NES supports Robinson's claim that "high-ranking" jobs, full-time by assumption, are increasing and improving the prospects for female employment.

8.4 Age profiles

The employees in the NES are getting younger, on average. Figure 8.7 uses the employment cohorts to build up a picture of the average age of employees in the dataset, and shows that the average age for both full-time and part-time workers for both sexes is declining over time. This figure masks two opposite influences. Robinson (1994, population census) and Johnes and Taylor (1989, DEG) both note that the participation rate for older males is decreasing but for older females is rising. However, the participation rate for young females is rising even faster (Johnes and Taylor(1989)) and this group is relatively numerous, thus lowering the average age of female employees.

Figure 8.8 displays the age profiles of the dataset in each year for all females and for male full-timers. Figure 8.8(a) appears to show a significant "cohort" effect on males aged 25-30 in 1975. There are significantly higher numbers employed in these cohorts than in either younger or older groups of males. These are the workers who joined the labour force before the early 1980s recession, which would seem to support the Phelps (1972) argument that there is an element of hysteresis in employment.

In contrast, figure 8.8(b), showing the age profiles for females working full-time, indicates a strong "life-cycle" effect: that employment is strongly influenced by family circumstances, including breaks from full-time employment to raise children. Peak employment in full-time jobs occurs between sixteen and twenty, whereupon the numbers employed fall until around thirty; finally, there is some return to full-time employment over the remaining working life, but not to the levels observed for teenagers. There is a wide body of theoretical and empirical

evidence supporting this outcome (see, for example, Dex and Puttick (1988); Elias and Main (1982); Joshi (1986); Main (1988a); Main and Elias (1986)). Note that there is no real cohort effect here; unlike the males, no group of cohorts has consistently higher rates of full-time employment than any other. The counterpoint of figure 8.8(b) is the numbers employed in part-time work, given in 8.8(c). This indicates that full-time and part-time work appear to be substitutes for women, as the numbers in part-time employment are high when those in full-time employment are low, and vice-versa. Again, the theoretical and empirical work on this area would predict this result, but the congruency between figures 8.8 (b) and (c) is pleasing.

The implication of these results is that women's participation in paid work is strongly influenced by the "life-cycle" effects mentioned above. In their early years most work is full-time; with the advent of families full-time work drops sharply to be replaced partially by part-time; however, from the early thirties up to retirement age there is a steady return to full-time work with a small but corresponding fall on part-time work. Thus women follow a much wider range of employment options suited to their circumstances, substituting between full-time and part-time employment as appropriate.

In contrast, males are more prone to "cohort effects", where being employed at the right time has a significant effect on future employment prospects. Finally, there is no significant substitution between part-time and full-time work for males[13].

These results are very similar to the "M-shaped" participation patterns for females described by Main (1988a) and Elias(1988) using the WES: women typically work full-time, leave employment in their twenties, and then return to full-time employment by way of part-time employment, with some increase in part-time employment around retirement. What is striking

13 The profile for part-time male employees is virtually zero except for the very young and very old.

is the complete absence of any change in the structure over time. While Main notes that apparent participation rates in employment have been shifting over time, the proportions of a cohort working full-time has barely changed over three decades (Main, pp43-46), reflected in figure 8.8b. Elias and Main both note some small increase in part-time employment, which is less clear in the NES data; however, figure 8.8c provides strong support for their contention that part-time employment is still dominated by family factors[14].

Figures 8.7 and 8.8 have detailed the "snapshot" distribution of ages for those currently working part-time or full-time. Using the data cohorts enables some analysis of individuals based on their whole job history. Figure 8.9 shows the average age at which each cohort joined the dataset, for each data group[15]. The initial fall in the starting ages is probably due to individuals who should have been included in the first cohorts having missing observations and so being included in later cohorts. By the early 1980s the average age of each cohort joining the dataset has settled down at around twenty-five to thirty, with full-timers joining at an earlier age. For female part-timers the starting age is much higher (in accordance with figure 8.8d) and falls steadily[16].

The stability of the starting ages is a promising finding, implying that inclusion in the dataset is primarily determined by the easily measurable characteristic of age (and simplifying the problem of selection bias arising from initial inclusion in the dataset). However, the *level* of joining ages is worrying. These ages are averages over all those joining the dataset in a particular year. Because the selection criterion is that everyone whose NI number ends with a certain combination should be included, the bulk of those joining the dataset should be those joining the labour market for the first time and so being allocated new numbers; all those

---

14 Elias does show that part-time employment in "low grade" occupations has been virtually constant.

15 The reason for the concentration on starting ages here is to remove the effect of ages rising over time.

16 It is not clear whether this reflects the age profile of part-timers in the labour market, or whether this is due to the construction of the dataset (such as the initial-observation effect or the possibility that younger-part-timers are now earning above the NI limit.)

already working should be included in the dataset. On the assumption that most workers begin employment between the ages of 16 and 23 (school- and University-leaving), the evidence of figure 8.9 is that these are outweighed by large numbers of much older individuals also joining the dataset for the first time. While young workers both change jobs more frequently (and thus are less likely to be traced by the NES) and are paid less (and so may fall below the NI limit), these seem poor explanations.

Ritchie (1995) breaks down figure 8.9 into separate starting ages for each year for those remaining in the data cohort, to show how the age profiles of the cohorts change. As would be expected, the average starting age of each data cohort falls as older workers leave the labour force; this is most significant in the data cohorts beginning in the late 1970s. This effect is most significant for part-time employees. The high rate of attrition may be due to the poor observation records for this group, but is more likely to result from the relatively high age of part-time female employees as the attrition rate does fall with the average age of the cohort.

Interestingly, the cross-over group shows almost no decline in average starting ages over time, indicating that observations of the members of this group are largely unrelated to age. Missing observations appear to be evenly spread over all ages. Again, this is consistent with the view that the cross-over group is composed of women who have taken or will take career breaks. Although there may be some loss of older workers due to retirement, there is a counterbalancing loss of younger women leaving the dataset for family reasons[17].

Figure 8.10 uses the data cohorts to construct an alternative age profile to that of figure 8.7. The contradistinction is that the former classifies females into part-time or full-time employees

17 This argument is consistent with the results in figures 8.8(c) and (d). However, until the cross-over group can be broken down into precise periods of part-time and full-time employment then the hypothesis of young women leaving full-time work and taking up part-time work after a break remains untested.

depending on their whole observation histories rather than a snapshot[18]. There are a number of effects to consider

- the age of those observed rises over time
- older cohorts retire, and younger ones join; the numbers may not be the same
- the average starting age of the data cohorts is constant or falling (figure 8.9)
- within cohorts, older members are more likely to leave (Ritchie (1995))

Figure 8.10 illustrates that the decline in average age for both part-time and full-time females of figure 8.7 has two different components. For those who only work part-time or full-time the average age is falling noticeably, whilst for those who hold both types of job the average age is rising. It has been noted that the age profile of this group is relatively flat, implying that disappearance rates are fairly constant for all ages. Thus these data cohorts age by one year every year, and the effect of workers leaving or joining the dataset is muted as new cohorts replace a range of ages and not just older workers. However, as the average age of those starting is less than the age of those currently in the dataset, this damps down the rise in average ages.

The fall in the average age of full-time and part-time female employees is to be expected given the net inflows into the dataset (table 8.1) and the fact that these new employees are young on average. The age profile for males only working full-time shows little variation over the period, indicating that the various influences are finely balanced: although those currently in the dataset are growing older, this is balanced by the oldest workers leaving the dataset and being replaced by young workers.

8.5 Wages and the cohort effect

[18] Figure 8.10 only has figures from 1977 onwards and excludes single observations.

We now consider levels and growth rates of wages in the NES, adjusted for a real-wage deflator[19]. Figure 8.11 gives the average (adjusted) wage observed for each group in each year, calculated from the employment cohorts. It is apparent that overall the relative wages of the four groups change very little over the period apart from a slight growth in the female full-time mean wage. Within cohorts, wage *growth* is almost always positive even using a wage deflator (Ritchie (1995)); this is to be expected, as those who remain in the dataset are likely to be the most successful in their experience of employment.

One interesting feature is that wages for part-time males appeared to rise in the mid-1980s. At a time of rising unemployment among males, the increase in supply of males might be expected to drive down part-time wages (Layard and Nickell (1987)). The rise in real wages and employment (measured by the LFS) suggest that demand for part-time male employees increased by more than supply during the mid 1980s. However, the number of men working part-time is small and so this result may be an aberration.

In figure 8.12 the age earnings profiles for the NES are presented. These are averaged over all years at the same age, to give, for example, the average wage for a sixteen-year-old in all years. This is a valid simplification because, apart from that of male part-timers, the profiles are astoundingly stable over time; the only significant variation is in the sixty-plus category. This in itself is remarkable, given the strong cohort effect on the numbers of men employed mentioned above; for women, this result is less surprising given the strong evidence for a stable "life-cycle" pattern of participation.

The numbers themselves indicate that males have a predictably smooth concave age-earnings profile, with relative wages peaking in the late thirties. For females in full-time employment the profile is also concave, but peaks in the late twenties. Moreover, for males relative

[19] DE Gazette, Table 5.3 January 1979/82/89/91. Figures are for April.

wages stay at a plateau for some years before tailing off; for females relative wages start falling almost immediately the peak has been passed. Note that the graph shows only those currently working full-time; this profile could therefore be lowered by some of the females who will be leaving full-time work. However, this suggests that the profile should then continue rising throughout the thirties as the "non-career" women leave the labour force or turn to part-time work, and this is clearly not evident here. Some analysis of the OHs suggests that those women who only engage in full-time work have profiles somewhere between those of female part-timers and male full-timers, as in figure 8.12. This implies that the profile for full-time work in 8.12 is the same for all women, irrespective of their career strategies.

At sixteen both sexes appear to be earning the same wage. From a human capital point of view, the implication is that women acquire human capital at a lower rate throughout their working life; that they invest in less human capital in total; and that relative human capital deteriorates faster than for men. The fact that both sexes appear to start from the same point but diverge immediately supports arguments that attitudes and attributes prior to and on entering the labour market are an important determinant of both initial employment and subsequent work history (Polacheck(1981); Dolton and Kidd(1994); Vella(1994)). It does not support Becker and Lindsay's (1994) claim that female age-earnings profiles should be steeper as women should bear more of the cost of firm-specific investment because of their higher quit rate (although if men and women invest in different amounts of human capital there is an identification problem).

For part-timers, it is interesting to note that both sexes again start work at the same pay. The concave profile for male workers is a little surprising, implying that age and/or experience is an important determinant of part-time earnings; this contrasts with a theoretical background that part-timers have little or no incentive to invest in human capital (for example, Theodossiou (1992)). The profile for part-time female employees is more in line with

theoretical and empirical results (for example, Main (1988b)). There is some wage growth in the teens, possibly indicating some basic training. However, the complete absence of growth (positive or negative) between the ages of twenty and sixty suggests that the benefit from further experience or training runs out very quickly.

As Murphy and Welch (1990) show, the typical age-earnings profile is poorly approximated by the popular quadratic form. In the case of the NES this is clearly the case, and for females any continuously differentiable specification is likely to perform badly. In subsequent chapters a flexible dummy-variable specification has been used.

8.6 Sector and agreement

This discussion of the cohort characteristics of the NESPD ends by looking briefly at two other areas of interest. Figure 8.13 shows the proportion of observations in the private sector for 1975 and 1990, from the employment cohorts. Three different trends are visible for the three different groups. For full-time male employees, employment in the private sector in 1990 is substantially higher than in 1975 for all ages. For females, all but the youngest part-timers spent more time in the public sector in 1990 than in 1975, whilst for full-timers there was little change for older age groups but the youngest became more likely to work in the private sector. The overall effect is that the proportion of employment in the private sector has increased between 1975 and 1990, as younger female employees are an increasingly large part of the NES; the lower public sector participation rates for this group reduces the overall proportion of female full-timers in the public sector (see figure 10.9).

Figure 8.14 shows the number of individuals whose wage bargains are affected by national collective agreements (but who are not necessarily union members). For males, the levels of coverage have fallen fairly uniformly across all ages, suggesting that the decline in union membership throughout the 1980s fell across existing employees rather than new workers

joining non-union recognised businesses[20]. For female full-timers, the largest decline in union membership appears to be in the young. For part-timers, there appears to be a small fall in the level of agreement coverage, with the young once again the least likely to be in positions covered by collective agreements. Again, the increasing prevalence of young female full-timers means that overall levels of coverage have fallen (see figure 10.7).

8.7 Quasi-complete cohorts

One issue touched upon earlier is how closely the NESPD relates to the UK labour market. As other evidence suggests those who are unemployed or who withdraw from the labour force have different characteristics from those continually employed (eg low wage growth; relatively young or old; working in declining industries, and so on), then a simple test of whether absence from the dataset can be reasonably associated with non-employment may be to compare those who appear in the dataset all the time with those who have occasional absences. The more similar the characteristics of the two, the more likely it is that absence from the dataset is a statistical error rather than a reflection of non-employment.

To this end, similar statistics to those above were calculated for "quasi-complete cohorts" (QCCs): that is, those who had no missing observations between the first observation and leaving the dataset for good. Unfortunately, the results do not clarify the issue. The wage profile of the QCCs is almost identical to that of the full dataset. On the other hand, the QCCs have a higher overall within-cohort wage, and lower total "snapshot" wage. Moreover, the QCCs tend to be much older and have a disappearance rate of around 30%-40% - twice as high as for the full dataset.

The overall suggestion is that those with missing observations do perform differently to QCCs;

---

20 This is a qualified result, as only national collective agreements (from a given list) are recorded every year. Information on local and company agreements has been collected twice.

however, this is a very tentative result. The reason is probably that there is not a simple employment-non-employment question: absence from the dataset could have a number of causes, which may all act in different ways and have different effects on those who do not go missing. To identify some of these effects requires rather more information than is available from the NESPD.

8.8 Summary

A few aspects of the NESPD have been discussed, and some difference between cohort and dataset characteristics has been noted. Although using these results to make inferences about the labour market as a whole is not necessarily justified, some sensible inferences may be drawn, particularly with respect to the participation of women. Some worries about the composition of the NES have emerged, most notably the high starting ages and the difficulty of determining how closely appearance in the dataset approximates to labour market experience. Many of these results are tempered by the difficulty of making assertions about the cross-over group, and must wait for a new set of OHs allowing for multiple destinations (full-time, part-time, unemployed, and so on).

With respect to transitions, there appear to be significant differences between both gender groups and cohorts. Bell and Ritchie (1993b) constructed hazard and survivor functions from similar observation histories which indicated very little difference between the sexes. This chapter suggest that these results need to be reconsidered in the light of the heterogeneity of the female groupings.

The age profile of the database was shown to vary significantly between the groups: the average age is constant for males, falls for female part-timers and full-timers, and rises for the cross-over group. Age-earnings profiles suggested the concave structure of human-capital theory. However, for women in particular there are clear indications that the returns to

human capital come to an abrupt halt and for part-timers, are negligible.

Some of the other information available was briefly discussed and it was noted that the widely reported fall in union membership over the 1980s appears to be spread over all male full-time workers but is chiefly found amongst the younger female employees.

Finally, a brief discussion of quasi-complete cohorts suggested that conditioning on a complete set of observations prior to finally leaving the dataset may lead to different results compared to the case where the whole dataset (including those who have intermediate missing observations) is used. This implies that selection bias is a significant issue for the NESPD.

Although cohort statistics over dataset figures have been emphasised, this does not imply that the former are necessarily any "better" than the latter. However, the view that some analysis of cohorts is necessary has been reinforced by results indicating that dataset statistics may not adequately reflect the characteristics of individuals. These results are supported in subsequent chapters which argue that the dynamic structure of the NESPD should not be ignored lightly.

Chapter 9

Male earnings 1977-1990: fixed-effects and varying coefficients

This chapter documents the results of estimating straightforward fixed-effect (FE) and cross-section wage equations for males using the NESPD from 1977 to 1990. These results are notable in a number of ways. First, they provide a complete set of cross-sectional estimates for the period 1977-1990 allowing the coefficients to vary over time. Thus it is possible to study the unrestricted evolution of the coefficients over fourteen years. More importantly, the fixed-effects estimator is used to generate panel estimates. Although some cross-sectional estimates have been made, these are the first true panel estimates from the disaggregated NESPD. The evidence to be presented here suggests that allowing for individual heterogeneity has a significant impact on the results.

Of some interest in their own right, these results are also used to compare the effect of differing estimators on the results obtained. This comparison takes two forms. Firstly, the results of the fixed-effects specification are contrasted with cross-section (CS) estimates (the models of sections 5.2 and 5.1). Secondly, the ability to let coefficients vary over time is used to consider the evidence for parametric stability in the UK labour market over the period.

9.1 Econometric issues

9.1.1 Functional form

The framework for estimation is the Mincer-type reduced form specifying log earnings as a function of "human capital" and other control variates; more specifically, for the FE estimator the unrestricted equation (5.45) is the basic specification:

$$w_{it} = x_{it}\beta_t + \lambda_t + \alpha_i + u_{it} \qquad (9.1)$$

This model has time-varying coefficients and fixed individual effects, and is referred to as the TVFE model. The cross-section model ignoring the individual-specific term α_i still has time varying coefficients, and henceforward is referred to as the TVCS model[1]. The models are estimated using the covariance methods of section 5.1 (TVCS) and section 5.2 (TVFE).

The nature of the reduced form specification means that the presence of the controls can be given a wide variety of interpretations, reflecting various theories of wage determination. For example, regional dummies might reflect compensating differentials or differences in the pressure of demand in local labour markets due to geographic immobility of the labour force. Because the reduced form is consistent with many structural models of the labour market it may be unable to discriminate between them. Therefore the results reported here are not interpreted with respect to any one particular hypothesis.

Given the size of the NES, the asymptotic properties of the estimated coefficients are an important consideration. As the TVFE estimator is fundamentally OLS, implicit assumptions of the results in this section are that

$$E(x_i'u_i) = 0 \qquad \underset{n\to\infty}{plim}\ x_i'u_i = 0 \qquad (9.2)$$

where $\mathbf{x}_i \equiv [\mathbf{x}_{i1}' .. \mathbf{x}_{iT}']'$ and $\mathbf{u}_i \equiv [u_{i1} .. u_{iT}]'$. The validity of these assumptions is debatable. It could be argued that all the variables in the NES are potentially endogenous[2]. However, the effect of this endogeneity is unknown. Therefore, in this particular study, the null hypothesis is that sufficient exogenous variables have been included to avoid omitted variable bias so that

---

1 The TVFE and TVCS specifications are the "unrestricted" equations (5.45) and (5.1), respectively. The terms TVFE and TVCS are used in future to avoid confusion over the use of "unrestricted" in describing the models.

2 For example, self-selection (ie labour supply) by the NES sample could involve wages and hours offered, overtime rates, occupation, region, industry, union status and the predilection for full-time work. Self-selection by employers in the NES (labour demand) could involve wages and hours desired, occupation, industry, union status, public/private sector, and so on. These two lists contain all the NES variable fields.

(with the allowance for individual heterogeneity) the assumptions in (9.2) hold. Although further work may refute this hypothesis, it seems a reasonable starting point[3].

The TVCS model estimates (9.1) as T separate cross-sectional analyses:

$$w_{it} = x_{it}\beta_t + \lambda_t + \eta_{it} \qquad \eta_{it} \equiv \alpha_i + u_{it} \tag{9.3}$$

Clearly, if α_i is non-zero, η_{it} will appear to be serially correlated and OLS estimates of (9.3) will be inefficient. More importantly, CS estimates of (9.3) involve three extra assumptions:

$$E(\alpha_i) = 0 \qquad E(x_{it}\alpha_i) = 0 \qquad \underset{n\to\infty}{plim}\ x_{it}\alpha_i = 0 \tag{9.4}$$

The first assumption is not important as the individual intercept can be split into a mean common for individuals and deviations from that mean. Thus a non-zero mean for the individual-specific effects is simply subsumed into the time-intercept for those appearing in a particular period. The second and third assumptions are more important, implying additional restrictions on the regressors. Any potential correlation between the job characteristics and the invariant characteristics of the individual must be discounted (see section 2.2).

In the context of the NES, independence of the characteristics of the individual and the job is a large assumption. A number of authors (for example, Chamberlain (1985); Hartog and Oosterbeek (1993); Jakubson (1991); Killingsworth (1986); Rees and Shah (1992)) have noted that individual heterogeneity may influence occupation, sector, location, and so on. A related argument which has seen more attention in the literature on female earnings is that significant determinants of labour market experience are pre-entry decisions and the initial job taken[4]. If the choice of first job is non-random and significant in determining future

---

3 In the following chapter the issue of endogeneity is considered in more detail.

4 Elliott(1991) pp404-407 discusses some aspects. Empirical analyses on pre-entry influences include Dolton and Mavromaras(1994) on expectations of career prospects; and Vella(1994) on sociological attitudes.

employment, this may lead to an additional selection bias in CS models[5]. Most importantly, a premarket factor influencing future employment is likely to be education, which the NES omits. Given that formal (certified) education is usually completed before employment, and is thus a time-invariant individual characteristic as far as employment history is concerned, assumption (9.4) should be treated with caution.

9.1.2 Hourly versus weekly wages

Both hourly and weekly wages have been used as dependent variables in labour market studies. It may be argued that weekly pay is a better indicator of the slope of the budget constraint faced by workers given that standard hours of work are usually fixed by the employer rather than through direct negotiation. If the marginal value of leisure hours is relatively constant over the week and there is little flexibility in the relationship between working time and wages (for example, there is little or no overtime premium), then demand and supply functions based on the weekly or annual earnings may be appropriate. This measure is most likely to be the case where working time is not a significant determinant of the labour supply decision (for example, for salaried employees or non-manual employees for whom overtime is not generally available).

However, the dependent variable used in this study is hourly wage. The main argument is that this is a better indicator of marginal benefits and costs, as that the wage rate and the number of hours worked form two separate (if not independent) choice criteria[6]. This allows for the joint nature of the income/effort decision by employees and the employee/working time decision by employers. Where the marginal value of leisure varies (for example, evening

---

5 The reason this can lead to a bias in CS and not FE models is that the initial job choice may manifest itself as a "one-off" influence which is constant throughout an individual's working life; in other words, a fixed effect. See Ridder (1990) or Verbeek and Nijman (1992a).

6 A number of authors have argued that the hours/wage decision is made simultaneously (see Killingsworth (1983), MaCurdy(1985) or Stern (1986) for surveys) which will not be considered here. The key point is that the wage rate determines participation levels rather than total income.

working requires more compensation than Saturday jobs) an hourly rate more accurately measures the leisure/income tradeoff at the margin; and where a range of working practices and incentives is available to the employer the hourly rate arguably represents the marginal cost of labour more truly.

Put more formally, in an individual-level study it is reasonable to assume that the individual maximises a utility function which contains both hours worked and leisure:

$$U = U(f(h_{max}-h), wh) \qquad U_w>0, \quad f_h \leq 0 \tag{9.5}$$

where h_{max} is maximum hours available for work, h is hours worked, w is the hourly wage rate and the function f(...) captures the utility of leisure time. In this context, using weekly wage (wh) is only appropriate if $f_h=0$ or if h is not a choice variable for individuals. In the latter case, which may be appropriate for non-manual workers, hourly and weekly wages only differ by the scaling factor h and weekly wages will capture the marginal value of work accurately. If, however, workers have some control over h, then only the hourly wage is appropriate.

However, the choice of *the* wage rate is problematic when all hours are not paid at the same rate. For example, if overtime and weekend premia are all available to the employee, then the marginal wage rate may differ dramatically from the average wage rate which is typically reported. When the wage function is significantly non-linear, then a simple mean wage averaged over all hours will not reflect the employee's labour supply function (Brown, Levin, Rosa, Ruffell and Ulph (1986)). Similarly, if untimed payments are made (such as production bonuses) then the allocation of these bonuses to wages may be arbitrary and unjustified. Thus the labour supply decision is likely to involve a range of possible "wages" at both weekly and hourly rates. Moreover, the difficulties of measuring hours of work for non-manual employees means that often the hours of non-manual workers are concentrated

around a standard hourly week and do not reflect the actual hours worked[7].

The choice is further complicated in the NES as the reported hourly wages are the actual hourly wage experienced during the survey week, defined as wages for that week divided by the number of hours actually worked. In other words, only weekly wages and hours worked are known. Information on overtime hours and bonus payments is available, but not on whether such payments are typical or atypical; thus, it is difficult to say whether the wage received represents the "normal" distribution of wage offers and therefore gives a "fair" view of the remuneration options open to the individual.

Despite these difficulties, the wage rate used here is the natural logarithm of actual observed hourly wages excluding overtime payments, adjusted for RPI. Given the nature of this study, to provide results from a new estimator on a familiar dataset, it takes the approach of the bulk of the literature (see Killingsworth (1983, especially tables 3.1, 4.1, and 5.1) for a comprehensive survey of earlier results).

However, it may be noted that using the weekly wage as the dependent variable makes little qualitative difference to most estimates. Bell and Hart (1995) study the question of hourly versus weekly wages and basic versus total compensation in some detail, and report that the choice of measure makes relatively little practical difference. This was in respect of one variable only, the "union markup", and so may not hold for other variables, but a cursory comparison of the TVCS study of Andrews, Bell and Ritchie (1993) and the TVCS results presented here suggests that the only notable difference between using weekly and hourly earnings is to be found in the sectoral coefficients (see section 9.2.8).

---

[7] Atkinson, Micklewright and Stern (1982) compare employee perceptions of hours worked (from the FES) with employer perceptions from the NES. While for manual workers the two are similar, for non-manual workers employees believe they work much longer hours than their employers think they do.

9.1.3 Attrition and missing data

A serious, but largely unrecognised problem with the NES is the large amount of missing data. Missing data has two effects: it reduces the precision of estimates; and, if correlated with the variables of interest, it can invalidate the estimation results. This problem is, of course, not unique to the NES or panels, but the nature of the dataset makes it difficult to counter.

A descriptive analysis of the missing data problem in the NES has been attempted in Bell and Ritchie (1994). The general conclusion of this work is that the likelihood of selection bias is large; the probability of individuals and observations being included in the dataset appears to be correlated with almost all the variables in the dataset. This is not surprising given the range of characteristics covered by the NES and the potential for complex variable relationships, but it is a source of concern. However, the *magnitude* of the effect of missing data is much harder to identify, especially as the size of any effect is specific to the particular equation being estimated.

As noted in chapters two and four, the construction of models of attrition for panels is complex in practice and requires strong assumptions. The standard econometric approach to dealing with non-random attrition in a two-period model was developed by Hausman and Wise (1979). Absence from, or presence in, the panel is modelled in a separate probit equation and a Mills ratio for each individual derived from this equation is included in the earnings relationship. Unfortunately, the Hausman and Wise solution is derived from the simplest possible panel model; in a multi-period model this procedure is computationally prohibitive, even if the size of the NES did not preclude non-linear estimation. It is also of doubtful value if a dynamic specification of attrition is desired.

Following a suggestion of Verbeek and Nijman (1992a), the effect of absence (from the panel

and the workforce) is linearly approximated by including as additional regressors variables which are related to the probability of attrition. This is a simple approach in the present context, since, for example, it is straightforward to calculate whether an individual was present in the previous year, how many previous years they had been present and so on.

This is not as ad hoc as it seems. Although some authors claim that invalid assumptions invalidate the selection mechanism, a number of recent studies have questioned the relative importance of the various methods of accounting for selection bias. In particular, Robinson (1989) and Vella and Verbeek (1993) have shown that there is a close relationship between the switching-regression adjustments for selection bias commonly used and linear instrumental variables approaches. More importantly, Lanot and Walker (1993a) consider several different estimators which attempt to take account of selection bias in a union membership equation, and find that their results are not very sensitive to the particular correction method used; the important thing is to include *some* sensible form of correction. This is not an unexpected result, given that the switching regression approach merely requires *consistent* estimates of the selectivity term, not efficient ones. Whilst the Lanot and Walker findings are specific to their model, this does support to some extent the findings of others that a better specification of the particular selection mechanism may not materially improve estimates[8].

The use of proxy variables is similar in principle to Heckman's correction mechanism, in that the selection bias is seen as a specification error due to omitted variable bias. The advantage of these two approaches (as opposed to the different emphasis in the IV solution) is that the coefficients on the omitted variables can give an indication of the importance of these effects. Thus testing for selection bias is a simple matter of hypothesis testing on estimated coefficients. As already discussed, the control function approach is not really appropriate for panels in general and the NES in particular. However, unlike cross-sectional data, a

---

8 For example, Maddala (1983, p.267) notes that two-stage estimation of truncated regressions are likely to be effective even though the truncation regression appears to be very poorly estimated.

longitudinal dataset makes available a ready source of proxy variables in the record of previous observations. This also avoids the problem of having to observe non-participation to estimate participation probabilities.

To summarise, some emerging evidence seems to suggest that the choice of selection correction is relatively unimportant. We have therefore used variables created from the past histories of the individuals as proxies for attrition; these are then added into the equation in a manner analogous to the control function approach. Hypothesis test on the coefficients can then be used to ascertain the importance or otherwise of attrition. Note, however, that the impact on the error structure is undefined. Whereas both the IV and control function methods lead to well-defined results on the necessary correction for standard errors, the use of proxy variables does not lead to any particular form for the error. Thus the estimates still need to be treated with some care.

9.2 Fixed-effects and cross-section results

In addition to the basic TVFE and TVCS models, "pooled" and "restricted" models (to use the terminology of chapter five) were also estimated for both the FE and CS models. This gives six basic specifications.

$$
\begin{array}{lll}
(a) & w_{it} = \lambda_t + x_{it}\beta_t + \alpha_i + u_{it} & \textit{TVFE} \\
(b) & w_{it} = \lambda + x_{it}\beta + \alpha_i + u_{it} & \textit{FE pooled} \\
(c) & w_{it} = \lambda_t + x_{it}\beta + \alpha_i + u_{it} & \textit{FE restricted} \\
(d) & w_{it} = \lambda_t + x_{it}\beta_t + u_{it} & \textit{TVCS} \\
(e) & w_{it} = \lambda + x_{it}\beta + u_{it} & \textit{CS pooled} \\
(f) & w_{it} = \lambda_t + x_{it}\beta + u_{it} & \textit{CS restricted}
\end{array}
\qquad (9.6)
$$

The pooled and restricted models are those discussed in chapter 5: (5.61) and (5.74) for the FE specification, and (5.20) and (5.27) for the CS. This section presents results of estimation on (9.6a) and (9.6d); the next section uses the restrictions to consider the question of parametric stability.

The explanatory variables used were: the attrition variables (AVs); occupation (CODOT grouping); region; industry; age; private sector; coverage by collective agreement; coverage by Wages Council; and job held for less than one year. The last four of these are single dummy variables; the others, save the AVs, are all categorical variables. Codes are given in an appendix to the chapter. The panel is unbalanced: all males "whose earnings are not affected by absence" in the NES phrase (for example, due to sick leave) are included.

