Myck, Michal; Reed, Howard

Working Paper

A review of static and dynamic models of labour supply and labour market transitions

IFS Working Papers, No. 06/15

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Myck, Michal; Reed, Howard (2006) : A review of static and dynamic models of labour supply and labour market transitions, IFS Working Papers, No. 06/15, Institute for Fiscal Studies (IFS), London,
https://doi.org/10.1920/wp.ifs.2006.0615

This Version is available at:
https://hdl.handle.net/10419/71593

A Review of Static and Dynamic Models of Labour Supply and Labour Market Transitions

Michal Myck
Howard Reed

THE INSTITUTE FOR FISCAL STUDIES
WP06/15

Labour Supply Estimation Project

Report 1

A REVIEW OF STATIC AND DYNAMIC MODELS OF LABOUR SUPPLY AND LABOUR MARKET TRANSITIONS

Michal Myck and Howard Reed *

This report has been co-financed by HM Treasury, the Inland Revenue, the Department for Work and Pensions, and the Economic and Social Research Council's Research Centre at the Institute for Fiscal Studies. The authors are extremely grateful to Jude Hillary and Victoria Mimpriss from HM Treasury, who managed the research project, and to other named and unnamed officials in HM Treasury, the Inland Revenue and the Department for Work and Pensions who gave useful comments at various stages of the project's development. We would also like to thank the participants of the IFS seminar organised in November 2002 for very helpful advice and comments and to Mike Brewer at the IFS for his help in the final stages of the project.

Howard Reed is now Research Director at the ippr, and Michal Myck is at Deutsches Institut fuer Wirtschaftsforschung (DIW, Berlin).

The authors remain responsible for all remaining errors and omissions. Data from the Family Resources Survey was obtained from DWP and is used with permission; the FRS is also available from the UK Data Archive. Data from the Labour Force Survey, available from the UK Data Archive, is used with the permission of the ONS, and crown copyright material is reproduced with the permission of the Controller of HMSO and the Queen's Printer for Scotland.

Introduction

This report forms the first phase of a project funded by HM Treasury, the Department for Work and Pensions, the Inland Revenue and the Economic and Social Research Council on the design and estimation of labour supply models. This phase – Part 1A of the project – aims to undertake a full review of the techniques and methods which have been developed by researchers to study labour supply and employment, unemployment and inactivity in the labour market. Progress in labour supply modelling in the last thirty years or so has been considerable. Firstly, the theory of labour supply has become much more sophisiticated; simple static-period models of the budget constraint and the hours decision have been augmented with new developments such as intertemporal optimisation, explicit treatment of the participation decision as distinct from the hours decision, and search theory. Secondly, the econometric techniques available to estimate these more advanced models on the data have expanded massively, along with increases in the amount and quality of data available and huge improvements in computing power. In this report we aim to provide a comprehensive survey of the state of the art in the field of labour supply estimation.

Part 1 of this review presents the theory and estimation of labour supply models with the focus of attention on the individual's choice of whether to work or not, and how many hours to work, given a 'budget constraint' which relates gross earnings to net disposable income. Broadly this type of models could be called 'static' as they refer to and rely on the assumption of equilibrium in the labour market. In Chapter 1 we show how these labour supply models are rooted in the concept of rational utility maximisation subject to constraints, and how they can be adapted to deal with labour supply decisions in a family context, and extended from a single-period analysis to an intertemporal optimisation framework. Chapter 2 presents a discussion of how these models are estimated, discussing the specification of the utility function in empirical work, dealing with non-convexities and kinks in the budget constraint, the selectivity issues raised by non-participation, accounting for fixed costs of work, modelling childcare costs, and the modelling of non-take up of benefits and tax credits. In Chapter 2 we look both at the estimation of both structural models described in Chapter 1 and at models which assume equilibrium in the labour market but are not underlined by any specific utility function, and can thus be called non- or semi-structural. The latter present an alternative way for evaluating the labour supply impact of policy reforms and include 'difference-in-differences' methods and methodologies based on random assignment. In Chapter 3 we look at some criticisms of the standard labour supply framework and ask how a theorist might respond to them.

Part 2 is designed to complement the labour supply analysis of Part 1 by focusing explicitly on the dynamics of the labour market – transitions into and out of work, the theories which relate them to financial incentives and other factors, and the way in which the models are estimated on empirical data. We

begin in Chapter 4 with a detailed examination of the various economic theories of work entry and exit and wage growth; human capital – based explanations, search and matching models, and the literature on 'deferred compensation'. Chapter 5 examines the empirical evidence on entry and exit wages, wage progression for those who move into work, the returns to experience and tenure, and the relationship between wages, quits and layoffs and the business cycle. Chapter 6 examines the empirical estimation of models which include an explicit role for labour market transitions in their analysis. These include hazard models, search models, dynamic programming models of schooling and career choice, and the entry-wage based analysis of Gregg, Johnson and Reed (1999).

In the second report from this project (Report 2), we discuss the significance of the theory and techniques which we have examined for the dynamic model of work entry and exit which we have designed in this project. The aim is to incorporate the best practice from current labour supply and transition modelling methodologies whilst recognising both the strength and the limitations of the data available to us.

Part 1 Static Labour Supply Modelling

Chapter 1. The theory of structural labour supply modelling

We start the discussion of different approaches to labour supply modelling by outlining the basic theory of rational choice. The theory is a direct application of the results of demand theory, relies on the assumption of individual rationality and develops a structural model of labour supply, i.e. a model underlined by a specification of a utility function which individuals are assumed to maximise. We first describe the basic principles of the theory in the context of individual labour supply. Extensions to cover joint labour supply decisions of people in couples are presented in section 1.2. Two important features of the structural framework are (i) testability of theoretical predictions and (ii) integrability/identification of the problem, i.e. the possibility of recovering the underlying unique parameters of the utility function. We discuss these with respect to individual and family labour supply in the respective sections.

The structural labour supply theory assumes that an individual's choice of whether to work or not, and the number of hours to work, is a result of utility maximisation in the space of consumption and leisure. Usually the price of consumption is normalised to one, while the individual wage determines the price of leisure. Variation in wages and the number of hours worked among people, under certain assumptions, allows the econometrician to estimate the parameters of the utility function, and the resulting labour supply function, from cross sectional data[1].

1.1 Individual choice on the individual budget constraint

The first section presents the most straightforward decision process, where a single individual chooses his/her optimal combination of consumption and leisure. We discuss the interpretation of three different demand systems: Marshallian, Hicksian and Frisch. The first two are considered both in static and dynamic contexts (sections 1.1.1 and 1.1.2), while the third applies only to choices involving dynamic optimisation (section 1.1.3).

[1] Modelling demand for goods, because of lack of price variation at a point in time requires time-series price information.

1.1.1 Model excluding inter-temporal decisions

Rational choice and restrictions on demand functions

In its simplest representation, the structural approach models individual decisions over hours of work under the assumption of utility maximisation at a point in time. This assumes away any inter-temporal optimisation but provides a useful starting point for the development of the theory. Individuals maximise a utility function over consumption and leisure at time t (the t subscript is suppressed below for notional clarity):

$$U_i(c_i, l_i, x_i) \qquad (1.1)$$

subject to a budget constraint:

$$c_i + l_i w_i = y_i + w_i(T) \qquad (1.2)$$

where w_i is the individual wage, c_i individual consumption, l_i leisure and y_i unearned income. T is the total available time and x_i is a vector of individual characteristics. The price of consumption is normalised to 1. The right hand side of the budget constraint is the so-called “full income” from which the individual purchases consumption and leisure.

Figure 1.1. Indifference curves chart

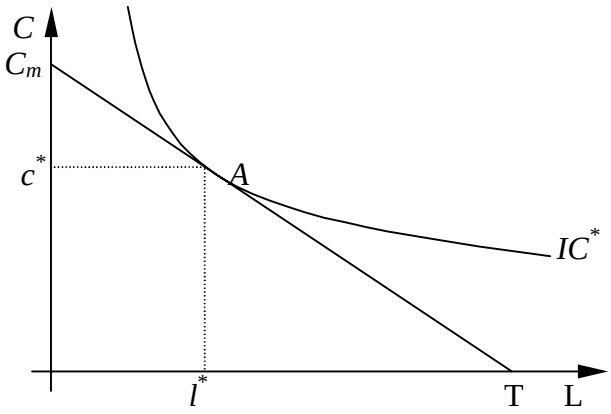


Figure 1.1 shows a graphical interpretation of the above problem. The individual wage determines the slope of the budget constraint (TC_m), while

the "height" of the constraint at zero hours of work (T hours of leisure) is determined by y_i (in Figure 1.1, y_i is zero). $IC*$ – one of the individual's indifference curves – represents a specific level of utility of the individual. This level of utility is achieved when the individual makes an optimal choice between leisure and consumption (c^*, l^*) at point A on the budget constraint. At point A, the marginal conditions are:

$$\frac{\partial U_i}{\partial c_i} = \lambda_i, \qquad \frac{\partial U_i}{\partial l_i} = \lambda_i w_i, \tag{1.3}$$

where λ_i is the marginal utility of money, and the marginal rate of substitution is equivalent to individual wage rate:

$$U_l / U_c = MRS_{Li} = W_i$$

Note, however, that these conditions necessarily hold only when the individual consumes less than T hours of leisure, i.e. if he/she participates in the labour market. In the case of non-participation conditions (1.3) become:

$$\frac{\partial U_i}{\partial c_i} = \lambda_i, \qquad \frac{\partial U_i}{\partial l_i} \geq \lambda_i w_i, \tag{1.4}$$

This is an important difficulty in modelling labour supply. Various results reported below refer to the case where an individual reports positive number of hours worked and thus where conditions (3) hold. We discuss the issue of non-participation in detail in sections (2.1.4 and 2.1.5).

The two marginal conditions (1.3) together with the budget constraint allow derivation of demand equations for consumption and leisure for a specified functional form of the utility function. These demand functions, relating prices of consumption and leisure to the quantity demanded, can be specified under the assumption of either constant non-labour income (uncompensated, Marshallian demand) or constant utility (compensated, Hicksian demand). While Marshallian demands are those actually observed, i.e. the amounts of leisure and consumption individuals choose to enjoy, Hicksian demands play an important role in the process of estimation of the parameters of the utility function and in welfare analysis of price and wage changes.

The axioms of consumer theory, which reflect the assumption of individual rationality, imply several requirements with regard to the two demand functions. These requirements are:

1) Marshallian demand functions are homogenous of degree zero in prices/wages and non-labour income (i.e. if price of consumption and wage change by factor k and non-labour income changes also by factor k, then Marshallian demands do not change)

2) Marshallian demand functions satisfy the 'adding up' property (i.e. the sum of "expenditure" on leisure and consumption equals the "full income")

3) Derivatives of the Hicksian demand functions are symmetric (i.e. $\frac{\partial DH_{ic}}{\partial w_i} = \frac{\partial DH_{il}}{\partial p}$. This results in the symmetry of the 'Slutsky matrix' – a matrix of price derivatives of Hicksian demands)[2]

4) The Slutsky matrix is negative semi-definite (which results from the fact that, given the axioms of consumer theory, for constant utility an increase in price never leads to higher demand)

These four are the complete set of requirements following from the assumption of 'rational' decision making of individuals regarding their optimal choice. If we therefore estimate a demand system which satisfies them, we can say that observed demands have been generated from a 'rational' set of individual preferences. One can therefore choose a demand system to be estimated without necessarily specifying the underlying utility function. On the other hand if a utility function is specified (and it fulfils all requirements of a 'rational' set of preferences, see MasColell (1995)) and we estimate a demand system which is implied by it, then failure to satisfy any of the four requirements might call into question either the assumptions which underlie our chosen functional form of the utility function or the 'rationality' of individuals as defined by the axioms of consumer theory. The four above requirements are thus extremely important in the choice of modelling strategy and the evolution of consumer theory.

Testability

The aim of "structural" estimation is to find parameters of the demand and/or utility functions. Knowing either of these allows us to calculate elasticities of demand which in turn make predictions of individual responses to changes in the prices of consumption and leisure possible. Results of these estimations are in many cases supposed to assist in an ex-ante assessment of effects of changes in the budget constraint, either as a result of changes in out of work/unearned income or in net wages. Both of these are affected by the design of fiscal policy, and therefore the correct estimation and interpretation of the labour supply elasticities are of crucial importance for ex-ante evaluation of reforms of the tax and benefit system.

Estimation of the demand function for leisure is based on the specification of an hours of work equation. The functional form of this equation is a direct consequence of utility maximisation, and precise specification of the hours

[2] Symmetry follows from the fact that price derivatives of Hicksian demands corresponds to the a matrix of second derivatives of the "expenditure function" in the dual representation of the optimisation problem (see: Deaton & Muellbauer, 1980)

equation is directly linked to a chosen specification of preferences. Alternatively we can interpret the hours equation as imposing restrictions on preferences[3]. As we mentioned above, establishing a direct relationship between the hours equation and the utility function is not necessary to test whether a demand system is a consequence of rational optimisation. However, it is important to establish such a relationship to facilitate our interpretation of the results, and our understanding of the assumptions underlying the estimation.

Given unearned incomes and wages, we can estimate the elasticity of labour supply. Taking the general representation of the hours equation to be:

$$H_i = H(w_i, y_i, x_i),$$

the wage elasticity of uncompensated, Marshallian, demand is:

$$\varepsilon_M = \frac{\partial \ln(H_i)}{\partial \ln w_i}$$

This is the overall effect of the change in net wage w_i on hours worked. Marshallian demands can be decomposed into income and substitution effects. This separation allows derivation of the basic tests of the underlying assumptions of individual rationality. At constant levels of utility, at which Hicksian demands are defined, because it is only the change in price that is taken into consideration, a relative price increase never leads to higher demand – in other words the substitution effects are always negative (thus the requirements placed on the Slutsky matrix, as shown above). In the labour supply example, an increase in the wage (i.e. increase in the price of leisure) should never lead to higher compensated (Hicksian) demand for leisure or lower supply of labour. Separation of the two effects of price changes gives us the compensated demand. The (wage) elasticity of this demand is:

$$\varepsilon_H = \varepsilon_M - \frac{w_i h_i}{y_i} \frac{\partial \ln(H_i)}{\partial \ln y_i} \qquad (1.5)$$

The share $w_i h_i / y_i$ is the size of earnings relative to non-labour income. This is weighted by the responsiveness of hours to changes in unearned income. The second term on the right hand side of equation 1.5 is the income effect of the change in the price of leisure. If the observed hours choices are based on individual rationality, the matrix of compensated demand price elasticities should be negative semi-definite (because of the sign of Hicksian demand elasticities) and symmetric. Therefore, the empirical estimation of labour supply functions does not end with simply finding the values of coefficients. Having estimated them, and knowing the overall "Marshallian" wage elasticity and income elasticity, one can test the two implications of the theory (requirements 3 and 4 above) using (1.5). These tests have found ample

[3] We discuss the relationship between various utility functions and hours equation in section 1.3.

application in microeconomic theory both in the context of demand for goods (consumer theory), and in labour supply.

Integrability

If an estimated demand system fulfils the four requirements imposed by consumer theory, one can recover the preferences that generate it. We can therefore recover the parameters of the utility function which underlies individual choices. Individual optimisation can be represented either as the maximisation of utility subject to a budget constraint or alternatively as the minimisation of expenditure for a given utility level. It is the latter representation of the problem which allows us to recover the parameters of the utility function. Observed Marshallian demands correspond to the derivatives of the expenditure function with respect to prices/wages:

$$\frac{\partial e_i(\mathbf{p},U_i)}{\partial p} = c_i(p, w_i, M_i)$$
$$\frac{\partial e_i(\mathbf{p},U_i)}{\partial w_i} = l_i(p, w_i, M_i)$$

where $e_i(\mathbf{p},U_i)$ is the expenditure function and $\mathbf{p}$ is the price vector (including the price of consumption and individual wage). This relationship, together with satisfaction of the four requirements on demand functions, allows recovery of the parameters of the expenditure and utility functions.

Elasticities in the static framework

One of the most common hours equation specifications is:

$$\ln h_i = \alpha \ln w_i + \beta Q_i + v_i \qquad (1.6)$$

where Q_i is a vector of control variables including an income variable, while v_i is a vector of unobserved random effects (for other specifications and utility functions they derive from, see section 2.1.1). Interpretation of the coefficient on log wages significantly depends on the variables included in Q_i. The most basic static specification of the hours equation includes a set of “taste-shifting” variables and a measure of non-labour income, y_i. In this context the α is the uncompensated substitution elasticity given income y_i.

This interpretation is valid under the assumption of static optimisation where the individual makes no inter-temporal decisions. Thus the assumption is either that the individual is extremely myopic and optimises only in one period at a time, or that he/she is not allowed to make inter-temporal decisions as to the allocation of income, for example due to inability to lend and borrow. All income received in period t is spent in this period.

Clearly in the context of labour supply, where inter-temporal optimisation is a natural way of thinking about people's choices, such a specification has significant drawbacks. When deciding whether to work or not and how much, individuals are likely to take into account the effect of additional schooling, of experience and training on future earnings. Their choices are likely to depend on expected future non-labour income, on the rate of return on assets, liquidity constraints, and so on, and will include allocation of assets in addition to intertemporal leisure/work decisions. We turn to extensions of the static framework to account for these below.