Note that one important feature of the TVFE specification is that dummy variables which do not change over time will still contribute useful information to the model. This is in contrast to a commonly-reported defect of panel transformations, that the remove variables which are constant over time and so place excessive emphasis on "mover" rather than "stayers". This is easily seen by comparing the effect on (5.129) and (5.130): when the explanatory variables are constant over two periods then they drop out of the regression completely in the time-invariant case.

Three AVs were employed:

InLast	1 iff in the panel last period; 0 otherwise
YrsIn	Number of years in the panel to date
CurrStay	Length of current continuous run in the panel

Regressions were run on the period 1977-1990 (1975-1990 to calculate the AVs). Because of the large amount of output (797 coefficients are estimated; fifty-seven variables over fourteen years less one time dummy), only a sample of output is given in table A9.2 in the appendix. The sample year (1984) is half-way through the period under review. Full results are available on request.

9.2.1 TVFE versus TVCS specifications

A first consideration is whether the TVFE and TVCS models are significantly different, for the TVFE model is computationally more involved than the TVCS model. As the TVCS is nested within the TVFE model, F-statistics can be constructed readily to test the null hypothesis that the TVCS restrictions are justified. Summing the fourteen TVCS RSSs and comparing with the TVFE RSS gives a test F-statistic of 12.22 with {194782, 907324} degrees of freedom, rejecting the null hypothesis (see table A9.1 for summary statistics). This is a reasonable result for, as noted in section 9.1.1, individual characteristics are likely to be related to both earnings and job characteristics.

Of more interest is how close the TVCS results are to the TVFE results, as the former are much easier to estimate than the latter. The estimation results to be presented here indicate that the *qualitative* features of the two models are similar, but the *scale* differs - dramatically in some cases. As the TVCS and TVFE results are generally consistent with theory and other research, the qualitative similarity is not surprising. However, the scale differences suggest that cross-section results for the NES are likely to be significantly biased.

TVCS results are biased away from zero in most cases: for all but two variables (agreement coverage and the attrition variable YrsIn), the FE model reduces the absolute coefficient values. In other words, after allowing for individual-specific heterogeneity, the variation in returns to a characteristic are much smaller. This supports the view of section 9.1.1 that the explanatory power of some variables in cross-sections is due to the correlation between earnings, unmeasured individual heterogeneity and job characteristics.

One other general comment on the TVFE/TVCS results is that the TVFE estimates tend to be much smoother over time. This is to be expected given the greater efficiency of the TVFE estimator and the susceptibility of the TVCS model to outliers in particular years.

The specific results are now considered in more detail.

9.2.2 Region (reg)

The reference category is Greater London. Figure 9.1 shows that the return to working in London as opposed to other regions increased steadily over the period. The largest differential was with northern England and Scotland, whilst those living in the south-east and East Anglia saw the smallest drop in their relative earnings. The CS model over-estimates the coefficients (compared to the FE model) by around 50% in the early years when inter-regional differences are relatively small.

Layard and Nickell (1987) argue that differences in inter-regional wage pressure lessened somewhat throughout this period in terms of regional unemployment-to-vacancies ratios, which should have led to a fall in the regional variation in wages, ceteris paribus. The apparent contradiction of figure 9.1 is because the Layard and Nickell do not take account of the regional characteristics, whereas the regression results are conditioned on industry, occupation, and sector and imply a "pure" regional effect which supports anecdotal evidence that the South prospered relative to the North over the 1980s. If the country is crudely characterised as a manual, manufacturing, low demand North and a non-manual, predominantly services-based high demand South, then these studies are consistent (especially as Layard and Nickell note a small rise in "industrial mismatch")[9].

9.2.3 Industry (div)

Figure 9.2 depicts the improvement in earnings in all industries relative to farming and fishing (FF) over the fourteen years. Although earnings in this industry start off around the average level for all industries, by 1990 FF has the lowest remuneration. One difference between the

---

9 The Layard and Nickell measure of wage pressure also only takes account of total unemployment, whereas many authors (including Layard and Nickell; note 21, p175) have argued that both the number and type of the unemployed affects wages; see Ham (1986) and Theodossiou (1992) for example.

FE and CS models is that the former puts FF at the bottom of the earnings heap almost immediately, whilst for the latter it is not until 1982 that FF really moves into the low-pay bracket.

More distinctive is the result for Banking, Finance and Insurance (BFI: group 8). The cross-section moves the increment to pay from 0.05 to 0.25, a relative move in line with other industries. However, the FE estimates show the increment moving from the bottom of the range (at -0.14) to almost the top (at +0.17).

Thus, once individual differences are allowed for, the BFI group has made a very large advance in relative earnings since 1977. The rise throughout the 1980s can be put down to increasing financial deregulation and the boom in financial services. More surprising is the relatively poor initial state. This may be due partly to a "cohort effect" - younger employees with high earnings and high earnings growth push down the relative wage of older cohorts. An alternative is that branch banking employees have relatively low wages and wage growth compared to those with similar qualifications in similar jobs; before the growth in financial services these constituted the bulk of employees in this sector. However, this is largely speculation without more detailed information on this group.

It may also be noted that, apart from the two groups mentioned, the other industries maintain much the same relative position over time. This result has also been reported in the US (Helwege(1992)), where there is a wide literature discussing the "efficiency wage" view of inter-industry differentials (see Krueger and Summers (1988), for example).

9.2.4 Occupation (kos)

The default occupational grouping is "clerical and related" - the basic non-manual group. Figure 9.3 gives the relative positions of the manual groups over this period. The results for

both models are similar in shape, but the cross-section shows a much larger absolute divergence and a relative upward shift in the position of the non-manual worker. Allowing for individual heterogeneity, the non-manual worker is comparable with the average manual worker, although his relative position has steadily improved since 1977.

The change in the relative position of security and protection (group 254) stands out. Although the cause of this change is not clear, both TVFE and TVCS estimates show a steady improvement until 1985, then a falling off for the rest of the period. As individual-specific fixed effects do not affect this shape, it may be that the change in the relative position of this group is due to demand rather than supply factors.

Robinson (1994) notes that "low-paid" (including clerical) jobs have declined in importance for men since the war in terms of numbers employed; employment growth has been in "high-paid" (professional and managerial) occupations. Figure 9.4 shows that, compared with other non-manual workers, clerical workers have become steadily worse off during the 1980s. Thus professional and managerial workers have not only improved their relative earnings but also their share of employment, which suggests that the increased returns to this group are due to increased demand. This is less clear in the CS figures because the scale is much larger. One reason for the huge difference in the size of the returns to occupations may be that non-manual occupations rely much more on "unmeasurable" characteristics: personality, motivation, talent, ability, and so on. If these characteristics stay roughly constant over the period of the survey, this could explain the disparity between the TVFE and TVCS findings.

A second significant factor in the differences between the estimates may be education. Greenhalgh (1980) notes the relationship between occupations and levels of education. For the reasons noted in section 9.1.1, education is likely to produce an individual specific element correlated with occupation which would be transformed out by the TVFE model. Thus the relatively poor performance of the TVCS model may be due to omitted variable bias.

Assuming that education and occupation are not significantly correlated (and so both the TVFE and TVFE are consistent and unbiased), this raises the issue as to which model gives the "better" result. In section 9.1.1 it was noted that the assumed mean of these effects is zero - implying that a non-zero mean is subsumed into the characteristics of the reference individual. The TVCS, by not transforming out this effect, may more accurately reflect the occupational returns due to the average individual. On the other hand, the TVFE model produces a "pure" coefficient and so gives the return to an occupation allowing for any individual characteristics. Thus the TVCS predicts overall returns in an occupation, whereas the TVFE is more appropriate for comparing occupational differences[10].

9.2.5 Age

Age is commonly included as an explanatory variable in Mincerian reduced forms as a proxy for a number of "human-capital" characteristics - experience, tenure, seniority, and so on. The age earnings profile is typically concave, reflecting the benefits of human capital accumulation at an early age[11]. This has led to the common adoption of a quadratic form for age or experience, a practice criticised by Murphy and Welch (1990) for under-estimating initial earnings growth and over-estimating the relative decline in wages of older workers.

The age profiles reported in chapter eight (figure 8.12) indicate that continuous specifications of the age-earnings profile are likely to perform badly. In this study a categorical variable was used to avoid imposing a specific functional form on the age profile. With enough dummy variables this is more flexible than a continuous form (although it does require more degrees of freedom, which may explain the rarity of this specification).

---

[10] Note that the TVFE will also transform out those who remain within the same occupation in all periods - this cannot be distinguished from individual heterogeneity. Thus the TVFE model places much greater weight on changes in occupation than does the TVCS.

[11] See Berndt (1991, pp152-158). This profile is of course consistent with a number of theories: for example, job-search/segmented labour markets, forced saving, class stratification (Theodossiou(1992), Neumark (1994), Bowles and Gintis (1975)).

The profiles are given in figure 9.5, with reference age 31-35. The cross-sectional results are very appealing, both theoretically and in the apparent stability of these effects over time. This latter result is echoed in the actual age-earnings profiles which show almost no shift over time (hence the aggregation over years used to produce figure 8.12). Note that the shape of the profiles support Murphy and Welch's (1990) contention that a quadratic form would produce an excessively flat profile for young workers and an overly steep one for older workers.

However, the TVFE results make little economic sense, suggesting that, for example, in 1990 a sixteen-year-old would earn twice as much as a thirty-five-year-old doing the same job. Although results for initial years are sensible, the profile appears to be rotating about the reference age over time. This result recurs in all the TVFE studies carried out so far and only those models, and is limited to the age variables; nor is it an issue with the "pooled" or "restricted" fixed-effect models which have time-invariant slope coefficients and the expected concave shape.

Bell and Ritchie (1995a) have argued that this effect is spurious, a hitherto unreported side-effect of some models with time-varying coefficients. The reason for this apparently nonsensical result is the collinearity of time dummies (and trend variables) with variables which advance or decline in constant steps; for example, age, experience, length of residency, age of youngest child, and so on. Incremental variables incorporate an implicit trend varaible which means that the effects of time and incremental variables may not be properly separated. Most importantly, Bell and Ritchie show that there is a problem of identification with categorical variables[12]. Experiments with the data seems to suggest that the coefficients on age <u>are</u> poorly identified, and thus interpretation of the coefficients in figure 9.5a is of dubious value.

---

12 If the incremental variables are cardinal, then the model is fully identified; however, uncovering the true coefficients still requires some manipulation of the regression results.

It may be argued that, in the light of these findings, TVFE models should exclude age variables. However, the TVCS model would clearly be subject to missing variable bias if the age variables were excluded, and so for comparability the age variables should be included in the TVFE regression (the curious age coefficients have little effect on the other variables, including the time dummies which are large relative to most age coefficients). Moreover, there is the small possibility that these coefficients are the genuine result of a "cohort effect", although this would require a remarkably large increase in the earnings potential of young workers which continued steadily throughout the period. Most importantly, the validity of these variables as regressors is not dependent upon being able to identify the true coefficients. Bell and Ritchie (1995a) show that although the coefficients on age are not the structural parameters, the set of age variables still contributes useful information to the model. Therefore the age dummies are included in TVFE specifications.

9.2.6 Union coverage (agt)

The effect of unions on wages is a large issue which is not tackled in detail here. However, figure 9.6 depicts the coefficient on a dummy variable, set to one if earnings are affected by collective agreement. This variable is thus much wider ranging than a dummy on union membership and should avoid unmeasured spill-over effects; on the other hand, only national collective agreements are considered for this question. The net effect is that the NESPD coverage variable approximates union membership (Booth (1995)) in the proportion of individuals covered[13].

Using this dummy as a proxy for "union effect" sidesteps a number of issues: selection bias in union presence and membership, contemporaneous membership/wage decisions, and so

13 Andrews and Bell (1995) have analysed the NES's coverage dummy using the information on local agreements collected in two years, and report that the NES coverage figures agree with other survey data when similar definitions are used.

on (see Elliott (1991); Farber(1986, section 5); Lewis(1986) for surveys of these issues). Most importantly, there is no interaction between the union dummy and the other variables; that is, the effect of union coverage is assumed to be a simple hike in the wages of those covered, with no effect on the return of other characteristics[14]. Clearly this may be violated; for example, preliminary studies on the NES using a union/sector interactive dummy suggest that there are sectoral differences in the impact of unions. Another obvious example is industry/union interactions, particularly in the light of the decline in manufacturing during the period in question. Thus these estimates need to be seen as the difference in the means of covered and uncovered workers conditional on all other features of the job or individual.

Whilst the coefficient on the dummy may be a crude measure of union influence, it has the useful feature that the effect can be followed over time; thus this variable can at least indicate the direction in which any union effect may be changing. This is of some interest given the changes in union legislation and membership since the mid-1970s but, as noted in Andrews, Bell and Ritchie (1993), studies constructing a time-series for the union effect are rare on the ground. The only other studies constructing micro time-series for the UK on a consistent basis appear to be Meghir and Whitehouse (1992) and Lanot and Walker (1993b), both using repeated cross-sections on the FES. The former found a union markup of around 10% over the period 1975-1983, and then a rise to around 17% for 1984-1986; Lanot and Walker report a markup on OLS estimates of 5% in 1978, rising to 12% by 1985. Both sets of results indicate the problems of snapshot estimates of the differential[15].

Figure 9.6 presents the TVFE and TVCS results. Bearing in mind the above qualifications, the "union markup" varies widely over the fourteen years. The decline in union membership

---

[14] Obviously, this is not unique to union status: a case can be made for interactions between a number of the variables: for example, industry/occupation, occupation/tenure, tenure/age, and so on.

[15] Booth (1995) notes that snapshot studies appear to show a relatively stable union markup, which is unexpected given the changes in the industrial environment. However, there are difficulties comparing separate studies which use different data, periods, and estimation methods.

over the period is reflected in the fall in the mean level of coverage. The fall in the coefficient from 1982 onwards may be due to this decline but may also reflect the anti-union legislation enacted over this period. The rise in the coefficient from 1979 to 1982 may indicate that unions were effective in maintaining wage levels over a period when the economy was undergoing a major restructuring. These results may suggest that the union effect is counter-cyclical (that is, the union wage gap is largest when demand for labour is low and smallest in a tight labour market), but the legislative and membership changes in the 1980s make this assertion difficult to prove using this data.

These results exhibit a pattern similar to those of Meghir and Whitehouse (1992) up to 1982 (if not for the subsequent three years), but the results bear little relationship to the steady rise in the coefficient from 1977 to 1985 found by Lanot and Walker (1993b).

Figure 9.6 suggests a relatively small "pure" union effect, and that it is related largely to individual ability. The cross-sectional result is exceptionally small, and is occasionally insignificant. Table 9.1 presents the t-statistics for this variable for both FE and CS results, with values significant at the 5% level in bold. In four years the agt variable is insignificant, which is unusual for this dataset where the large number of observations tends to produce high t-statistics.

Table 9.1 T-statistics for agreement variable

	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	Pool.	Rest.
TVFE	**14.954**	**9.102**	**6.525**	**9.105**	**15.148**	**18.504**	**14.689**	**11.957**	**10.383**	**8.983**	**7.494**	**5.768**	**5.512**	**2.063**	**31.127**	**32.853**
TVCS	**8.415**	1.345	1.841	**2.433**	**6.701**	**9.142**	**3.139**	**3.347**	**5.059**	**3.282**	**3.112**	-0.190	1.183	**-3.644**	**7.931**	**16.767**

Together, the results in table 9.1 and figure 9.6 contrast strongly with other findings that unions have a large effect on income of between 10% and 20%[16]. One feature of this

[16] For example, Lewis (1986) surveys and tries to assess on a consistent basis US studies, and places most of the studies (including panel studies) within this range. Stewart (1987) using WIRS finds differentials around 10%, but this is very dependent on the characteristics of the workplace, not the worker; the differential falls

study is the relatively high number of workplace variables, particularly in the breakdowns of occupation and industry, and it may be that the larger union markups in other are due to correlation between the variables rather than a "pure" effect[17]. Alternatively, it could be argued that the relatively small effects described here are due to collinearity between union status and, for example, occupation. In the absence of interactive dummies these competing hypotheses cannot be tested.

A final reason for the small coefficients is that the coverage variable is restricted to national agreements. Blackaby, Murphy and Sloane (1991) found that the coverage measure had a notable effect on estimates. Stewart (1987) showed that both union and employer associations had a marked effect on the apparent union wage gap, with the maximum gap being 21% but for several types of bargaining arrangement no significant effect at all. Andrews and Bell (1995), using the information on local agreements collected in 1978 and 1985, found that the inclusion of these bargains raised cross-sectional estimates by around 8%. If the lack of local agreements has this effect in all years, then the coefficients in figure 9.6, while still relatively small, are more in line with other studies.

Unusually, the TVFE model produces larger coefficients than TVCS. The implication is that the union has a larger effect when we allow for differences between individuals. This result is unexpected: Lewis (1986) argues that CS studies should represent the upper bounds on the union effect as higher union wages should lead to higher quality employees - which the TVFE model should detect. Jakubson (1991) also puts the case for a positive bias, claiming that a CS wage gap of 20% may be reduced to 5% by allowing for

---

to zero for some firms. Barth, Naylor and Raaum (1995) find a similar result. Murphy and Sloane and Blackaby (1992) using SCELI reported gaps of 7%-13% depending on the type of worker and the allowance made for selection bias. Finally, Blackaby, Murphy and Sloane (1991) using GHS record a union gap of 28% but note that "coverage" and "membership" give substantially different results.

[17] Although Stewart (1987), using the establishment-level data in WIRS, finds that increasing the number of industry variables seems to make relatively little difference, reducing the union wage-gap by 0.5%.

individual heterogeneity. Finally, Booth (1995) notes that the exaggeration of classification errors under the covariance transformation means that, theoretically at least, FE estimates cannot be larger than CS ones.

However, this argument for the "upper bound" of CS estimates ignores the potential for other errors in estimation, most importantly the potential correlation between the explanatory variables and the error terms (including the individual heterogeneity). The reverse result here seems to suggest a negative CS bias in that individuals of relatively low "ability" are attracted to union positions through a form of adverse selection: individuals of high "ability" are encouraged to strike individual deals with employers rather than joining unions[18]. This is consistent with the stylised facts that the union effect tends to be larger for manual workers and that unions tend to have an equalising effect when different skill levels are covered (see, Farber (1986) and Lewis(1986)).

There may also be an element of selection bias. Recent studies by Lanot and Walker (1993a, 1993b) have shown that the estimated markup can vary wildly if a selection mechanism is introduced; but the effect is always to increase the markup. If the selection probability is relatively stable over time, then the TVFE model will transform this out. Thus the TVFE model may, more by accident than design, be taking account of an element of selection bias and so uncovering the true differential.

9.2.7 Wages Council coverage (wbc)

Figure 9.7 shows that the effect of being in a position covered by Wages Council (WC) regulations is negative. Although this seems to imply that WC coverage reduces wages, the causation probably runs the opposite way: that the very lowest paid jobs, all other things

[18] The characterisation of unmeasured heterogeneity as "ability" here is for convenience. Similar arguments hold if "ability" is replaced by "motivation", "productivity", "nice shoes", or a number of other qualities.

being equal, are those most likely to be covered by the WCs. The absolute coefficients from the FE model are small for most of the time, rising at the end of the period. Those from the CS model are much larger, but fall sharply in 1982 and continue to drop until 1988 whereupon differences between the models become relatively small. The increasing wage gap between those covered and the rest of the working population is consistent with the increasing inequality in the UK labour market at the end of the 1980s (Bell(1995); Bell, Rimmer and Rimmer (1994)).

These results should be treated with suspicion, as this dummy only applies to about 5% of those included in the NESPD. The FE model drops individuals with only one observation and this will reduce the number still further, which may explain the minimal coefficients for this model for most of the time. It may be that the increasing computerisation of records has been accompanied by a steady rise in the number of employees earnings under the NI limit (and possibly covered by the WCs). This could explain the increasingly strong results for the FE model, were it not for the declining mean of those covered. Moreover, the time path of the CS coefficient displays little consistent trend[19]. In short, these results do not afford a clear interpretation of the effect of WCs.

9.2.8 Public sector working (sector)

The proportion of individuals in the private sector increased steadily throughout the 1980s, as can be seen in figure 9.8; the relative returns to working in the private sector rose at a corresponding rate. This is consistent with the view that public sector wages tend to be counter-cyclical; that is, the private sector improves its relative pay during prosperous times (Ehrenberg and Schwarz(1986); Holmlund and Ohlsson(1992)). The TVFE model produces smaller absolute coefficients.

19 Sudden leaps in 1982 occur in several NES statistics for no readily apparent reason.

Interestingly, the results indicate that it is only in recent years that the private sector has become relatively better paid than the public sector - a reversal of the standard argument that public sector employment compensates for lower wages with more job security (see Ehrenberg and Schwarz(1986)). However, Rees and Shah (1992) and Bell and Ritchie (1993a) have also found a public sector premium in the hourly wage rate, for males and females respectively. Rees and Shah argue that the public sector employees work significantly fewer hours (around 7%), which could produce a private sector premium if the difference in hours is not recognised; for example, Andrews, Bell and Ritchie (1993) using a CS method on similar data to this study but with weekly wages find a public sector discount for the government sectors.

This does not explain why public sector positions should earn a higher hourly rate. Bell and Ritchie (1993a) argued that this public sector premium occurs largely in the government sector; there may be a small discount in public corporations. Given the lack of comparable "governmental" jobs in the private sector and the DE's rather idiosyncratic classification system (for example, *all* teachers are classified as "public sector"), this may be evidence of a misspecified occupation/industrial characteristic rather than a pure "sector" effect. There may also exist compensating differentials other than the job security issue (Ehrenberg and Schwarz (1986, pp1246-1251)) which can have a negative effect; for example, the disamenity experienced by dustmen or street cleaners. Finally, some authors (Hartog and Oosterbeek (1993); Rees and Shah (1992)) have argued for "comparative advantage" in the choice of sector: that people are inherently public or private sector workers by nature or early training[20]. Unfortunately, the reduced form model as specified in (9.1) is consistent with all these hypotheses and so sheds little light on the causes of the public sector premium.

9.2.9 Length of time in the job (j12m)

20 This does raise the possibility of another source of selection bias.

The NES does not collect a tenure measure on a yearly basis, but it does record as a binary variable whether an individual has been in a job for at least twelve months. Figure 9.9 shows the effect of holding a job for less than one year, and it is clear that there is a discount on earnings. This is in line with human capital theory, the argument being that tenure and, by implication, experience increases (possibly firm-specific) human capital and is thus rewarded in higher wages (Mincer (1974); Coleman(1994)). However, as for the sector variables, this is not the only interpretation that can be put on these results; for example, this result would be expected from job-search models, increased tenure being associated with improved job matching and consequent increments to earnings (Theodossiou (1992)).

Although this variable is always significant, the effect is relatively modest, with the effective discount on earnings peaking at 3% in 1987 (TVFE specification) and generally around 2%. However, this dummy is unable to distinguish between those who have moved into jobs from unemployment and those who move from one job to another (possibly in the same company). Empirical evidence shows that those who move between jobs tend to take up a position with a higher wage, which may imply that those who stay in one post may earn less than those who move up the pay scale by changing jobs regularly. The small size of the coefficient may therefore reflect the conflation of these two opposing effects.

Note that the TVCS estimates are almost twice the size of the TVFE coefficients. This supports the view that individual characteristics influence whether people change jobs regularly, although these results cannot definitely ascribe this to heterogeneity (Cripps and Tarling (1974)) or some form of state-dependence (Phelps (1972)).

9.2.10 Attrition variables

Assuming the AVs are adequate proxies for selection, then ideally the coefficients on these variables should be small for the TVFE specification at least. If both the TVFE and TVCS

coefficients are insignificant, then selection bias is unlikely to be an issue; if the FE estimates are much smaller then heterogeneity plays a large part in selection, which simplifies the correction process. However, Figure 9.10 suggests that, while unmeasured characteristics may have a part to play, there remains a significant element of serial correlation and/or state dependence in selection. The coefficient on YrsIn is always very significant; the coefficients on the other two usually are. It should be noted that these attrition variables are affected by the collinearity problem discussed in 9.2.5, and also are highly correlated when the individual is in the dataset for long periods of time; they should be thus treated with some care.

The coefficients on InLast and CurrStay are almost negligible and (given that these variables are only proxies) can probably be ignored. However, although the coefficient on YrsIn appears small, the mean value of this variable is large relative to the others in the dataset and thus the total effect is relatively large[21]. The TVFE estimator appears to place a higher value on YrsIn compared to the TVCS model, suggesting that, after allowing for individual differences, complete observation history to date is even more important than recent experience.

9.3 Structural shifts in the UK labour market

In this section two restrictions are placed upon the time-varying characteristic of the TVFE and TVCS models. The first is that all coefficients are constant over time (the pooled model of (9.6b) and (9.6e)); the second is that only the intercepts vary over time (the restricted model of (9.6c) and (9.6f)). Although the qualitative results on the coefficients are the same for all six models (apart from the TVFE age coefficients), regressions over the whole period unambiguously reject the hypothesis of constant slopes and/or intercepts (see table A9.1 in the

[21] For example, an individual with 14 years of observations in 1990 can expect a premium of around 15% over someone making their first appearance, ceteris paribus. This would amount to a substantial premium for, for example, a woman re-entering the labour force after raising a family compared to a woman remaining in employment throughout the period.

chapter Appendix).

One question of interest is whether choosing a different time frame might uncover structural stability in the labour market. Identifying parameter stability with structural stability, this amounts to finding periods when the pooled or restricted hypotheses cannot be rejected.

A problem with testing this hypothesis using the TVFE model is that it needs separate means matrices to be generated for all the desired combinations of "stable" periods. However, combinations of years are easily tested in the CS framework. Therefore, despite the drawbacks described in section 9.1, the TVCS model is used as a rough guide to the potential for structural shifts in the UK labour market.

Given 14 periods of observation, structural changes in the labour market can be represented in around 2^{13} ways in a simple same-not same framework. A consideration of the UK labour market simplifies this with some sensible assumptions. A wide body of evidence indicates a shift in the UK labour market at the beginning of the 1980s (see Robinson (1994), for example). The path of unemployment, peaking in 1986, also suggests a change in direction for employment prospects. Since the end of the data period (1990) does not fully reflect the economy's shift into recession (particularly with regard to expectations) a reasonable suggestion would be to look for a three-stage pattern in the UK labour market, with one shift in the early 1980s (into a period of rising unemployment and a sharp decline in manufacturing) and one in the late 1980s (with unemployment generally falling and a boom in services).

The method employed is the Chow test; that is, to run several TVCS estimates and construct F-statistics to indicate whether restrictions on the model seem justified. Allowing for structural breaks in the years {1980, 1981, 1982} and {1985, 1986, 1987} gives nine date combinations and fifteen periods to test. Table 9.2 gives the resulting F-statistics for both the

pooled (all coefficients constant over the period) and restricted (intercepts allowed to vary) models, represented by "U v P" and "U v R" respectively. Degrees of freedom have not been given as in all cases these are extremely large (>200 for the numerator, >300000 for the denominator). So to test for stability over 1977-1980, for example, the pooled f-statistic is 15.45; for 1981-1986 it is 14.82.

Table 9.2 F-statistics from specification tests (cross-section)

	U v P	U v P	U v P	U v P	U v R	U v R	U v R	U v R
	80	81	82	90	80	81	82	90
77	15.45	16.85	18.12	44.08	13.02	14.85	16.45	26.65
85	19.87	11.56	6.93	17.48	12.39	9.74	5.33	8.71
86	16.80	14.82	10.75	10.18	12.96	10.43	6.48	6.90
87	20.38	18.47	14.56	6.93	13.76	11.24	7.54	5.48

The results in Table 9.2 clearly reject all the hypotheses. As the TVFE models generally give larger F-statistics, it is a fairly safe assumption that the TVFE model would also reject tests of parameter stability for all the above combinations of periods.

The rejection of the pooling hypothesis is predictable given that the model is adjusted for RPI rather than wage inflation; some growth in wages remains and so the time dummies are increasing over the period. The F-statistics for the restricted model are smaller, arising from the more general specification of this model; however, this hypothesis is also rejected by the data.

The F-statistics fall with the length of the period under review, as would be expected. However, even taking two years at a time the hypothesis of parameter stability is rejected (although the F-statistics can fall as low as 2-3). While the tests are only strictly applicable to the particular equation under review (and dependent upon the assumed normality of the error term), the suggestion is that parameter stability is a non-starter - even over short periods.

An important implication of this parameter instability is that CS studies may produce "better" models than simplistic panel studies - especially those employing differencing transformations. Compare the TVCS model of section 5.1 with the differencing approach of section 5.4. For the latter, it was remarked that the estimator forces slope parameters to be the same over the two periods being differenced. Thus the underlying assumptions of the two models are

$$\begin{array}{ll} \textit{cross-section} & y_{it} = \lambda_t + x_{it}\beta_t + u_{it} \\ \textit{differencing} & y_{it} = \lambda + x_{it}\beta + \alpha_i + u_{it} \end{array} \qquad (9.7)$$

for any two-period estimation. The CS model restricts the error term but allows parameters to vary over time, whereas the differencing model can account for individual heterogeneity but forces coefficients to be the same over any two periods. Thus there is no guarantee that the differencing panel estimator will not produce worse results than the CS model. If the restriction on parameters is carried over to more than two periods, then the validity of the differencing model is likely to decline sharply[22].

This should not be seen as a criticism of differencing models; the key point is that the parameter variability may be more important then individual heterogeneity. This issue was raised in section 2.1, when it was noted that although a panel model cannot be less efficient then a comparably specified CS model, a poorly specified panel model may perform worse than a cross-section which has different assumptions - as is the case in (9.7). This issue is peculiar to the equation being estimated, and so general conclusions about the virtues of parameter constancy versus heterogeneity cannot be drawn. However, the results presented above suggest that structural stability is not something to be assumed without some testing, and that simple CS models may be a better choice than panel specifications with ad hoc restrictions.

9.4 Summary

22 This does not invalidate the differencing approach in general; for example, the differencing model in section 5.3 (allowing for fully-varying coefficients) will out-perform the CS model.

In this chapter the first panel estimates on the NESPD have been presented and compared with cross-sectional estimates. These comparisons have supported the view that individual heterogeneity is correlated with the occupation, industry, sector et cetera of an employee, although as the TVFE model is both more efficient and has a smoothing effect on the estimates the difference between the two models cannot be ascribed wholly to heterogeneity.

The results generally make economic sense. The wage premium on working in the public sector is unusual but is consistent with other studies using basic hourly wages. Two results are particularly noteworthy. First, the rise in private sector wages throughout the 1980s relative to the public sector seems to support a counter-cyclical hypothesis. Secondly, the decline in the union premium over the same period is consistent with the view that anti-union legislation reduced the bargaining power of unions, although this effect could also be ascribed to changes in the macroeconomy.

Finally, a crude attempt to find periods of parametric stability rejected this hypothesis in all the test cases. This has important implications for the bulk of panel models on the labour market which habitually assume structural stability over time. It may be that cross-sectional models will perform better than poorly-specified panel models which impose constant slope coefficients on the model.

Appendix A9 Regression details and summary regression results

Summary statistics for the unrestricted, pooled and restricted regressions are given in Table A9.1a (FE) and Table A9.1b (CS). Separate results for the fourteen unrestricted cross-sections are not given here, as this involves fourteen sets of summary statistics. Instead, the results for a sample year (1984, chosen for being halfway through the period) are given.

Table A9.1a Fixed-effects summary statistics 1977-1990

Fixed-effects	Unrestricted TVFE (9.6a)	Pooled (9.6b)	Restricted (9.6c)
F-test for general significance (degrees of freedom)	1011.9922 (797,907324)	11587.3083 (56, 908065)	9520.7224 (69, 908052)
R^2 Adjusted R^2 (adjustment factor)	0.4706 0.4701 (907324, 908121)	0.4168 0.4167 (908065, 908126)	0.4198 0.4197 (908052, 908121)
TSS ESS RSS Estimated variance σ^2	53436.987 25147.602 28289.385 0.031	53436.987 22270.845 31166.142 0.034	53436.987 22431.156 31005.831
Observations Restrictions Variables	1103018 194897 797	1103018 194897 56	1103018 194897 69
F-tests for models			
vs Pooled (degrees of freedom)	124.5156 (741, 907324)	-	-
vs Restricted (degrees of freedom)	119.8410 (727, 907324)	335.3523 (14, 908051)	-

Table A9.1b Cross-section summary statistics (part: 1984)

Cross-section	Unrestricted TVCS (9.6d)	Pooled (9.6e)	Restricted (9.6f)
F-test for general significance (degrees of freedom)	1573.4970 (56, 75845)	23631.9566 (56, 1102961)	22166.0247 (56, 1102948)
R^2 Adjusted R^2 (adjustment factor)	0.5374 0.5371 (75845, 75901)	0.5454 0.5454 (1102961, 1103017)	0.5925 0.5295 (1102948, 1103004)
TSS ESS RSS Estimated variance σ^2	15742.463 8460.315 7282.149 0.096	232080.576 126582.291 105498.285 0.096	221647.203 117364..033 104283.170 0.095
Observations Restrictions Variables	75902 1 56	1103018 1 56	1103018 14 56
F-tests for models			
vs Pooled (degrees of freedom)	44.0800 (728, 1102220)	-	-
vs Restricted (degrees of freedom)	26.6462 (714, 1102220)	917.9622 (14, 1102934)	-

Table A9.2 gives the results of the FE regression pertaining to the sample year 1984. All coefficients are relative to the representative categorical variables. The constant term is relative to the intercept in 1977.

Table A9.2 Time-varying fixed-effect regression results (part:1984)

Variable	NES code	Mean	Coefficient	Error	T-value	T-prob
Constant		1.000	0.0074	0.014	0.524	0.600
InLast		0.832	0.0106	0.003	4.260	0.000
YrsIn		7.202	0.0282	0.001	40.991	0.000
CurrStay		5.302	-0.0014	0.000	-3.760	0.000
reg	45	0.169	-0.0684	0.003	-24.779	0.000
reg	48	0.034	-0.0873	0.005	-17.008	0.000
reg	55	0.075	-0.1115	0.004	-28.495	0.000
reg	60	0.095	-0.1164	0.004	-30.745	0.000
reg	66	0.068	-0.1097	0.004	-26.983	0.000
reg	70	0.089	-0.1153	0.004	-29.639	0.000
reg	74	0.113	-0.1041	0.004	-28.878	0.000
reg	79	0.056	-0.1254	0.005	-26.830	0.000
reg	88	0.044	-0.1239	0.005	-24.385	0.000
reg	98	0.095	-0.0943	0.004	-22.384	0.000
agt	998	0.469	0.0229	0.002	11.957	0.000
wbc	248	0.045	-0.0173	0.004	-4.006	0.000
j12	2	0.128	-0.0282	0.002	-12.053	0.000
sec	0	0.658	-0.0175	0.003	-6.234	0.000
div	1	0.052	0.1664	0.010	17.032	0.000
div	2	0.062	0.0972	0.010	10.255	0.000
div	3	0.189	0.0561	0.009	6.096	0.000
div	4	0.113	0.0483	0.009	5.197	0.000
div	5	0.079	0.0362	0.010	3.810	0.000
div	6	0.119	0.0004	0.009	0.045	0.964
div	7	0.105	0.0668	0.010	7.050	0.000
div	8	0.080	0.0241	0.009	2.567	0.010
div	9	0.186	0.0223	0.009	2.415	0.016
age	16	0.005	-0.2778	0.015	-18.646	0.000
age	18	0.028	-0.1336	0.009	-14.890	0.000
age	20	0.042	0.0158	0.007	2.210	0.027
age	22	0.047	0.0087	0.006	1.494	0.135
age	24	0.050	-0.0079	0.005	-1.529	0.126

age	26	0.049	-0.0150	0.005	-3.262	0.001
age	30	0.096	-0.0070	0.003	-2.057	0.040
age	40	0.128	-0.0012	0.003	-0.387	0.699
age	45	0.103	-0.0074	0.004	-1.994	0.046
age	50	0.102	-0.0078	0.004	-1.859	0.063
age	55	0.100	-0.0095	0.005	-2.039	0.042
age	60	0.086	-0.0112	0.005	-2.182	0.029
age	120	0.042	-0.0056	0.006	-0.921	0.357
kos	100	0.012	0.1793	0.007	24.550	0.000
kos	122	0.075	0.1411	0.004	38.835	0.000
kos	147	0.044	0.0963	0.005	19.298	0.000
kos	156	0.009	0.1294	0.008	15.440	0.000
kos	189	0.088	0.1026	0.004	28.337	0.000
kos	211	0.065	0.1307	0.004	34.138	0.000
kos	246	0.040	0.0397	0.005	8.549	0.000
kos	254	0.030	0.0940	0.005	17.535	0.000
kos	281	0.037	-0.0968	0.005	-20.232	0.000
kos	295	0.022	-0.0625	0.008	-8.197	0.000
kos	327	0.034	0.0345	0.005	6.891	0.000
kos	385	0.052	0.0344	0.004	7.743	0.000
kos	462	0.185	0.0416	0.003	12.768	0.000
kos	477	0.041	0.0285	0.005	6.233	0.000
kos	503	0.044	0.0216	0.005	4.511	0.000
kos	533	0.108	-0.0190	0.004	-5.451	0.000
kos	540	0.019	-0.0235	0.006	-3.904	0.000

Table A9.3 lists the categorical variables used. Reference categories are marked by an asterisk.

Table A9.3 Dummy variable categories and description

Variable	Category	Description
Region		
reg	33 *	Greater London
reg	45	South East
reg	48	East Anglia
reg	55	South-west
reg	60	West Midlands
reg	66	East Midlands
reg	70	Yorkshire and Humberside
reg	74	North-west

reg	79	North
reg	88	Wales
reg	98	Scotland
Agreement		
agt	998	Earnings affected by collective agreement
agt	999 *	Earnings not affected
Wages Council board		
wbc	248	Job is covered by Wages Council regulations
wbc	249 *	Job not covered
Length of job		
j12	1 *	Current job held for over one year
j12	2	Current job held for less than one year
Sector		
sec	0	Job is in private sector
sec	3 *	Job is in public sector
Industry		
div	0 *	Farming and fishing
div	1	Energy and water supply
div	2	Other mineral and ore extraction
div	3	Metal goods, engineering and vehicles
div	4	Other manufacturing
div	5	Construction
div	6	Distribution and hotels
div	7	Transport and communication
div	8	Banking, finance and insurance
div	9	Other services
Age		
age	16	<=16 years old
age	18	17-18
age	20	19-20
age	22	21-22
age	24	23-24
age	26	25-26
age	30	27-30
age	35 *	31-35
age	40	36-40
age	45	41-45
age	50	46-50

age	55	51-55
age	60	56-60
age	120	>60 years old
Occupation		
kos	100	General management (including directorial)
kos	122	Professional and related supporting management and admin.
kos	147	Professional and related in education, welfare and health
kos	156	Literary, artistic, and sports
kos	189	Professional and related in science and engineering
kos	211	Managerial (excluding general management)
kos	238 *	Clerical and related
kos	246	Selling
kos	254	Security and protection
kos	281	Catering, cleaning, and hairdressing
kos	295	Farming and fishing
kos	327	Materials processing (excluding metal)
kos	385	Making and repairing (excluding metal)
kos	462	Process, making and repairing (metal and electrical)
kos	477	Painting, assembling, inspecting, and packaging
kos	503	Construction and mining
kos	533	Transport operating, materials moving and storage
kos	540	Miscellaneous

Chapter 10

Female earnings 1977-1990 and the wage gap

This chapter follows on from the previous one in providing some fixed-effects estimates of wages for female full-time workers. The estimation method is the same, and the conclusions drawn regarding cross-sections versus fixed-effects and on parametric stability still apply, so this chapter mainly presents the unrestricted fixed effects (TVFE) results. Again, these are the first panel estimates on female earnings from the NES[1]. After discussing these results the endogeneity of occupational choice is considered. The chapter ends with a discussion of the male-female earnings gap, which the TVFE specification allows us to trace and break down over time.

This chapter limits itself to females in full-time employment to provide a comparison with male earnings in the pervious chapter. As was demonstrated in chapter eight, females experience a much wider range of employment patterns than men, substituting between part-time and full-time employment. A fuller description of the determinants of women's earnings should take into account this variety of employment options. However, in this paper we are interested in a comparison of the competing returns to men and women in full-time work and investigating "discrimination" in the market for full-time employees. We are therefore conditioning on a subset of the working female population. However, it should be noted that by doing so we are ignoring a potential source of differential treatment.

10.1 Econometric issues

The basic equation is once again the Mincerian human-capital model used in the previous

[1] This chapter is based on Bell and Ritchie (1995b). An earlier paper, Bell and Ritchie (1993a), presented cross-section results.

chapter with identical regressors included[2]; that is, the TVFE model:

$$w_{it} = x_{it}\beta_t + \alpha_i + \lambda_t + u_{it} \qquad (10.1)$$

An important feature of these reduced forms concerns the relationship between $\mathbf{x}_{it}$ and $\mathbf{u}_{it}$. It is quite plausible that some of the elements of $\mathbf{x}_{it}$ are endogenous, and thus the assumption that $\text{Cov}(\mathbf{x}_{it}\mathbf{u}_{it})=0$ is invalid. In the literature on female earnings, one of the more significant sources of endogeneity is occupational choice. It has been argued that women self-select into "women's jobs", and this crowding forces down earnings in those occupational groupings and so lowers relative female wages; thus the coefficients on occupation should not be seen as the result of a random allocation to an occupation (see Miller (1987)).

There are two well-known methods of dealing with this problem. The first of these, the control function (or switching regression) approach, was popularised by Heckman (1979). It usually proceeds by augmenting the standard regression with a correction factor from a probabilistic model of the self-selection process. In the case of unionisation, for example, this may be a probit including as regressors those characteristics most likely to encourage individuals to join a union. The second approach is the instrumental variables method. This replaces those regressors which are correlated with the disturbances with instruments which are purged of such correlation. Thus, the unionisation dummy would be replaced by a fitted value from some auxiliary regression, which again models the factors influencing whether an individual will join a union or not.

As was discussed in section 9.1.3 in the context of the attrition variables, these two approaches are closely related to each other and to the proxy-variable method (Robinson(1989); Vella and Verbeek(1993); Lanot and Walker (1993a)). As before, separate probit models to take account of absence (and endogeneity) are not feasible, and the

2 Two occupational categories had to be dropped (top management and mining) which had no or alsomst no observations each year.

linear IV and proxy-variables approaches are used.

In the case of the participation decision, the AVs may be expected to capture some of the effects of 'home-time' on female earnings. Note, however, that the existence of part-time work as an alternative to full-time work is much more important for women and past experience of part-time work may well affect future earnings opportunities (Main(1988b)). As the AVs are calculated on the basis of both part-time and full-time work counting as valid employment, then we might expect the response to these variables to differ between men and women.

Wright and Ermisch (1991) argue that occupational variables should not be included in studies of female earnings as the crowding of women into occupations is a part of the 'discrimination' which such studies may hope to measure[3]. If occupational dummies are included then any results are conditioned on the choice of occupation and an important source of discrimination has therefore been side-stepped.

The omission of occupation implies that its defining characteristic is a way of categorising workers, with no inherent features that affect wages except through discriminatory practices. However, "occupation" captures a number of job characteristics which are important in the determinant of wages; coefficient estimates thus reflect more than just the supply-side impacts of occupational crowding. Omission of occupational variables may therefore lead to inefficient and biased estimates. This approach therefore means that a potentially significant explanatory variable has been left out. In addition, it can be argued that occupational choice is determined outside of the model under review, in which case conditioning on the occupational distribution of female workers is an appropriate response. Therefore, categorical variables for occupation are included in these estimates, although in section 10.4 the effect

[3] Blinder (1973) also left out these characteristics, but this was in an attempt to distinguish between "reduced" and "structural" forms rather than making any explicit claim for the endogeneity of the variables.

of leaving out these variables is considered.

10.2 Fixed-effects results

This section reviews the results from the TVFE model; subsection 10.2.10 discusses the effect of instrumenting the occupational dummies. The dependent variable is the natural logarithm of hourly earnings; the sample is all females in the NES employed full-time whose earnings are not affected by absence from work.

As before, to save space all results for all years are not presented here. Instead, table 10.1 at the end of the chapter shows results for the sample year 1984, the middle of the period. The variables are those listed in table A9.3 in the appendix to chapter nine; the default categories remain the same except for the two omitted occupations.

10.2.1 Region

Figure 10.1a shows the fixed effects results for females, and figure 10.1b compares the coefficients in 1990 for males and females. The regional coefficients for females are rather larger than those for males although they show the same general pattern: females in the 'south' generally earn a significant premium over women in the 'north' with otherwise identical observed characteristics. Joshi (1986) argued that women's ability to search for paid work is much more constrained than men's options, due to family factors such as partner's income or children at school. The wider differentials for females may therefore be an indication of greater geographical immobility.

10.2.2 Industry

Figure 10.2 displays the coefficients by division and compares those for males and females

in 1990. The effect of industry on female earnings appears relatively less stable over the period in question (compare figure 9.2a) although for males there is a noticeable fall in the earnings in farming and fishing over this time. Note that the remarkable improvement in the returns to banking, finance and insurance for males is much smaller for females. Overall, the implication of figure 10.2 is that, in 1990, inter-industry differentials are larger for males than for females. Moreover, these differentials appear to be fairly constant apart from the fluctuations in the reference category.

As for occupation, it may be argued that the industry is endogenous: flexibility in working hours, time off may be appealing to female working looking for a variety of job characteristics. If these features are unevenly distributed among industries, then the earnings associated with these industries will reflect a non-random choice of employer. However, it seems more likely that these characteristics are more closely related to occupation: a cleaner in a manufacturing company faces much the same conditions as a cleaner in a service industry. As we have a relatively large number of occupations, it seems likely that occupational endogeneity is the more probable source of bias. However, because certain jobs are closely related to industry, in a later section we consider the effect of excluding *both* industry and occupation.

10.2.3 Occupation

Figure 10.3a shows occupational coefficients for female manual workers, and figure 10.3b those for non-manual workers. The position of junior clerical workers has improved slightly in comparison with manual workers, but fallen sharply against other non-manual workers. This is much the same result as for males. Robinson (1994) places clerical work for females in the "middle-paid" sector of the economy, rather than the "low-paid" sector of male clerks, but, as for males, notes that this sector has been losing shares of employment to professional and managerial occupations. For females, this rise in the "professional classes" is particularly

notable in the 1980s (Robinson (1994, table 12)). Again, the increase in women in professional and managerial positions, and the higher associated earnings, imply a significant demand effect[4].

Figure 10.4 compares coefficients in 1990 for both sexes. Non-manual occupational differentials are much the same for both sexes, but for females the returns to manual occupations vary much less than for males. However, even ignoring the endogeneity issue, interpretation of coefficients on occupation is problematic. As has been noted, occupation embodies a wide range of characteristics including flexibility in working time and the importance of employment experience. The fact that neither these characteristics nor the employment and family history of employees are directly observed may affect estimated coefficients and may bias the allocation of male-female differentials between explained and unexplained components. This issue is discussed further below.

10.2.4 Age

Figure 10.5 reproduces cross-sectional (rather than fixed-effects) age coefficients, for the reasons of collinearity discussed earlier in section 9.2.5. Figure 10.5b reproduces figure 9.5a for comparison. In general, the overall dataset profiles in figure 8.12 appear to be a good approximation for the estimated profiles, in that females have a lower and flatter age-earnings profile compared to males; however, the estimated coefficients do suggest that the relative wages for the young rise more quickly than the overall figures indicate.

These profiles raise three issues. First, there has been very little shift in the coefficients over time apart from some small variation at the far ends of the range where numbers of

[4] Sloane and Theodossiou (1994) argue that a significant proportion of the improvements in female earnings since the mid-1970s has been due to increased demand for female labour. The evidence here suggests that this increased demand is concentrated in the "senior" occupations.

observations are small[5]. The profiles for females show marginally more variation over time.

Secondly, the variation in age effects on the earnings of full-time females is much smaller than that for full-time males. The female coefficients have a range of 0.6 while males vary by 0.9, and the difference is most marked in the steep age-earnings profile for young males. For males, the age variable is likely to be a reasonably proxy of years of experience in the labour market. For females, due to career interruptions, this is less likely to be true. Thus, one might expect the profile of age coefficients for females to be somewhat flatter than that for males.

Becker and Lindsay (1994) argue that the age-earnings profile for young females should be steeper than for young males in some firms, as expected interruptions in work patterns lead to females bearing more of the risk of firm-specific investment through lower starting wages. The evidence here seems to refute this, presumably because such schemes would be likely to fall foul of equal pay legislation[6].

Third, female wage rates peak before those of males. Allowing for the implicit coefficient of zero on the default 31-35 age group, the age coefficients for full-time females reach a peak in the 26-35 range while those of males reach a maximum which is broadly constant over the age range 36-45 (although it should be noted that coefficients around the reference age tend to be insignificant). The overall profile in figure 8.12 reflects this difference in peak earnings fairly accurately. The earlier peak of female age coefficients is consistent with their accruing less labour market experience during their working life.

10.2.5 Union coverage

---

[5] Bell and Ritchie (1993a) found a significant difference between the 1976 and 1990 cohorts for both sexes. In the light of Figure 10.5, this suggests that the earlier result is more probably due to the peculiar characteristics of the 1975/1976 data, when the dataset was being set up.

[6] If the amount of firm-specific investment varies between men and women, then there is a problem of identifying the true effect from this result.

Figure 10.6 displays coefficients and mean levels of coverage for both male and female full-timers. An interesting implication is that females derive more benefit than males from national collective agreements. For both sexes, there was a decline in the benefits from coverage during the late 1970s from around 3% of mean wages to around 1.5%; some recovery during the early 1980s, peaking in 1982 at 5% for females; and then a considerable reduction to less than 2% in the subsequent period.

Figure 8.14 showed that, whilst for males the level of union coverage has fallen for males across the age range by 30%-50%, for females this fall has been most marked amongst the young; for older workers the fall in coverage has been less than 10%. As a significant part of the growth in the female labour force has been due to increasing numbers of young women entering the labour force, then the fall in the mean coverage of figure 10.6 would seem to be the result of demographic factors rather than an across-the-board move away from unions. However, Robinson (1994) notes that the rise in female employment has occurred in the service industries where the membership of unions and the associated returns are typically smaller. This implies that union coverage has fallen because young females are increasingly participating in non-union occupations. Without information on coverage levels and age structures within industries we cannot identify which of these two hypotheses is the cause and which the effect. Thus these results are conditioned on particular markups in occupations and industries, whose variation over time may or may not be the result of the same demographic factors.

These results are also conditioned by the sector to which the individual belongs, and it is worth noting the contrast between the numbers of females covered by collective agreement in the public, as opposed to the private sector. In the government sectors, the proportion covered never falls below 80%[7]. In contrast, and recalling that the absence of information on local

[7] For both the public sector as a whole and for government and public corporations separately; see Bell and Ritchie (1993a).

agreements will lead to a downward bias on coverage statistics, particularly for the private sector, the proportion of females covered in the private sector had fallen below 10% by 1990.

It may be noted that the cross-section estimates for females exceed the fixed-effects estimates. Thus allowing for individual differences depresses the union effect for females, in contrast to males (see section 9.2.6). The implication is that unions have an *equalising* effect on the wages of males of differing 'ability' (or whatever the unobserved heterogeneity represents) and a *discriminatory* effect on female earnings.

These estimates are much smaller then other studies which have mainly used cross-sectional data. For example, Nickell (1977) estimates a "union effect" of 20% using aggregated data; Yaron (1990) a 10% gap for manual workers in the 1983 General Household Survey; and Main and Reilly (1992) using the 1986 SCELI dataset find union gaps of around 15%. Booth (1995) reports that FE studies are usually smaller than cross-sectional results. Jakubson (1991) finds that moving from a cross-section to a fixed-effects specification can have a large (15%) effect on union coefficients, but cross-section studies on this data indicate the union effect to be a fairly constant 4% higher for most years; thus the cross-section results are also lower than comparative estimates.

As for males, one possibility for the difference is that the equation used here has more detailed industry and occupational variables, both of which may be correlated with the level of union coverage. Further, Andrews and Bell (1995), using male data from the NES, found that the inclusion of local bargains increased cross-sectional estimates by around 8%. If this holds true for females, then these results are comparable with the other studies. One final reason for the difference may be that the TVFE model makes no explicit correction for the selection mechanism which Main and Reilly (1992) and Lanot and Walker (1993a), for example, found to be significant.

One important issue raised by figure 10.6 is that the estimated coefficient varies greatly over the period (all coefficients are significant at the 5% level), suggesting that the period of measurement is important to the interpretation (see also Meghir and Whitehouse (1987), Lanot and Walker (1993b)). For example, the TVFE models estimates that the union markup in 1982 is over three times that in 1979; the latter figure indicates that the markup is negligible, while the former indicates that it is at least as important an influence as sector.

10.2.6 Wages Council coverage

The coefficients for Wages Councils (WCs) in figure 10.7 suggest that those who are covered by such arrangements should expect, ceteris paribus, to receive substantially lower wage rates. As for males, this is likely to indicate that those whose wages are determined by such arrangements have a very weak bargaining position in the labour market which is not fully offset by the effects of the WCs, rather than suggesting that these bodies directly reduce workers wages. Both sexes experience much the same effect, and the increasing disparity between those affected by WC agreements and other workers is consistent with the increasingly dispersion of wage rates over this period (see Bell (1995)).

10.2.7 Sector

Figure 10.8 indicates how employees of both sexes in the private sector fared relative to the public sector during the period, ceteris paribus. There is a general upward trend in the coefficients for both males and females, and this evidence certainly supports the view that private sector workers improved their hourly wage rates relative to public sector workers during the 1980s. Nevertheless, even in 1990, most public sector workers' hourly wage rates were above those of otherwise similar workers in the private sector. Bell and Ritchie (1993a), using weekly wages in cross-section, note that this public sector premium largely arises amongst those working in government; for females in public corporations the premium is

small and generally negative. As for males, the lack of "governmental" jobs in the private sector may mean that this premium is a misspecified occupational characteristic rather than a pure response to employment in the public sector; in addition, there may be some form of non-wage compensation in the private sector not picked up by the variables used here.

When the model is re-estimated using weekly earnings, higher levels of overtime, piecework and bonus payments in the private sector raise the private sector weekly wage for males by relatively more (although this is not usually enough to eliminate the public sector premium). For females, there is relatively little change. It is noticeable that there is much greater variation between fixed-effects and cross-sectional estimates when weekly wages are used, suggesting that individual characteristics have a significant influence when deciding the amount of "effort" to put in.

Until the mid-1980s, the premium for employment in the public sector was larger for females but by the end of the period was much the same for both sexes. At least three interpretations can be put on this finding. First, higher wage rates in the public sector may reflect lower levels of discrimination against females than in the private sector. The decline in the public sector premium could then be due to the delayed impact of the equal pay legislation (in that, for example, the public service bodies tend to implement social directives before the private sector).

Second, the impact of coverage by collective agreement may vary by sector. At low levels of coverage, bargaining power is likely to be low and the consequent returns relatively small. The decline in the public sector premium may reflect the privatisation of large, highly-unionised public corporations (with, in some cases, private sector counterparts) throughout the 1980s. In the absence of interactive dummies, the sector coefficients may partly reflect differences in the effectiveness of collective bargaining.

Third, Sloane and Theodossiou (1994) argue that the relative increase in women's wages throughout the period was largely due to demand pressure. In this view, the fall in the public sector premium is due to the rapid growth in private sector service industries providing (according to Robinson(1994)) well-paid jobs.

10.2.7 Length of time in the job

Figure 10.9 calibrates the effect on wage rates of having spent less than 12 months in the present job. Until 1983, the disadvantage of a short tenure was broadly equivalent for both sexes but since then there has been a marked divergence. By 1990, the impact on female earnings of not having been in a specific job for twelve months is broadly neutral, while for males the discount persists. This perhaps suggests that females are increasingly participating in jobs where skill acquisition on the job is not particularly important; it may be that for females, human capital and experience is less closely tied to years of service. However, the coefficients are small (discounts of around 2% on the hourly wage) and this variable does not distinguish between acquiring and changing jobs, so only limited inferences can be drawn.

10.2.8 Attrition variables

As for males, AVs are included to take account of absence from the panel. It may be expected that females, with their more variable participation rates and longer absences from the labour market, would experience different effects on wage rates from males (see, for example, Joshi(1986, pp225-227); Main(1989); Mincer and Ofek(1982)). It could also be argued that the women face a greater variety of work options than men (in that full-time work, part-time work, and home working all form a large part of the typical female employment history), and so the likelihood of selection bias is higher than for males (Ermisch and Wright (1993)).

However, the coefficients shown in figure 10.10 indicate little difference from those for males given in figure 9.10. The coefficient on YrsIn indicates that both men and women gained a premium for additional years in the panel. Clearly, this result is consistent with the view that this variable is a proxy for labour market experience, albeit rather an imperfect one. The declining coefficient reflects the increasing mean of the variable.

The coefficient on CurrStay is very small and rarely significant; as YrsIn is always large and significant, this seems to contradict Main (1989) who found that immediate employment history was a more important component of earnings than general employment experience. However, this finding must be treated with some suspicion. Firstly, observation in the panel cannot be directly related to employment, and CurrStay is more affected by errors in the construction of the panel than YrsIn. Secondly, these two variables are affected by the TVFE collinearity problem (section 9.2.5), which may be indicated by the fact the CurrStay is often insignificant. Cross-section estimates give higher values for CurrStay and lower ones for YrsIn, although the latter still dominates. Thirdly, part-time and full-time observations were used for the AVs, and Main(1989) argues that it is the *type* of experience that matters (but see Main(1988b)).

10.2.9 Instrumental variable estimation

This section discusses an instrumental variable (IV) approach to the endogeneity issue. It was noted in the previous chapter that a case can be made for the endogeneity of any and all regressors, but here the focus is on the occupational groupings. The reason is twofold; firstly, the 'crowding' of females into certain occupations implies that the dummy coefficients do not necessarily represent the effect of being randomly assigned to an occupational grouping

(P.W. Miller (1987))[8]. For example, C.F. Miller (1993) argues that to a significant degree occupational choice depends on initial career decisions and life cycle patterns of labour market participation, which may be influenced by earnings.

Secondly, an occupation can embody a number of unmeasured characteristics which influences the choice of job: flexibility on working hours, compensating differentials, social influences, and so on. Helwege (1992) notes that there is a strong correlation between occupation and industry; however, occupation (rather than industry) appears more likely to embody the particular characteristics which will lead to the selection of a position.

Following Bowden and Turkington (1984, Ch. 2), an admissible instrument to counter the endogeneity in occupational choice should be an exogenously determined probability of observation[9]. Assuming a vector of exogenous variables is available, one possibility is to use discriminant analysis to generate the probability that individual i will select occupation j. The probabilities are normalised to sum to one. A simpler alternative, the approach taken, is to use the actual proportion of women in each occupation as a probability instrument. The proportion of women in each occupation is relatively stable over time, and so the proportions over the whole sixteen years of the NES was used.

Use of these instruments makes almost no difference to the estimated coefficients or standard errors; even the coefficients on occupation are typically affected by less then 5%. A Wu-Hausman test failed to reject the null hypothesis of no endogeneity. One possible explanation is that the instruments are not statistically independent of the residual. As the Wu-Hausman test actually compares OLS and IV specifications rather than testing for endogeneity directly,

---

[8] Sloane and Theodossiou (1994, note 8) report that in 1982 4% of females work in 'female-only' occupations, whereas 22% of males work in 'male-only' posts. This result is reflected in the necessity to drop some occupations from the female regressions due to lack of observations.

[9] An alternative is to calculate binary variables, setting $p_1^j=1$ for the most likely occupational choice and $p_1^j=0$ for all other j. As the occupation data is already categorical, the probabilities must be used as instruments.

the test is not appropriate where the instruments are still correlated with the error term. Given the nature of the instruments this seems unlikely. However, because it is only a test for the *relative* validity of a specification, the test is biased towards the acceptance of the null as two inefficient estimators will drive the test statistic (6.27) towards zero.

It may be that individual heterogeneity manifests itself in a non-random choice of occupation; for example, it has been argued that occupational choice is significantly affected by educational and social choices made before entry into the labour market (see Dolton and Kidd(1994); Polacheck (1981); Vella (1994); and Elliott (1991, pp404-409) for a more general discussion). However, cross-sectional results, which cannot pick up individual heterogeneity, are also unaffected by the use of instruments.

These results seem to imply that the categorical variables for occupation are strictly exogenous; that is, the "crowding" of female occupations is due to factors not reflected in the Mincerian wage equation. For example, the significant influence of gender "attitudes" in occupational choice claimed by Vella (1994) may persist throughout an individual's working life. This still leaves open the question of whether wage variations in occupations are due to crowding or the differing characteristics of workers in different posts.

A number of authors have argued that inter-occupational differentials are less important in explaining wage differentials than intra-occupational differences[10]. This may explain why the categorical dummies used to determine occupational intercepts appear to be exogenous. This could be tested by increasing the number of occupational groupings from eighteen (there are over four hundred in the NES); however, the results presented here are already much more detailed than the other studies, and the coefficients are highly significant. A further refinement of the categories is unlikely to change the result markedly.

---

10 For example, Dolton and Kidd(1994) use UK data; Harland and Sahellarion(1993) for Canada; Lucifora (1993) for Italy; Reilly (1991) on Irish data.