1.1.2 Marshallian and Hicksian elasticities of labour supply in dynamic context - the two-stage budgeting problem[4]

Many empirical applications of the structural model of labour supply have used the extended, more realistic version of the structural model which takes account of inter-temporal optimisation. Considering choices across time involves one fundamental assumption which makes the problem tractable and solvable. Only *levels* of utility at time t are assumed to have an impact on the allocation of leisure and consumption at time $t+s$. The *combination* of leisure and consumption which leads to a given utility level at time t is assumed to be irrelevant for the choice at time $t+s$. This assumption of *intertemporal separability* leads to the following life-time utility function[5]:

$$U_{it} = U\left(U^{t}(c_{it}, l_{it}, x_{it}), U^{t+1}(c_{it+1}, l_{it+1}, x_{it+1}), ..., U^{\tau}(c_{i\tau}, l_{i\tau}, x_{i\tau})\right) \qquad (1.7)$$

Apart from this assumption the simple structural dynamic model assumes away credit constraints, allowing the individual to borrow freely against future income. This allows the following specification of the inter-temporal budget constraint, represented by a time path of assets:

$$A_{t+1} = (1 + r_{t+1)})(A_t + B_t + W_t H_t - C_t), \qquad (1.8)$$

where
A_{t+1} is the real value of assets at the beginning of period $t+1$,
r_{t+1} is the real return on assets,
B_t represents unearned-non-asset income.

Because of the assumed separability of the utility function, the marginal conditions for utility maximisation are the same as in the static model (eq. (1.3) and (1.4)), and in case of participation the condition

$$U_{lt} / U_{ct} = MRS_{Lit} = W_{it} \qquad (1.9)$$

4 This and the following section draw on Blundell and MaCurdy (1999).
5 Dropping the assumption of intertemporal separability from the model would allow interaction between levels of leisure and consumption in different periods leading to a more general utility function: $U_{it} = U(c_{it}, l_{it}, x_{it}, c_{it+1}, l_{it+1}, x_{it+1}, ..., c_{i\tau}, l_{i\tau}, x_{i\tau})$

still holds.

The solution follows from a two-stage process, where in the first stage individuals optimally allocate their full life-time income to each period and then maximise utility at each point as in the static case. The first stage allocation allows estimating the level of consumption and leisure at each point given the marginal condition (1.9).

Full income in period t in this dynamic context is again equal to consumption in period *t* plus the value of leisure at time t. This however no longer needs to equal the total income received in period t. The difference in the definition of full income between the static case and the two-stage budgeting problem follows from allowing inter-temporal allocation of income. Full income M_t is the sum of consumption and leisure at time t, but this is now dependent on decisions concerning the allocation of assets over time:

$$M_t = c_t + w_t l_t = r_t A^*_{t-1} + \Delta A^*_t + B_t + w_t T$$

where $r_t A^*_{t-1}$ is the real interest income available for expenditure on consumption at the beginning of period t and ΔA^*_t is the adjustment of the level of real assets by the end of period t.

Because of inter-temporal decision making, consumption (and thus the allocation of assets in period t) will depend on all future (expected) values of wages and unearned income which influence the change in real assets ΔA^*_t. M_t will therefore be a function of:

$$M_t = M(A^*_{t-1}, r_t, w_t, B_t, x_t, Z_t)$$

where Z_t represents the future (known or expected) values of w, B, x, and r.

Non-labour income in period t is now:

$$y^c = r_t A^*_{t-1} + \Delta A^*_t + B_t$$

which is equivalent to:

$$y^c = c_t - w_t h_t \qquad (1.10)$$

(1.10) can be calculated if we know consumption and hours of leisure/work. Note, however, that y^c will be a function of expectational variables Z_t, and some specification of these variables will be necessary to complete the model.

Interpretation of the wage equation (1.6) in the two-stage budgeting case differs from that in the static framework. The wage elasticity coefficient is now

conditional on initial allocation of income and consequently on y^c. It therefore captures the effect of anticipated changes in wages through time but does not capture unanticipated changes in the overall life-time wage profile. This is because changes in the wage profile will have an impact on hours worked also through changed allocation of y^c.

Note, also, that because y^c is a function of leisure it will be endogenous to hours and appropriate estimation will require the use of instrumental variable techniques. Because c is a function of future wages and unearned income it will no longer be exogenous, as it will respond to changes in wages via the first stage process of optimal allocation of income.

1.1.3 Frisch labour supply function - dynamics assuming constant marginal utility of wealth.

Apart from Marshallian and Hicksian demand functions for leisure, the third representation of the labour supply problem is in the form of Frisch functions. While Marshallian demand functions assume a constant level of non-labour income, and Hicksian demand functions keep utility at a constant level, Frisch functions take account of the dynamic nature of the problem by keeping constant the marginal utility of wealth.

As in the two-stage model, the solution relies on the assumption of strong separability in preferences and availability of credit to facilitate optimal allocation of assets. In this model the marginal utility of wealth serves as the sufficient statistic which gives the solution of current-period's maximisation problem.
Individuals are assumed to maximise a value function:

$$V(A_t, t) = \max[U(c_t, L_t, X_t) + \kappa V(A_{t+1}, t+1)]$$

subject to the time path of assets (1.8). κ is the individual's discount factor. The solution to this dynamic programming problem gives the usual first order conditions and an Euler equation for the marginal utility of wealth:

$$\lambda_t = \kappa(1 + r_{t+1})\lambda_{t+1} \qquad (1.11)$$

The Euler equation reflects the optimal distribution of assets. The solution to this problem is a pair of demand equations for consumption and leisure which are functions of the wage and other usual control variables as well as λ_t. Future values of wages, other income, etc. affect consumption and labour supply only through their effect on the value of marginal utility of wealth. As we can see from equation (1.11) the path of λ_t depends only on the individual discount factor and the interest rate, and is independent from wage. λ_t will not change if the individual's wage changes as expected. His/her labour supply on the other hand might respond to it. This means that the Frisch elasticity of

labour supply is the correct elasticity for analysing the impact of changes of wages through time.

Empirical studies of labour supply look usually not at changes in individual wages through time, but at differences across individuals. Because of the lack of suitable panel data with a long enough time-series dimension in the UK in particular, in many cases the empirical studies do not observe specific wage profiles of individuals through time, but instead a cross-section of results of assumed individual optimisations which take these individual wage profiles into account. For each individual these profiles certainly influence the value of λ which means that for the Frisch elasticity to have economic meaning we have to account for the effect of the full wage profile on λ. We take up this issue below.

Frisch elasticity

We can show that λ_t can be represented as a combination of an individual-specific fixed effect and a common time path. If we assume that interest rates and individual discount rates are the same across time, taking logs in eq. (1.11) we can specify λ as:

$$\ln \lambda_t = b^* t + \ln \lambda_0$$

Using the hours equation specification (1.6), we can now condition wages on λ in the following way. Assuming the utility function to be:

$$U_t = G(c_t, x_t) - (\exp(\varphi x_t - \nu_t))(h_t)^{\sigma}$$

where G is an increasing function of c and $\sigma > 1$ is a time invariant parameter common across consumers, equation (1.6) becomes:

$$\ln h_i = \alpha \ln w_i + bt + \varphi x_t + \alpha \ln \lambda_0 - \alpha \ln \sigma + v_i \qquad (1.12)$$

Because the time path is common to everyone, this allows estimation of the demand equation in first differences. Equation (1.12) then becomes:

$$\Delta \ln h_i = \alpha \Delta \ln w_i + b + \varphi \Delta x_t + \Delta v_i \qquad (1.13)$$

Equation (1.13) can be estimated to yield a value of α - the Frisch elasticity which is interpreted as the intertemporal elasticity of substitution. It represents the hours response to expected “evolutionary” changes in wages.

In the example above we assumed perfect certainty about individual wage profile in the future. A similar analysis can be conducted in the case where future outcomes are uncertain, although in this case the interpretation of the error terms changes and additional assumptions on the utility function are needed (see Blundell & Walker (1986), Blundell & MaCurdy (1999)). The

introduction of uncertainty is undoubtedly crucial if we want to extend our analysis to include responses to unexpected shocks to wages or unexpected changes in whole wage profiles. One reason for changes in the latter are changes in the design of the tax and benefit system. Individuals' responses to these changes are of obvious interest from the perspective of policy evaluation. To account for unanticipated changes in individual's wage profiles we need an empirical specification for λ. This is also necessary for the Frisch elasticity to be interpretable as the effect of wage variation across individuals in a cross-sectional estimation.

Responses to changes in wage profiles

Let us specify an approximation for $\ln \lambda_0$ as:

$$\ln \lambda_0 = \sum_{j=0}^{\tau} \phi_{0j} E_0(\ln w_j) + D_0 \delta_0 + \theta_0 A_0 + e_0 \qquad (1.14)$$

where
D_0 - a vector of demographic characteristics observed or anticipated at time 0,
e - an error term.
The individual is assumed to know to work up to period τ.

This can be incorporated into (1.12) to give:

$$\ln h_i = (\alpha + \alpha\phi_{0t}) \ln w_i + bt + \varphi x_t + \alpha \left[\sum_{j=0, j\neq t}^{\tau} \phi_{0j} E_0(\ln w_j) + D_0 \delta_0 + \theta_0 A_0 \right] - \alpha \ln \sigma + \omega_i \qquad (1.15)$$

This hours equation specification conditions wage responses on determinants of λ. Estimation of (1.15) requires an appropriate specification of the individual's expected future wages and initial wealth. Given these, it allows us to interpret the coefficient on wage $(\alpha + \alpha\phi_{0t})$ as the response to an evolutionary change in the individual's wage rate at period t in a cross sectional analysis. The estimation also yields an estimate of the response to a shift in the entire wage profile which is equal to:

$$\overline{\alpha} = (\alpha + \alpha \sum_{j=0}^{\tau} \phi_{0j})$$

$\overline{\alpha}$ is the parameter of interest for the estimation of labour supply responses to tax and benefit policy changes. This estimation can be conducted in differences using panel data information, or on cross sectional data with appropriate instruments and control variables.

1.1.4 The "wage elasticity"

As we saw in the sections above "wage elasticity" is by no means a straightforward concept. Its value and interpretation will differ depending on what control variables we choose in the hours equation. Without an appropriate specification, the coefficient on wage may completely lack economic interpretation. The most important distinction in interpreting elasticities is that between within-period elasticities and life-cycle elasticities. The first group includes compensated and uncompensated elasticities in the static and two-stage budgeting models, while the second group comprises elasticities corresponding to responses to evolutionary and parametric wage shifts. The inter-temporal elasticity of substitution measures responses to evolutionary changes along an *individual* wage profile. This differs fundamentally from responses to unexpected changes involving shifts in the *entire* wage profile. Analysis and interpretation of the coefficients on the 'wage' in any labour supply model must thus be extremely careful and take into account the conditioning variables in the hours equation.

1.2 Individual choice on the "joint" constraint – the labour supply of couples

Modelling the labour supply of a couple introduces complexities absent from the application of the standard theory for single individuals presented in the previous section. Analysis of the labour supply decision of two people living together involves additional complexity, due to the question of the distribution of the arguments of the utility function – leisure and consumption – between the partners, with possible externalities in both arguments. The question who should pay for the provision of "public goods", such as expenditure on children, the electricity bill, etc. further complicates the labour supply story.

Below we start the discussion of family labour supply modelling by outlining a model which would be a natural extension of the methodology used for individual labour supply modelling. Following the discussion in section 1.1 we restate what requirements such an extension would have to satisfy. We describe the most general "collective" model that in principle could fulfil these requirements and then point to the difficulties which the theory is faced with when applied to the available data.

Following this, we describe the traditional "unitary" approach to modelling family labour supply. In the unitary model, the household choice is presented as a result of maximisation of a single aggregate utility function. Next, we contrast the unitary model with recent developments in modelling labour supply of people in couples in a "collective" framework where the preferences of the individuals in the couple are treated individually rather than being

aggregated. We finish by presenting the implications and relative advantages of these two approaches to modelling the labour supply of couples. On the one hand we present the usual criticisms of the unitary model, and on the other the difficulties involved in applying the collective model to the available data.

The dynamic issues surrounding family labour supply are far more complex than those surrounding individual labour supply, as they include decisions to marry, separate, timing of children, etc. Because of this the theory of family labour supply has so far only focused on the static case corresponding to section 1.1.1. Below we present an extensive summary of this literature.

1.2.1 Labour supply of individuals in couples - the general framework

Neoclassical labour supply theory is based primarily on the methodological principle of individual rationality. In extending the theory to the labour supply decisions of couples we would expect this principle to continue to hold. Further formal requirements we would expect the theory to fulfil are the same as those mentioned in reference to individual labour supply and refer to the ability of the models to generate testable predictions and to allow the recovery of preferences underlying the choices and determining the observed outcomes.

We would therefore expect both partners to maximise their individual utility function, allowing for complementarities and substitutability in leisure and consumption between them, subject to the overall family budget constraint. Allowing for consumption of private (c_m, c_f) and public goods (q) the individual utilities would be:

$$U_i\left(c_f, c_m, l_f, l_m, q\right) \quad , i = m, f \tag{1.16}$$

At this point it is important to stress that in the discussion which follows, the definition of types of goods is absolutely crucial. In specification (1.16) the distinction between private and public goods, along the usual lines, becomes almost superfluous. This is because person i's consumption of "private" goods (c_m, c_f) directly enters the utility function of person j. Below we introduce a different distinction between goods, the applicability of which will become clear when we discuss alternative specifications of preferences (see: Browning, et. al. 1994).

Because in the general formulation of the household's behaviour shown above we do not aggregate the preferences of each member of the couple into a single utility function of the couple, we have to account for the process that governs the allocation of leisure and consumption between them. That is, we have to identify what distinguishes two people living together from two separate single individuals. The outcomes of such a process could either be co-operative or non-co-operative. Several attempts have been made to derive

results in the non-co-operative game-theoretic framework, assuming in most cases Nash solutions to the bargaining process (e.g. Manser and Brown (1980), McElroy and Horney (1981)). Here, following Chiappori (1988, 1992) we focus on co-operative games, which are believed to be a more appropriate framework for modelling the decisions of people living in couples. This literature starts with the assumption that rational individuals living in long-term partnerships would arrive at Pareto efficient solutions to their bargaining process – that is, the outcome would be such that it would be impossible to make one person better off without making the other worse off[6]. The assumption of Pareto efficiency does not impose any restrictions on the distribution of consumption between the individuals. Efficient outcomes can be fully egalitarian or extremely unequal.

Let the household budget constraint be:

$$c + l_f w_f + l_m w_m = y_f + y_m + y_s + w_f(T) + w_m(T) \qquad (1.17)$$

where c is total household consumption (the price of which is normalised to 1), $w_{m,f}$ are wages, T is total available time, $y_{m,f}$ are individual non-labour incomes assignable to members of the couple, and y_s is non-labour income which cannot be assigned to the members individually. Total household non-labour income is then: $y = y_s + y_m + y_f$.

If λ denotes the welfare weight of the woman in the couple and $(1-\lambda)$ the weight of the man, then, if individual utility functions are strictly concave and if the household budget constraint is convex, by application of the second theorem of welfare economics we know that Pareto efficiency will be achieved as any solution to the following maximisation problem:[7]

$$\textit{Max} \qquad \lambda * U^f(c_f, c_m, l_f, l_m, q) + (1-\lambda) * U^m(c_f, c_m, l_f, l_m, q) \qquad (1.18)$$

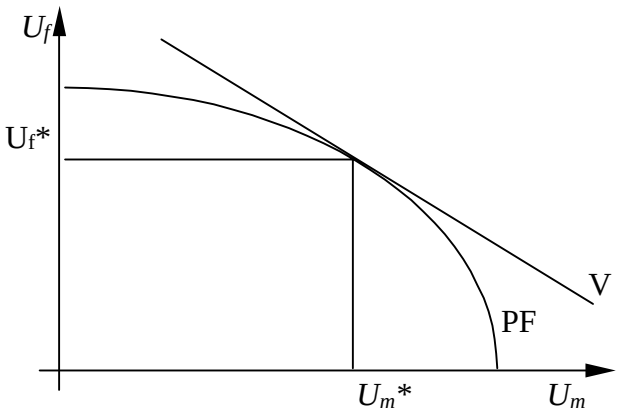
subject to (1.17).

The determinants of the utility weights form a crucial part of this framework. Figure 1.2 shows an example of an optimal choice by a couple. The Pareto frontier is determined by the household unearned income, earnings and individual preference parameters. At each point of the frontier, one member of the couple maximises his/her utility given the utility level of the other partner. The welfare weight λ determines the slope of line V representing a certain welfare level of the household. It is the highest possible level of welfare achievable on the Pareto frontier and thus the point which is chosen as the solution to (1.18).

[6] Note that this would seem to rule out certain behaviour observed in couples where one member of the couple obviously becomes worse off while the other becomes no better off (e.g. domestic violence) unless we assume (distastefully) that the aggressor or the victim in the couple derives positive utility from such activities.

[7] Budget constraint non-convexities, which result from various social security programs and different forms of negative taxation, may lead to non-convex utility possibility set. This non-convexity is sometimes questioned on the ground that rational agents could ensure convexity by randomization between different points of the Pareto frontier.

Figure 1.2. Pareto efficient decisions of individuals in couples.



Notes: PF – Pareto frontier, V – household welfare function corresponding to λ. The Pareto frontier is determined by the household unearned income, earning potential and parameters of individual utility functions. Problem (1.18) corresponds to choosing an 'optimal' combination of U_m and U_f. The 'optimality' is determined by λ.