10.3 Male-female differentials

In this section the method of Blinder (1973) and Oaxaca(1973) is used to study the differences between male and female earnings. Differences in the male-female hourly wage-rate can be decomposed into that which can be explained by systematic differences in identifiable characteristics and that which appears to result from differences in returns to the same characteristics. Specifically,

$$\overline{\ln(w_{mt})} - \overline{\ln(w_{ft})} = \bar{x}_{mt}\hat{\beta}_{mt} - \bar{x}_{ft}\hat{\beta}_{ft} = (\bar{x}_{mt} - \bar{x}_{ft})\hat{\beta}_{mt} + \bar{x}_{ft}(\hat{\beta}_{mt} - \hat{\beta}_{ft}) \qquad (10.2)$$

where w_{ft}, w_{mt} are wages of females and males at time period t, respectively, $\mathbf{x_{ft}}$, $\mathbf{x_{mt}}$ are the mean values of the regressors for males and females respectively, and $\boldsymbol{\beta}_{\mathbf{ft}}$ and $\boldsymbol{\beta}_{\mathbf{mt}}$ are values of the female and male regression coefficients at time t respectively. The first term on the right-hand side of (10.1) is the contribution to the difference in average male and female characteristics of the mean wage differential, while the second term provides a measure of the difference in returns between the sexes[11]. Figure 10.12 plots these components. It shows the total earnings gap, the "explained" component (due to differences in characteristics) and the "unexplained" component (due to differences in returns).

The raw wage gap peaked at around 27% of mean male wage in 1979, but had fallen to 21% by 1990. However, while the "explained" component also fell, the "unexplained" component rose until it exceeded the total gap. Note that the estimates of the "unexplained" component are comparable with those of Wright and Ermisch (1991) from the WES, although they omit industry and occupation from the set of explanatory variables.

The implication of Figure 10.12 is that, after allowing for differences in the characteristics of employees (including individual heterogeneity as these are fixed-effect estimates) the

---

[11] The Oaxaca decomposition suffers from the usual index number problem; Oaxaca and Ransom(1994) offer alternative formulations. However, in the present example, the differences between the various methods are minimal and so are not considered here.

expected earnings for females are higher than for males in 1990; that they are actually lower is entirely due to the lower value placed upon those characteristics in the labour market. However, it can be shown (Bell and Ritchie (1995c)) that this negative differential is largely due to the age coefficients, and so this result should be taken with some caution. On the other hand, the other explained components (apart from the AVs) are also declining over time, which would seem to indicate a general narrowing of the explained differentials between males and females.

Bell and Ritchie (1995c) discusses the TVFE and TVCS decompositions in more detail. Around two-thirds of the unexplained component is made up by the residual difference in the constant term. Of the rest, region has been a steady but significant contributor, but the major change has been in the large unexplained difference due to the different returns to industry. The industrial differential varies greatly, but appears to move cyclically with male employment prospects; from the early 1980s it contributes positively to the unexplained differential. It is interesting to note that cross-section studies (the TVCS model) allocate a significant part of the unexplained component to occupation and agreement, and almost none to industry; moreover, the explained component due to occupation has moved significantly in women's favour.

There are two important caveats when considering the decomposition results. First, it should be noted that omitted regressors (such as a direct valuation of experience or some measure of compensating differentials) may result in some of the differential being wrongly allocated to the unexplained, rather than the explained category. Second, the Oaxaca decomposition assumes that both sexes have the same responses to the variables used, but this is not necessarily the case, especially for "human" variables. For example, a dummy for 'married' may lead to a fundamentally different option set for males and females, and this appears to be reflected in the positive coefficients for males and negative ones for females typically found in empirical work. As most of the variables relate to job characteristics, it seems reasonable

to assume that the response to measured variables is the same for both sexes in the absence of any unmeasured effects.

The Oaxaca decomposition has also been criticised for its reliance on mean differences rather than distributive effects (see Dolton and Makepeace (1985); Munroe (1988); and Jenkins (1994)). Blau and Kahn (1992) and Juhn Murphy and Pierce (1993) both allow for differences in earnings variance (although, curiously, they both assume identical coefficients for both sexes). A number of papers have recently appeared using variations on the Lorenz curve methodology; for example, Jenkins(1994); and Sloane and Theodossiou (1994). While all these are valid alternatives to the means-based decomposition, a distributional breakdown of gender differences is a long paper in itself and is therefore left out of this work.

10.4 Regression *sans* occupation, *sans* industry

Wright and Ermisch (1991), and to some extent Blinder (1973), argued that occupational variables should be omitted from the regression as the occupational crowding should be interpreted as an element of discrimination in wage differentials, not an explanatory factor. For the reasons outlined in section 10.1, occupational dummies have been included in this regression, but here the results of leaving out these variables are discussed.

If occupation is omitted as a regressor, then industry also should be omitted, due to the close relationship between the two. The TVFE model was then run on the remaining variables in table 10.1.

The omission of industry and occupation has little effect on the age, tenure, or attrition variables. It does reduce the public sector premium, but this premium now appears to be increasing for women, from zero in 1977 to around 7% by 1990. It also has a significant effect on the regional coefficients, increasing them by around 20%-30%, which is to be

expected given the regional differences in industrial structure.

The most notable difference is on the agreement coefficients. For males and females, omitting industrial and occupation variables increases the "union effect" by around 2% and 1% respectively in the late 1970s. However, this effect falls, and although the peak in 1982 is similar in both regressions, the decline in the union effects is larger and faster when no industrial/occupational dummies are used. This result may be due to the concentration of unions in declining industries, particularly manufacturing workers. The end result is that, by 1990, there appears to be a *discount* of 1%-2% for being covered by a collective agreement.

The effect on the Oaxaca decomposition is that the explained component of the differential is almost halved (see Bell and Ritchie (1995c)). The main increase in the unexplained component is found in the residual, due to differences in the constant terms, although there is some increase in the regional and sectoral components. Overall, these results would seem to support the view that industry and occupation are both significant explanatory variables and exogenous in the models estimated.

10.5 Conclusion

In this chapter the fixed-effects estimator has been applied to the female data in the NES, and the results compared with those of males obtained earlier. The impact of coverage by collective agreement on hourly wage rates for full-time females fell during the 1980s as did the premium for belonging to the public sector. Age coefficients suggested a much flatter earnings profile for females and that earnings were likely to rise more sharply for young males than for females. The TVFEIV estimator failed to show evidence of endogeneity in occupational choice, but the basic model suggested that occupational differences in wages are concentrated in the manual sector. There is some evidence that females have been moving

up the occupational ladder.

A decomposition of the male-female differential suggested that, while the overall wage difference has fallen, this has been largely due to changes in the characteristics of females in the labour market - especially the younger age profile of women. The unexplained differential, which includes unmeasured effects and individual heterogeneity as well as 'discrimination', has been rising steadily since the late 1970s.

Finally, regression without occupation and industry dummies seemed to support the TVFEIV findings that the variables used here are exogenous. There is some effect on the regional and agreement coefficients, but the decomposition of the differential suggests that these variables are absorbed into the differences in the means.

Table 10.1 Time-varying fixed-effect regression results (females, part: 1984)

Variable	Code	Mean	Coefficient	Error	T-value	T-prob
Constant		1.000	-0.2096	0.033	-6.354	0.000
InLast		0.813	0.0101	0.003	3.303	0.001
YrsIn		6.151	0.0304	0.001	35.341	0.000
CurrStay		4.710	-0.0022	0.001	-4.250	0.000
reg	45	0.165	-0.1269	0.004	-36.324	0.000
reg	48	0.030	-0.1752	0.007	-25.638	0.000
reg	55	0.072	-0.1845	0.005	-36.513	0.000
reg	60	0.091	-0.1777	0.005	-35.063	0.000
reg	66	0.064	-0.1793	0.005	-33.430	0.000
reg	70	0.080	-0.1844	0.005	-35.712	0.000
reg	74	0.118	-0.1666	0.005	-35.151	0.000
reg	79	0.054	-0.1920	0.006	-30.681	0.000
reg	88	0.041	-0.1855	0.007	-27.118	0.000
reg	98	0.107	-0.1731	0.005	-31.831	0.000
agt	998	0.502	0.0371	0.003	14.032	0.000
wbc	248	0.109	-0.0244	0.004	-6.315	0.000
j12	2	0.168	-0.0210	0.003	-7.808	0.000
sector	0	0.574	-0.0205	0.004	-4.636	0.000
div	1	0.018	0.1462	0.022	6.659	0.000
div	2	0.030	0.0770	0.021	3.606	0.000
div	3	0.083	0.0577	0.021	2.761	0.006
div	4	0.115	0.0521	0.021	2.497	0.013
div	5	0.011	0.0158	0.022	0.705	0.481
div	6	0.153	0.0161	0.021	0.775	0.438
div	7	0.045	0.0892	0.021	4.203	0.000
div	8	0.140	0.0664	0.021	3.184	0.002
div	9	0.402	0.0248	0.021	1.188	0.235
age	16	0.008	-0.1751	0.015	-11.986	0.000
age	18	0.054	-0.0675	0.009	-7.498	0.000
age	20	0.091	0.0032	0.008	0.426	0.670
age	22	0.094	-0.0023	0.007	-0.352	0.725
age	24	0.082	-0.0120	0.006	-1.977	0.048
age	26	0.067	-0.0218	0.006	-3.862	0.000
age	30	0.090	-0.0071	0.005	-1.499	0.134
age	40	0.094	-0.0113	0.005	-2.477	0.013
age	45	0.091	-0.0010	0.005	-0.187	0.852
age	50	0.092	0.0030	0.006	0.516	0.606

age	55	0.085	-0.0043	0.007	-0.662	0.508
age	60	0.055	-0.0076	0.008	-1.014	0.311
age	120	0.013	-0.0180	0.011	-1.686	0.092
kos	122	0.032	0.0966	0.005	18.514	0.000
kos	147	0.178	0.0718	0.004	19.792	0.000
kos	156	0.007	0.1014	0.012	8.681	0.000
kos	189	0.015	0.0949	0.008	11.969	0.000
kos	211	0.023	0.1015	0.006	16.300	0.000
kos	246	0.060	0.0088	0.005	1.817	0.069
kos	254	0.004	0.2853	0.015	19.141	0.000
kos	281	0.090	-0.0507	0.004	-12.643	0.000
kos	295	0.002	-0.0085	0.024	-0.348	0.728
kos	327	0.015	0.0028	0.008	0.336	0.737
kos	385	0.043	-0.0108	0.006	-1.853	0.064
kos	462	0.015	0.0251	0.008	3.031	0.002
kos	477	0.054	0.0271	0.005	5.423	0.000
kos	533	0.008	-0.0068	0.011	-0.648	0.517
kos	540	0.003	0.0171	0.019	0.911	0.362

Chapter 11

Conclusion

The basis for this thesis has been the desire to make efficient use of very large datasets in general and one in particular, the NESPD. For most of its life, the importance of the NES to UK economists was largely due to use of the published NES rather than analyses of the underlying data. Since the micro-data in the NESPD was made available to researchers at the end of the 1980s, the amount of analysis carried out on the data has grown considerably, although this is still relatively small compared to the potential of such a large and comprehensive dataset and utilisation of other surveys like the FES, GHS and WIRS.

This thesis has described a way in which effective use can be made of the data. The core of the work has been to show how linear estimation methods can be made fast, practical, and free from the constraints of having to work at arm's length with data only available at the DE. The validity of this approach is indicated by the fact that, in recent years, a score of papers using OLS regressions on the NESPD have been produced at the Universities of Stirling and Manchester.

The value of cross-sectional studies on the NES is high. In terms of numbers surveyed and the range of work variables defined, it far exceeds any other survey of the UK labour force. It also allows for estimation over a long period (sixteen years worth of data was available for this thesis, and a further five years is due shortly) and the construction of time-series of estimated coefficients. This is a feature of the data which has seen little interest outside the confines of the Universities involved in this project.

While the ability to perform time-varying cross-sectional analyses quickly and efficiently is a significant step forward, linear regressions have been run by other researchers. With respect to cross-sectional analysis, the NES is similar to many other surveys except in its size. However, the ability to construct fixed-effects estimators which allow for individual

heterogeneity is almost unique amongst the large UK surveys. The allowance for individual heterogeneity goes some way to making up for the lack of personal and educational variables in the NES; indeed, the fixed-effects assumption may be a better way to treat education than to employ categorical variables which lump different types of education together. Moreover, the fixed-effects assumption also takes into account characteristics which are constant over time but essentially unmeasurable, such as "motivation", "preference", "ability", "attitude", and so on. This is only feasible on a panel dataset, and the NESPD is the only major UK survey constructed as a panel with several years worth of data[1]. The importance of allowing for individual heterogeneity was shown in chapter nine, where the results from the TVCS and TVFE models were compared. Although the qualitative results were similar, the scale of the estimates produced by the two models can differ dramatically. Moreover, even if the individual heterogeneity is not significant, or is not constant over time (and so cannot be removed by the panel transformation), the use of the pooled dataset rather than separate cross-sections makes for inherently more efficient estimators.

Chapter two described the theoretical advantages of using panel models, and the practical aspects of panel estimation. The basic literature has been well established for some years now, the significant work being Hsiao's (1986) monograph. However, one potential advantage of using panel models has been almost completely overlooked: the ability to let coefficients vary over time and so to create time-series estimates of coefficients. The usefulness of time-varying coefficients is apparent from chapter ten, where a time-series of the decomposition of wage differentials was created. Bell and Ritchie (1995c) have used this to show how the various components of the differential have changed over time. This gives a dynamic aspect to the analysis of male-female wage differentials.

The models presented in this thesis are unusual in that the basic specification of all models (save to some extent the unbalanced differenced model) includes the assumption that

[1] The British Household Panel Survey is a relatively new dataset constructed as a panel and containing a range of information on individuals, not just work charactersitics. The third wave of the data is now available.

coefficients are not constant over time. A brief consideration of the UK labour market should suggest that this is a sensible basis for modelling; and from a theoretical position the general-to-specific school of econometric methodology would argue that the "restricted" models of this thesis (the standard "fixed-effects" models of most applied work) should be tested as special cases of a more general hypothesis. The validity of the restricted models was tested to some extent in chapter nine, and it was rejected fairly comprehensively. This is not a very surprising result but has been overlooked by a large number of researchers.

Chapter five described how cross-section, fixed-effects, and differencing estimators could be created, and the mathematics underlying the efficient analytical routines. This is the crux of the practical aspect of the thesis. Although some other programs (such as STATA) can run regressions on cross-product matrices, the intention to use the cross-product matrix as the basic data type allows both extraction and analysis programs to be tailored to produce particular results - the fixed effects estimator being one such outcome. The basic assumption of coefficients varying over time will also affect the calculation of cross-product matrices, and it was shown that the construction and manipulation of matrices to achieve this requires some forethought.

The end result of this has been the first (indeed, only - so far) estimates to make use of the NESPD's distinguishing feature - its panel nature. Chapter nine showed that the effect of using panel estimators may be considerable, and suggests a significant area for further study. One such possibility is the implementation of minimum-distance estimators to allow for flexible definitions of the error term - another estimator with theoretical advantages which is rare in applied work but whose practical problems are by no means insurmountable in the cross-product framework.

Chapters nine and ten produce a number of results on male and female earnings, largely using the TVFE estimator. Whilst most of the results are in broad agreement with other applied work, there are some surprises; most notably the very small coefficients on collective

agreement, and their path over time. Several of the coefficients illustrate the importance of constructing time-series of the coefficients, where possible: estimates of the effects of public sector working, unions, and region, for example, are very dependent on when the estimates are made.

Finally, one other aspect of collecting and using data efficiently has been described, although it has had a marginal effect on this thesis. This is the observation history, an extremely compact way of storing information about states. Currently this is the subject of some interest in the construction of multi-destination transition matrices, but in this thesis it has been used largely to provide an alternative view of the labour market by comparison with the more usual employment cohort. The descriptive analysis of the labour market presented in chapter eight generally ties in with other studies of the labour market. However, the employment and data cohorts illustrated that the dynamic effects are both important in explaining the characteristics of the labour market and the differences between the sexes.

In summary, the final contribution of this thesis has been twofold. Firstly, to provide some efficient techniques for retrieving and analysing data from one of the largest datasets of its kind in the world; secondly, to provide the first analyses on this dataset allowing for individual heterogeneity and a flexible parameter structure. And as a side effect, it is to hoped that the benefits of allowing for time-varying coefficients when estimating on repeated micro-data have been demonstrated convincingly.

Bibliography

Adams, M and J Owen (1989) "The New Earnings Survey panel dataset", Employment Market Research Unit Working Paper No. 1, Department of Employment

Andrews, M and DNF Bell (1995) "Union coverage differentials: some panel estimates for the UK using the NES", University of Manchester, mimeo

Andrews, M, DNF Bell, and F Ritchie (1993) "The New Earnings Survey and the structure of male earnings in Great Britain 1975-1990: a varying-coefficents fixed-effects model", University of Manchester, mimeo

Arellano, M and S Bond (1991) "Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations" Review of Economic Studies 58:277-297

Ashenfelter, OC and R Layard (1986) *Handbook of Labor Economics*, Amsterdam: North-Holland

Atkinson, AB, J Micklewright and NH Stern (1981) "A comparison of the FES and NES 1971-1977: Part I: Characteristics of the sample", Social Science Research Council Programme on Taxation, Incentives and the Distribution of Income Working Paper No 27

Atkinson, AB, J Micklewright and NH Stern (1982) "A comparison of the FES and NES 1971-1977: Part II: Hours and earnings", Social Science Research Council Programme on Taxation, Incentives and the Distribution of Income Working Paper No 32

Barth, E, R Naylor, and O Raaum (1995) "Union wage effects: does membership matter?", University of Oslo Department of Economics Memorandum No.9

Becker, E and CM Lindsay (1994) "Sex differences in tenure profiles: effects of firm-specific shared investment", *Journal of Labour Economics*, v12:1 pp98-118

Bell, DNF (1995) "Earnings inequality in Great Britain: some new evidence", *Scottish Journal of Political Economy* v42

Bell, DNF and RA Hart (1995) "Wage rates, working time and collective agreements" University of Stirling discussion paper 95/2

Bell, DNF, R Rimmer and S Rimmer (1994) "Earnings inequality in Great Britain 1975-1990: the role of age", *Review of Income and Wealth*, v40 pp166

Bell, DNF and F Ritchie (1993a) "Female Earnings in Great Britain 1975-1990: some evidence from the New Earnings Survey" University of Stirling, mimeo

Bell, DNF and F Ritchie (1993b) "Transitions, non-response and earnings in the NESPD", University of Stirling, mimeo

Bell, DNF and F Ritchie (1994) "Transitions, non-response and earnings in the NESPD", University of Stirling, mimeo

Bell, DNF and F Ritchie (1995b) "Female earnings in Great Britain 1977-1990: some evidence from the New Earnings Survey" University of Stirling Discussion Paper 95/6

Bell, DNF and F Ritchie (1995c) "Decomposing male-female earnings differentials in Great Britain 1977-1990", University of Stirling, mimeo

Bell, DNF and F Ritchie (1996) "Identification of incremental effects in varying-coefficient estimators" University of Stirling, mimeo

Berndt, ER (1991) *The Practice of Econometrics, Classic and Contemporary*, Reading, Mass: Addison-Wesley

Biorn, E (1992) "Panel data with measurement errors", in Matyas and Sevestre (1992), pp152-195

Blackaby, DH, PD Murphy and PJ Sloane (1991) "Union membership, collective bargaining coverage and the trade union mark-up for Britain", *Economics Letters*, v36 pp205-208

Blackburn, ML (1990) "What can explain the increase in earnings inequality among males?", *Industrial Relations*, 29:441-455

Blau, FD and LM Kahn (1992) "The gender earnings gap: learning from international comparisons", *American Economic Review*, v82:2 pp:23-28

Blinder, A (1973) "Wage discrimination: reduced form and structural estimates", *Journal of Human Resources*, v8:4 pp436-455

Blundell, R and I Walker (1986) *Unemployment, Search and Labour Supply*, Cambridge: Cambridge University Press

Booth, A (1995) *The Economics of the Trade Union*, Cambridge: Cambridge University Press

Bound, J, C Brown, GJ Duncan and WL Rodgers (1994) "Evidence on the validity of cross-section and longitudinal labor market data", *Journal of Labor Economics*, v12:3 pp345-368

Bowden, RJ and DA Turkington (1984) *Instrumental variables*, Cambridge: CUP

Bowles and Gintis (1975) "The problem with human capital theory - a Marxian critique", *American Economic Review*, v65:2 pp74-82

Brown, CV, EJ Levin, PJ Rosa, RJ Ruffell, and DT Ulph (1986) "Payment systems, demand constraints and their implications for research into labour supply" in Blundell and Walker (1986) pp190-214

Brown, R, M Moon and BS Zoloth (1980) "Incorporating occupational attainment in studies of male-female earnings differentials" *Journal of Human Resources*, vol 15, no 1, pp 3-28

Card, DI (1994) "The effect of unions on wages: redistribution or relabelling?", NBER Working Paper Series no 4195

Chamberlain, G (1984) "Panel Data" in Z Griliches and MD Intriligator (eds) *Handbook of Econometrics*, Amsterdam: North-Holland pp1247-1318

Chamberlain, G (1985) "Heterogeneity, omitted variable bias, and duration dependence" in JJ Heckman and B Singer (eds) *Longitudinal Analysis of Labor Darket Data*,

Cambridge: Cambridge University Press pp3-38

Chiplin, B and PJ Sloane (1988) "The effects of Britain's anti-discrimination legislation on relative pay and employment: a comment", *The Economic Journal*, v98 pp833-838

Coleman, JS (1994) *Earnings-Tenure Profiles in the UK Public and Private sectors*, unpublished PhD thesis, University of Stirling

Cripps, T and R Tarling (1974) "An analysis of the duration of male unemployment in Great Britain, 1932-1973", *Journal of the Royal Statistical Society series B*, v24 pp406-24

Dex, S and E Puttick (1988) "Parental employment and family formation" in Hunt (1988), pp123-149

Diechartz, J (1988) "Changing and random coefficient models: a survey" in P. Hackl (ed.) *Statistical Analysis and Forecasting of Structural Change*, Springer-Verlag

Dolton, PJ and MP Kidd (1994) "Ocucpational access and wage discrimination", *Oxford Bulletin of Economics and Statistics*, v pp457-474

Dolton, PJ and GH Makepeace (1985) "The statistical measurement of discrimination", *Economic Journal*, v95 pp391-395

Dolton, PJ and GH Makepeace (1987) "Marital status, child-rearing and earnings differentials in the graduate labour market", *Economic Journal*, v97 pp897-922

Dolton, PJ and KG Mavromaras (1994) "Intergenerational occupational choice comparisons:

the case of teachers in the UK", *Economic Journal*, v104:2 pp841-863

Ehrenberg, RG and JL Schwarz (1986) "Public sector labor markets" in Ashenfelter and Layard (1986), vII pp1219-1268

Elias, P (1988) "Family formation, occupational mobility and part-time work", in A Hunt (1988), pp83-104

Elias, P and BG Main(1982) "Women"s working lives: evidence from the National Training Survey", University of Warwick Institute for Employment Research report

Elliott, RF (1991) *Labour Economics*, Maidenhead: McGraw-Hill

Ermisch, JF and RE Wright (1993) "Wage offers and full-time and part-time employment by British women", *Journal of Human Resources*, v28:1 pp111-133

Farber, HS (1986) "The analysis of union behaviour" in Ashenfelter and Layard (1986), vII pp1039-1089

Freeman, RB (1984) "Longtidunal analysis of the effectof trade unions", *Junral of Labor Economics* v2:1 pp1-26

Gourieroux, CA, E Montfort, E Renault and A Trognon (1987) "Generalized Residuals", *Journal of Econometrics*, vol 34, pp 5-32

Greenhalgh, C (1980) "Male-female wage differentials in Great Britain: is marriage an equal opportunity?", *Economic Journal*, v90:4 pp751-775

Groshen, EL (1991) "The structure of the male-female wage differential: is it who you are, what you do, or where you work?", *Journal of Human Resources*, v26:3 pp 457-472

Hall, A (1993) "Some aspects of generalized methods of moments estimation" in GS Maddala, CR Rao and HD Vinod (eds) *Handbook of Statistics*, Amsterdam: Elsevier Science Publishers vII pp393-417

Ham, J (1986) "On the interpretation of unemployment in empirical labour supply analysis" in Blundell and Walker (1986) pp121-142

Harland, J and C Sahellariou (1993) "Wage discrimination, occupational segregation, and visible minorities in Canada", *Applied Economics*, v25 pp1409-

Hartog, J and H Oosterbeek (1993) "Public and private sector wages in the Netherlands", *European Economic Review* v37 pp97-114

Hausman, JA (1978) "Specification tests in econometrics", *Econometrica* v46 pp1251-71

Hausman, JA and Wise, DA (1979) "Attrition bias in experimental and panel data: the Gary Income Maintenance Experiment", *Econometrica* v47 pp455-473

Heckman, JJ (1979) "Sample selection bias as a specification error", *Econometrica* v47 pp153-161

Heckman, JJ (1981a) "Statistical models for discrete panel data" in CF Manski and D McFadden *Structural Analysis of Discrete Data with Economic Applications*, Cambridge, Mass.: MIT Press

Heckman, JJ (1981b) "The incidental parameters problem and the problem of initial conditions in estimating a discrete-time discrete-data stochastic process" in CF Manski and D McFadden *Structural Analysis of Discrete Data with Economic Applications*, Cambridge, Mass.: MIT Press

Helwege, J (1992) "Sectoral shifts and interindustry wage differentials", *Journal of Labor Economics*, v10:1 pp55-84

Holmlund, B and H Ohlsson (1992) "Wage linkages between public and private sectors in Sweden", *Labour*, v6:2 pp3-17

Hsiao, C (1986) *The Analysis of Panel Data*, Cambridge: CUP

Hsiao, C (1992) "Random coefficients models" in Matyas and Sevestre (1992) pp72-94

Hunt, A (1988) *Women and Paid Work: Issues of Equality* (ed) Basingstoke: Macmillan

Jakubson, G (1991) "Estimation and testing of the union wage effect using panel data" *Review of Economic Studies* v58 p971-991

Jenkins, SP (1994) "Earnings discrimination measurement: a distributional approach", *Journal of Econometrics*, v61 pp81:102

Johnes and Taylor (1989) "Labour" in MJ Artis (ed.) *The UK Economy* London: Wiedenfield and Nicolson pp288-350

Johnston, J (1984) *Econometric Methods*, Maidenhead: McGraw-Hill

Joshi, HE (1986) "Participation in paid work: evidence from the Women and Employment Survey" in Blundell and Walker (1986) pp217-242

Joshi, HE and M Newell (1985) "Parenthood and pay differentials: evidence from the MRC National Survey of the 1946 birth cohort", Centre for Population Studies, University of London

Juhn, C, KM Murphy and B Pierce (1993) "Wage inequality and the rise in returns to skills", *Journal of Political Economy*, v101:3 pp410-42

Killingsworth, M (1983) *Labor supply*, Cambridge: Cambridge University Press

Killingsworth, M (1986) "A simple structural model of heterogeneous preferences and compensating wage differentials", in Blundell and Walker (1986) pp303-317

Krueger, AB and LH Summers (1988) "Efficiency wages and the inter-industry wage structure", *Econometrica*, v56:2 pp259-293

Lambert, SF (1993) "Labour market experience in female wage equations: does the experience measure matter?", *Applied Economics* v25 pp1439:1449

Lancaster, T (1990) *The Econometric Analysis of Transition Data*, Cambridge: CUP

Lanot, G and I Walker (1993a) "Alternative estimators of the union/non-union wage differential: UK pooled cross-section evidence", University of Keele, mimeo

Lanot, G and I Walker (1993b) "The union/non-union wage differential: an application of semi-parametric methods", University of Keele, mimeo

Layard, R and S Nickell (1987) "The labour market" in R Dornbusch and R Layard (eds) *the Performance of the British Economy*, Oxford: Oxford University Press

Lewis, HG (1986) "Union relative wage effects" in Ashenfelter and Layard (1986), vII pp1139-1181

Lucifora, C (1993) "Inter-industry and occupational wage differentials in Italy", *Applied Economics*, v25 pp1113-1124

MaCurdy, TE (1985) "Interpreting empirical models of labor supply in an intertemporal framework with uncertainty" in JJ Heckman and B Singer (eds) *Longitudinal Analysis of Labor Darket Data*, Cambridge: Cambridge University Press pp111-155

Main, BG (1988a) "The lifetime attachment of women to the labour market" in A Hunt (1988), pp23-51

Main, BG (1988b) "Women's hourly earnings: the influence of work histories on rates of pay", in A Hunt (1988), pp105-122

Main, BG (1989) "Women's hourly earnings over the life cycle", University of Warwick Institute for Employment Research

Main BG and P Elias (1986) "Women returning to paid employment", *International Review of Applied Economics*, v1:1 pp86-108

Main, BG and B Reilly (1992) "Women and the union wage gap", *The Economic Journal* v102:410, pp 49-66

Matyas, L and P Sevestre (1992) *The Econometrics of Panel Data: Handbook of Theory and Applications*, Dordrecht: Kluwer

Meghir, C and E Whitehouse (1992) "The evolution of wages in the UK: evidence from micro data", Institute for Fiscal Studies and University College London, mimeo

Micklewright, J and C Trinder (1981) "New Earnings Surveys 1968-80: Sampling methods and non-response", Social Science Research Council Programme on Taxation, Incentives and the Distribution of Income Working Paper No 31

Miller, CF (1993) "Part-time participation over the life-cycle among married women who work in the market", *Applied Economics*, v25:1 pp91-99

Miller, PW (1987) "The wage effect of the occupational segregation of women in Great Britain", *The Economic Journal,* vol 97, pp 885-896

Mincer, J (1974) *Schooling, Experience and Earnings*, New York: Columbia University Press/National Bureau of Economic Research

Mincer, J and H Ofek (1982) "Interrupted work careers: depreciation and restoration of human capital", *Journal of Human Resources*, v17:1 pp3-24

Mundlak, Y (1978) "On the pooling of time series and cross section data", *Econometrica* 46:69-85

Munroe, A (1988) "The measurement of racial and other forms of discrimination", University of Stirling Discussion Papers in Economics no. 148

Murphy, KM and F Welch (1990) "Empirical age-earnings profiles", *Journal of Labor Economics*, v8:2 pp203-229

Murphy, PD , PJ Sloane and DH Blackaby (1992) "The effects of trade unions on the distribution of earnings: a sample selectivity approach", *Oxfrod Bulletin of Economics and Statistics* v54 pp517-542

Neumark, D (1995) "Are rising earnings profiles a forced-savings mechanism?", *The Economic Journal*, v105:1 pp95-106

Neyman, J and EL Scott (1948) "Consistent estimates based on partially consistent estimates", *Econometrica* 16:1-32

Nickell, S (1977) "Trade Unions and the Role of Women in the Industrial Wage Structure", *British Journal of Industrial Relations*, vol 15, pp 192-210

Oaxaca, RL (1973) "Male-female wage differentials in urban labour markets" *International Economic Review*, v14 pp 693-709

Oaxaca, RL and MR Ransom (1994) "On discrimination and the composition of wage differentials", *Journal of Econometrics*, v61 pp5-22

Oi, W (1983) "The fixed employment costs of specialist labor" in JE Triplett (ed.) *The Measurement of Labor Cost*, Chicago: University of Chicago Press/National Bureau of Economic Research pp63-122

Ogaki, M (1993) "Generalized methods of moments: econometric applications" in GS Maddala, CR Rao and HD Vinod (eds) *Handbook of Statistics*, Amsterdam:

Elsevier Science Publishers vII pp 455-488

Pagan, A and F Vella (1989) "Diagnostic tests for models based on individual data: a survey", *Journal of Applied Econometrics*, Vol 4, pp 29-59

Page, ES and LB Wilson (1983) *Information Representation and Manipulation using Pascal*, Cambridge: Cambridge University Press

Phelps, E (1972) *Inflation Policy and Unemployment Theory: the Cost Benefit Approach to Monetary Planning*, London: Macmillan

Polacheck, SW (1981) "Occupational self-selection: a human-capital approach to sex differences in occupational structure", *Review of Economics and Statistics*, v58 pp60-69

Raj, B and A Ullah (1981) *Econometrics: A Varying Coefficients Approach*, London: Croom Helm

Rees, H and A Shah (1992) "Work and employment in public and private sectors", University of Bristol Discussion paper 92/327

Reilly, B (1991) "Occupational segregation and selection bias in occupational wage equations: an empirical analysis using Irish data", *Applied Economics*, v23:1 pp1-7

Ridder, G (1990) "Attrition in multi-wave panel data" in J. Hartog, G. Ridder and J. Theeuwes (eds) *Panel Data and Labour Market Studies*, Elsevier, North-Holland

Ritchie, F (1994) "Felit II: The revenge of the incidental parameters", paper presented to

the Scottish Doctoral Programme winter seminar, Crieff, January

Ritchie, F (1995) "Sex, age and earnings - the cohort effect", paper presented to the Scottish Doctoral Programme winter seminar, Crieff, January

Robinson, C (1989) "The joint determination of union status and union wage effects: some tests of alternative models", *Journal of Political Economy*, v97 pp639-667

Robinson, P (1994) "The British labour market in historical perspective: changes in the structure of employment and unemployment", Centre for Economic Performance, London School of Economics Discussion paper no. 202

Sevestre, P and A Trognon (1992) "Linear dynamic models" in Matyas and Sevestre (1992)

Sloane, PJ and I Theodossiou (1994) "A generalised Lorenz curve approach to explaining the upward movement in women"s relative earnings in Britain", *Scottish Journal of Political Economy*, v41 pp464-476

Stern, N (1986) "On the specification of labour supply functions" in Blundell and Walker (1986) pp143-189

Stewart, MB (1987) "Collective bargaining arrangements: closed shops and relative pay", *Economic Journal*, v97:1 pp140-156

Taylor, WE (1980) "Small sample consideration in estimation from panel data", *Journal of Econometrics* 13:203-23

Theodossiou, I (1992) *Wage Inflation and the Two-Tier Labour Market*, Aldershot:

Avebury

Topel, R (1991) "Specific human capital, mobility and wages: wages rise with job seniority", *Journal of Political Economy*, February 1991, Vol 99, No 1, pp 145-176

Vella, F (1994) "Gender roles and human capital investment: the relationship between traditional attitudes and female labour market performance", *Economica*, v61 pp191-211

Vella, F and M Verbeek (1993) "Estimating the impact of endogenous union choice on wages using panel data", Center for Economic Research, Tilberg University, mimeo

Vella, F and M Verbeek (1993) "Estimating and testing simultaneous equation panel data models with censored endogenous variables" Rice University, mimeo

Verbeek, M (1990) "On the estimation of a fixed-effects model with selection bias", *Economics Letters* 34:267:270

Verbeek, M (1992) "Pseudo-panel data" in Mayas and Sevestre (1992), pp303-315

Verbeek, M and T Nijman (1992a) "Incomplete panels and selection bias: a survey", in Matyas and Sevestre (1992)

Verbeek, M and T Nijman (1992b) "Testing for selectivity bias in panel data models", *International Economic Review* 33:681:703

Wansbeek, T and A Kapteyn (1989) "Estimation of error-components models with incomplete panels", *Journal of Econometrics* 41:341:361

Wright, RE and JF Ermisch (1991) "Gender discrimination in the British labour market: a reassessment", *The Economic Journal*, May 1991, Vol 101, pp 508-522

Yaron, G (1990) "Trade unions and women"s relative pay: a theoretical and empirical analysis using UK data", Applied Economics Discussion Paper Series, No 95, Institute of Economics and Statistics, University of Oxford