Interpretation, identification and testability of the theory relies on what enters the function λ. It is usually assumed that the distribution of the weights assigned to individual utilities is a function of their actual or potential wages, non-labour incomes, and the so-called "distribution factors", z (Bourguignon, et al., 1994):

$$\lambda = g(w_m, w_f, y_m, y_f, z) \tag{1.19}$$

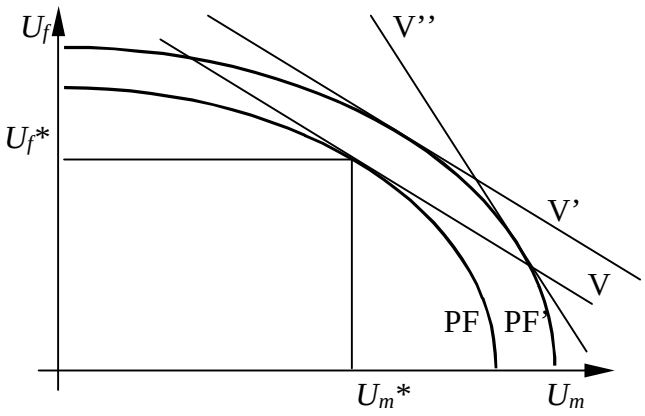
The "distribution factors" are factors which determine the welfare weights within the household but do not have a direct effect on either preferences or the budget constraint.[8] If we consider a situation in which the budget constraint of the household changes, the new optimum may be determined not only by the resulting change in the Pareto frontier. If factors affecting the budget constraint also determine the distribution of resources between partners (λ), a change in the gross wage of one partner, or a reform affecting net wages, will shift the Pareto frontier *and* change the household welfare function V.

Let's consider an increase in wage of one of the members of the couple. As a result of this change, the utility possibility (Pareto) frontier shifts out since the "full income" of the couple increases. If wages do not affect the distribution of resources the welfare weights λ and (1- λ) of the two members are unaffected. The slope of the welfare function (the slope of line V on Figure 1.2) in this

[8] Examples of distribution factors could be individual unearned incomes, divorce laws, alimony payment laws - see Bourguignon, et. al, 1993, 1994.

situation therefore does not react to the change in wage. The optimal choice would be made at the point of the new Pareto frontier (PF' on Figure 1.3), at the higher level of welfare represented by the point of tangency between PF' and V' on Figure 1.3. In the world in which a change in wages alters the welfare weight λ we would observe a change in the slope of the welfare function. The new solution is therefore made up of two effects: a change in the Pareto Frontier and a change in the distribution of welfare among the partners. On Figure 1.3 this is represented by the tangency of PF' and line V''. Line V'' represents the new level of the couple's welfare.

Figure 1.3. A change in wage in the unitary and collective model.



1.2.2 Testable restrictions of the general collective framework

In the formulation of the household utility function (1.18), wages enter not only as prices of leisure but also directly by determining the weights assigned to every member of the couple as shown in (1.19). As demonstrated by Pollak (1977), with such price dependent preferences the restrictions of classical demand theory are no longer implied by rational choice. The Slutsky matrix does not have to be either symmetric nor negative semidefinite, and thus the standard results which facilitate tests and allow integrability in the individual demand framework no longer apply.

A crucial result which allows testability of this general framework has been developed by Browning and Chiappori (1998). Equivalent tests to the symmetry and negative-semidefiniteness of the Slutsky matrix[9] are derived regardless of the specification of the individual preferences, and regardless of how consumption and leisure enter the utility function. However, the tests can only be carried out for cases where we observe demand for more than four parameters in each of the utility functions (more than four goods in the case of demand modelling). This obviously presents difficulties for models of labour

[9] Browning and Chiappori (1998) demonstrate that in the collective setting the equivalent of the Slutsky matrix is a sum of a symmetric and negative semi-definite matrix and a matrix that has at most rank one.

supply, where we only observe demand for three “goods” – household consumption, male leisure and female leisure.

Other tests of the most general collective framework rely on the “distribution factors” as defined above. They provide the possibility of testing the collective model without any additional restrictions on preferences or nature of “goods”, and unlike the tests in Browning and Chiappori (1998) they do not require observability of five demands. This is a significant advantage from the perspective of modelling family labour supply.

Since the tests outlined above are tests of the most general formulation of the collective framework, first and foremost they allow rejection of the overall collective approach. Rejection of other tests in cases where more restrictions are made on preferences or “goods”, without rejection of “symmetry and rank one” (Browning and Chiappori, 1998) and/or “distribution factor proportionality” (Bourguignon, et al., 1994), would then question only the specific additional restrictions without necessarily casting doubt on the overall framework.

Unfortunately, despite the possibility of testing the general framework of the collective model as outlined above, this set up does not allow us to recover the underlying parameters of the two utility functions. The difficulty lies in the already mentioned implication of this representation of preferences for the nature of “goods”. The “public” character of the consumption and leisure of both partners renders the identification of the model impossible. This is one of the principal reasons behind the formulation of the traditional “unitary” model of family behaviour and of the “sharing rule” interpretation of the “collective” approach (described below). Additional assumptions made in both of these allow recovery of the underlying preferences and formulation of testable restrictions of the theories.

1.2.3 Families in the “unitary” framework

The traditional framework of modelling demands (or labour supply) within couples assumes that the individual preferences of the two members of a couple are combined into a single utility function:

$$U = u\left(c, l_m, l_f\right), \tag{1.20}$$

A special case of the above model, in which consumption and leisure combinations of each partner are separable in the joint utility function, would be equivalent to assuming λ in (1.18) to be constant. Such a model, however, is unidentified because of lack of variation in λ. All known to us applications of the unitary model rely on the formulation of a single utility function.

For obvious reasons the distinction between private and public goods no longer applies in this case as the couple is treated as a single decision maker.

The arguments in this utility function are household consumption and the leisure of both partners. This ad hoc transformation of two individual utility functions into one family preference ordering finds no justification in economic theory and indeed has been a major reason for criticism of the unitary model. The model has also been criticised for departure from the methodological foundations of micro-economics in its departure from individual rationality, and attempts to justify such a departure (by for example Samuelson (1956) and Becker (1974))[10] have proved unsuccessful.

The unitary framework treats the family as a "black box" – processes within the family are considered either irrelevant to the policy-maker or optimal relative to the policy-maker's welfare function. Therefore welfare analysis can only be applied to inter-household distribution and intra-household processes are therefore absent from considerations on policy reform. Such an approach to family behaviour is also incapable of shedding light on such important issues as family formation and divorce.

Of course the unitary model does have some advantages. The framework is very close to that used for modelling behaviour of individuals (see section 1.1.1), and can hence be used to deal with problems of non-participation, fixed costs of work and non-linear budget constraints by simply re-using the econometric methods which already exist in the single-person case. A further methodological advantage of the unitary approach over the general framework (1.16) presented above is that it allows derivation of additional testable restrictions. Because prices no longer enter the utility function directly the standard Slutsky restrictions apply and can be tested empirically. Moreover, the model implies that the effect of increases in non-labour income on demand for leisure and consumption will be the same regardless from which of the partners receives it:

$$\frac{\partial l_i}{\partial y_s} = \frac{\partial l_i}{\partial y_f} = \frac{\partial l_i}{\partial y_m}, i = f, m; \qquad (1.21)$$

This is referred to as the "income pooling" hypothesis.

However, the advantage of testability of the "unitary" model tends to turn sour when confronted with the data, as the restrictions have been rejected in numerous studies which looked at labour supply of couples and in several studies of consumption demands. The "income pooling" implication has been tested by Lundberg, (1988), Schultz (1990), Thomas (1990), Bourguignon et al. (1993), Fortain and Lacroix (1997). It has been rejected in all of the above with the exception of Lundberg (1988) when applied to couples with pre-

[10] Samuelson (1956) assumes that λ depends on some factors independent of the environment (and so factors such as prices, wages and incomes), but this has generally been agreed to be unsatisfactory. According to Becker's "rotten kid" theorem (1974), a joint household utility can be rationalised as a result of the existence of an altruistic member of a family who decides on the distribution of resources. However, this theorem applies only to instances where utilities are transferable between members of the household and therefore cannot be reconciled with the usual ordinal approach to preferences (Bergstorm, 1989).

school children[11]. Some demand studies on household data provide further evidence of problems with the "unitary" approach. For example, Browning and Meghir (1991), and Blundell, Pachardes, and Weber (1993), Browning and Chiappori (1998) reject the symmetry of the Slutsky matrix implied by the theoretical framework. The fact that implications of the unitary model are rejected when tested on actual empirical data may lead us to doubt the predictions and recommendations made using simulations based on the unitary framework. A recent project comparing the predictions of the unitary and collective models on a generated data set with collective features (Laisney, ed., 2002) suggests that if data is generated as a result of a collective decision process and one estimates reforms using the unitary model, there may be important divergencies between simulated and actual response. On the other hand, though, simulations of the WFTC reform based on the unitary model (e.g. Blundell, Duncan, McCrae and Meghir, 2000) have been broadly correct in anticipating the effects of the reform in terms of the labour supply response. One possible explanation of this may be the fact that the reform in question affected only families with children. As already mentioned above Lundberg (1988) demonstrated using Swedish data that, although the income pooling hypothesis was rejected for people with older children, for those which pre-school children the unitary model could not be refuted.

1.2.4 The "sharing rule" approach to the "collective" model

Chiappori (1988, 1992) demonstrated that under several additional assumptions it is possible to derive more easily testable restrictions of the 'collective' framework presented earlier in this section. These would moreover allow recovery of the underlying preferences and of the parameters of the distribution process (up to an additive constant). The crucial assumptions which make identification of two utility functions possible are (1) that of egoistic preferences (although the model easily extends to Beckerian caring individuals as outlined in Becker, 1974) and (2) that of dependence of the distribution process on individual wages and/or unearned income. The framework has been developed in the absence of public goods, and has so far failed to incorporate them into the theory. In the egoistic framework we can formulate the household's problem as:

$$\max(1-\lambda)[u_m(c_m,l_m)]+\lambda[u_f(c_f,l_f)] \qquad (1.22)$$

subject to (1.17).

where λ is a function of individual wages and/or unearned income.

In the Beckerian "caring agents" interpretation, the individual utility function becomes a weakly separable function of two "egoistic" utilities:

$$W_i=\varpi_i\{u_f(c_f,l_f),u_m(c_m,l_m)\},i=f,m; \qquad (1.23)$$

[11] The fact that income pooling hypothesis was not rejected when applied to couples with young children is not equivalent with rejecting all other models.

where ϖ is strictly increasing in both arguments. The maximisation problem of (1.22) then becomes:

max. $(1-\lambda)[\varpi_m(u_m, u_f)] + \lambda[\varpi_f(u_f, u_m)]$, subject to (1.17)

Individuals thus care about the level of utility of their partner but are indifferent as to what combination of consumption and leisure contributes to achieving this level. In the "caring agents" representation of the equivalent problem, as with the representation of equation (1.1.6) earlier, we can no longer distinguish between private and public goods in the usual fashion. Because the consumption level and leisure of person i indirectly enter the utility function of person j (through the level of i's utility), individual consumption and leisure can both be interpreted as having some of the features of public goods. However, because of the separability of utility functions, it is possible to define goods as *exclusive* and *non-exclusive*. Exclusive goods enter only one utility function directly, and thus in representation (1.23) both male and female leisure and consumption are treated as exclusive[12]. As all available information about family consumption is collected at family level and not at the individual level, the "sharing rule" interpretation of the "collective" model of family labour supply relies on making leisure an exclusive good.

As demonstrated by Chiappori (1988, 1992), the household maximisation problem of (1.22) is equivalent to a two stage process, in which the partners first decide on the allocation of unearned income according to some "sharing rule", and then maximise their own utility function subject to their individual budget constraints (here presented with "caring" rather than egoistic preferences):

max. $\varpi_i\{u_f(c_f, l_f), u_m(c_m, l_m)\}$, subject to:

$$\begin{aligned} &c_i + l_i w_i = w_i(T) + \phi_i(w_f, w_m, y_f, y_m, y, z), \\ &i = f, m; \\ &\phi_m = y - \phi_f \end{aligned} \quad , \qquad (1.24)$$

The advantage of the "sharing rule" approach is that person i's wage only has an income effect on the demands of person j through the process of sharing of unearned income, and conversely, person i's wage, or otherwise the price of i's leisure, therefore has no direct effect on the choice between consumption and leisure of person j. It is this feature of the model which allows identification. This of course implies that there are no externalities in the consumption of leisure by the two partners. Leisure is enjoyed equally

[12] In terms of consumption, exclusiveness reflects the possibility of assigning specific consumption items or groups of goods to particular individuals. A term often used in the "collective" literature is "assignability". For example if clothes can be assigned between partners, men's clothes and women's clothes are two "exclusive" goods. Note also, that "public" goods in general will not be exclusive, though it is possible to find examples where goods usually thought of as "public" could be assigned to specific individuals (e.g. telephone).

whether consumed together or separately and there are no benefits to partner j from partner i's non-working time. This excludes any joint benefits from household production, for example.

Thus, assuming that family decisions are Pareto efficient, and allowing for leisure to be an exclusive good, it is possible to present the problem as two separate utility maximisations. Because person i's wage only has an income effect on person j's choice, changing i's wage is equivalent to changing j's unearned income. This allows identification of the sharing rule up to an additive constant (see for example Chiappori, 1988, 1992, or Browning et al. 1994). Since now wages no longer enter the utility function directly, the standard Slutsky restrictions continue to hold and this makes possible the identification of preference parameters of both individual utility functions also up to an additive transformation). We therefore arrive at a model which not only fulfils the requirements of methodological individualism, but also provides relatively easily applicable tests. These on the one hand allow the theory to be validated by observed behaviour, and on the other make possible the identification of underlying preferences and the decision process.

This means that, although if we know only the overall consumption of the household then we cannot learn about the distribution of consumption within the couple, we are nonetheless readily equipped to answer questions relating to the effect of changes in wages and the budget constraint on the original distribution[13]. The framework therefore allows us to extend welfare analysis to include intra-household distributional issues. Although the model is still at its early stage and its weaknesses are far from trivial, it provides a base for the development of a consistent and applicable framework of family labour supply. Indeed since the early 1990s it has been extended in several important ways.

1.2.5 Extending the sharing rule interpretation of the collective model

Household production and public goods

In the unitary representation of family labour supply decisions, an additional assumption of concavity of the household production function is sufficient to allow identification of the model in the usual way. This is regardless of whether the household produced good is private or public. However, because separability of non-market time is such a crucial assumption of the "sharing rule" representation of the collective model, introducing household production into the collective framework implies important limitations.

If the household-produced good is exclusive, marketable and is observed, then the separability assumption still holds. Labelling the home-produced good as G_f and G_m, individuals then maximise:

[13] The sharing rule can only be recovered completely if we know the allocation of all non-public goods (Bourguignon et. al, 1994)

$$W_i = \varpi_i\{u_f(c_f, l_f, G_f), u_m(c_m, l_m, G_m)\}, i = f, m; \quad (1.25)$$

and we can recover individual preferences in the usual way.

Identification of the model is no longer possible if either we cannot observe consumption levels G_f and G_m, or if the good is public. In the first case non-observability of levels of consumption of the home-produced good in each sub-utility function upsets the assumption of separability if we have to specify the production function of G_f and G_m in terms of individual wages. In the latter case, when G is not exclusive, or in the standard terminology a public good, individual utility functions are:

$$W_i = \varpi_i\{u_f(c_f, l_f, G), u_m(c_m, l_m, G)\}, i = f, m;,$$

which creates the same problems as presence of any public (unassignable) good in the household. As mentioned earlier the theory has been derived under the assumption that there are no public goods. Strictly speaking it extends to include public goods, but only under the assumption of weak separability of public goods in each partner's sub-utility function. Labelling the vector of public goods as Q, this implies the following form of the sub-utility function: $u_i((c_i,l_i),Q)$. A particular problem that the theory encounters is in modelling preferences and the sharing rule of couples with children. On the one hand we could think of spending (time and money) on children as public expenditure. This could be dealt with, though arguably in less than satisfactory fashion, by assuming separability. The additional complication that children introduce into the model is that they are likely not only to influence preferences, but also to affect the sharing process and thus should be included in the sharing rule.

Chapter 2. Empirical Estimation of Labour Supply Models

In this section we relate empirical modelling strategies to the modelling framework outlined in Sections 1.1 and 1.2. We first deal with estimation methodologies developed for structural models but supplement the discussion with an outline of how non-structural or reduced form models, which do not rely on assumption of a specific utility function, can be used to estimate determinants of behaviour on the labour market.

The discussion of estimation methodologies focuses on models defined as 'static' in the sense that they are looking to compare one equilibrium labour supply outcome for a population (e.g. the number of people in work before the Working Families Tax Credit was introduced in the UK) with an alternative outcome under a different budget constraint for workers and potential workers. The change in the budget constraint could be due to tax and benefit policy (for example an increase in generosity of in-work benefits, or a cut in income tax), or due to some other feature of the labour market (e.g. a change in wage levels). Whilst this much is common to all the methods of estimation of the models we look at in this section, there is a large variation in specific empirical strategies.