Figure 8.1 Numbers observed

Employment cohorts

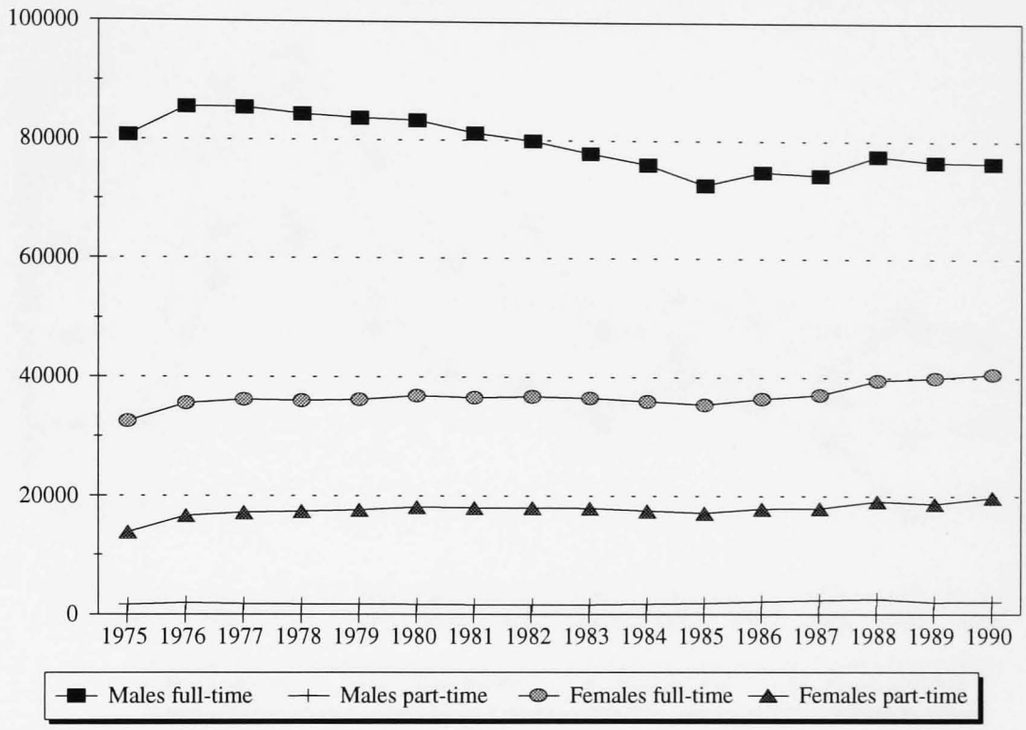


Figure 8.2 Percentages of observed

Employment cohort

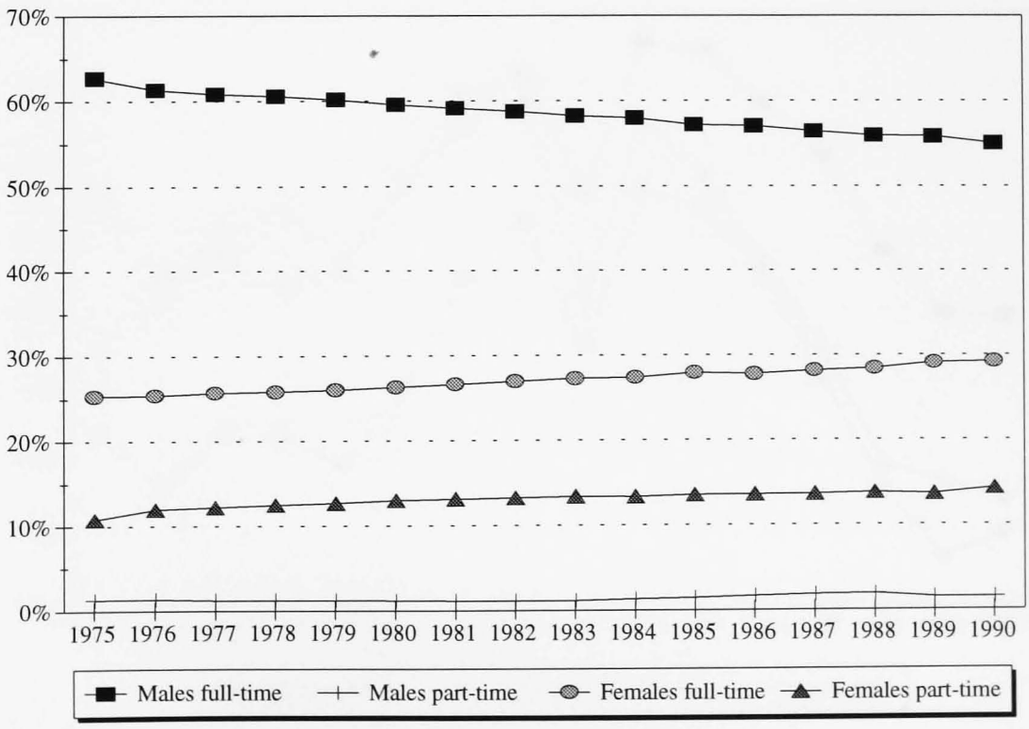


Figure 8.3 Disappearance rates

Data cohorts

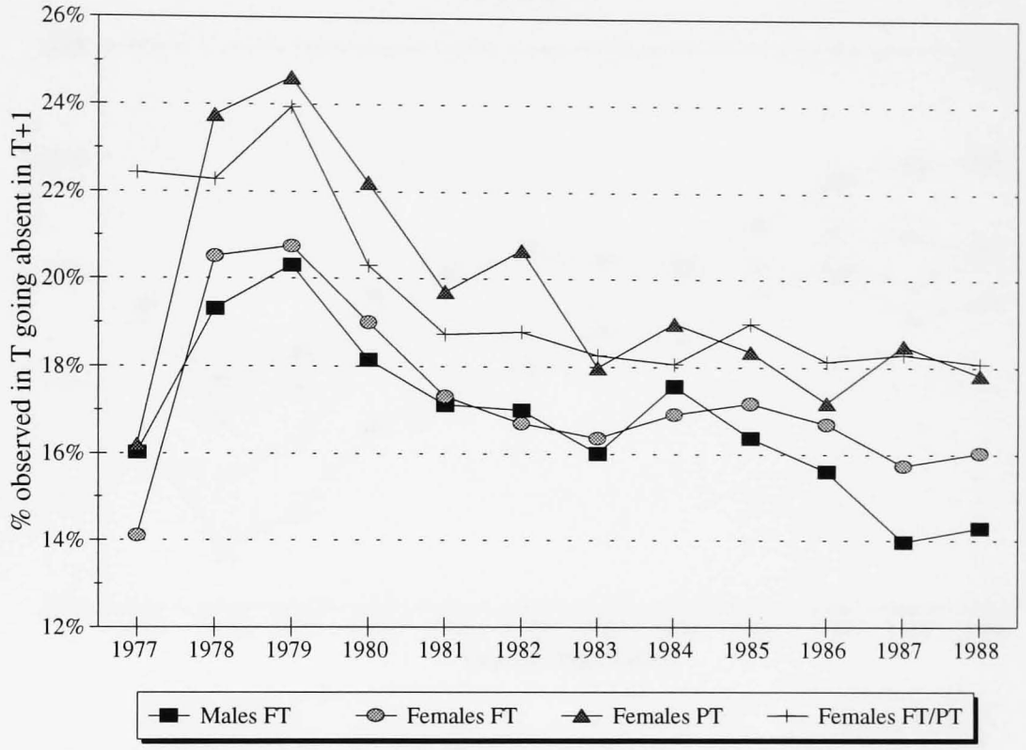


Figure 8.4 Job held for >1 year

Employment cohort

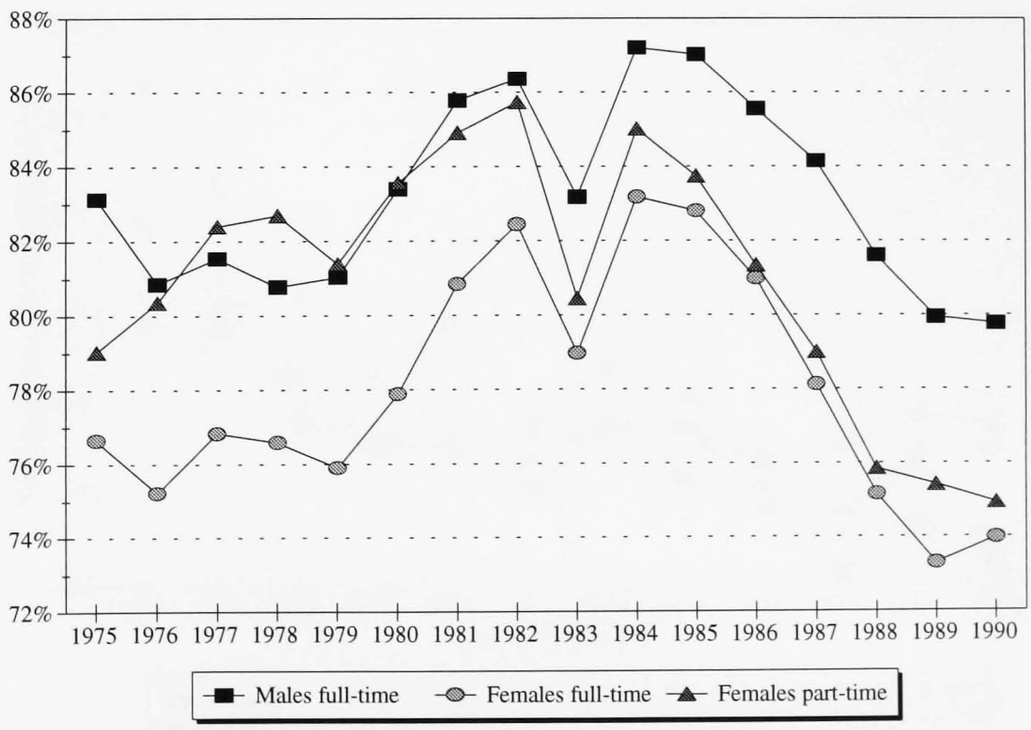


Figure 8.5a Proportions observed

In-in-in

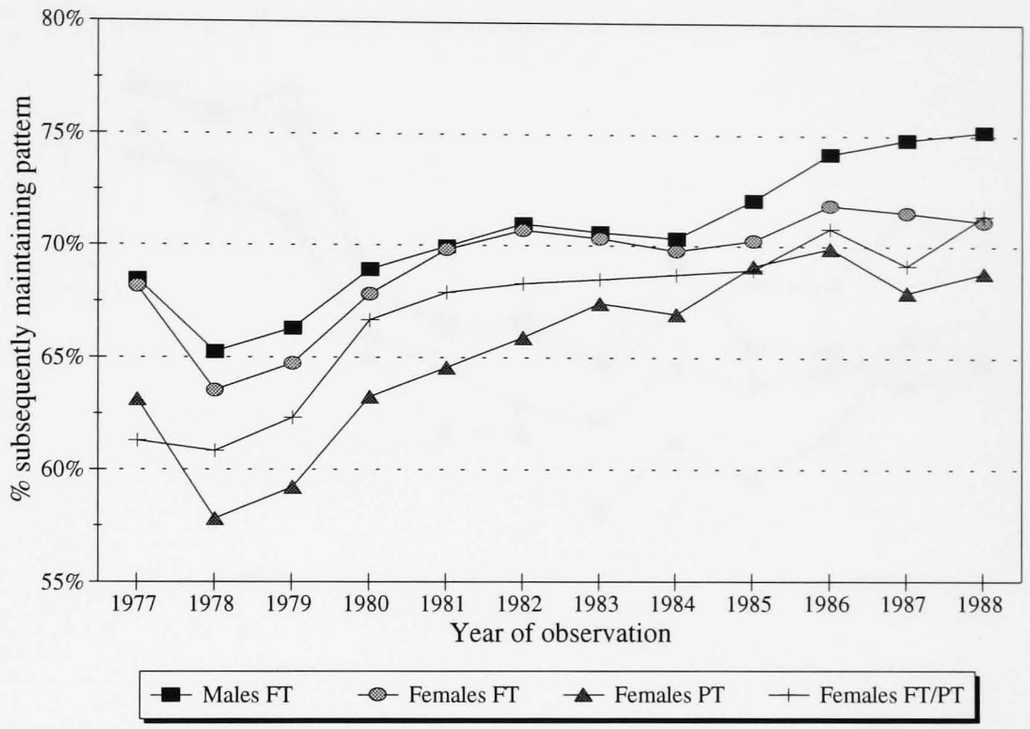


Figure 8.5b Proportions observed

In-in-out

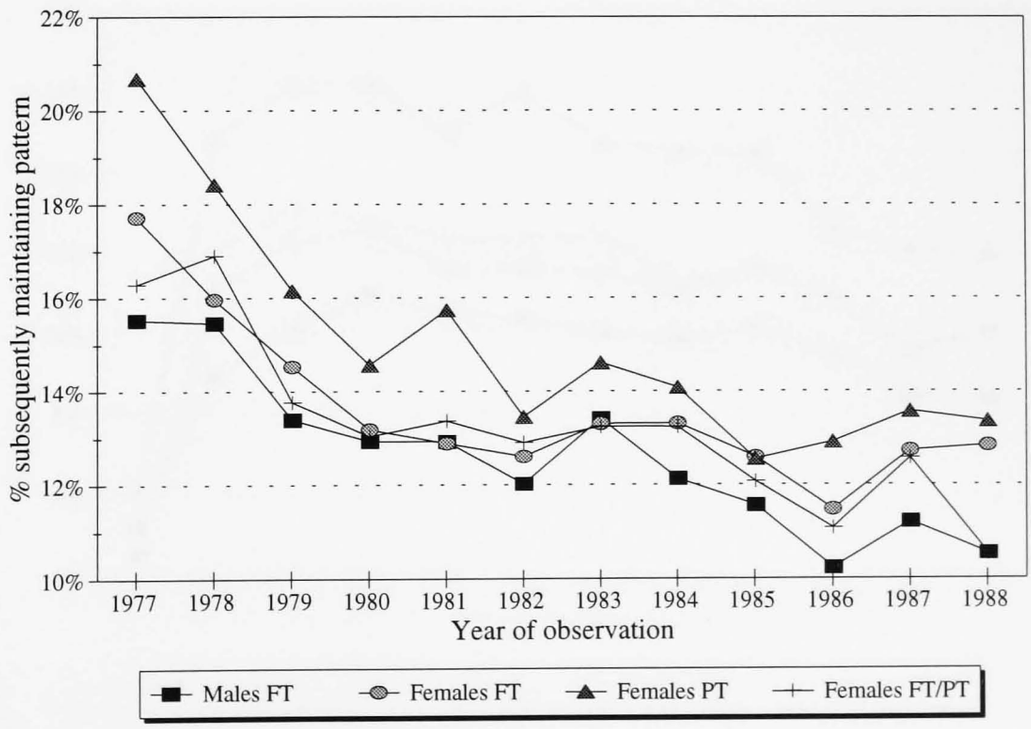


Figure 8.5c Proportions observed

In-out-in

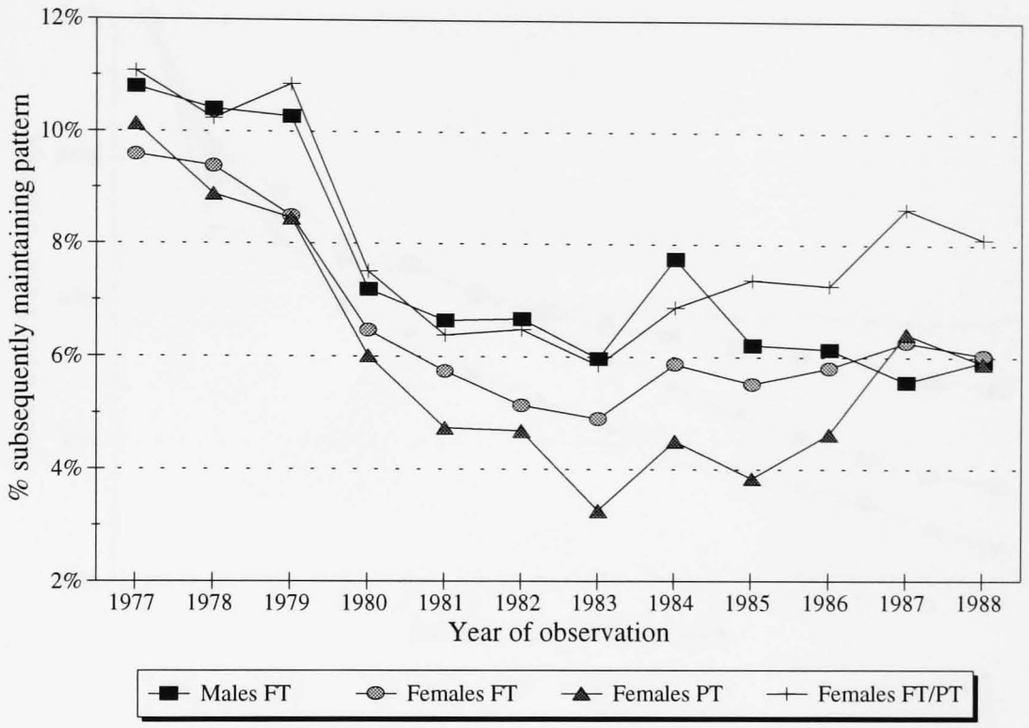


Figure 8.5d Proportions observed

In-out-out

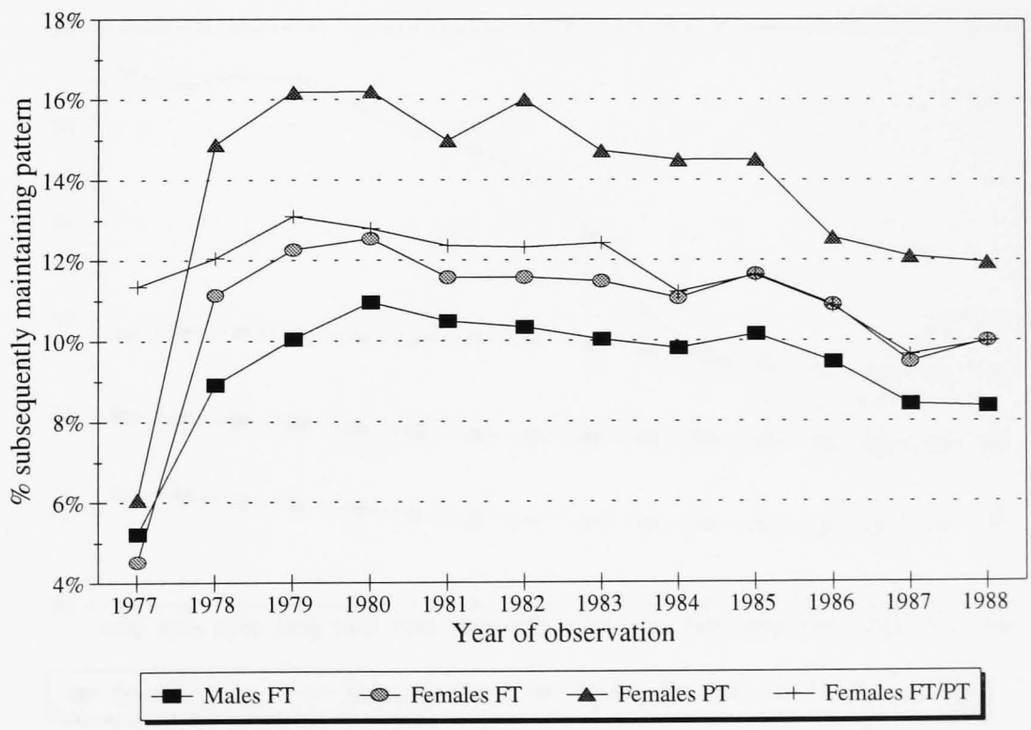


Figure 8.6 Cohort observation rates

averaged over all data cohorts

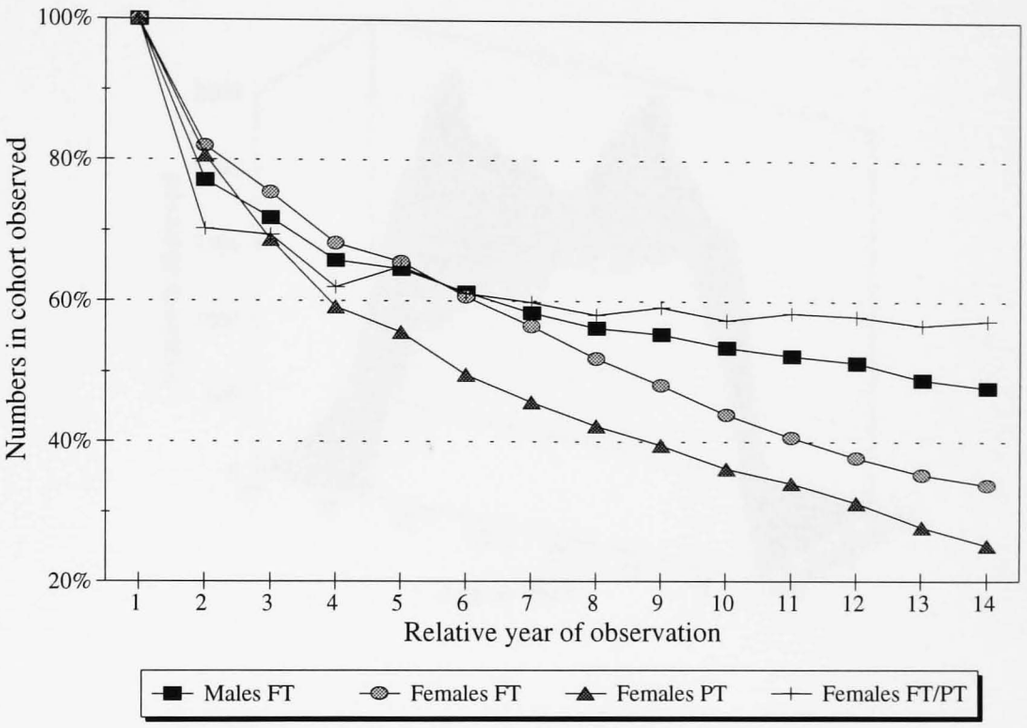


Figure 8.7 Average age in the NES

Employment cohort

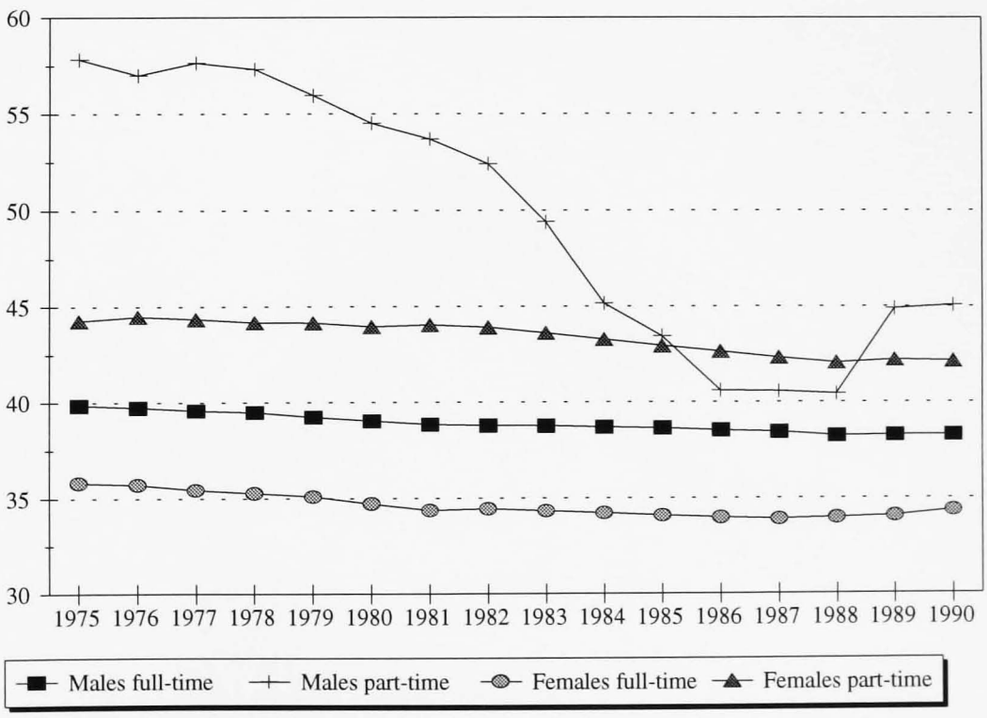


Figure 8.8a Age profiles

Males full-time

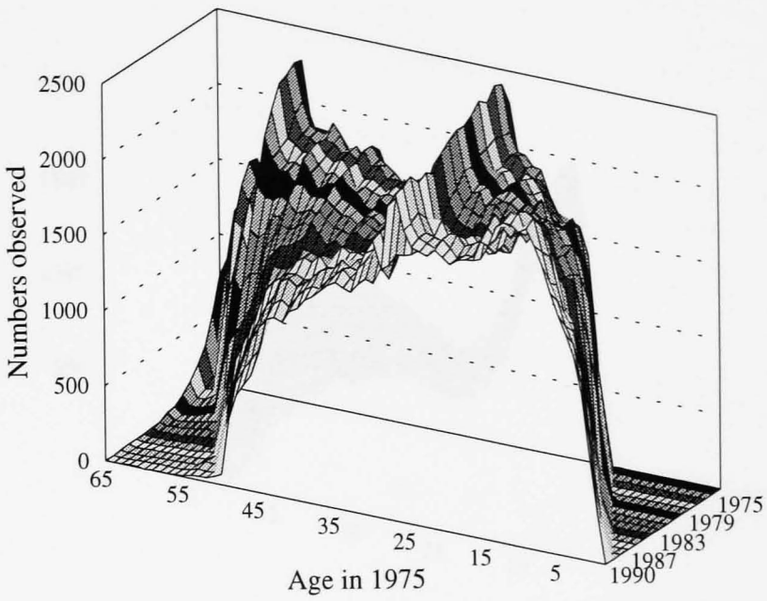


Figure 8.8b Age profiles

Females full-time

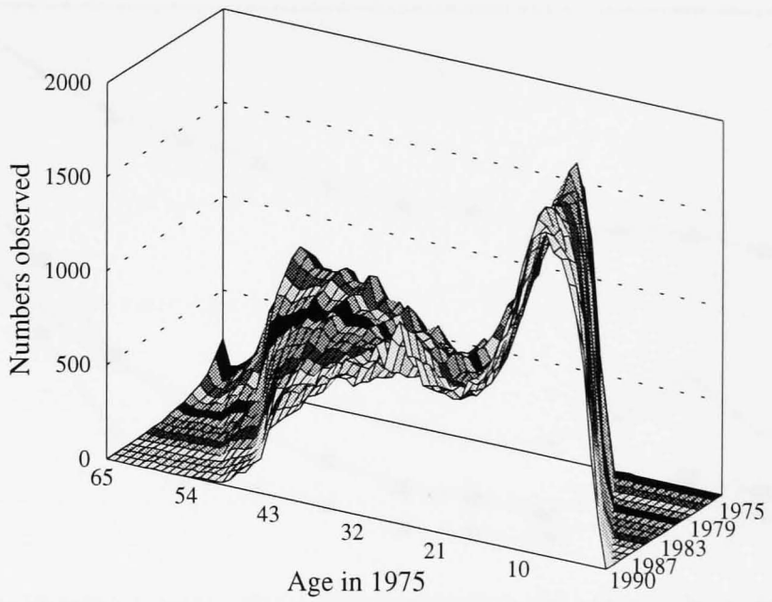


Figure 8.8c Age Profiles

Females part-time

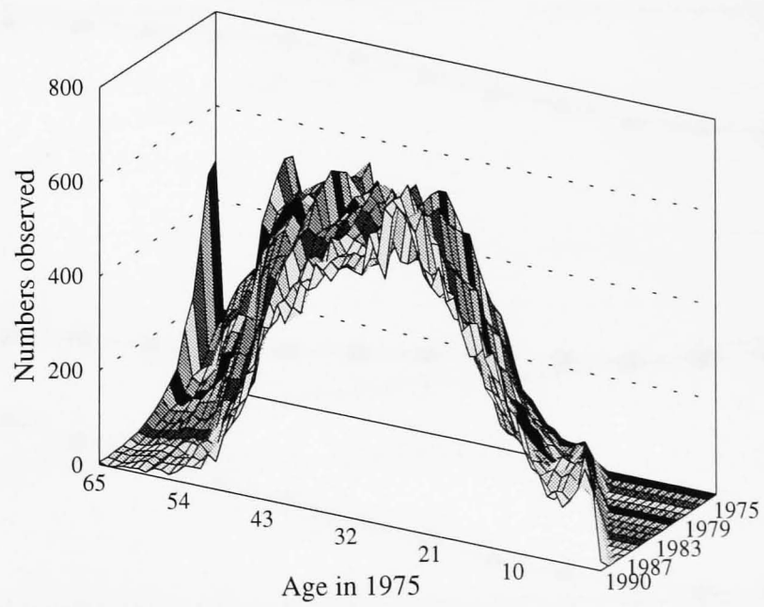


Figure 8.9 Starting ages

Data cohorts

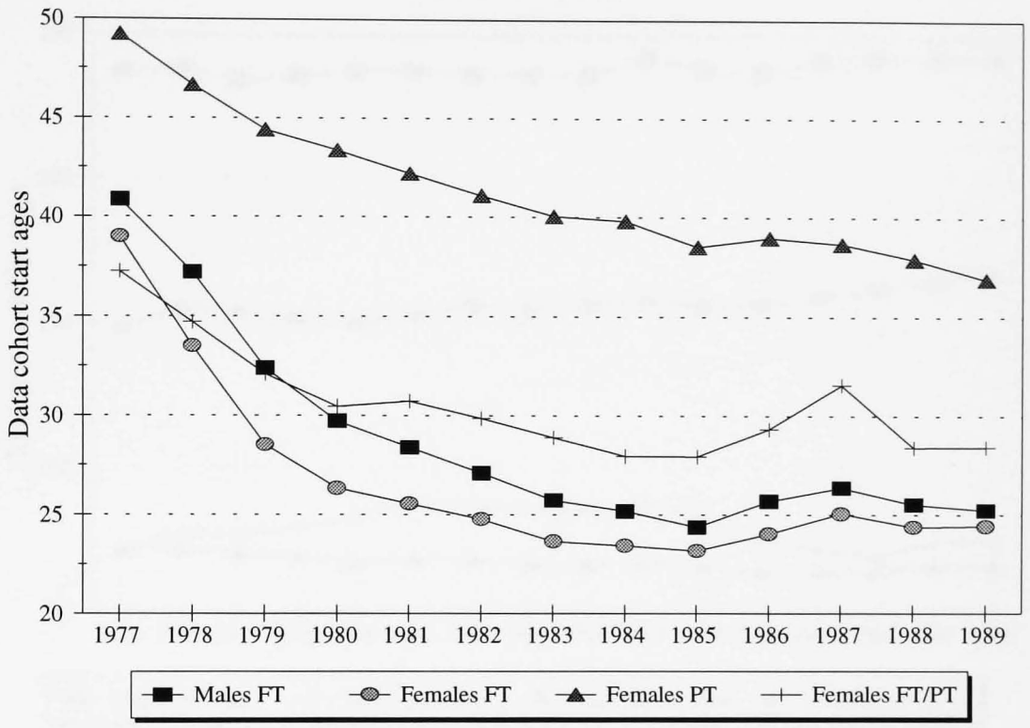


Figure 8.10 Average age in the NES

Data cohorts

Excluding single observations

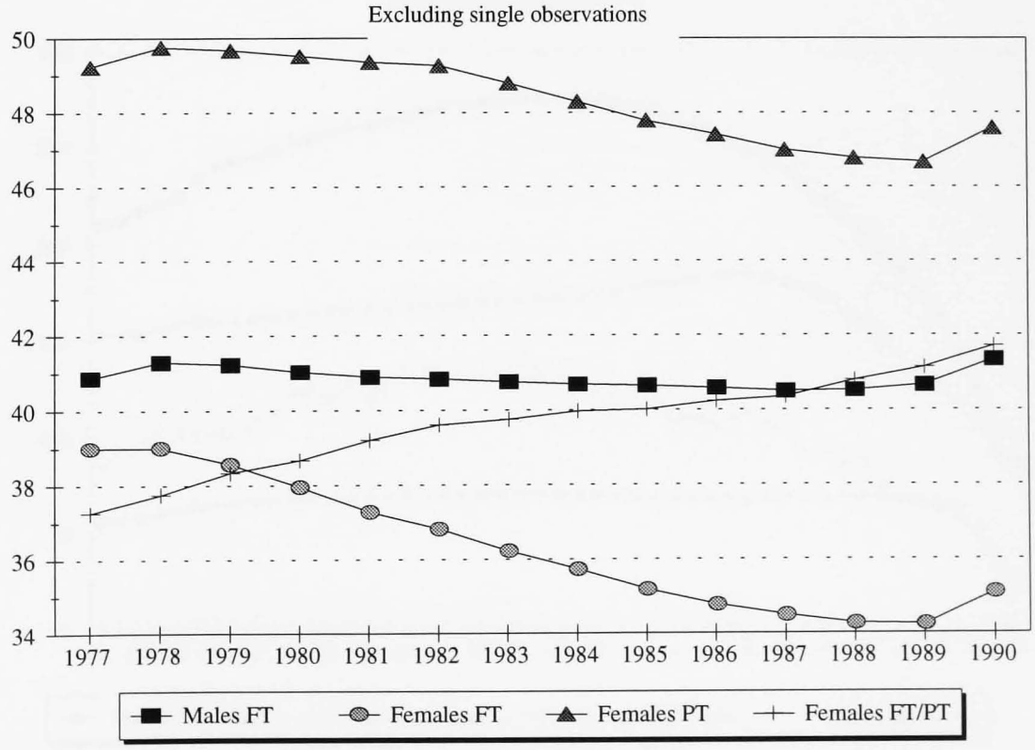


Figure 8.11 Average wages

Employment cohort

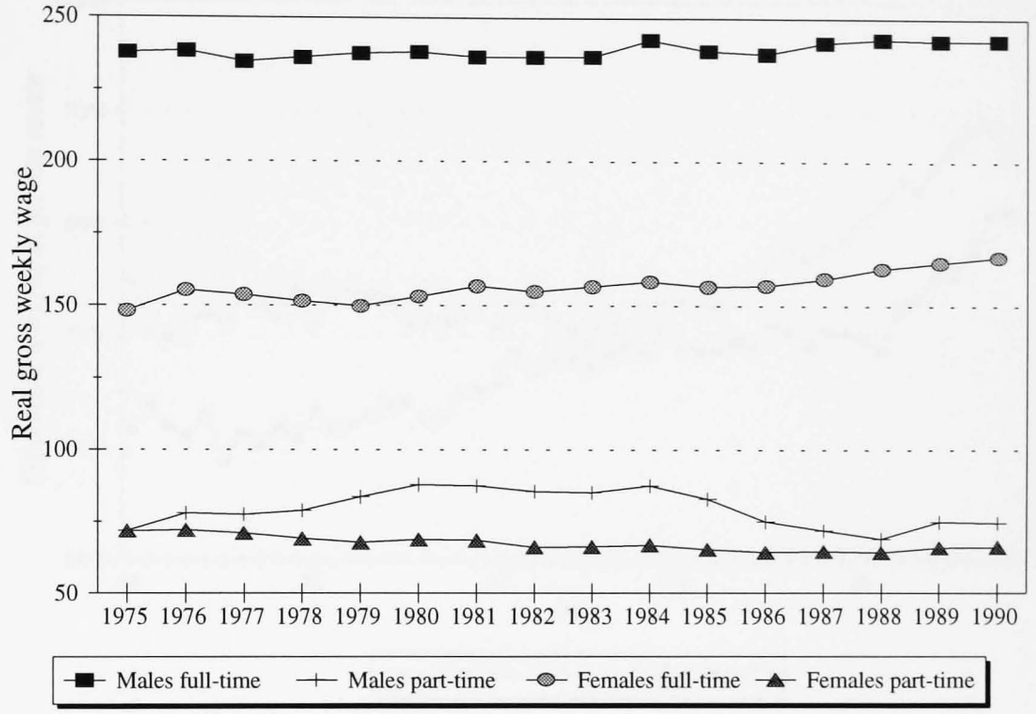


Figure 8.12 Wage profiles all years

Employment cohort

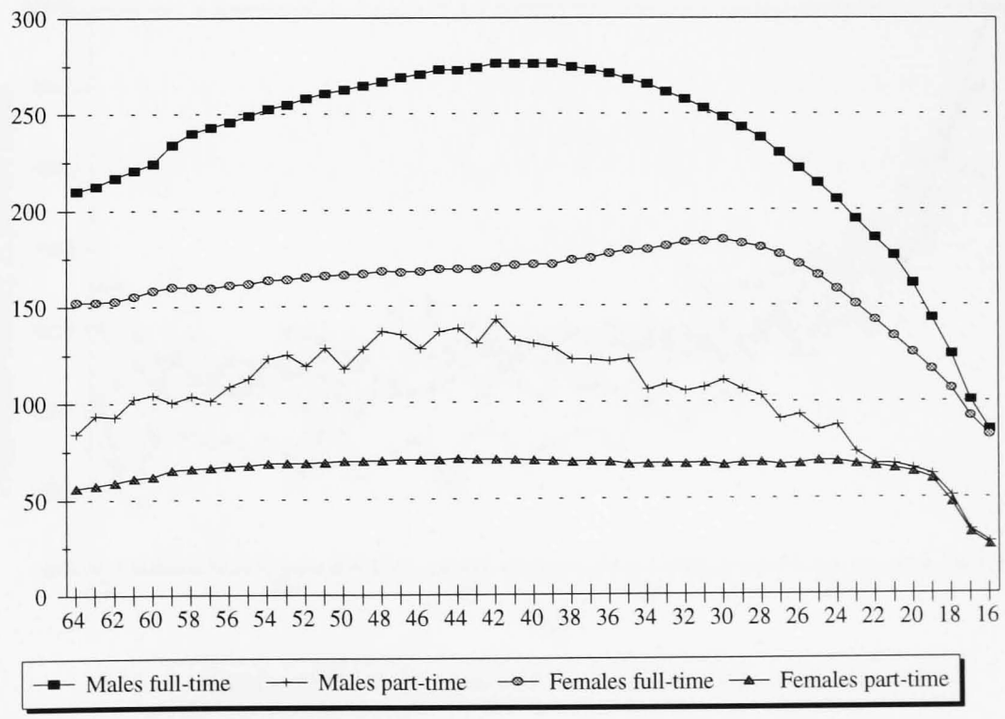


Figure 8.13a Private sector working

Males

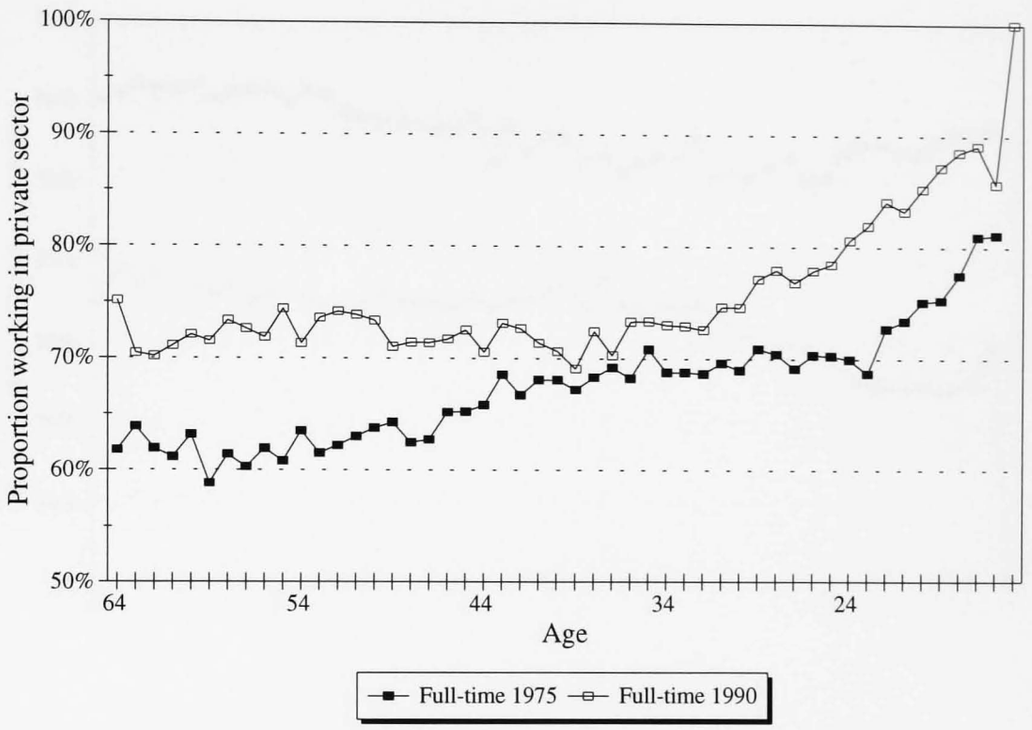


Figure 8.13b Private sector working

Females

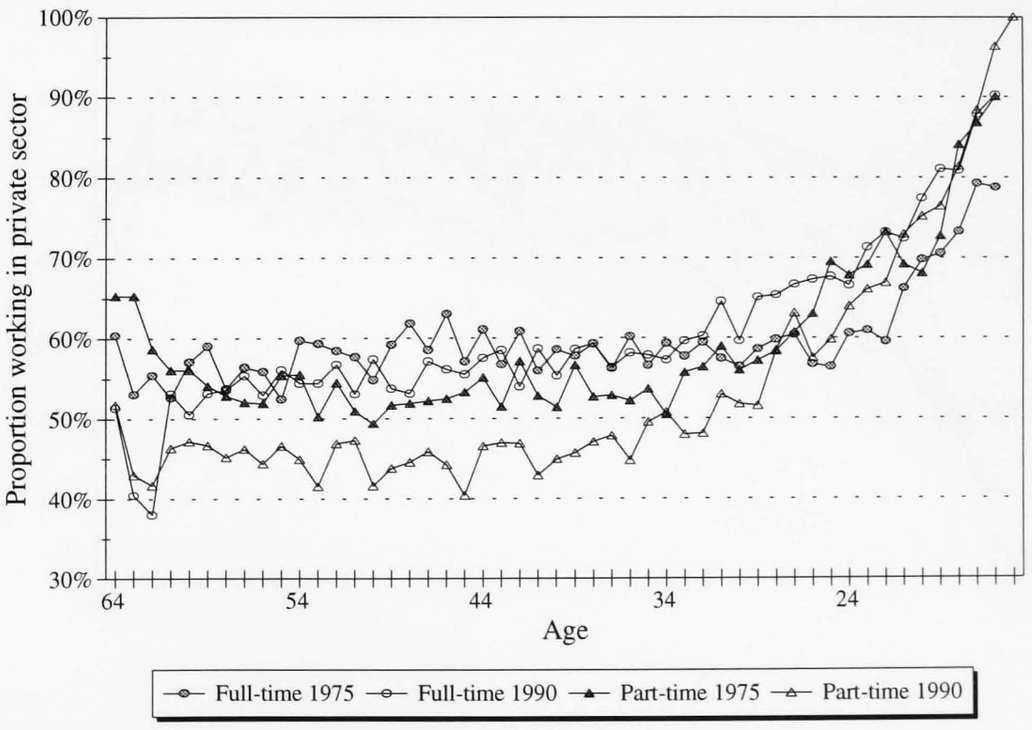


Figure 8.14a Covered by agreement

Males

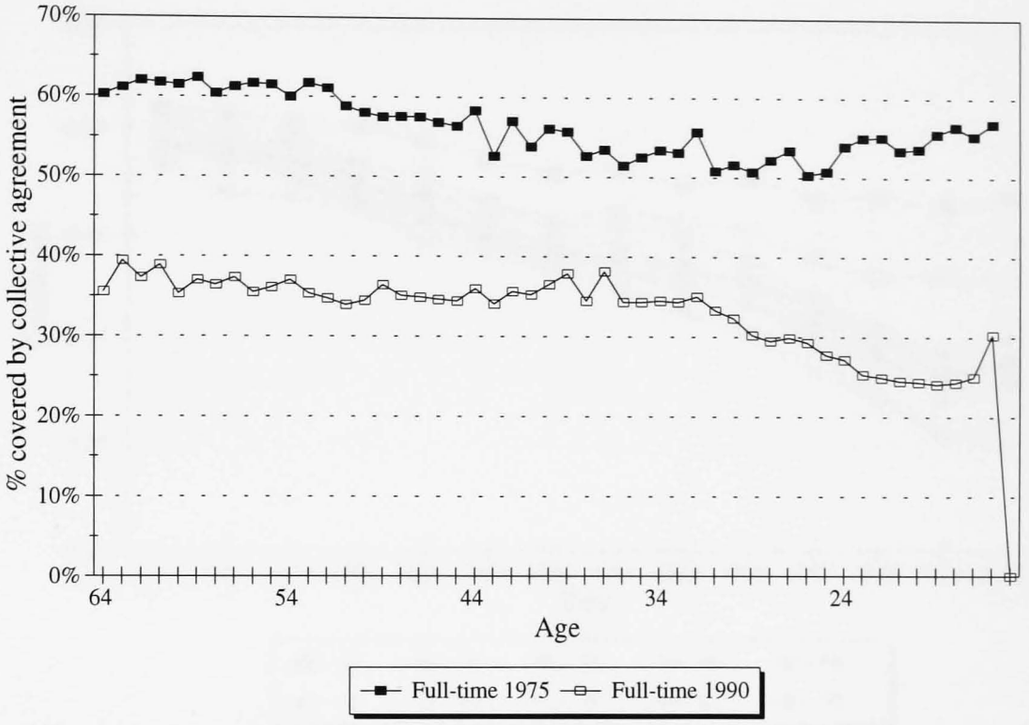


Figure 8.14b Covered by agreement

Females

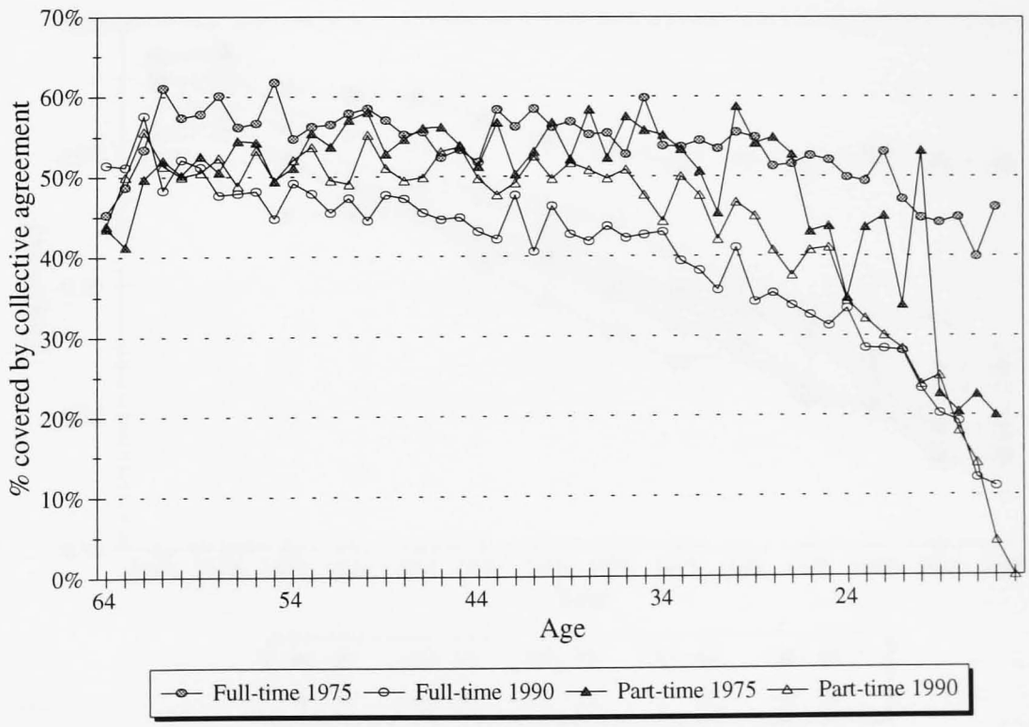


Figure 9.1a Region (default=London)

Males fixed-effects

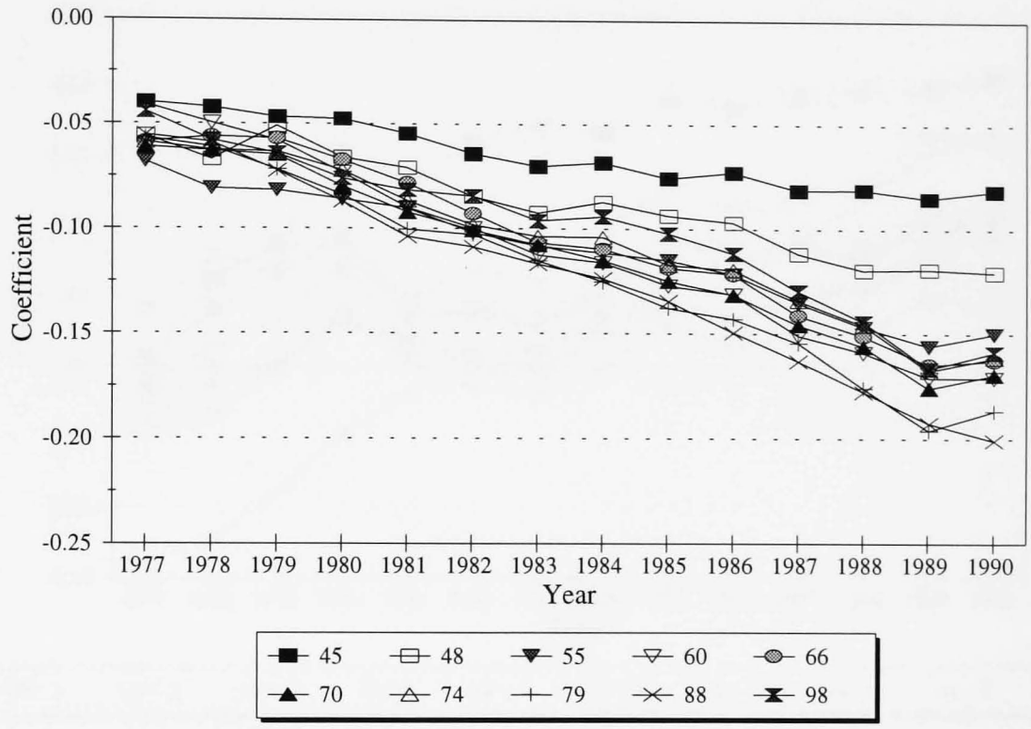


Figure 9.1b Region (default=London)

Males cross-section

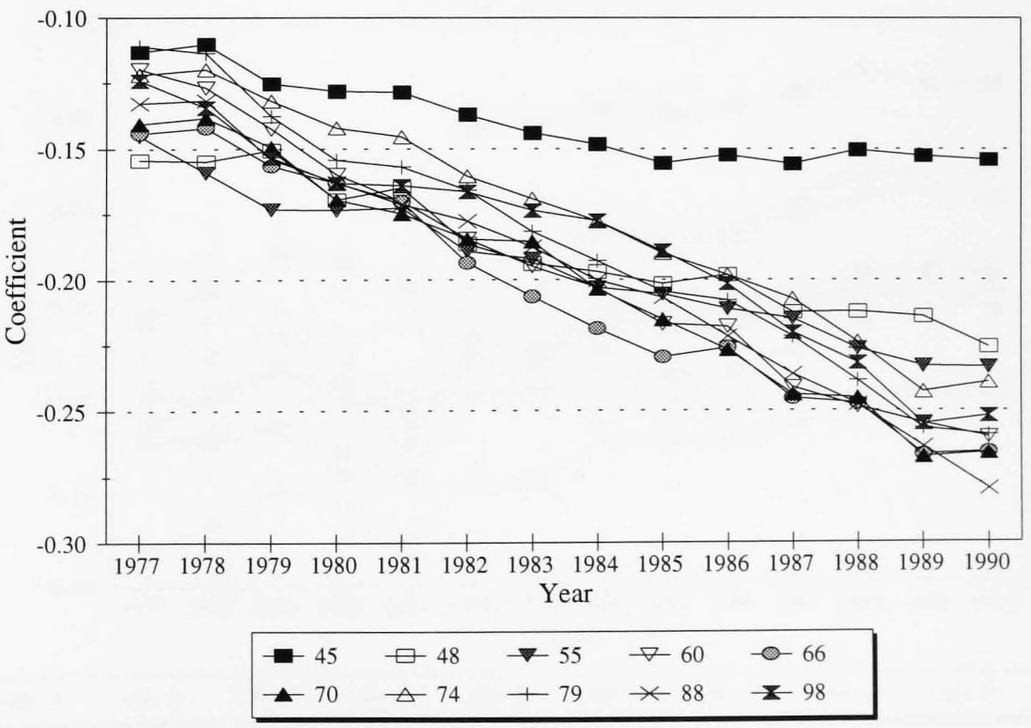


Figure 9.2a Division (default=FF)

Males fixed-effects

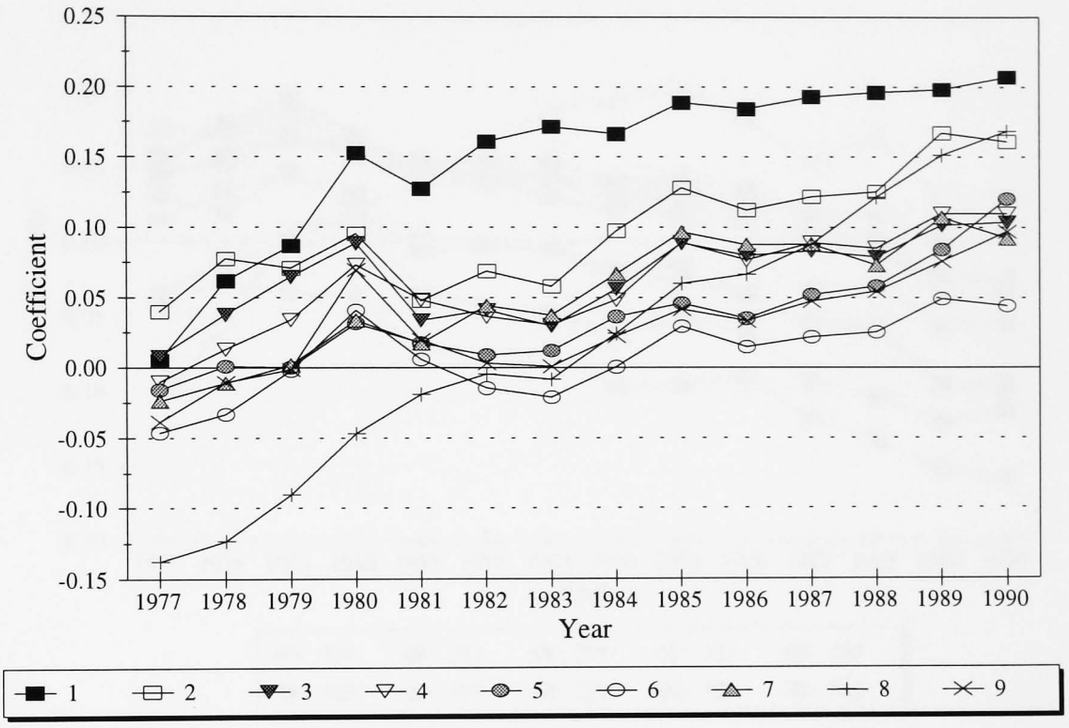


Figure 9.2b Division (default=FF)