To structure the discussion here, we have classified labour supply models according to their *function* on one hand and according to their *technical implementation* on the other. The function can be of two sorts:

- **Predictive** models attempt to use existing data to produce predictions of what hypothetical (or future planned) changes to the budget constraint would do to labour supply. For example, if income tax were to drop by 2%, how much would employment and hours worked change?

- **Evaluative** models attempt to use data on a specific policy which has already been implemented to assess the impact of the policy. Sometimes evaluative models use data specifically collected in the course of implementing a program (e.g. Card et al (1998)), whereas in other cases they simply use existing data sources (e.g. Eissa and Leibman (1996)).

The two functions are not mutually exclusive and some studies (e.g. Bingley and Walker, 1997) combine elements of both. Focusing on the completely predictive approach versus the completely evaluative approach makes the exposition easier.

The technical implementation can be of three main sorts:

- **Structural** models provide a direct implementation of the theory shown in Sections 1.1 and 1.2 (subject to what the data will allow). These constitute parametric or semiparametric models of labour supply incorporating maximisation of some utility function. Structural models are typically used

for either prediction or evaluation, although a structural approach lends itself particularly well to predictive analysis.

- **'Reduced form'** models use less assumptions than the structural approach and do not attempt to uncover the parameters of the underlying utility function from the data. Instead, estimation methods designed to uncover the effect of a policy on labour supply which rely on the minimum of economic assumptions (such as 'difference in differences', explained later) are used. The inverted commas around 'reduced form' here are intentional, as whilst the approach is often called 'reduced form' estimation, this is misleading; a certain amount of structural assumptions are invariably necessary in practice to identify the econometric model or interpret the results, and perhaps 'minimal structure' estimation, or some similar term, would be more appropriate. These methods are most often used for evaluation, although in some cases it may be possible to produce predictions based on hypothetical scenarios.

- **Experimental** approaches are purely evaluative. These use 'randomised trial' methods to isolate the pure impact of a policy.

We now provide details of the implementation of each type of model with reference to some recent well-known examples in each field. Each approach has strengths and weaknesses and in many cases the type of data which are available for empirical work will dictate the model that can be used, meaning that no one approach can be seen as 'the best' for every given task.

2.1 Structural modelling

2.1.1 From utility functions to hours equations

This section presents the relationship between utility functions and hours equations derived from them. The latter are results of utility maximisation subject to the budget constraint. In some cases it is more convenient to use indirect utility functions rather than the usual direct representation. Indirect utility functions are expressed in terms of prices/wages and income rather than as direct functions of hours of leisure and consumption. Indirect utility functions are derived by substituting expressions for Marshallian demand for hours and consumption from the utility maximisation problem. Marshallian demands can be derived from indirect utility functions using Roy's identity, which in case of the hours equation is:

$$H = \frac{\partial v(w, y)}{\partial w} \div \frac{\partial v(w, y)}{\partial y} \tag{1.26}$$

where v is the indirect utility function.

The hours equation used in the sections above:

$$\ln h_i = \alpha \ln w_i + \beta y_i + v_i \tag{1.27}$$

corresponds to several utility functions. One of them is the additive exponential form of the indirect utility function:

$$v(w,y) = \frac{w^{\alpha+1}}{\alpha+1} - \frac{e^{-\beta y}}{\beta e^{-v}} \qquad (1.28)$$

Table 2.1 below presents several corresponding utility functions and hours equations which will help to illustrate the restrictions that are imposed on labour supply elasticity and preferences. As we discussed in section 1.1, the interpretation of structural models relies on the conditioning variables used in the estimation. This applies to all the example functions presented in Table 2.1.

Table 2.1. Examples of hours equations and utility functions.

Hours equation:	Direct or indirect utility function:
Constant elasticity labour supply	
$\ln h = \alpha \ln w + \beta y + \pi$	$v(w,y) = \frac{w^{\alpha+1}}{\alpha+1} - \frac{e^{-\beta y}}{\beta e^{-\pi}}$
The sign and value of the wage response is restricted to be invariant with hours. The Marshallian wage elasticity of labour supply is thus constant. Income has a constant proportional effect on hours.	
Linear labour supply	
$h = \alpha w + \beta y + \pi$	$v(w,y) = \exp(\beta w) * \left[y + \frac{\alpha w}{\beta} - \frac{\alpha}{\beta^2} + \frac{\pi}{\beta} \right]$
The wage and income responses are assumed to be constant throughout the hours range. Marshallian wage elasticity always has positive sign. Income elasticity is restricted to have a constant sign as well.	
Semi-log labour supply	
$h = \alpha \ln w + \beta y + \pi$	$v(w,y) = \frac{\exp(\beta w)}{\beta}(\beta y + \pi + \alpha \ln w) - \frac{\alpha}{\beta}\int_{\beta w} \frac{\exp(\beta w)}{\beta w} d(\beta w)$
The Marshallian wage elasticity declines with hours but is constrained to be positive. Income response is constant throughout the hours range.	
Linear Expenditure System (LES)	
$h = (1-\beta)\gamma_h - \frac{\beta y}{w} + \frac{\beta \gamma_c}{w}$	$u(h,c) = [\beta \ln(\gamma_h - h) + (1-\beta)(\ln(c-\gamma_c)]$
Allows positive and negative wage response, Marshallian wage elasticity can therefore be positive and negative. Income response is restricted to be have constant. Direct utility function is explicitly additive in hours and consumption.	

Notes: π – observed and unobserved heterogeneity. Table based on Blundell & MaCurdy (1999).

2.1.2 Empirical estimation in the linear labour supply framework

The aim of the straightforward 'single-period' model is to estimate labour supply responses to specific changes in the budget constraint facing individuals. This can be done either through estimation of an hours equation or by direct estimation of the utility function. The most straightforward situation to analyse, which would boil down the estimation to a simple linear regression of hours worked on explanatory variables including wages, would have to satisfy the following five conditions:

1. each individual commands a given hourly wage w_i when in work.
2. full labour force participation (so that everyone is assumed to be in work, and the only choice variable is hours of work).
3. linear income taxation and no (income-dependent) transfers to the household (i.e.linear budget constraint)
4. no fixed costs of work.
5. the individual's choice set is defined only over a single period (i.e. there are no intertemporal effects).

If these five conditions were met, the most important choice would involve the form of the hours equation bearing in mind the restrictions each of them imposes. In reality few of the above conditions are met. Below we discuss the ways in which modern research relaxes these restrictions. We begin the discussion with an account of how one should model wages if these are not observed. The analysis then focuses on the question of non-participation, non-linear and non-convex budget constraints, fixed costs of working, childcare costs, and partial take-up of benefits and tax credits.

2.1.3 Imputing wages for non-workers

Whether one models the hours equation or the utility function directly analysis is impossible without information on the price of leisure, i.e. individual wage. This presents a problem if, as is almost always the case, there are non-participants in our sample, for whom we have no information on their wage. There are two main methods in the literature of dealing with this problem and imputing wages for non-participants.

Heckman-style selectivity adjusted wage equations

The most common technique relies on a procedure pioneered by Heckman (1974, 1979). First, assume that wages are related to observable characteristics such as age and educational attainment by a human capital earnings function:

$$\ln W_i = \alpha' Q_i + \varepsilon_i \qquad (1.29)$$

we assume that wages are also affected by observable factors which may be correlated with skill so that $E(\varepsilon \mid Q) \neq 0$. This means that Ordinary Least Squares estimation of (1.29) yields biased estimates. The Heckman model amends the wage equation by adding an extra term

$$\ln W_i = \alpha' Q_i + \beta\lambda_i + \varepsilon_i \ , \qquad (1.30)$$

where the additional term $\lambda_i = \phi(\hat{\gamma}' Z_i) \Big/ \Phi(\hat{\gamma}' Z_i)$, the predicted inverse of Mills' ratio from a participation equation, specified as:

$$\Pr(P_i = 1) = \Phi(\gamma' Z_i + u_i) \qquad (1.31)$$

(where Z_i is a vector of observable factors determining participation) is the additional regressor. The Heckman model uses the additional λ_i term in the wage equation to control for correlation between unobservable factors in (1.30) and (1.31). If the decision to work or not is driven by the financial return to work, then, *ceteris paribus*, individuals with a low return to work will be less likely to be in work in the data we observe. If factors *unobservable* to the econometrician which help determine the choice of hours also help determine wages, then the predicted wage for a current non-participant with certain observable characteristics is likely to be lower than the predicted wage for a participant with similar characteristics. To put it another way, u_i and ε_i are likely to be positively correlated. The Heckman procedure allows us to control for this correlation. The Heckman model can be estimated in two stages by running the participation equation, finding λ_i and including it in the wage equation (whilst adjusting the standard errors of the wage equation to take account of the fact that λ_i is a generated regressor). However, in practice maximum likelihood methods are normally used to estimate a model based on the two-equation system (1.30, 1.31) to obtain maximum efficiency in the estimation procedure.

Whilst the two-equation system of (1.30) and (1.31) is technically identified due to differences in functional form between the two equations, for reliable identification it is desirable for there to be at least one variable in Z which is not in Q. This variable is assumed to affect participation but not the wage conditional on participation. Examples of this instrument which have been used by various researchers include family demographics and the predicted level of household income out of work.

Entry wage measures

Another imputation method which has been tried in the literature is to use the 'entry wage' - i.e. the wage at which individuals with certain characteristics Q first enter work. For this to be possible, data on entry wages has to be available. Gregg, Johnson and Reed (1999) used data on the wages of people who had entered work during the five-quarter UK labour force survey panel to derive a distribution of entry wages. Depending on their observable characteristics, unemployed and inactive individuals were assigned probabilities of receiving a wage offer at different decile points in the distribution according to an ordered probit on observable characteristics. So, for example, a person with high educational attainment had a better chance of receiving a wage offer from the high end of the entry wage distribution than a person with low educational attainment. The estimation procedure for the probability of moving into work over the five-quarter period used an 'expected gain to work' variable derived from the entry wage distribution and the budget constraint as a regressor (more details of this procedure are given in Section 2.4.4).

Which imputation method is best?

Given that imputation via Heckman-style selectivity adjustment on one hand, and via entry wage information on the other, are very different techniques, it makes sense to ask which works better. The Heckman approach attempts to estimate an average wage measure from the entire distribution of wages in the economy, but subject to a (downwards) selectivity adjustment – i.e. a person of characteristics X who is not in work will be predicted to have a potential wage which is the same as a person of exactly the same (observable) characteristics but who is already in work. This method has at its heart a reservation wage condition (as discussed in section 2.2.2 on search models). The idea is that, controlling for all the observable factors which influence participation and wages, the reason that people who are observed to be out of work are not going into work is that their potential in-work wages are lower than comparable people who enter work.

One can imagine circumstances under which the Heckman method would not provide an accurate estimate of what currently non-working people would be able to earn in work. For example, if there were large differences in reservation wages between people in work and out of work which were not picked up by anything observable in the data – for example, if out-of-work income was much larger for the out-of-work group and this was not properly controlled for – then the group of out-of-work people might be just as (potentially) productive as individuals who are already in work. One way of checking whether this is the case is to use a model where the income available to individuals or couples when out of work is used as the extra regressor in the participation equation (1.31) which is not in the wage equation (1.29). This is the method used by Blundell, Reed and Stoker (2003) in a wage equation based on Family Expenditure Survey data, and the model

seems to produce a significant and correctly signed selectivity adjustment term.

The Heckman approach makes sense if we feel that an important factor which people take into account when deciding whether to enter work is the 'average' wage for someone of their characteristics (corrected for selection). By contrast, the prediction from an entry wage equation gives the average wage which individuals of certain characteristics actually earn when moving into work, and this corresponds directly to an accepted 'offer wage' in search theory (as explained in Chapter 2). As Gregg, Johnson and Reed (1999) and Gregg and Wadsworth (2000) show, the entry wage distribution has a much lower mean than the overall wage distribution. The former study shows that predicted entry wages are substantially lower than predicted wages from the FRS even after running a selectivity adjustment on the FRS wages. This means that focusing on the entry wage will result in estimated gains to working being significantly smaller than if Heckman-adjusted overall wages are used.

A third possible wage measure is to use entry wages combined with a selectivity correction, on the basis that observed entrants are likely to have higher wages conditional on observed characteristics than people who currently stay unemployed. We discuss the use of this measure in Section 2.4.4 which examines the Gregg-Johnson-Reed report in more detail. A selectivity corrected entry wage measure would presumably be even lower than the uncorrected entry wage measure.

On the other hand, the entry wage measure can be criticised because it takes no account of wage progression. We discuss the growth of wages on the job and across jobs, and mobility in the wage distribution, in Chapter 2, but at this point it is just worth saying that to take account of wage progression we may want to try to estimate some measure of the 'net present value' of taking a job, which accounts for the estimated growth in wages over time for entrants of different types, along with the possibility of exiting the job or moving to a different job later on. The Heckman wage measure may capture some 'long run' equilibrium wage for a worker of a certain type, but it would be useful to have a more explicit treatment of wage progression. In particular, if entry jobs tend to be short-lived and offer little opportunity for wage progression, then the 'long-run' wage may never be reached. In this situation, an entry wage measure would be more appropriate.

Interestingly, the only attempt so far in the literature to compare the usefulness of Heckman-style wages and entry wages for evaluating labour supply estimates comes from Blundell et al (2000), who compared predicted labour supply responses to the introduction of the Working Families Tax Credit estimated using imputed entry wages with predictions from a selectivity adjusted model of the overall wage distribution. They found that the estimated employment effects of the WFTC were slightly bigger using the entry wage measures; for single parents, the WFTC was predicted to increase the participation of single parents by 3 percentage points using entry wages, compared with 2.2 points for Heckman-style wages (for other affected groups

the differences were much smaller). They rationalise the results on the grounds that 'higher wages among those who have chosen not to work tend to imply less elastic preferences and less responsiveness to changes in incomes in work'. This would suggest that using entry wage measures combined with a selectivity adjustment would magnify the predicted employment effects even more. In later stages of this project we do some comparisons between the predicted employment effects in our model using each of the different wage assumptions that we have discussed in this subsection.

2.1.4 Accounting for non-participation

Once we have wages or wage estimates for all people observed in the sample we can move on to the estimation of the response of labour supply to changes in the budget constraint. Taking as an example one of the functional forms of the hours equation (see Table 1.1), one could try to estimate the following log-linear specification:

$$H_i = a \ln w_i + bQ_i + v_i \qquad (1.32)$$

where H_i is hours worked by individual i, Q_i is a vector of control variables and v_i is a normally distributed error term. All variables are measured at time t; the t subscript is suppressed in our discussion of the single-period model for notational clarity. Log wages are typically used because the hourly wage distribution is roughly lognormal.

A basic set of controls would normally comprise

$$bQ_i = \rho X_i + \theta Y_i \qquad (1.33)$$

where Y_i is a measure of unearned income from all sources, and X_i is a vector of other factors which determine the extent of work: e.g. age, region, family characteristics etc.

Clearly, however, equation (1.32) has a censored dependent variable, in that hours of work cannot be negative. Therefore any estimation which simply takes zero as the number of observed hours for non-participants will result in biased coefficients. One could of course leave non-participants out when estimating the labour supply model, as is often done when participation is theoretically difficult to account for (e.g. Chaippori (1992)). This, however, also most likely biases the coefficients (as non-participation is not random with respect to the right-hand side variables) and omitting the non-participants also makes it impossible to say anything about whether changes in the budget constraint will increase or decrease the number of people who are in work. However, if alternatively we include non-participants in the analysis, we have to account for the possibility that there may be factors which affect the decision *whether to work or not* but not the decision how much to work conditional on being in work (i.e. choice of H_i given $H_i > 0$). If we take the participation equation (1.31), it is likely that the vector of observable factors

determining participation, Z, will include factors that are not in X. This would particularly be the case if there were tastes for work which did not depend on the amount of hours worked, or fixed costs of working, for example.

Non-participation and the hours equation

When modelling the hours equation, one can account for the participation choice by explicitly modelling the process of selection into the sample of working people. This can be done in a similar way to that described above regarding the wage equation. For example, Blundell & Walker (1986) in their (unitary) model of labour supply of couples account for female non-participation by estimating the hours equation only for participants but correcting the selection bias on the coefficients. Only two-earner households are selected for the estimation of male and female hours equations, but the authors account for the female participation decision by correcting the estimates for the selection process.

Non-participation and estimation of the utility function

When a direct utility function estimation is conducted, one can relate the levels of consumption available given the number of hours worked to the number of hours of leisure. For example assuming a simple Linear Expenditure System we would have (as in Table 2.1):

$$u(h,c) = [\alpha \ln(\gamma_h - h) + \beta(\ln(c - \gamma_c)]$$

One can then estimate the parameters of the utility function either assuming that the individuals only face the choice of whether to work (full-time) or not, or allowing for greater flexibility by increasing the possible choices of the number of hours worked. Focusing for the time being on the simple binary version of the model and allowing utility from working/not working to vary stochastically, one can specify a logit model of the following form:

$$\Pr[U_j \geq U_k \mid 0,1] = \frac{\exp U(h_j, c_j : \alpha, \beta)}{\exp U(h_0, c_0 : \alpha, \beta) + \exp U(h_1, c_1 : \alpha, \beta)} \qquad (1.34)$$

where subscripts j and k represent the choice of working or not working and (1.34) describes the probability of choosing j. The model can allow for observed and unobserved heterogeneity. Below we discuss this model in more detail. The individual's choice set can of course be extended to allow for working at several levels of hours (for example 0, 20, 40, 60 per week). The simple version presented here has been used quite frequently, though, especially in modelling male labour supply where the choice is often presented as a binary one.[14] As we shall see below the model can be

[14] For example Blundell et. al. (2000) use this approach.

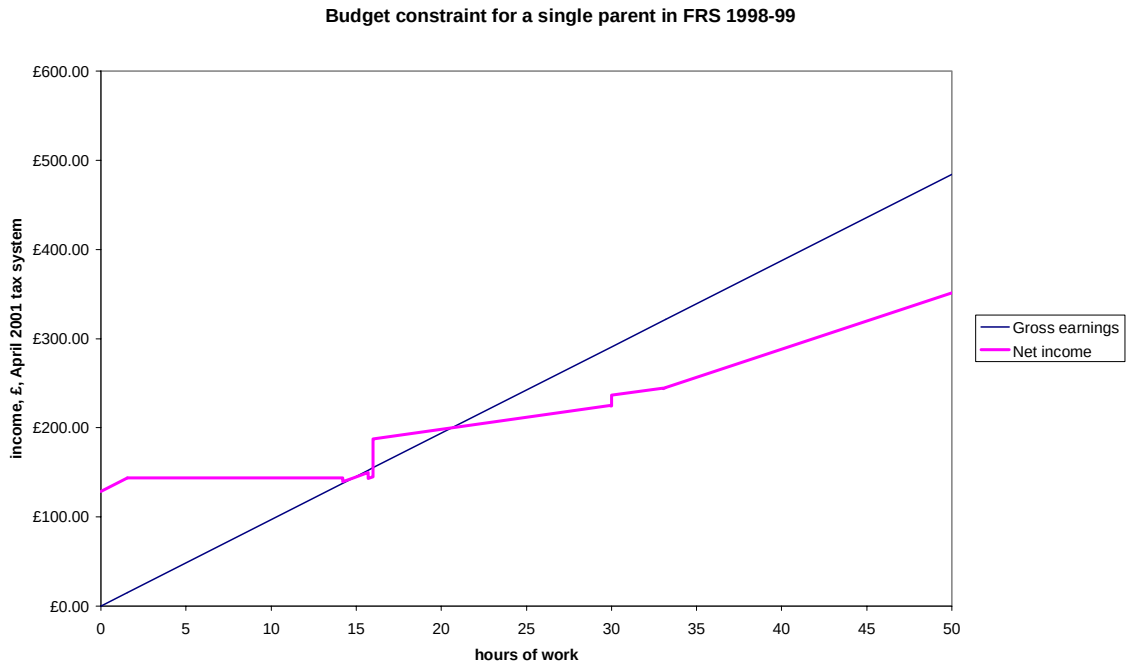
extremely useful for dealing with non-linear and non-convex budget constraints and fixed costs.

2.1.5 The budget constraint

Figure 1.3 below shows a typical budget constraint for a single parent in the 1998-99 FRS data. The thin line shows gross earnings (£9.69 per hour in this case – a relatively high wage rate for a single parent). The dotted line represents earnings net of taxes and National Insurance contributions, i.e. the budget constraint in the absence of any government transfers and means-tested support. The thick line is the actual budget constraint once all transfers are taken into account.

The Figure shows that the tax system introduces important non-linearities, while the system of means-tested support on top of that makes the budget constraint non-convex. At the extreme left of the picture net income rises due to the earnings disregard in Income Support. Once the disregards are reached Income Support is withdrawn one for one with rises in earnings, meaning that the net income schedule is completely flat until around 14 hours of work, where IS entitlement is exhausted (loss of entitlement to free school meals, which count towards net income in the IFS tax-benefit model, creates a slight drop in net income here). At 16 hours, Working Families Tax Credit entitlement begins, leading to a discontinuity with a large rise in net income. In this example, the 55% net income taper on WFTC kicks in immediately at 16 hours, meaning that the budget constraint is shallower until WFTC entitlement is exhausted, at which point the marginal rate falls to 32% (equal to the income tax rate plus the national insurance employee contribution rate). Another discontinuity arises at 30 hours due to the full-time bonus in WFTC. If the budget constraint were extended further out, or the hourly rate of pay were higher, we would see a reduction in the marginal rate to 22% above the upper earnings limit for national insurance contributions, followed by an increase to 40% above the higher rate threshold.

Figure 1.3



The budget constraint and the hours equation

The above-mentioned features of the budget constraint introduce another difficulty to the estimation of the wage equation. Because the tax systems in almost all developed countries are non-linear, net wages can't be calculated simply as (gross wage * (1-tax)). Depending on the level of earnings (and thus on the number of hours worked) the tax rate will be different and therefore the marginal wage which the individual gets for working an additional hour will differ also. For example, in the current UK system someone earning a gross wage of £12 per hour who works twenty hours a week would receive £8.16 pounds for an extra hour of work. Increasing the hours to sixty changes the marginal net wage to £7.20. As a result the wage is *endogenous* to the number of hours worked which means that using a simple gross wage measure, or the net wage at a given hours point, in the estimation would lead to a biased estimate of individual wage response.

Moreover non-convexities at least in theory allow more than one combination of consumption and leisure to give the same level of utility, and therefore make identification of preference parameters difficult. In the case of family labour supply non-convexity of the budget set implies a non-convex Pareto frontier. Because a "rational" couple in the "unitary" or in the co-operative "collective" framework would choose a point on the frontier, non-convexity again produces estimation difficulties

One response to these problems is to use instrumental variables techniques to correct the bias. This involves finding an instrument which affects hours worked but not the wage conditional on hours worked. Similar concerns

regarding the validity of instruments arise here as with the Heckman participation model examined earlier (for a recent empirical study in this vein see Blundell, Duncan and Meghir, 1998). Another option for dealing with the endogeneity problem, which is more common in the recent literature, is to model the nonconvexities in the budget constraint explicitly. We discuss this approach in more detail below.

Estimation of the utility function when modelling a non-convex budget constraint

In direct estimation of the utility function it is absolutely crucial to use a fair representation of individuals' alternative options from which they choose the one we actually observe. For example in our binary choice model discussed above we observe individuals either in work or out of work. In their observed situation we usually have information on their income, both earned and unearned. For their alternative choice, however, their income needs to be imputed. The imputed income will depend on their gross wage, income from investment and savings, and the tax and benefit system, which determines their net earnings and unearned income including various government transfers. The tax and benefit system often depends on age and family circumstances and all these factors will have to be taken into account when calculating income for the unobserved alternatives. To account for the complexity of the tax and benefit system one needs to use a microsimulation model (such as the Institute for Fiscal Studies' TaxBen, or the Cambridge Microsimulation Unit's Polimod) to generate budget constraints (or at least points along the budget constraints) for individuals using suitable microdata on income and family circumstances (such as the Family Resources survey) combined with a detailed knowledge of the rules of the tax and benefit system.[15]

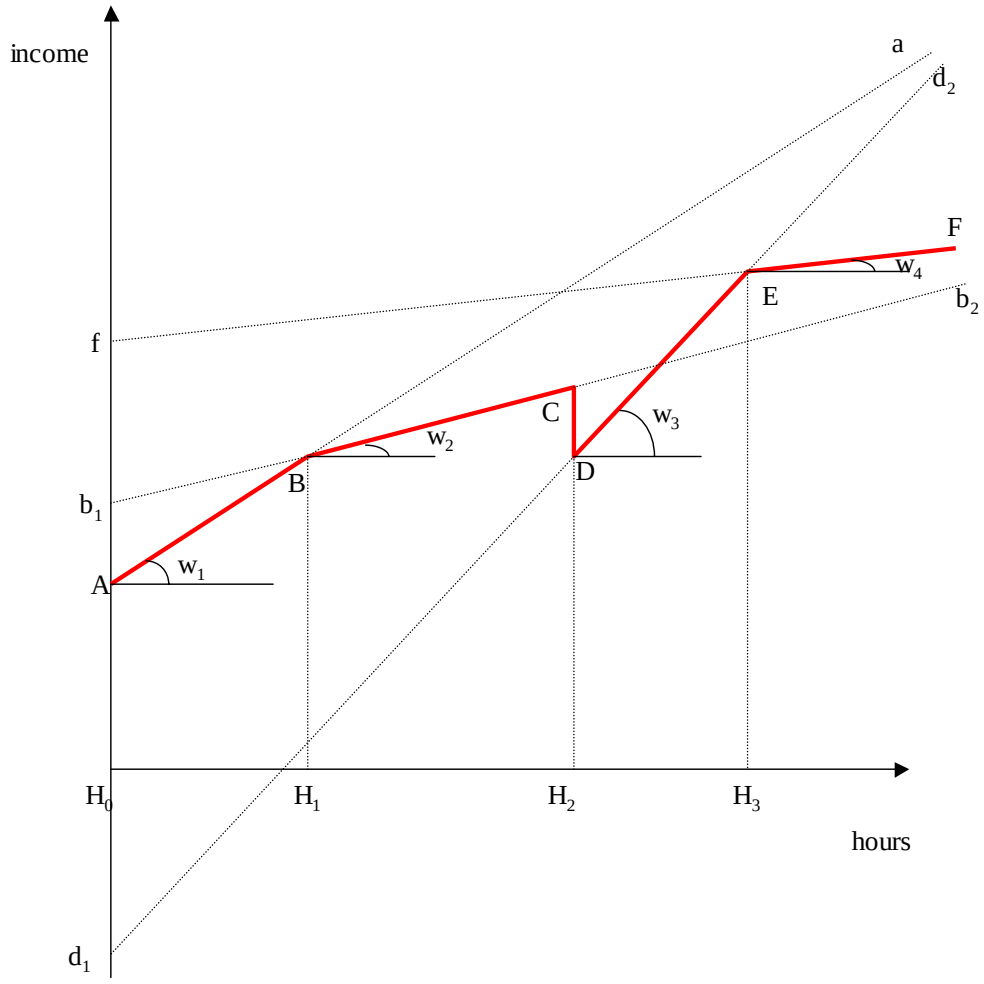
Below we present two methods for dealing with non-linear and non-convex budget constraints when the utility function is estimated directly: piece-wise linear estimation and hours discretisation. The latter is an extended version of the binary choice model mentioned above.

[15] Such calculations might sometimes also be necessary in cases of hours equation estimation. Because unearned income is a crucial explanatory variable in the wage equation, its appropriate level needs to be included in the estimation. When information on unearned income seems inadequate or is incomplete it needs to be imputed.

Piecewise Linear Estimation

A piecewise linear approach takes a budget constraint such as that shown in Figure 1.5 below and constructs a constrained utility maximisation problem characterised as follows:

Figure 1.5. Budget constraint for piecewise linear estimation



$$\text{Max } U(Y,h) \text{ subject to :}$$

$$\begin{aligned} Y &= A && \text{if} && h = 0 \\ & w_1 h + A && \text{if} && H_0 < h < H_1 \\ & w_2 h + b_1 && \text{if} && H_1 < h < H_2 \\ & w_3 h + d_1 && \text{if} && H_2 < h < H_3 \\ & w_4 h + f && \text{if} && H_3 < h \end{aligned} \qquad (1.35)$$

Estimation of the utility maximising point on this budget constraint proceeds in two stages:

i. The choice of h given location on the line segments (AB, BC, DE, EF) is calculated.

ii. Utility is calculated for each of the points on the line segments chosen in (i) and also for the kink points A (the lower limit), B, C and E.

The piecewise linear approach provides a means of estimating labour supply responses whilst taking into account the exact shape of the budget constraint and with a parameterised utility function, exact predictions of responses to changes in the budget constraint can be made. Estimation of the hours equation in this model has to be conducted using instrumental variables to account for the fact that marginal wages are endogenous to the choice of hours as a result of non-linear budget constraint.

The precision of the piecewise-linear model results in some interpretational difficulties. As Blundell and MaCurdy (1999) point out, the piecewise linear methodology assumes that both the researcher and the individuals in the sample have exact knowledge of the entire budget constraint that is relevant for the worker in question. Errors are permitted neither in the worker's perceptions or knowledge of the budget constraint, nor in the researcher's construction of it. This is an extremely stringent criterion and one which it is probably impossible to meet in the real world. For example, if hours and wages are measured precisely and individuals maximise the utility function exactly, we should expect to see bunching of hours of work at kink points such as C in figure 1.5. However, for the overwhelming majority of data sources currently used in the literature, only a trivial number of individuals, if indeed any at all, report hours of work at interior kink points. So, measurement error in *h* needs to be introduced to avoid the model being rejected on the grounds that very few observations are reported at exact kink points. Adding a continuously distributed measurement error ε to the model makes it consistent with a continuous hours distribution, but measurement error of hours of course implies measurement error in *wages* (since in many cases the hourly wage is calculated as reported weekly earnings divided by reported hours of work). This calls into question the piecewise linear approach's original contention that the budget constraint is perfectly measured. Thus this approach to estimating labour supply equations is of somewhat fragile robustness.

Discretisation of the hours choice

A more robust, if less accurate, approach to accounting for non-linear budget constraints involves treating the labour supply decision as if individuals are choosing from a discrete set of hours points rather than optimising over a continuous distribution of hours. This was originally proposed by van Soest (1995) and has been applied for example by Bingley and Walker (1997), Duncan and Weeks (1997) and Blundell *et al* (2000). The model implies that peoples' actual choices of hours (and their actual incomes) are in most cases

different from the points of estimation. Such an approach allows for optimisation errors by individuals in a more sensible manner than the piecewise linear approach presented earlier.

As in the binary choice model presented in the discussion of non-participation, people are allocated to specified discrete "hours points" where the hours point corresponds to hours being observed in a certain "hours bracket" in the data. For example, in the simplest case we could specify a set of two hours points as [0,40], where [0] corresponds to a zero hours 'bracket', and [40] corresponds to positve hours. A more complex set of hours would be, for example, [0,10,20,30,40,50], corresponding to hours brackets such as (for example) [(0),(1-15),(16-24),(25-34),(35-44),(45 or more)]. Thus the continuous hours distribution is collapsed into a set of discrete points. Given people's wages and their other characteristics, incomes are calculated for the set of hours points corresponding (via the hours brackets) to the hours actually observed for individuals in the data as well as for other hours points which are selected as possible options. Choices over the hours points can be made flexible by allowing utility from different hours choices to vary stochastically over individuals. In a Linear Expenditure System we would then have:

$$u(h_{ji}, c_{ji}) = [\alpha \ln(\gamma_h - h_{ji}) + \beta(\ln(c_{ji} - \gamma_c)] + \varepsilon_i \qquad (1.36)$$

where for every hours point j the individual i's utility function is determined by the level of leisure $(\gamma_h - h_{ji})$ and consumption corresponding to income earned when working j hours and determined by the budget constraint. The model allows unobserved heterogeneity among individuals only through the random parameter ε_j. However, preferences can vary with observed individual characteristics by allowing parameters α and β to differ between different groups:

$$\begin{aligned} \alpha &= \alpha_0 + \alpha'_x x \\ \beta &= \beta_0 + \beta'_x x \end{aligned} \qquad (1.37)$$

If the random disturbances ε_j have an extreme value distribution then choices across discrete hours points can be written as a conditional logit model:

$$\Pr[U_j \geq U_k \mid \text{all } k] = \frac{\exp U(H_j, Y_j : \alpha, \beta)}{\sum_k \exp U(H_k, Y_k : \alpha, \beta)} \qquad (1.38)$$

where subscripts j and k represent discrete hours points. (1.38) describes the probability of choosing hours point j (or in fact, the hours bracket corresponding to j) as the actual hours worked.[16]

The multinomial logit model and the independence of irrelevant alternatives

One of the major restrictions of the multinomial logit model is the fact that it imposes the assumption of independence of irrelevant alternatives (IIA). This means that the ratio of probabilities of any two events is independent of the alternatives which are not considered. For example, if we consider the following choice of the number of hours: [0, 10, 20, 30, 40, 50], the ratio of the probability that 20 is chosen to the probability that 40 is chosen is assumed to be the same as in the case when we consider 0, 10, 20, 30, 40, 50, and 60 hours. This is a very strong and rather unrealistic assumption, and two methods of relaxing it have been developed in the literature: random parameter logit (McFadden and Train, 2000) and mass point estimation (Wiliamson-Hoynes, 2001). Both rely on allowing the parameters of the utility function to be random variables. So in our example:

$$\begin{aligned} \alpha &= \alpha_0 + \alpha'_x x + \nu_\alpha \\ \beta &= \beta_0 + \beta'_x x + \nu_\beta \end{aligned} \qquad (1.39)$$

Unobserved preference heterogeneity among individuals is represented by disturbances ν. These are random and assumed to be normally distributed with mean zero and different variances. This makes unobserved individual heterogeneity correlated with individual characteristics and thanks to ν_α and ν_β, the random parameter logit does not exhibit the IIA property. To account for the unobserved heterogeneity represented by the ν terms, it is necessary to integrate over the range of the ν variables. Simulation methods are necessary for this; a large number of draws, usually over 100, are taken over the two equations in (3.7), assuming a multivariate normal distribution for the ν variables. This is used to construct a simulated likelihood for (1.36) which can then be used for maximum likelihood estimation of (1.37).[17]

2.1.6 The problem of fixed costs

The standard representation of the choice between consumption and leisure allows for non-participation because of people's strong preference for leisure relative to their preference for consumption, or because their income while out of work is high enough to rationalise non-participation. Another interpretation of the optimality of non-participation is the fact that apart from financial gains