Males cross-section

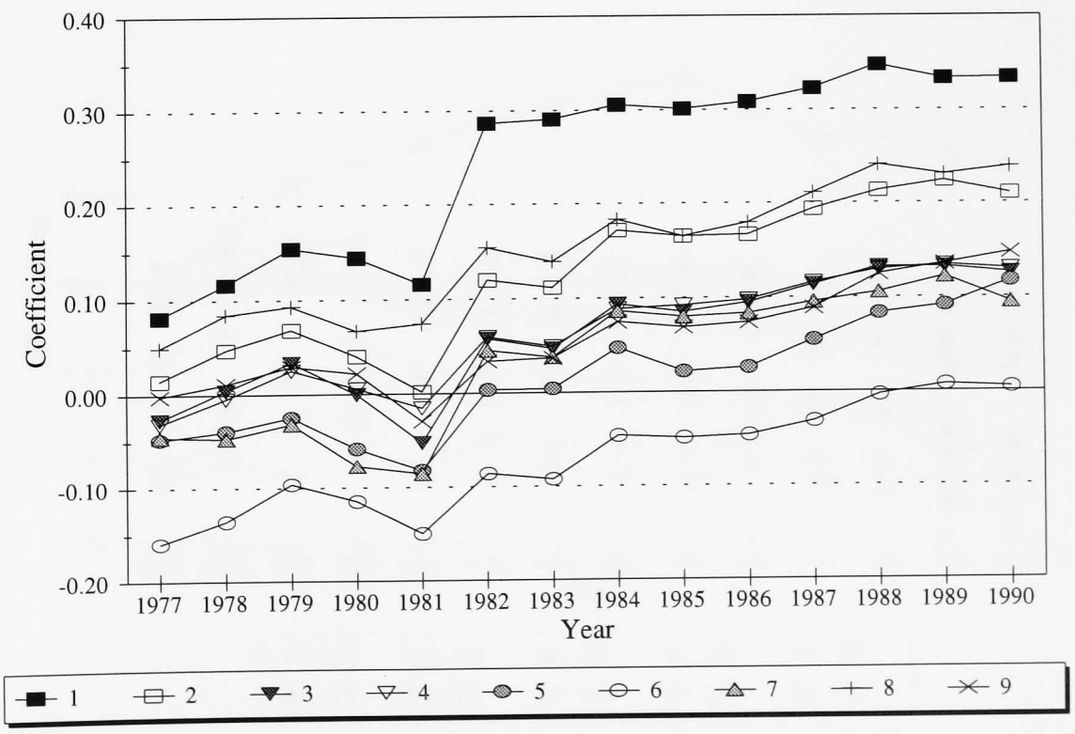


Figure 9.3a Manual occupations

Default=clerical; Males fixed-effects

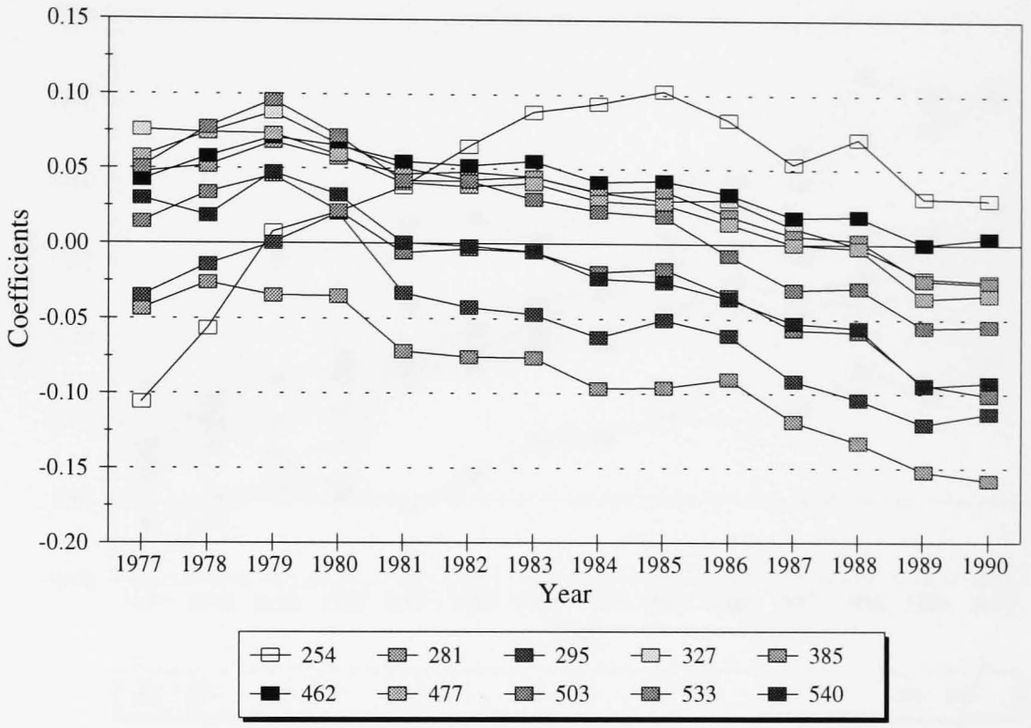


Figure 9.3b Manual occupations

Default=clerical; Males cross-section

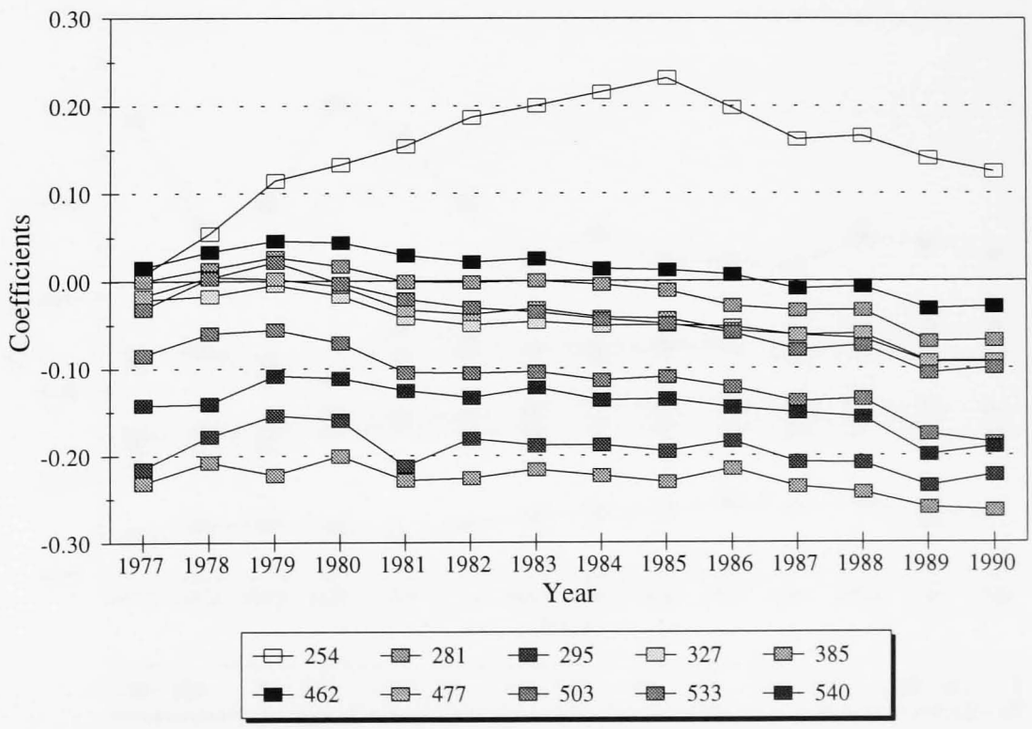


Figure 9.4a Non-manual occupations

Default=clerical; Males fixed-effects

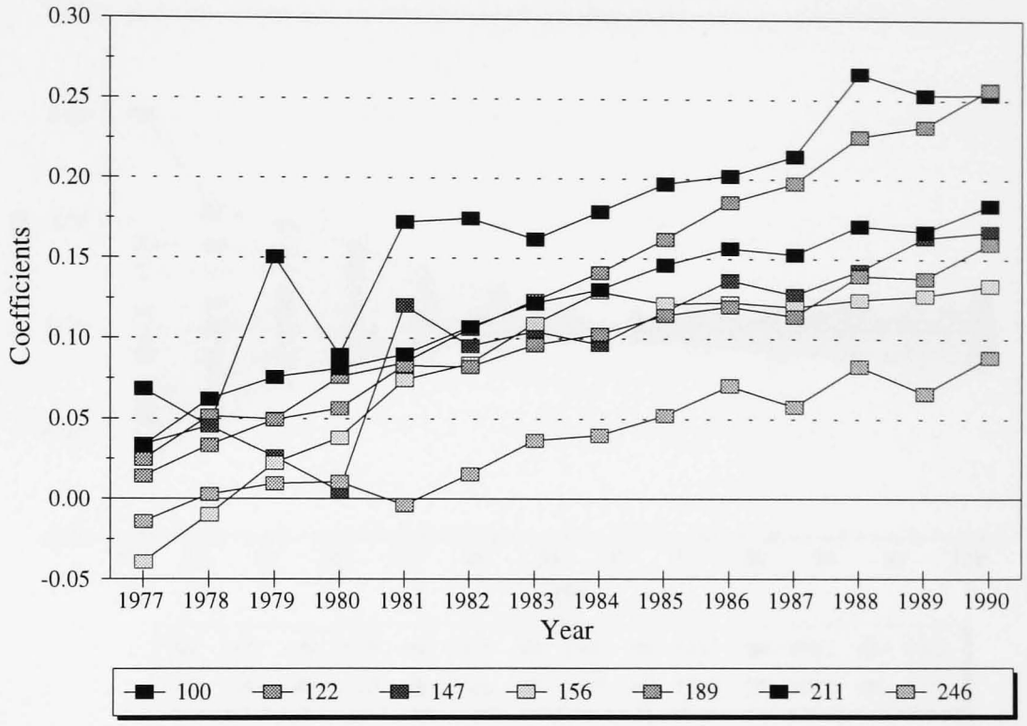


Figure 9.4b Non-manual occupations

Default=clerical; Males cross-section

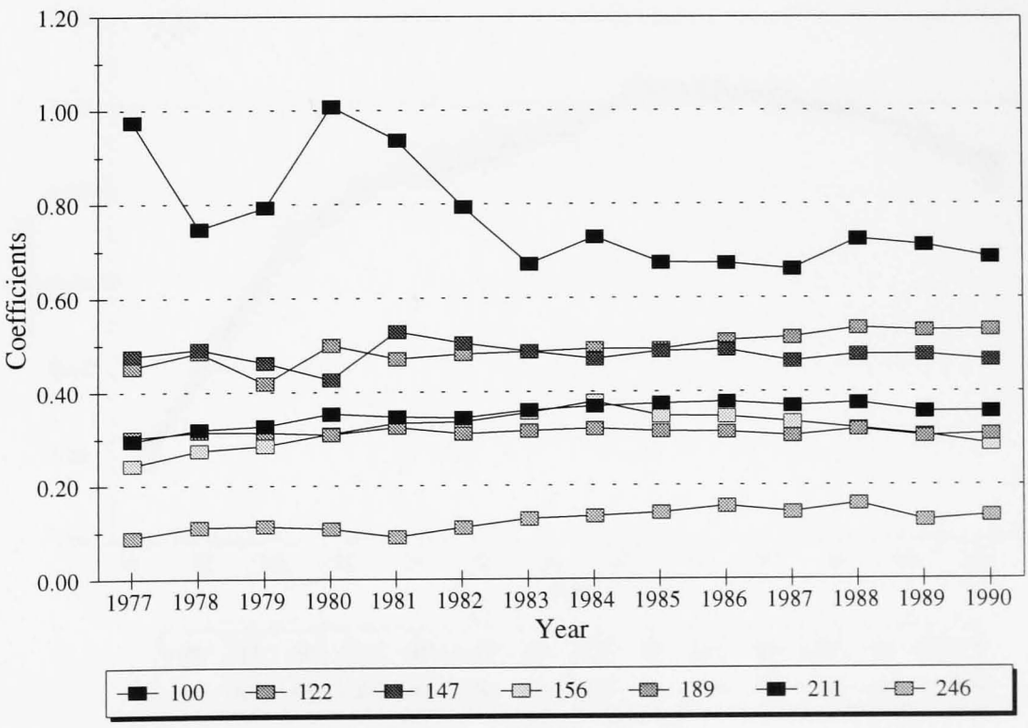


Figure 9.5a Age profiles(default=35)

Males fixed-effects

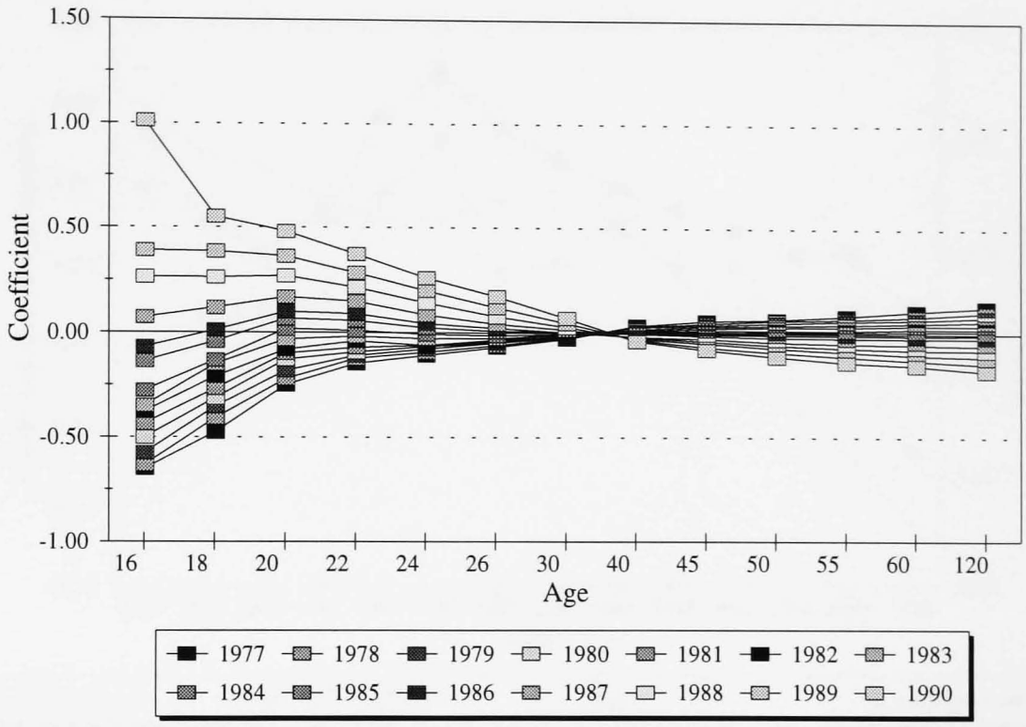


Figure 9.5b Age profiles(default=35)

Males cross-sections

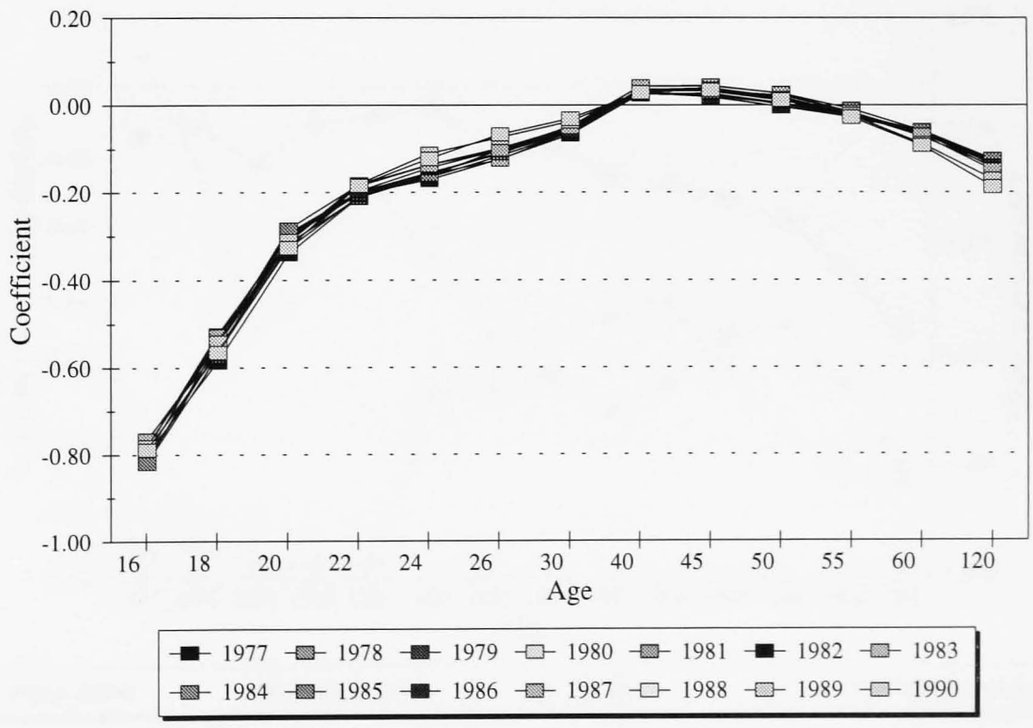


Figure 9.6 Effect of union coverage

Males TVFE and TVCS

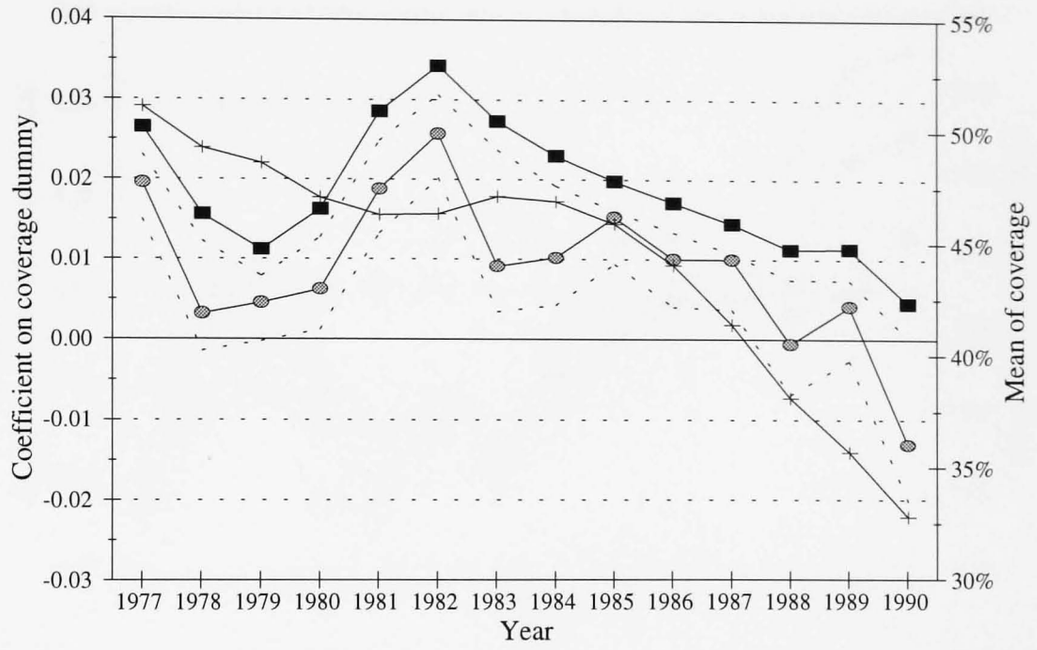


Figure 9.7 Wages Council coverage

Males FE and CS

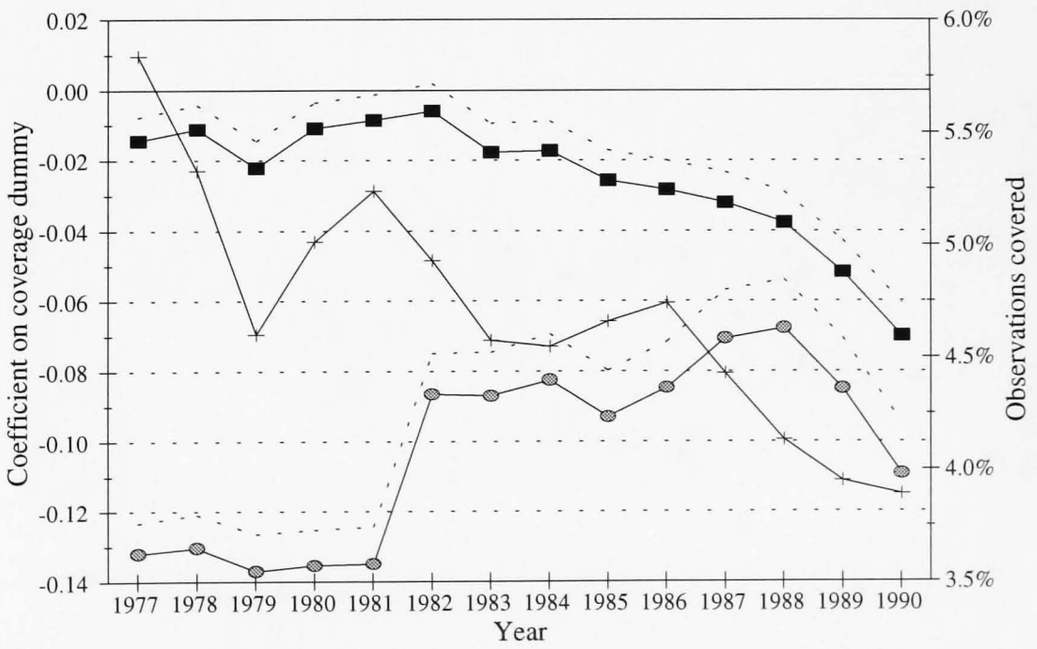


Figure 9.8 Sector (default=Public)

Males TVFE and TVCS

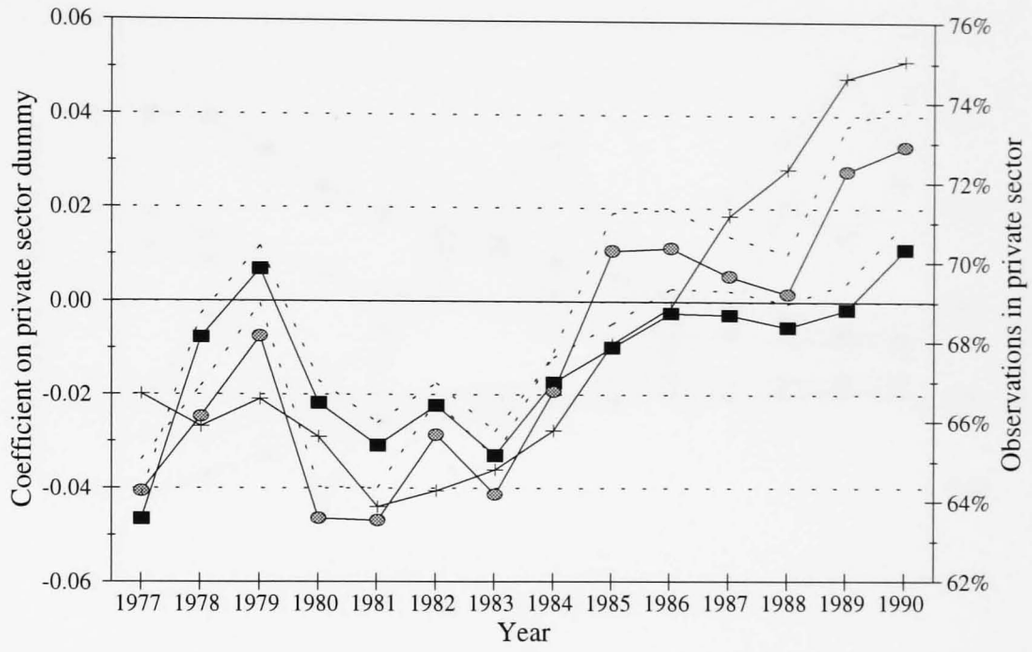


Figure 9.9 Effect of tenure

Default=job held for > 1 year; males

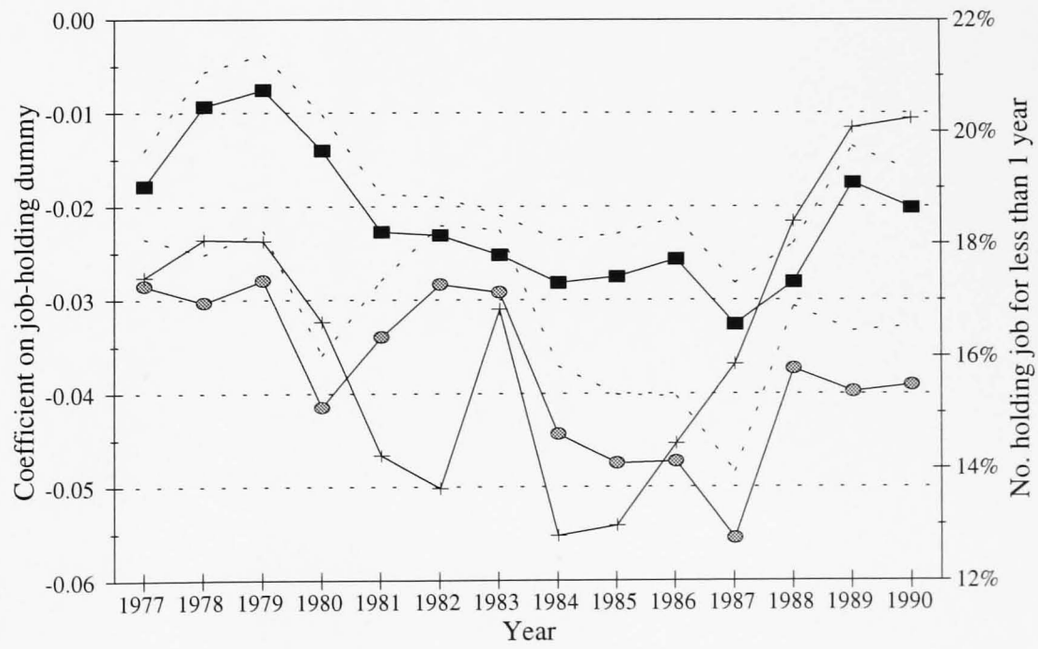


Figure 9.10 Attrition variables

Males TVFE and TVCS

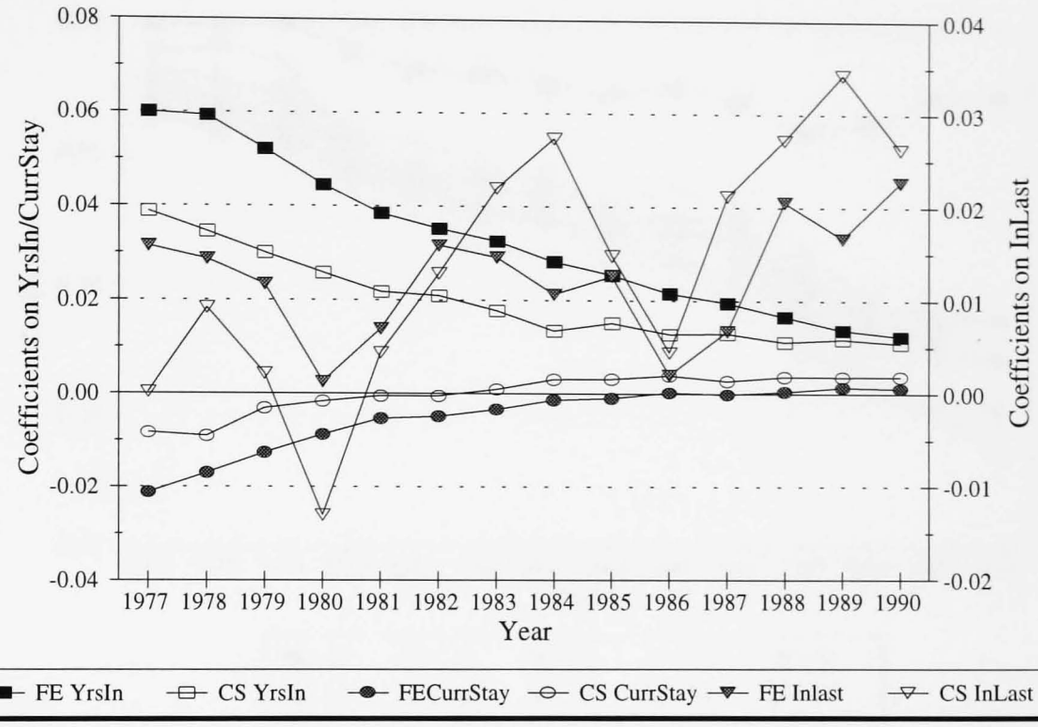


Figure 10.1a Region (default=London)

Females fixed-effects

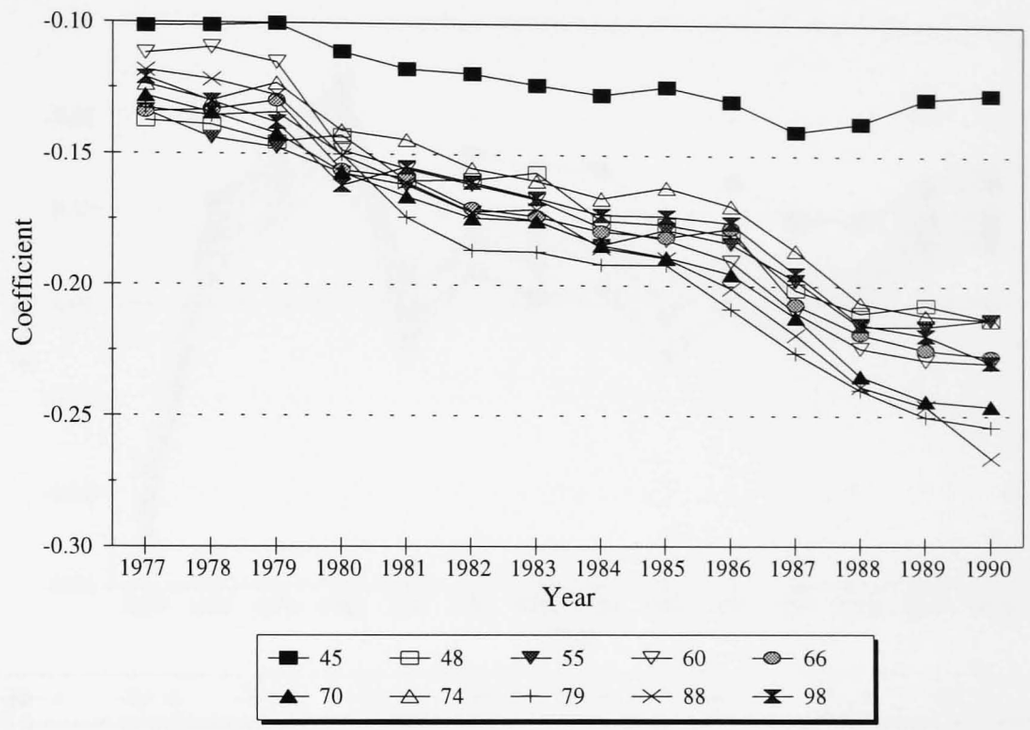


Figure 10.1b Region in 1990

Fixed-effects

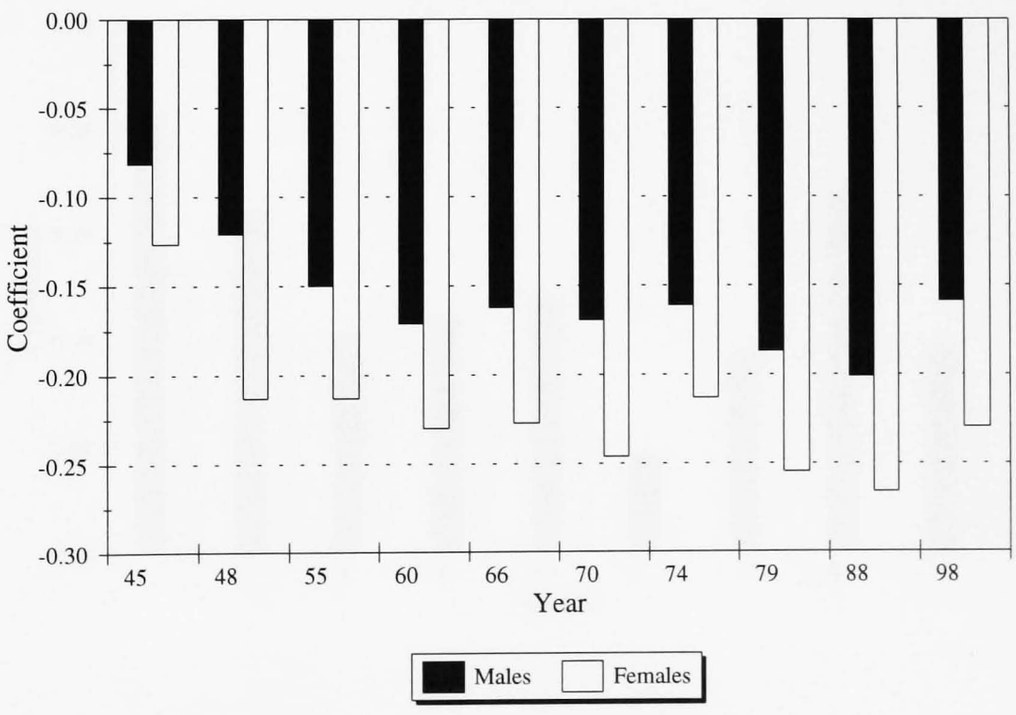


Figure 10.2a Division (default=FF)