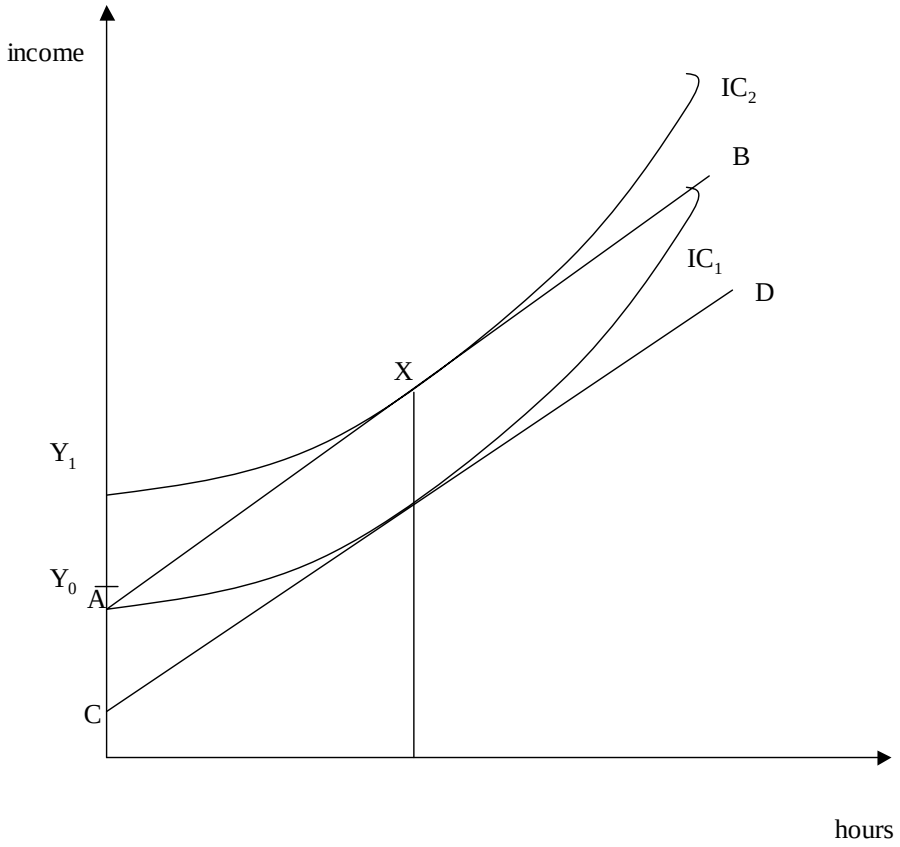
[16] Van Soest (1995) applies this model (although assuming a different form of the utility function) to estimate preferences of people in couples on Dutch data assuming the "unitary" process of intra-household decision making.

[17] For further technical details of the simulated likelihood procedure in the context of labour supply, please see Duncan and Weeks (1997).

from work there may be certain financial costs of participation, the so called ‘fixed costs’ - factors which influence participation and do not influence hours choice.

Among many examples of fixed costs are expenditure on travel to work, work clothing, etc. The presence and size of these ‘fixed costs’ can seriously influence the decision whether to work or not, as shown in Figure 1.4 below. In the presence of fixed costs of size AC, the budget constraint shifts from AB to ACD. In the absence of fixed costs, the individual prefers to work h hours, as shown by the tangency of IC_2 to AB at X. However, with fixed costs of AC, the individual would prefer inactivity (a corner solution where IC_1 hits point A). While in some cases we will have some information in the data on fixed costs for those who work, for non-participants this information will be absent. For this reason, as well as in the case where no information on fixed costs is available in the data, additional structure will need to be imposed on the model to allow for these costs.

Figure 1.4: Fixed costs and the budget constraint



Discretisation of hours choices also allows us to include fixed costs in the model, as the level of income at every choice of positive hours can be assigned a specified fixed cost "penalty". This is done for example in Blundell et al (2000) who model fixed costs as a one-off weekly cost and subtract them from net income for choices that involve work. If we have data on the size of the fixed cost then it is modelled as a simple deduction from net income. More often this is not possible, and so the fixed cost has to be modelled as conditional on observable factors and an unobservable heterogeneity variable as above. If we account for fixed costs, individual income is then:

$Y - F$, if $h > 0$ and

Y, if $h = 0$. (1.40)

If F is not directly observed, the estimation of the labour supply function must explicitly consider two 'regimes': working and not working. As in the case of discrete hours points considered above, estimation proceeds by finding the maximum utility under each regime and then comparing these to determine which option will be chosen. The fixed cost variable is treated as an extra

parameter which is estimated along with the rest of the model. For example, in a model where the utility function in the absence of fixed costs is defined as

$$U = U((Wh + Y - \tau(Wh + Y)), h) \quad (1.41)$$

where W is gross hourly earnings, Y is unearned income, h is hours and τ is the average tax rate on total weekly income, introducing fixed costs can be done as shown:

$$U = U((Wh + Y - \tau(Wh + Y) - \nu), h) \quad (1.42)$$

where ν is the fixed cost parameter. In the 'no-work' regime, utility is given by $U(Y - \tau(Y), 0)$; in the 'work' regime, by $U((Wh + Y - \tau(Wh + Y) - \nu), h)$, where $h > 0$. The hours choice in the work regime has to be accounted for using either the piecewise-linear or hours discretisation methods detailed earlier but with the added complication of the parameter ν. Positive fixed costs mean that there is likely to be a 'hole' in the hours distribution between zero (not working) and some positive value of hours – we are unlikely to see individuals working small numbers of hours because the new gain from work is insufficient to offset the fixed costs.[18]

Important contributions to the empirical literature on estimation of labour supply models in the presence of fixed costs include Cogan (1981), Moffitt (1983) and Bourgignon and Magnac (1990). It should be noted that all of these approaches treat fixed costs as essentially *exogenous* – i.e. they are a factor which the individual has no control over. For fixed costs like travel to work or the housing costs associated with location near to certain types of job, or costs associated with being at work such as clothing or equipment (if not paid for by the firm), the exogeneity assumption is probably valid in the short run. However in the long run we might expect agents to adjust their location or occupational choice to exploit the appropriate trade-off between the level of fixed costs and the wage or other benefits of the particular job they do. Accounting for this kind of long-run optimisation requires a much more life-cycle, panel-data orientated approach to the labour supply decision which we discuss in detail in our review of dynamic models of labour supply in Chapter 2 of this review.

2.1.7 Modelling childcare costs and labour supply

In addition to fixed costs of work, there can be costs which vary with the number of hours worked. An important example of a variable cost is childcare. Clearly, if childcare is paid for by the hour, then its cost will vary with the number of hours that the primary carer in the household (often the mother) is

[18] Note, however, that the fact that few people are observed working very small numbers of hours could also be caused by very high marginal deduction rates at low hours of work – for example, means-tested benefits for people working less than 16 hours a week in the UK are withdrawn pound for pound as earnings rise (bar a small earnings allowance), which gives no financial incentive to work at those hours levels, even before considering fixed costs of work.

working. Clearly a failure to take childcare costs into account is likely to lead to biased estimation and incorrect inferences being drawn from labour supply models. However accounting for childcare in labour supply models is complicated due to several factors:

- Families may use childcare for reasons unrelated to work, e.g. for babysitting, or simply to give the parents a break from caring for the child. Hence we cannot automatically assume that survey data on number of hours of childcare use actually corresponds to number of hours of paid childcare used for labour supply purposes unless the data specifically indicate this.
- The price of childcare can vary widely, either because of local market conditions, non-market childcare arrangements (such as care by relatives, where the childcare may be provided free of charge in many cases) or variations in the quality of childcare purchased.

Strategies for modelling childcare costs in the labour supply model

i) Joint modelling of childcare and labour supply

These kinds of considerations suggest that, data permitting, the ideal approach to modelling childcare in a structural framework is to treat it as completely endogenous (a choice variable) and to estimate a joint model of labour supply and childcare use. Recent examples of this kind of model include Blau and Hagy (1998), Blau and Robins (1988), Hotz and Kilburn (1994) and Ribar (1992). We focus below on a recent model for the UK by Duncan, Paull and Taylor (2001).

Duncan, Paull and Taylor (DPT hereafter) estimate a single-period model of the choice of formal childcare among women with at least one dependent child below school age. The following assumptions are made concerning childcare price and availability, and the structure of labour supply:

1. the price paid for formal childcare reflects heterogeneity in childcare quality. Mothers are confronted with a market-determined price for some 'base quality' of childcare, but can select differing quality levels over a range of prices. This assumption therefore rules out variation in the base price within the local childcare market (e.g. due to imperfect competition in the market), although prices can differ across localities.
2. Hours of work must be less than or equal to non-maternal childcare time. This reflects the requirement that there be somebody to look after the child(ren) when the mother is at work (conditional on the father also working of course). However, childcare time can be *greater* than hours worked in this model, which allows childcare to be used for non-work reasons.
3. The labour supply of the male partner in the household is taken as given.
4. The level of *informal* childcare (i.e. free care provided by close relatives or friends) is also taken as given (exogenous).

5. A single aggregate of formal care is modelled (rather than splitting formal care into different types, e.g. nursery, babysitter, etc.) This makes estimation of the model simpler but means that differences in families' preferences over different types of care cannot be allowed for (which may be unduly restrictive).

With the father's labour supply treated as fixed, the mother in each houshold makes decisions over labour force participation and childcare so as to maximise the value of a preference function:

$$U = U(C, L, Q \mid Z_1) \tag{1.43}$$

where *C* is private non-childcare consumption, *L* is time not in employment ('leisure'), *Q* is overall childcare quality and Z_1 is a vector of exogenous explanatory characteristics. *U* is assumed increasing in all three arguments. Overall childcare quality is described by some 'production process' of the form

$$Q = F(K_M, K_I, qK_F \mid Z_2) \tag{1.44}$$

where K_M is the mother's time devoted to the care of her children, K_F is the time spent in formal care and K_I is the time spent in informal care, and *q* is the quality of formal childcare, normalised at 1 for the 'base quality', with $q > 1$ meaning better than base quality and $0 < q < 1$ meaning worse than base quality. The model includes a budget constraint:

$$C + pqK_f = wH - T(w, H, V, pqK_F) + V \tag{1.45}$$

where *p* is the hourly price of the basic quality of childcare, *w* is the hourly wage of the mother and *V* represents exogenous household income, including the income of the partner. Hence, total expenditure on formal childcare (pqK_M) depends on the exogenous base price and the chosen quantity and quality levels. There is also the time constraint that $T = K_F + K_I + K_M$, i.e. that the children are cared for at all times. Assumption 2 above implies additionally that

$$K_F + K_I \geq H \tag{1.46}$$

DPT estimate an econometric model of employment and childcare use which incorporates the quantity constraint in (1.46). Underlying the model is the premise that observed choices of work (h_w) and childcare (h_c) are driven by a person's underlying 'propensities' to work and to consume formal childcare (denoted h_w^* and h_c^* respectively). These propensities are assumed to depend on a series of factors x and z respectively (including price, wage and incomes variables, individual exogenous characteristics and demographics, and local childcare and labour market conditions), and take the following general form:

$$h_w^* = x\beta_w + \varepsilon_w$$

$$h_c^* = z\beta_c + \varepsilon_c \qquad (1.47)$$

where β_w and β_c are parameters, and ε_w and ε_c represent factors influencing work and childcare which are not related to observables. We assume that the hours choice conditional on working divides into 'full-time' and 'part-time' hours. The constraint from assumption 2 means that households cannot use less than the required hours of childcare if working, leads to a six-state model of childcare use:

Table 2.2. A six-state model of childcare use and labour supply

Mother's hours of work	Use more than required hours of formal childcare?	
	Yes	No
Full-time $(h_w \geq 30)$	$h_w^* \geq \gamma_2, h_c^* > h_w^*$ $\varepsilon_w > \gamma_2 - x\beta_w$ $\varepsilon_c > (x\beta_w - z\beta_c) + \varepsilon_w$	$h_w^* \geq \gamma_2, h_c^* \leq h_w^*$ $\varepsilon_w > \gamma_2 - x\beta_w$ $\varepsilon_c \leq (x\beta_w - z\beta_c) + \varepsilon_w$
Part-time $(0 < h_w < 30)$	$h_w^* \in (\gamma_1, \gamma_2), h_c^* > h_w^*$ $\varepsilon_w \in (\gamma_1 - x\beta_w, \gamma_2 - x\beta_w)$ $\varepsilon_c > (x\beta_w - z\beta_c) + \varepsilon_w$	$h_w^* \in (\gamma_1, \gamma_2), h_c^* \leq h_w^*$ $\varepsilon_w \in (\gamma_1 - x\beta_w, \gamma_2 - x\beta_w)$ $\varepsilon_c \leq (x\beta_w - z\beta_c) + \varepsilon_w$
No paid work $(h_w = 0)$	$h_w^* \leq \gamma_1, h_c^* > h_w^*$ $\varepsilon_w \leq \gamma_1 - x\beta_w$ $\varepsilon_c > (x\beta_w - z\beta_c) + \varepsilon_w$	$h_w^* \leq \gamma_1, h_c^* \leq h_w^*$ $\varepsilon_w \leq \gamma_1 - x\beta_w$ $\varepsilon_c \leq (x\beta_w - z\beta_c) + \varepsilon_w$

DPT estimate this model using data from the UK Family Resources Survey, which features the drawback that consistent childcare use information is only collected for households where mothers are working. Hence the two cells in the bottom row of Table 2.2 must be merged into one 'childcare unspecified ' cell with just the conditions $h_w^* \leq \gamma_1$ and $\varepsilon_w \leq \gamma_1 - x\beta_w$. For estimation, ε_w and ε_c are treated as joint-normally distributed with unit variances and correlation ρ. To keep matters as simple as possible whilst capturing the essential features of the model, we present here the likelihood function for a three state model (not distinguishing between part-time and full-time work), which can be expressed as

$$\ln L(\beta_w, \beta_c, \rho) = \sum_{i=1}^{n} [I_i^{nw} \ln(P_i^{nw}) + I_i^{wmc} \ln(P_i^{wmc}) + I_i^{wc} \ln(P_i^{wc})]$$

$$\begin{aligned} &= \sum_{i=1}^{n} I_i^{nw} \ln(1 - \Phi(x_i\beta_w)) \\ &+ \sum_{i=1}^{n} I_i^{wmc} \ln\left(\Phi(x_i\beta_w) - \Phi_2\left[x_i\beta_w, \tfrac{z_i\beta_c - x_i\beta_w}{\sqrt{2(1-\rho)}}; \tfrac{\rho - 1}{\sqrt{2(1-\rho)}}\right]\right) \qquad (1.48) \\ &+ \sum_{i=1}^{n} I_i^{wc} \ln\left(\Phi_2\left[x_i\beta_w, \tfrac{z_i\beta_c - x_i\beta_w}{\sqrt{2(1-\rho)}}; \tfrac{\rho - 1}{\sqrt{2(1-\rho)}}\right]\right) \end{aligned}$$

where:

- I_i^{nw}, I_i^{wmc} and I_i^{wc} are indicator variables for the states of non-work, working with minimum childcare and working with more than minimum childcare respectively;
- P_i^{nw}, P_i^{wmc} and P_i^{wc} are the probabilities of individual i being in each of the respective states;
- Φ represents a univariate normal cumulative density function and Φ_2 a bivariate normal density function.

Clearly a structural model which endogenises childcare use increases the complexity of the estimation process considerably. In addition, accurate data on the costs and quantity of childcare used by each household are needed. Duncan, Paull and Taylor (2001) estimate the baseline childcare price using a approach which assumes that the price of a 'baseline' quality of childcare is fixed within a local market (UK local authority area in this case), and hence that differences in hourly childcare price within the local market reflects differences in quality. Such an assumption, although restrictive, is essential in the absence of direct survey information on childcare quality.

ii) Including an estimated measure of childcare costs in the labour supply equation as a deduction from disposable income when working

An alternative to estimating a full joint model of childcare choice and labour supply is to include an estimated childcare expenditure variable as an extra determinant of the budget constraint in the labour supply equation. Clearly, a positive cost of childcare when working will reduce the net financial incentive to work (on the assumption that childcare is used when working but not when out of work). However, in-work subsidies which include allowances for childcare costs (such as the childcare component of the UK Working Families Tax Credit) will mitigate this to some extent. Blundell, Duncan, McCrae and Meghir (2000) estimate a model which takes this approach in their work on the labour market impact of the Working Families Tax Credit in the UK. This strategy views childcare costs as simply an alteration to the budget constraint and is hence less flexible than the full joint model of childcare and labour supply which allows households to choose different amounts of childcare for a given amount of hours of work – here the childcare cost is treated as variable across the hours of work distribution, but as fixed for each level of hours of work. However this methodology is easier to estimate than the joint model and is usable even if the data on childcare use and cost are of poor quality or very aggregated. For example, it could be implemented (albeit crudely) using data on the average amounts of childcare used and the average cost of childcare for full-time and part-time workers. However, the Blundell et al. approach is more sophisticated than this as they use the Family Resources Survey data, as used by Duncan, Paull and Taylor (2001) in the previous example.

2.1.8 Modelling take-up

Much of the labour supply modelling literature assumes that benefits are fully taken up if households are eligible for them. This makes modelling the budget constraint easier but is very unrealistic. For example, research on the Working Families Tax Credit in the UK estimated that in 2001, only 78 percent of potential WFTC expenditure was actually taken up, and only 67 percent of eligible families actually claimed WFTC (McKay, 2001).

By 'non-take up' we mean a situation where someone does not claim a benefit to which they are entitled. Empirical studies of non take-up typically compare data on receipt of benefits as recorded in household surveys (or matched administrative data sources) with data on entitlement to benefits produced by a microsimulation model operating on data on household characteristics from the same household survey.

Why would someone not take up a benefit to which they are entitled? Possible explanations include:

- imperfect information on the part of the (potential) claimant about what benefits exist and/or what the eligibility criteria are;
- the 'hassle' costs of applying for benefits;
- stigma costs.

Several labour supply studies have sought to model take-up simultaneously with labour supply behaviour, to improve the accuracy of labour supply estimation. These include Moffitt (1983), Hoynes (1996), Bingley and Walker (1997) and Keane and Moffitt (1998).

Other features which econometricians are interested in are:

- allowing for modelling errors when estimating take-up;
- estimating a value (in some metric) of the stigma costs of receiving benefits.

A standard framework for thinking about non take-up

Duclos (1995) presents a commonly-used framework for thinking about how to model take-up of benefits, which we draw on in the following discussion. There are three agents involved in a study of non-take up:

1. the individual
2. the government agency which administers the benefit
3. the researcher

The government agency has an information set Ω_g (information from the benefit application form and the agency's knowledge of the benefit rules). The analyst has the information set Ω_a (data from a household survey).

Define $B*$ to be the true entitlement of a family as determined by the existing means-tested benefit rules. The actual level of benefit received by a family

could differ from this, either because the agency has failed to determine correctly all the variables necessary to calculate $B*$ or because it has failed to apply the rules correctly. Call the actual benefit amount received B^g where:

$$B^g = E\{B^* \mid \Omega^g\} \tag{1.49}$$

In general, $B*$ differs from B^g because the agency may have measured household characteristics with error, and/or the agency may make errors when interpreting the benefit rules laid down in legislation (as indicated by, e.g. successful appeals against WFTC decisions in tribunal cases).

The analyst typically uses a microsimulation model and data from a household survey to produce an estimate of the entitlement B^a. This will differ from $B*$ because of:

- measurement error or sampling defects leading to mismeasurement of household characteristics
- changes in circumstances between the time of the survey and the time when an individual's benefit claim was last assessed (this is particularly important in the case of WFTC where claims are fixed for 6 months)
- errors by the analyst in interpreting the benefit rules laid down in legislation

So:

$$B^a = E\{B* \mid \Omega^a\} \tag{1.50}$$

As $\Omega^g \neq \Omega^a$ the analyst's estimate is in general not the same as the agency's.

This implies that the population can be partitioned into four groups:

- eligible recipients
- eligible non-recipients (put off by stigma or 'hassle' or badly informed)
- non-eligible recipients (fraudsters or mistakes on the part of the agency)
- non-eligible non-recipients.

Economic models of non take up

The framework first outlined in Moffitt (1983) for analysing non take up suggests that people do not take up benefits if the disutility of claiming and receiving the benefit outweighs the utility gain of the extra income. Sources of disutility could include: information costs (awareness of the scheme, complexity of forms), process costs (time requirements), job-search costs (if search is mandatory whilst on the program), and/or outcome costs (stigma). A simple formulation of the take-up condition is:

$$\Pr(T=1) = \Pr[U(y + B*(y;X);X) - C(y;X) > U(y;X)] \tag{1.51}$$

where y is income from other sources when not working, X is a vector of family characteristics, $C(y,X)$ is the utility cost of claiming and receiving the benefit, and we assume that families costlessly know their entitlement and

there are no errors (i.e. $B^* = B^g$). This leads to the following set of 'observation rules':

	Individual: no claim ($T = 0$)	Individual: claim ($T = 1$)
Not entitled ($B^g = B^* = 0$)	No award, not entitled	Never happens in this framework
Entitled ($B^g = B^* > 0$)	Would be an award, entitled (genuine non take-up)	Award, entitled

This model can be generalised along two further dimensions:

- **allowing for errors in the agency's calculation of entitlement:** this means that $B^g \neq B^*$ is possible and introduces the possibility that someone may get an award if not truly entitled, and vice/versa (expanding the figure above to four rows instead of two).
- **Allowing for errors on the analyst's part:** this introduces twice as many rows again, in that B^a - the analyst's assessment of whether an award is made – may diverge from B^g .

Allowing for these two extra dimensions gives a model with four dimensions of varation:

1. Is the benefit taken up? ($T = 0$ or $T = 1$)
2. Is the individual truly entitled? (if all information were known by the agency), ($B^* = 0$ or $B^* > 0$)
3. Does the agency actually award the benefit? ($B^g = 0$ or $B^g > 0$)
4. Does the analyst observe take-up? ($B^a = 0$ or $B^a > 0$)

This gives a total of $2^4 = 16$ possible states. In practice, the data available make it impossible to distinguish between these states accurately. B^* is never observed and B^g is only observed if there has been a successful claim. What appears to be non-take-up due to high stigma costs may look identical – given our data – to non-take up due to modelling errors or poor expectations.

Modelling labour supply and take-up jointly

The discussion above assumed that pre-transfer income y was exogenous, but in general, take-up needs to be modelled jointly with labour supply behaviour. This is because:

1) entitlement to means-tested benefits will generally depend on labour supply behaviour, and labour supply incentives will be altered by the value of means-tested benefits. This simultaneity implies that, even if preferences for working and claiming benefits are uncorrelated, individuals working and claiming some in-work benefit will have lower

propensities to work than those working and not claiming, ceteris paribus. Conversely, it means that those observed not working must have relatively high stigma costs, ceteris paribus. The group of individuals observed working and claiming an in-work benefit have self-selected themselves into that group, and so inferences based on them may not hold for the population.

2) Preferences for working and preferences for receiving benefits may be correlated (i.e. the marginal utilities of leisure, income and stigma may be correlated).

The estimation strategy for joint modelling of labour supply and take-up employed by most researchers is to write down an observation rule as we did above, specify a utility function or a function for the net utility of receiving a benefit (some function of the terms $U(y, X)$ and/or $C(y, X)$) and perhaps the process of forming expectations about benefit levels, and then to use maximum likelihood techniques to estimate the parameters. Below we give examples in more detail. Throughout what follows, let I_b be an indicator for receipt of benefit b, which is worth $B*$.

Moffitt (1983)

This study looks at female-headed households in the US and take-up of AFDC. The direct utility function (suppressing characteristics X) is:

$$\log U(y, h, B, I) = -\log(\beta - \delta h) - \frac{\delta(h - \alpha - \delta(y + \gamma B))}{\beta - \delta h} - \phi I . \qquad (1.52)$$

This model allows income from AFDC to be valued differently from other income (through γ, with $\gamma < 1$ implying that benefits are not valued as highly as other income) and for there to be some fixed stigma, ϕ. Wages are modelled so that non-workers can be included. Having assumed or averaged away the non-linearities in the tax and welfare system, and assumed additive normal error terms in the (not shown) hours equation and take-up model, the model gives a simple form for hours h (Tobit) with endogenous take-up. The study finds $\phi > 0$ and $\gamma > 1$, the latter suggesting a mis-specification (such as omitting the value of food stamps).

Bingley and Walker (1997)

This study looked at lone parents in the UK, and the decision to take up Family Credit. The utility function as estimated is identified only relative to the utility of non-participation (in other words, they estimate the utility differences between part-time work and non-participation, part-time work, Family Credit and non-participation), which makes direct interpretation of the coefficients hard. The model allows for involuntary unemployment to prevent the stigma term on family credit from having to explain all observed unemployment/non-participation.

Moffitt and Keane (1998)

This study looks at female-headed households in the US and their decision to claim AFDC, Food Stamps and subsidised housing ($b = 1,2,3$). The direct utility function is a flexible form quadratic in its arguments:

$$U(y,H,\{B_b\},\{I_b\}) = \alpha h + y - \beta_{hh}h^2 - \sum_{b=1}^{3}\Psi_b I_b + \sum_{b=1}^{3}\sum_{c>b}\phi_{bc}I_b I_c + \beta_{hy}hy - \sum_{b=1}^{3}\delta_b hI_b - \sum_{b=1}^{3}\eta_b yI_b \quad (1.53)$$

where b indexes the three benefits, and $y = y(\{B_b\},\{I_b\})$ is income including whatever combination of benefits is claimed. This gives eight possible combinations of benefit take-up, combined with three choices of hours of work, and they are able to estimate the model by adding an extreme value error term to the direct utility function. Personal characteristics are allowed to affect preferences for work, income and take-up. Identification of the stigma term arises because some households are not eligible for these benefits. Blundell et al (2000) essentially use a similar utility function for a single benefit (WFTC) with stigma identified through non-eligibility.

Valuing the stigma costs

Any model of non-take-up that directly models the utility function is able to quantify in some way the magnitude of the stigma costs (models where the stigma costs vary with observable characteristics can also value the additional stigma costs arising through changes in observable characteristics).
For example, if we write a general utility function (not separable in stigma costs) as : $U(y,I;X)$, where y includes B as appropriate and I indicates take-up, then an obvious measure of the stigma costs in utility terms is $C(X) = U(y,I=0;X) - U(y,I=1;X)$ at some y. This can then be converted into a monetary metric using a compensating or equivalent variation approach (see Brewer, 2002).

2.2 'Reduced Form' Models of Labour Supply Response

2.2.1 Difference-in-differences models

The 'difference-in-differences' approach to estimating labour supply responses imposes much less structure on the labour supply decision than the utility-maximisation models covered in the previous subsection. Whilst difference in differences (DID) is often called a 'reduced form' estimation procedure, this is somewhat misleading; some structural assumptions are maintained, as will be explained shortly. Also, whereas the fully structural

utility-maximising specification can be used to estimate labour supply responses to a broad range of changes in the labour market, DID is geared more towards a specific type of policy reform – one which affects a group of people with certain observable characteristics (e.g. single mothers) whilst not affecting a different group of people with reasonably similar observable characteristics (e.g. married mothers, or single women without children).

The technical characterisation of DID is as follows: assume that a labour market policy reform (such as an increase in the Working Families Tax Credit, for example) occurs which affects one group (the 'treatment' group, known as τ) but has no direct effects on another group (the 'control' group, known as c). The outcome variable which the researcher is interested in is some measure of labour supply (we will focus on the participation rate, denoted below by P, although hours of work is another possible outcome variable). The DID model works by comparing the outcome variable before and after the policy reform for both groups (for this reason, DID is sometimes known as a 'before-and-after' estimator). The basic DID model makes two crucial assumptions:

1. Any 'time effects' on the outcome variable (e.g. macroeconomic shocks to labour demand, secular trends in labour market participation for subgroups of the population, and so on) are assumed to affect the control and treatment areas equally.

2. It is assumed that the composition of the control and treatment groups does not change substantially over the 'before' and 'after' periods. If the period between the data used for the 'before' regime and the data used for the 'after' regime is reasonably short, this assumption should be uncontroversial.

Denoting the point in time before the reform at which snapshot labour market data are taken as $t=1$ and the point after the reform as $t=2$, with the proportion of employed people in subgroup τ and c at time t as P_t^τ and P_t^c respectively. The DID estimator of the effect of the policy is given by

$$\hat{\delta}^\tau = (P_2^\tau - P_1^\tau) - (P_2^c - P_1^c) \qquad (1.54)$$

The assumption of common macro trends can be relaxed if an earlier comparison period can be found in which the macro-economy behaved similarly to the period $t=1$ to $t=2$ (say a similar point in the cycle). Call the previous period "period ρ to period $\rho+1$". The 'trend adjusted' difference-in-differences' estimator is given by

$$\hat{\delta}_{TA}^\tau = (P_2^\tau - P_1^\tau) - (P_2^c - P_1^c) - [(P_{p+1}^\tau - P_p^\tau) - (P_{p+1}^c - P_p^c)] \qquad (1.55)$$

The idea here is that any differential macro trends between the treatment group and the control group which have a tendency to occur at a particular point in the business cycle should be captured between time ρ and time

$\rho+1$, before the treatment was given. This estimate of the differential trend is then netted off the standard DID estimate to give the trend-adjusted DID. Obviously the trend adjustment may be misleading if there have been large changes in the composition of the treatment and control groups between period *p* and the present period.

The DID estimator is very popular in recent applied empirical work as an estimator which makes as few structural assumptions as possible. This simplicity is sometimes overstated; as Blundell and MacCurdy (1999) point out, the DID model basically decomposes the change in the outcome variable between periods 1 and 2 into a fixed effect (specific to the control or treatment group being looked at) and an effect common across both groups, with the differencing process stripping out the common effect. Looked at like this, it is identical to a stripped-down 'structural' model with fixed effects terms. Nonetheless, it remains the case that DID estimates a programme effect under the minimal (but very important) set of assumptions shown earlier.

The strength of the DID approach is that there is nothing constraining the effects of a program to go in a certain direction. For example, we might expect a reduction in benefits for non-working people to have a positive impact on labour market participation. A structural analysis which stipulated that individuals maximised well-behaved utility functions derived from neoclassical labour supply theory might well estimate a model where the labour supply response to a benefit cut was *compelled* to be either zero or positive. In the DID model however, a *negative* labour supply response to a benefit cut is entirely possible. This would strike some researchers as odd, in that it is not predicted by the standard static model of labour supply. However, the predictions of more complex dynamic models such as the search-matching models considered in section 2.1 are much more ambiguous. From this perspective, the DID approach is likely to appeal to two main groups of researchers:

(a) those who are sceptical about the extent to which the labour supply theory outlined in Sections 1.1 and 1.2 applies in the real world;
(b) those who believe that more complex theoretical models (e.g. search/matching) are correct) but that empirical structural models are currently too simplistic to capture the important features of the labour market.

The weakness of the DID approach compared with the structural approach is that we have much less to go on when trying to interpret the results. In a model where structural parameters are estimated it is easy, for example, to decompose the effect of a policy which alters the budget constraint into income and substitution effects. With the DID model, there are no estimated structural parameters, so we are left in the dark as to what features of the economy led to the measured labour supply effects. The problem worsens when we consider attempting to predict labour supply effects in other circumstances. The structural approach gives researchers a set of estimated parameters which they can in principle apply to new situations and completely different samples. The DID results, by contrast, cannot be divorced from their

particular temporal and spatial setting. Hence the only way to get generalised results for the impact of labour market programs from the DID methodology is to do a separate evaluation of each programme. For this reason, DID sits squarely in the camp of 'evaluative' rather than 'predictive' labour supply models if we follow the schema used earlier.

2.2.2 Grouping estimators

We have focused heavily on the difference-in-differences technique in this section, reflecting its popularity as a technique for the evaluation of labour market policy. However, grouping estimators, which can be used to evaluate policy reforms when treatment and control groups are not as clear-cut as they are in the standard DID case, are also popular, and are very similar in terms of the methodology and the assumptions used. The grouping approach uses a discrete grouping variable or set of variables G which allocates individuals into $g = 1,..., J$ groups of size N_{gt} in each period $t = 1,...,T$. For an individual i in group g, the grouping model can be expressed as

$$y_{it} = \gamma\delta_{it} + \beta' X_{it} + \theta_g + \eta_i + m_t + \varepsilon_{it} \qquad (1.56)$$

where y_{it} is the outcome variable of interest (e.g. labour market participation), δ_{it} in this case is a 'treatment' dummy variable for whether the policy being evaluated was administered to the group or not, X_{it} is a vector of control variables, θ_g is a group-specific fixed effect, η_i is an individual-specific fixed effect, m_t is a time specific 'macro' effect and ε_{it} a randomly distributed error term. If we average y_{it} within groups, e.g.

$\bar{y}_{gt} = \dfrac{\sum_{i \in g} y_{it}}{N_g}$ and do the same for the variables on the right hand side of (1.40) we derive the grouped specification

$$\bar{y}_{gt} = \gamma\bar{\delta}_{gt} + \bar{X}_{gt} + \theta_g + m_t + \bar{\varepsilon}_{gt} \qquad (1.57)$$

which can be estimated with a full set of time and group dummies. Effectively this is a fixed effects panel model with the group as the primary unit of observation.

(1.57) represents a grouping estimator which is closest in spirit to the difference-in-difference model – where there is a clearly identifiable treatment δ_{it} which some groups are affected by and some not. However, similar specifications can also be used to examine labour supply responses when the 'treatment' and 'control' groups are not as clear cut. For example, Blundell et al (1996) work with a specification which is a modified version of (1.57) where $\bar{y}_{gt}$ is the proportion of each group in work, the $\bar{\delta}_{gt}$ term is omitted, and an additional term $\bar{R}_{gt}$, a function of the net return to working at various hours levels given an hourly wage imputed from a wage equation, is used instead.

This has the advantage that the $\overline{R}_{gt}$ variable can in principle be computed for out-of-sample data and hence the approach can be used for quantitative prediction in a way that is not usually possible in pure difference-of-difference studies. The grouping variables used are age bands, region of residence, the presence or absence of children in the household, and the level of educational attainment. Gregg, Johnson and Reed (1999) also work with a grouping estimator, which will be discussed further in section 2.3.4. If financial incentive variables are used in a grouped model the implicit assumption is that labour supply responses depend on quantitative work incentives in a systematic way, which begins to move us back towards the structural models shown in Section 2.1.