Females fixed-effects

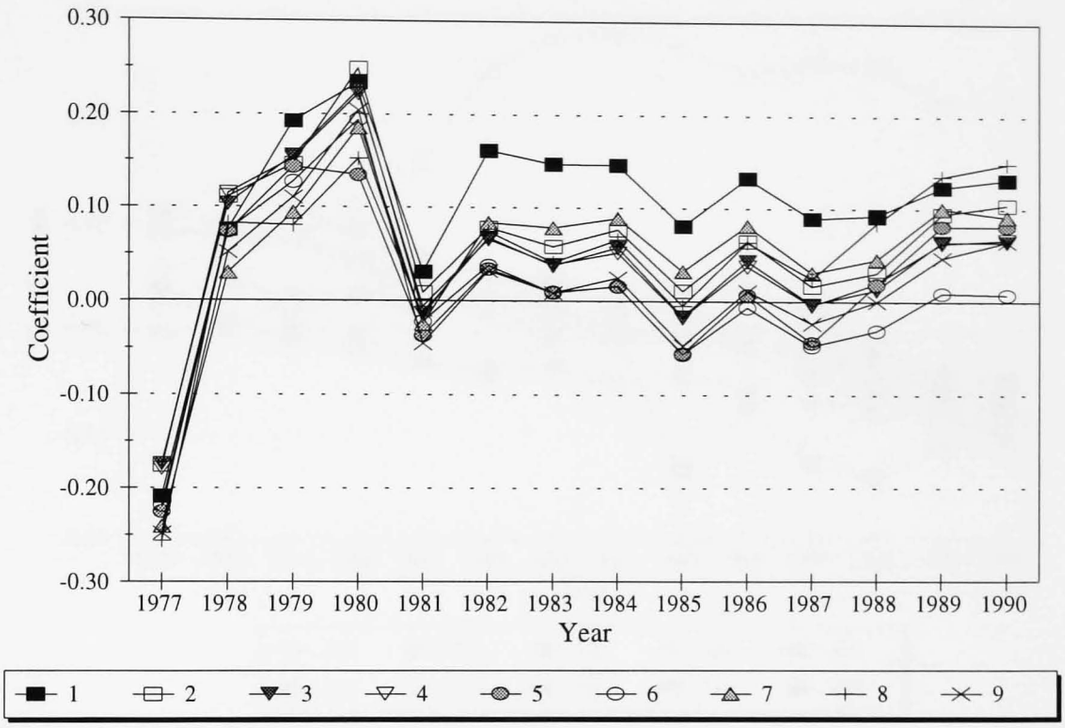


Figure 10.2b Division in 1990

Fixed-effects

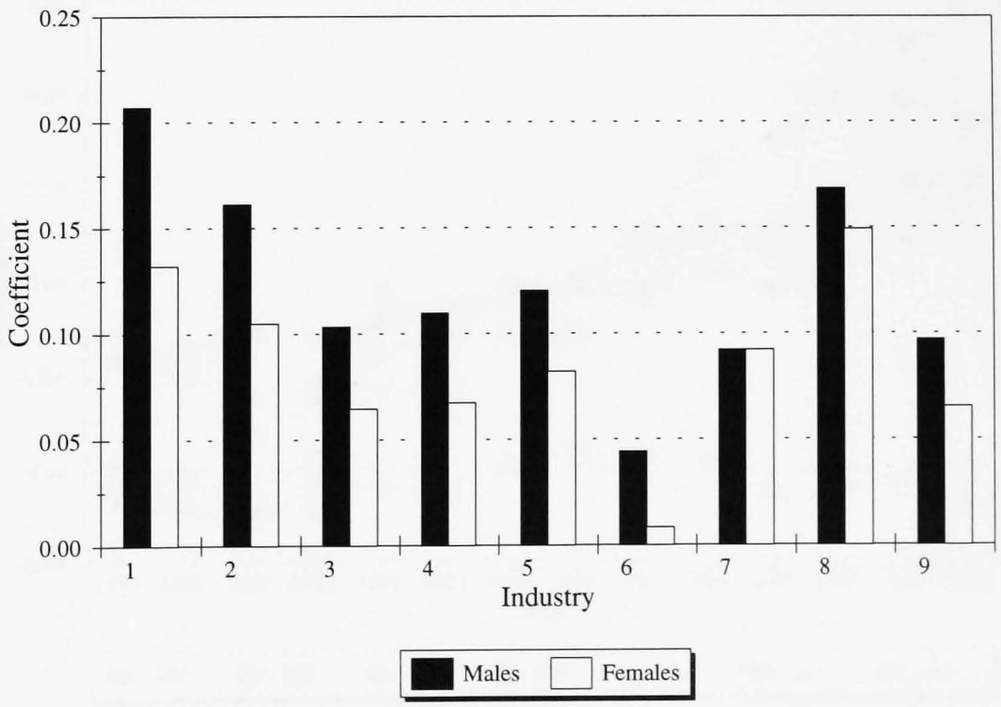


Figure 10.3a Manual occupations

Default=clerical; Female fixed-effects

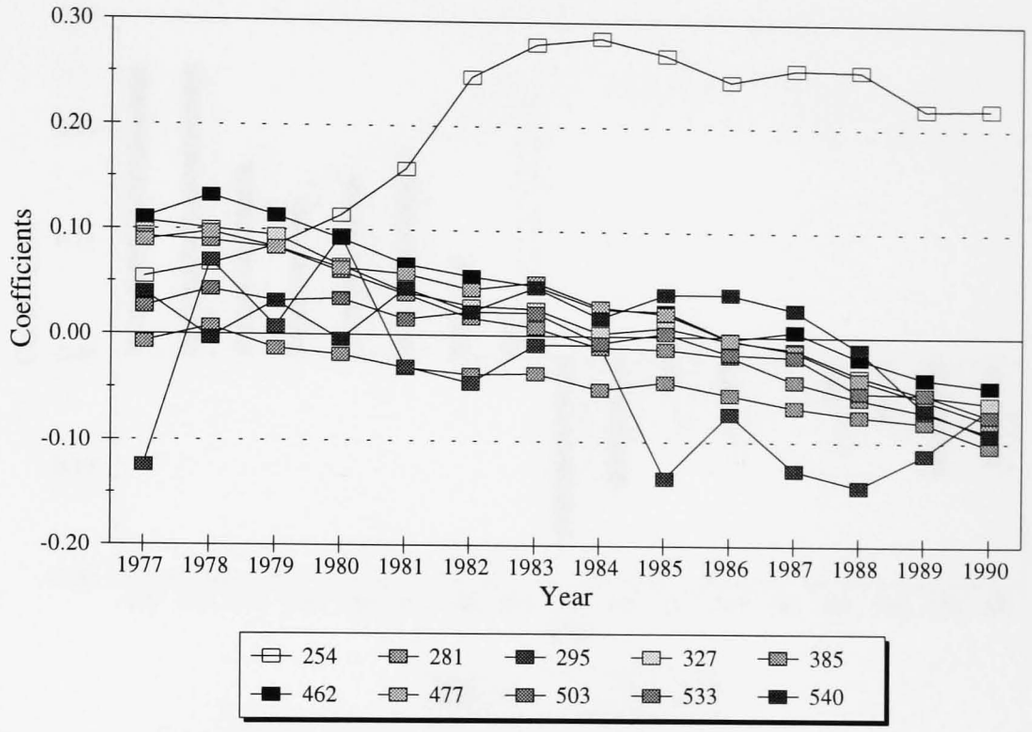
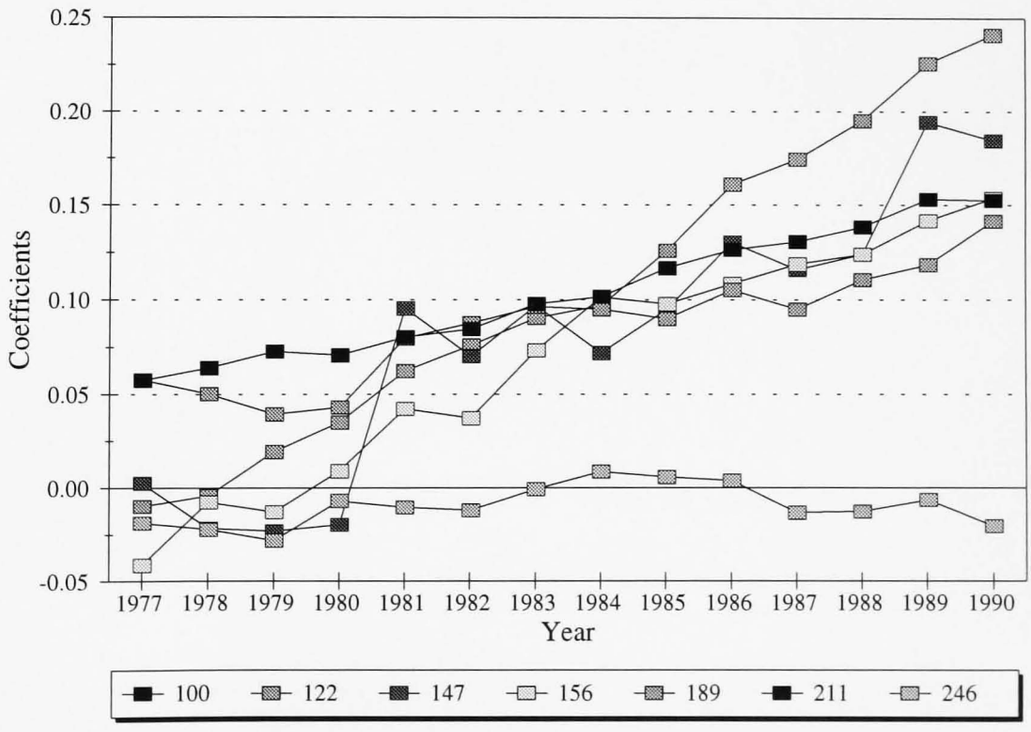


Figure 10.3b Non-manual occupations

Default=clerical; Female fixed-effects



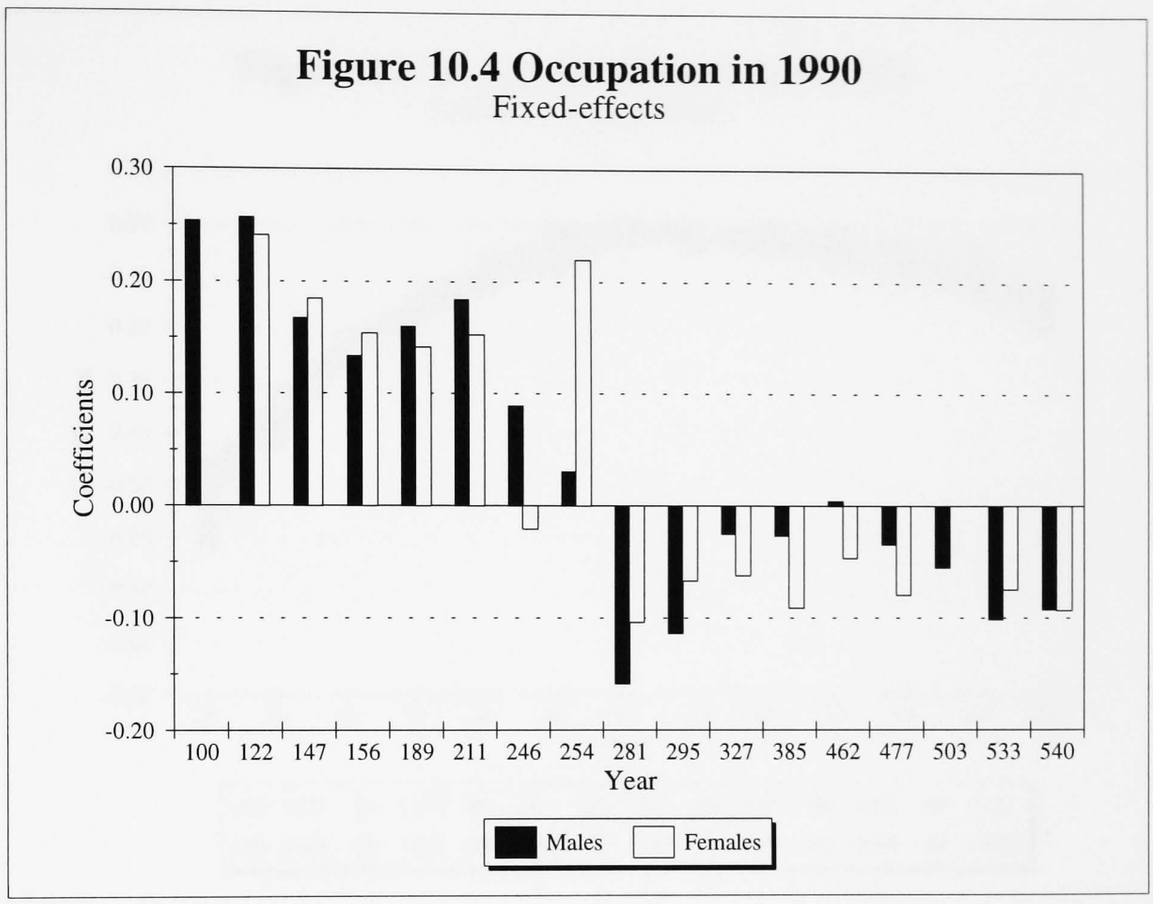


Figure 10.4 Occupation in 1990
Fixed-effects
0.30
0.20
0.10
0.00
-0.10
-0.20
Coefficients
100 122 147 156 189 211 246 254 281 295 327 385 462 477 503 533 540
Year
Males
Females

Figure 10.5a Age profiles(default=35)

Females cross-sections

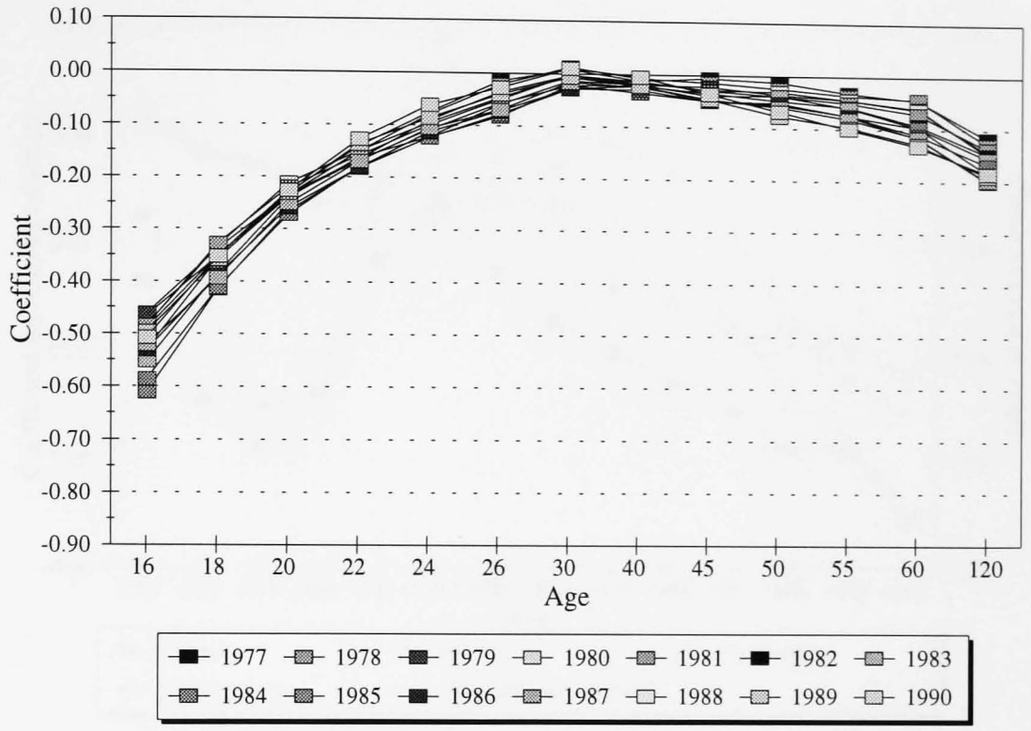


Figure 10.5b Age profiles(default=35)

Males cross-sections

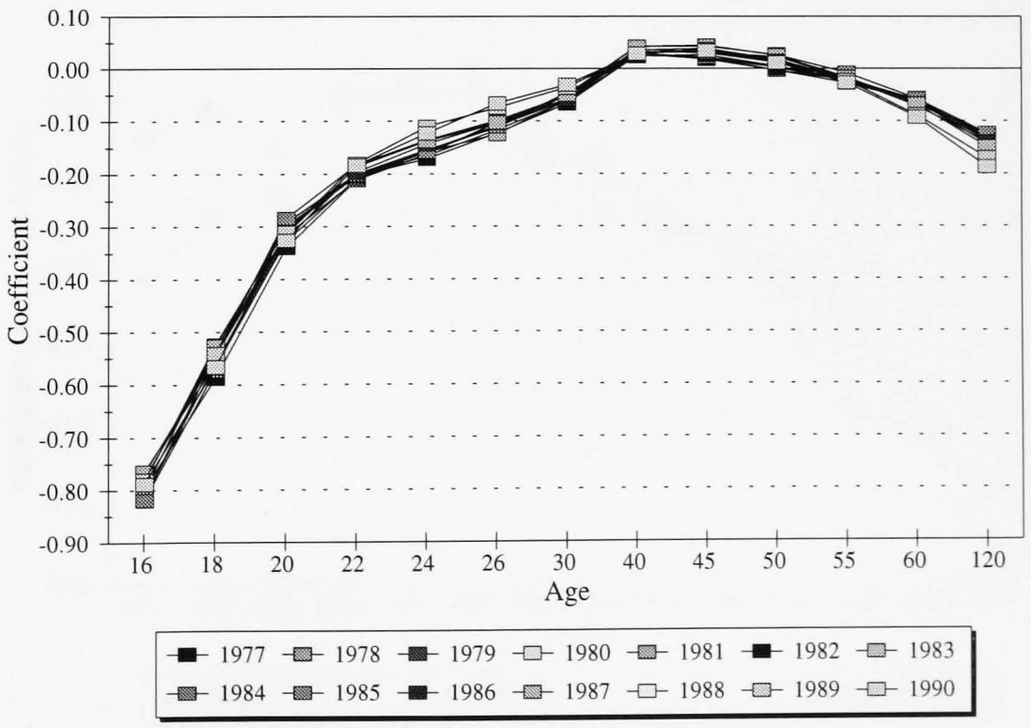


Figure 10.6 Effect of union coverage

Fixed-effects

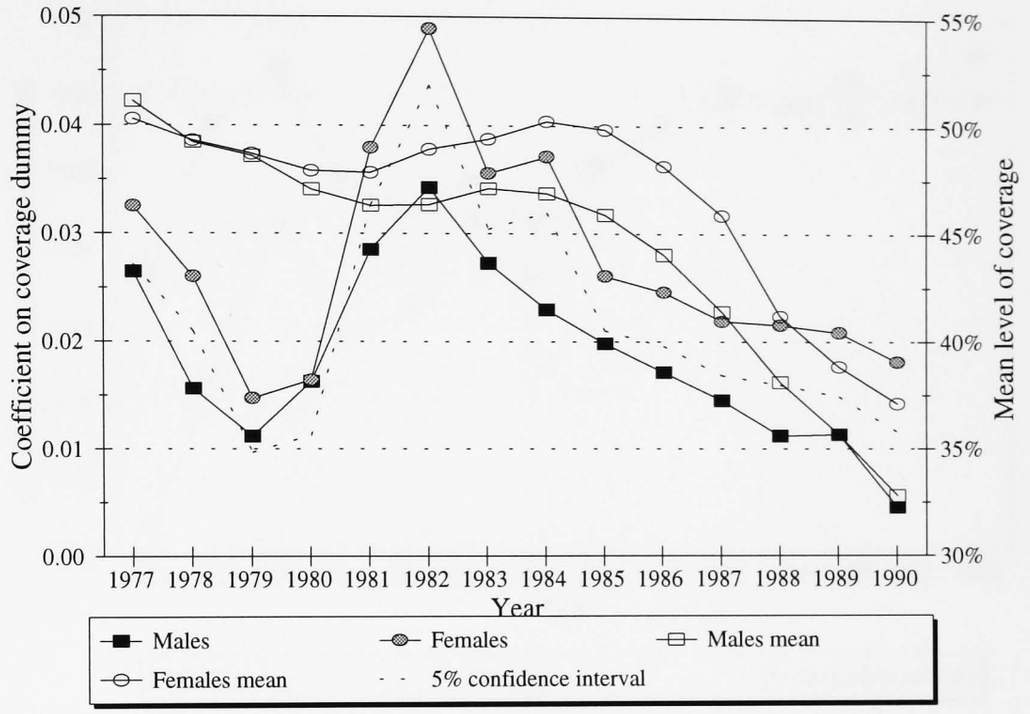


Figure 10.7 Wages Council coverage

Fixed-effects

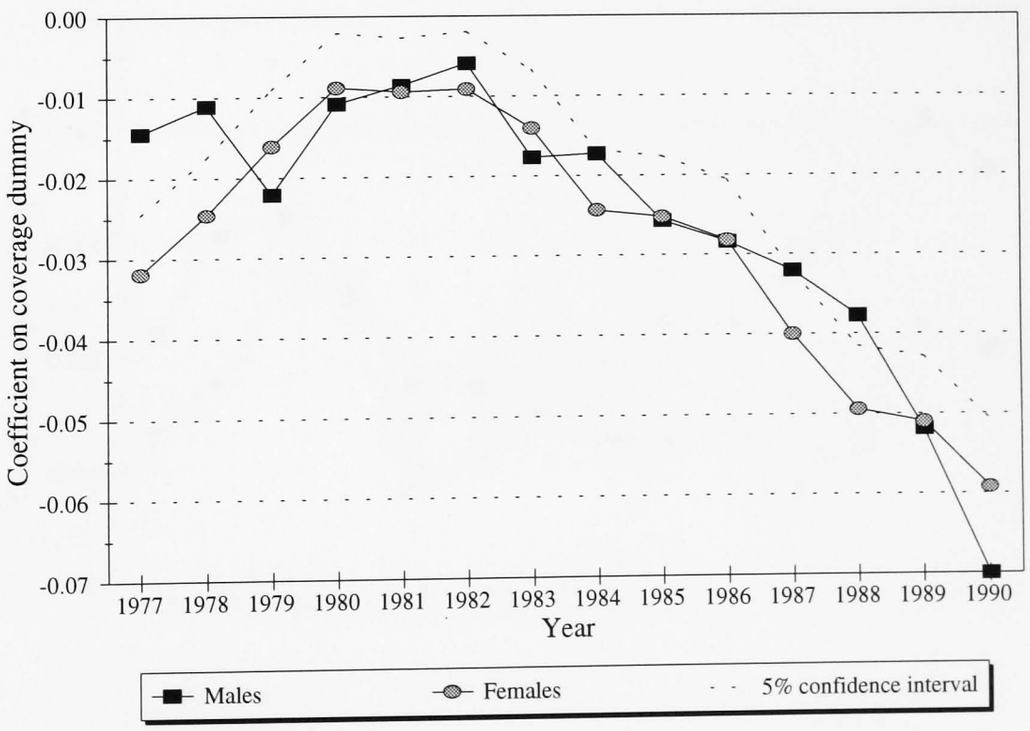


Figure 10.8 Sector (default=Public)

Fixed-effects

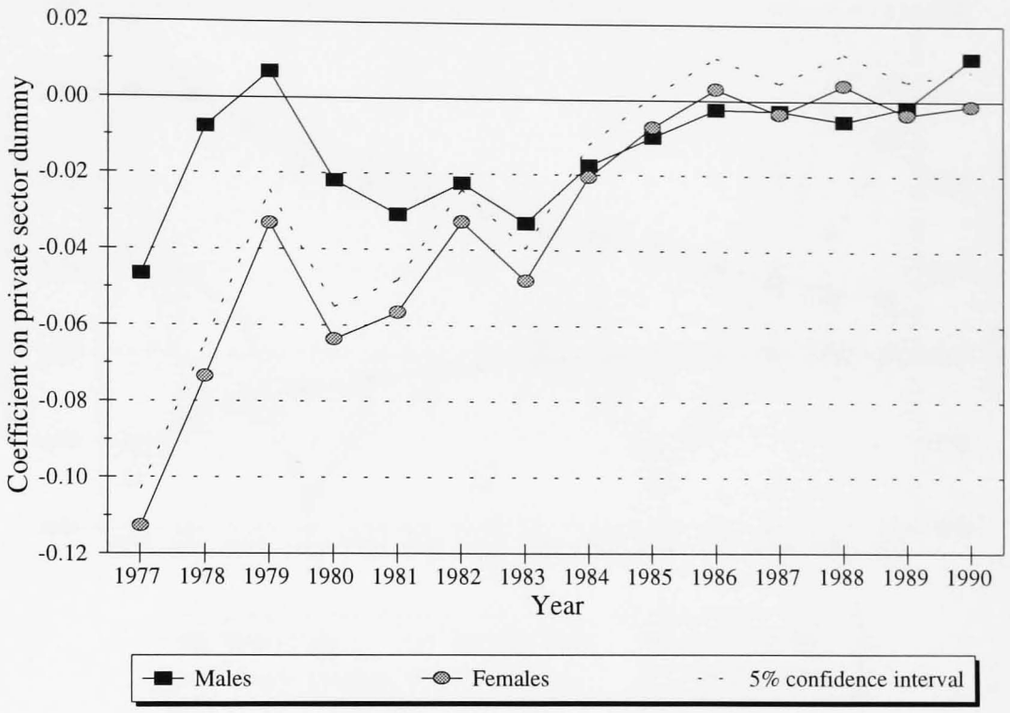


Figure 10.9 Effect of tenure

Default=job held for > 1 year

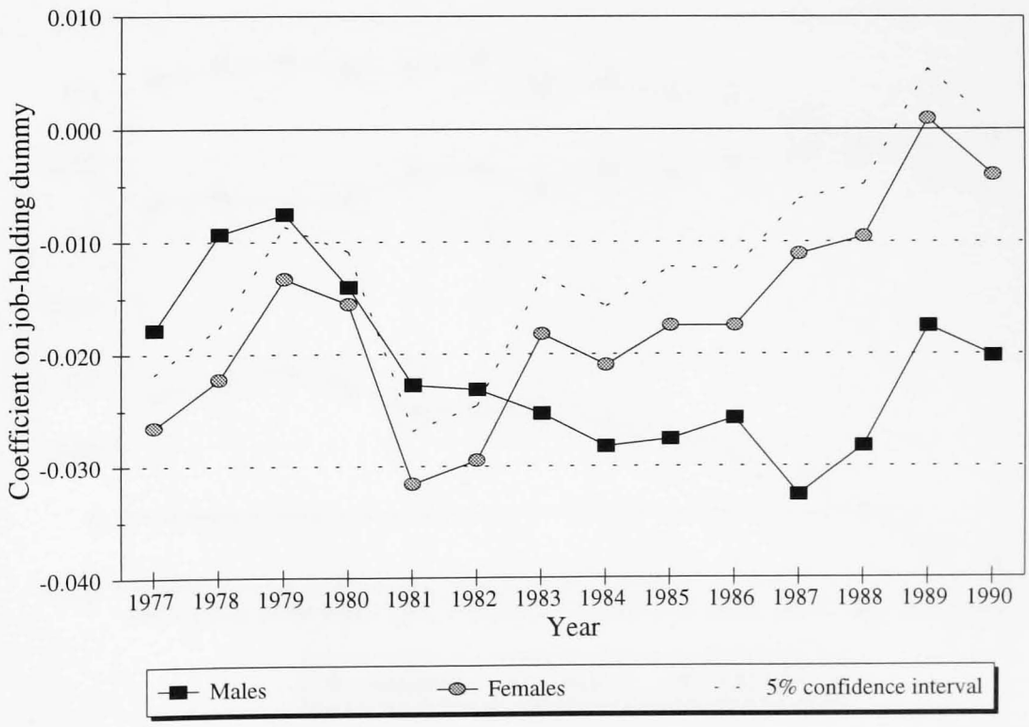


Figure 10.10 Attrition variables

Fixed-effects

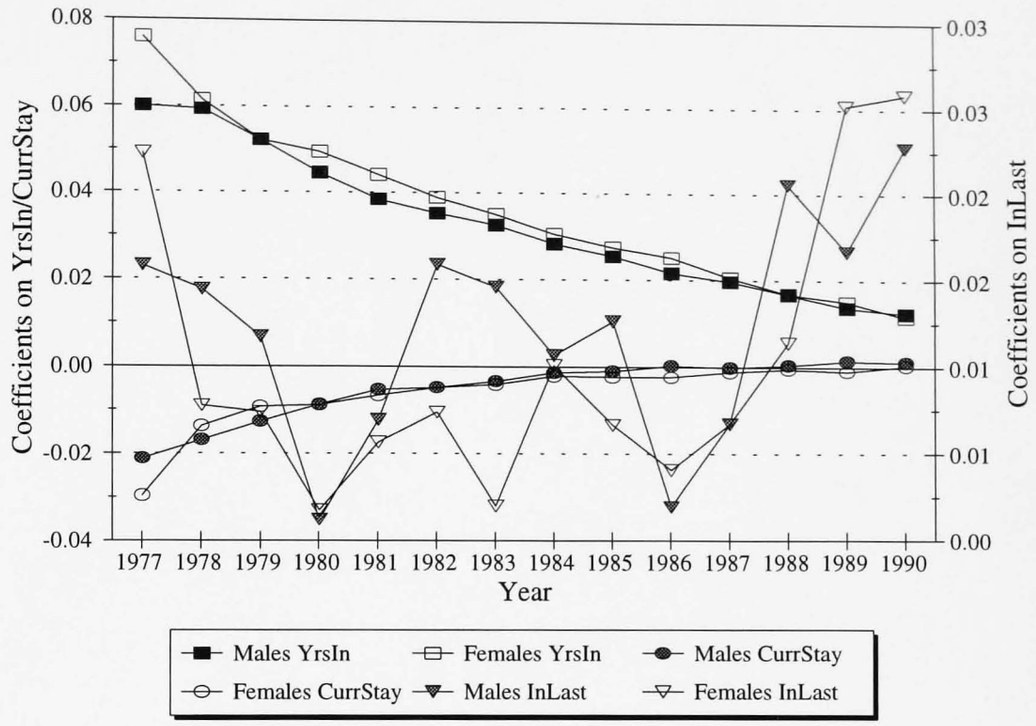


Figure 10.11 Oaxaca breakdown

Fixed-effects; (xm-xf)bm + xf(bm-bf)