Grouping estimators are also popular as a means of controlling for unobservable individual heterogeneity which may be correlated with the outcome variable *y*. To the extent that such heterogeneity occurs *within* each group, the grouping procedure averages it out of the model. This is similar to an instrumental variables model where the within-group means of the variables are used as an instrument for the individual values. However, unobservables correlated *across* groups will not be eliminated by the grouping procedure.

2.3 Experimental methods and the Random Assignment methodology

A third type of empirical study relies on *random assignment* being incorporated in the design of a policy intervention. The theory behind random assignment works as follows: imagine an outcome variable *Y* (e.g. labour market participation) which is distributed amongst the population according to an equation such as:

$$Y_i = f(X_i, Z_i) \qquad (1.58)$$

for individual i, where X_i are factors observable to the econometrician and Z_i are unobservable factors such as motivation, innate ability, etc. Now imagine a labour market intervention q. Denote the outcome for individual i in the presence of the intervention as Y_i^q and without the intervention as Y_i^0. Under random assignment, a treatment group (τ) is given the labour market intervention but a proportion ρ is randomised out, i.e. they are placed in the control group (c). Under assumptions to be discussed below, the average treatment effect $(\overline{Y}^q - \overline{Y}^0)$ is given by simply comparing the average outcome for group τ with the average outcome for group c. The assumptions are:

1. that assignment to the groups τ and c is *truly random*, i.e. there is no correlation between the probability of being randomised out and any of the regressors X or Z.

2. that there is no contamination of the control group c, i.e. that their actions and labour market outcomes are unaltered from what they would have been in the absence of the policy being implemented.

If these assumptions hold then random assignment can be seen as an idealised form of the difference-in-differences methodology where the treatment group are not just similar, but *identical* (on average) to the control group in all characteristics. This makes it a very powerful evaluative tool, but obviously it has the drawback that it can only be used for studies for which random assignment has specifically been built in to the design of the evaluation.

2.4 How our model fits into the framework

The model which we are planning to estimate in this project (outlined in detail in Section 3.2) is not a fully structural model in the sense that we are not planning to estimate the parameters of a utility function. However, it has more structure than a 'pure' difference-in-differences model in that we are planning to relate movements into and out of work to the financial incentives which people face. This procedure is useful because the labour supply effects of alternative reforms to the tax and benefit system can then be simulated by using a tax and benefit model to calculate the changed work incentives which individuals would face under a reformed system, then predicting the extent of movements into and out of work under the new system. To an extent, the source of identification of our model is similar to DID models in that we rely on labour market reforms over the sample period having an impact on groups who were affected by the reform whilst having no impact on groups who were not affected. Our model is in no sense an experimental model as we use data from the labour market as a whole, with no opportunity for random assignment to be used. In Chapter 3 we say more about the identification of our model and its strengths and weaknesses versus a fully structural approach.

Chapter 3. Criticisms of the standard labour supply theory

The framework explained above attracts criticism from some economists, and indeed other social scientists, on several grounds. Rather than simply accepting the framework as a *fait accompli*, we feel it is important to examine criticisms, and where possible to formulate some response to them.

3.1 Criticism of the rational utility-maximising framework

Criticism is often directed at the whole framework of the rational utility-maximising agent postulated in the neoclassical labour supply model (see, for example, [reference from intro to Backhouse (ed) *New Directions in Economic Methodology*, 1997]. Many observers from outside the economics profession (and indeed some from within the profession) are sceptical of the kind of intertemporal models which require individuals to maximise expected utility over an extremely long time horizon – 50 to 60 years into the future in the case of someone starting out in the labour market. In some cases this hostility to the utility-maximisation framework is based on a misapprehension. Most economists are very careful *not* to claim that people actually go round with 'utility generators' in their heads. The utility function should be seen as a mathematical representation of the process of deciding whether to work or not and for what number of hours. In its most abstract theoretical form, economic theory does not prescribe what variables should enter the individual's utility function. Whilst it is true that much of the economic labour supply literature starts with the assumption that individuals are self-interested, like leisure and do not like work, and prefer more income to less, in principle it would certainly be possible to formulate models in which individuals, for example, derived positive utility from work over a certain number of hours, whilst still being rational utility maximisers. We have seen above, for example, that labour supply functions for individuals in couples often take account of altruism between husband and wife. The framework is also capable of taking account of the limitations of human beings as decision makers. Rationality may be 'bounded', and individual decisions may be made under deeply uncertain conditions. Thus a rejection of the entire utility-maximising framework as 'wrong' seems too harsh, as the framework in itself is extremely flexible. Rather, critics are often objecting to a specific usage of the framework (e.g. a particular feature of the models we have been looking at in this chapter). We examine more specific criticisms below.

3.2 Criticism of the way people are assumed to choose from the budget constraint

The budget constraint is normally represented as a well-determined and unique, if complex, function for each individual which is fully known to the

individual. In other words the standard theory assumes that individuals looking to enter work know exactly what the net return to working a certain number of hours will be. A study by England et al. (1996) criticised this assumption on the grounds that interviews with people actually looking to enter work revealed that in many cases they did not even know what benefits they would be entitled to, let alone what the net gain from those benefits would be. If these findings are correct then one could argue they present a major problem for the economist's view of the budget constraint. On the other hand, it may be the case that people 'learn by doing', i.e. if they are entering work for the first time or are re-entering the labour market after a long break then they may be initially unaware of what the net gain to work for them may be, or of what benefits they are entitled to. However, once they are in work and can see what their net gain is, they will presumably be able to make an informed choice about whether to stay in work or not. Thus while the economist's model of the budget constraint under perfect information may not be applicable in the short run, it should be applicable in long-run equilibrium once the individual has acquired the necessary information. Also, whilst it is true that we do observe individuals who do not take up benefits they are entitiled to even in the long run, there are economic models which have been postulated to account for this, focusing on the stigma attached to benign on benefits, or the 'hassle factor' involved in claiming (see for example Besley and Coate (1992), Fry and Stark (1992), and Bingley and Walker (1997)). And if benefits are not taken up, the budget constraint can be altered to account for this.

A related criticism is that the budget constraint which economists use to evaluate people's labour supply decisions misses out on too many factors which affect the 'real' return to work to be informative. Some costs and benefits which potentially affect the return to work are tangible but hard to measure, e.g. fixed costs of work due to travel, clothes or equipment, payments in kind and 'perks of the job', bonus payments, and so on. Some are intangible, e.g. on the positive side the gain in self-esteem from being in work, the access to social networks from being in touch with the workplace, and benefits from a comfortable working environment; and on the negative side, work-related stresses, the risk of industrial accidents, and poor working conditions. Certainly many of these are hard, perhaps impossible to measure. But fortunately, we are able to estimate econometric models which can allow for the presence of certain intangible costs in the estimation process (see section 1.3.1.7 on fixed costs modelling).

3.3 Equilibrium and the market-clearing assumption

One characteristic of the standard labour supply model is that it is rooted firmly in the notion of the labour market being in equilbrium (in the sense of the system being 'at rest' and not liable to change unless something external changes). In the static model of choice on the budget constraint at one point in time this is obviously the case. In an intertemporal setting, where the individual chooses an optimal path of consumption and labour supply over the

life cycle, this can certainly change over time but does so in a way that is fully planned from the start. If further 'shocks' to the economy occur then the individual's plans may be altered, but the alteration to the new equilibrium is assumed to occur instantaneously and without a period of disequilibrium in between. Part of the reason for the focus on equilibrium models in the labour supply literature is probably that equilibrium is a much easier state of the world to describe than disequilibrium. By its very nature, disequilibrium is a process which is likely to be unstable and in flux, and by definiton, the set of restrictions which can be relied on to hold in equilbrium in order to generate a tractable theoretical model will not hold in disequilibrium. Nonetheless, or view as economists of phenomena such as unemployment is likely to differ very much according to whether we think they are an equilibrium or a disequilibrium phenomenon. There are a number of different responses to the criticism that standard labour supply is overly concerned with equilbrium. One is to assert that the labour market as a whole is always in equilibrium because if it were not then forces would come into motion which would restore equilibrium (for example, by changing the market wage). However, this seems to ignore the fact that in certain markets (for example the labour market) speeds of adjustment may be slow (because of there being imperfect information, or wages being set according to long-term contracts which do not instantly adjust to clear the spot market, for example). Clearly if markets do not adjust instantaneously then some way of characterising a labour market which is moving towards equilibrium, but not necessarily there yet, may be needed. This is part of the justificiation of our focus on models which explore the dynamics of the decision to move into work and separations from work (such as the search/matching framework shown in chapter 2). We woud argue that models which explicitly incorporate the dynamics of labour market transitions *can* account for a system which is not in equilibrium at each and every time in a way which the traditional static analysis may find difficult.

3.4 Criticism of the idea that there is a free choice of whether to work or not, or the hours of work one works

A criticism of standard theory often heard in the macroeconomic literature in particular is that it relies too heavily on the assumption that the labour market 'clears'. Here we take market clearing as the assumption that there is no 'involuntary unemployment' in the sense that there are no unemployed individuals who would like to work but cannot secure a job at the 'prevailing real wage'. The concept of involuntary unemployment is notoriously difficult to pin down, because it can always be argued that someone who cannot secure a job at a given wage level should nonetheless be able to secure work at a lower wage; some economists would argue that for this reason all unemployment can be viewed as voluntary.[19] However, constraints on

[19] In addition, the macroeconomic literature on voluntary versus involuntary unemployment has tended to suffer from the gross oversimplification of assuming that workers in the economy are paid at a uniform

available jobs can be crudely modelled in the static framework by setting the wage to zero. Also, as we will see in chapter 2, search theory explicitly describes a situation where individuals are involuntarily unemployed in the sense that they are actively seeking work and (given a certain wage) would rather be in work than unemployed. So we would argue that the validity of the mechanics of the labour supply model does not hinge on labour market clearing per se.

Some researchers, whilst accepting that there may be a free choice between working and not working, have questioned the assumptions of the standard model over choice of hours. Whilst the distribution of hours worked for women is very wide in empirical data, the vast majority of men appear to work full-time (defined as 35 hours or more a week) – there are very few part-time male workers. Of course, in itself this fact does not prove that men face constraints on their hours choices. It may just be that men overwhelmingly prefer to work full-time or not at all. However, if constraints do exist (as, for example, Stewart and Swaffield (1997) suggest) then they can be incorporated in the model by replacing the continuous choice of hours on the budget constraint with a series of discrete points at the available hours levels. This dovetails neatly with the discrete-choice approach to structural labour supply estimation of van Soest (1995) and others, shown in section 1.3.

3.5 Problems specific to family labour supply

Apart from all the above there are some additional features of the structural model specific to the family labour supply estimation. In the discussion of the "unitary" model we already mentioned the consequences of assuming a single utility function for the couple. This is equivalent to treating family income from all sources as having the same effect on the labour supply of either of the partners. The model suggests that "who gets what" is irrelevant for decisions of allocation of spending and leisure. It neglects the within-household distribution of resources and makes the discussion of "wallet-to-purse" redistribution irrelevant. On the other hand at its current stage of development the "sharing rule" interpretation of the "collective model" includes features which might also seem very simplistic. Because it assumes "exclusiveness" of leisure and consumption the model rejects any possibility of extra utility as a result of spending time or consuming together. The assumption also implies, for example, that time spent on household production by one member of the couple has no direct effect on the utility of the other person.

The additional difficulty of modelling couples' behaviour is that it is hard to account for the dynamics of the decision making process. There are two dynamic issues one would wish to include in the model. First of all decisions concerning issues such as separation or divorce and decisions regarding having children, which undoubtedly influence labour market choices. Neither

---

real wage level; whereas in fact there is a huge range of wage levels, even conditional on all observable factors (see Gosling, Machin and Meghir, 1999).

the “unitary” nor the “sharing-rule collective” model account for these issues. Secondly as in the case of modelling choices of individuals we would want to account for allocation of leisure and consumption over time, and therefore address the questions of saving and use of assets.

All methods of accounting for allocation of leisure and consumption over time which we discussed in section 1.1 apply to the “unitary” model, since we model the family as if it were a single individual. This no longer applies for the “collective” set up, including the “sharing rule” interpretation of it. To account for accumulation and depletion of assets of couples in the “collective” model we would have to include the process of “sharing” of these assets over time. Questions as to who saves and who has access to assets become extremely important and there is no information to address such issues in the data. The collective model, in its current form at least, is therefore limited to modelling of “myopic” couples, or has to assume extreme limitations on allocation of resources over time.

Part 2 The dynamics of the labour market

The theory of labour supply in a static setting outlined in Part 1 continues to be developed as a means of explaining participation and hours of work choices amongst the working age population, and is very useful in a number of contexts. But of course, individuals do not make a once-and-for-all choice whether to work or not; statistics for the UK and other industrialised countries show that large numbers of men and women move into and out of jobs every month. For example, according to ONS statistics, in October to December 2000 there were around 15.2 million men aged 16 to 64, and around 12 million women aged 16 to 59, in work. An estimate from Labour Force Survey data for the same period shows that around 1.7% of men and 2.3% of working age women left jobs and moved into unemployment or a state of non-participation in these three months. This means that approximately 260,000 men and 280,000 women left employment in autumn 2000. A similar exercise for transitions from unemployment shows that of a stock of 940,000 ILO unemployed men and 600,000 unemployed women in autumn 2000, about 230,000 men and about 180,000 women moved into work by January 2001. There were also transitions of around 140,000 men and around 190,000 women into work from inactivity. Clearly, the aggregate employment and unemployment levels used for policy analysis by many commentators in the media and elsewhere do not reveal the full extent of the transitions into and out of work which go on in the labour market.

Any sensible theory of the operation of the UK labour market has to take into account the dynamic nature of the market. Whereas naïve interpretations of the static model might lead us to believe that the working age population is split between relatively static 'blocks' of employees and non-participants, with movement occurring only at the margins, in fact every year sees a huge amount of transitions into work, out of work, from non-work seeker to work seeker, and movements between jobs. The UK Labour Force Survey (LFS) provides the best source of data on short-term movements within and between employment states in Britain, due to its five quarter 'rolling panel' structure.[20]

Gregg and Wadsworth (1999) use data from the LFS to describe the flows into and out of employment, unemployment and inactivity. They find that between 1992 and 1998, the percentage of economically inactive working age people entering employment in a three month period was fairly constant at around 5 percent. Over the same period outflows from *unemployment* (i.e. defined by the ILO as not in work but actively seeking and available to start work)

[20] The British Household Panel Survey (BHPS) provides information over a number of years for each individual in the survey, whereas the LFS only runs on each survey member for 15 months. But the BHPS has a much smaller sample size. The New Earnings Survey (NES) provides annual panel data for a 1% sample of employees who pay National Insurance contributions, but it contains very little auxiliary information on participants to use for analysis.

increased from 9.8 percent to 14.5 percent. Total outflows from employment reduced from 3 percent to 2.6 percent over the period.

Of course it is impossible to tell directly from these figures what the changes in the *stock* of employment, unemployment and inactivity were over the 1990s. A very small percentage increase in outflows to work from inactivity could result in a large increase in employment if the stock of inactive people was very large. In fact, we know from aggregate statistics that the year average employment rate increased from 70.4% of working age population to 74.6% between 1993 and 2000, while unemployment fell from 10.7% to 5.7% and inactivity fell slightly from 21.2% to 20.9% over the same period.

The theory of intertemporal optimisation introduced in chapters 1 and 2 provides a dynamic element in the static theory of labour supply. In intertemporal labour supply models (for example, MacCurdy, 1983), individuals move in and out of work in response either to pre-planned 'lifetime participation profiles' designed to maximise discounted lifetime utility, or to shifts in those profiles caused by unforeseen changes in the wage distribution and other features of the labour market. However, in this chapter, we will take the dynamics as the starting point of the theory rather than as its final development. Introducing transitions and dynamics into the model leads to a whole new set of theoretical and empirical issues and problems which we will also be discussing in detail here.

Part 2 of Report 1 begins with Chapter 4, which looks at different theories of the processes by which workers move in and out of jobs, and the related question of how and why wages change over time within and across jobs. We cover a number of different areas, including the human capital model of wage determination, search and matching models of the labour market, deferred compensation theories, and macroeconomic models of job creation and destruction. Chapter 5 covers the empirical evidence on wage and work dynamics, looking at wage mobility over time, the returns to experience and tenure, comparisons between entry wages, exit wages and overall wages, the penalty to labour market displacement, and patterns of labour turnover. Finally, Chapter 6 considers the empirical estimation of labour market models in a dynamic context, covering hazard or transition models, structural models of labour market search, general equilibrium search/matching models, life cycle models of the choice between work, schooling and other activities, and the previous UK work by Gregg, Johnson and Reed (1999).

Chapter 4. Theories of labour market dynamics: work transitions and wage progression

A simple static theoretical perspective views individuals as choosing whether to work or not at a given wage. If the labour market is perfectly competitive the wage will be equal to their 'marginal revenue product', i.e. the value of what they contribute to the productive process. We will say more about the importance of the competition assumption later but for the moment, let us assume that the market is competitive. In this case, if each individual entering work were paid a fixed wage which did not vary (in real terms) over the course of his or her working life, the static model would be completely adequate for capturing labour market dynamics.

However, there is a good deal of evidence which indicates that individual wages *do* change. For example, in the UK Gosling, Johnson, McCrae and Paull (1997) report that for a sample of working-age men in the British Household Panel study between 1991 and 1994, whilst 35% were paid a 1994 hourly wage that was within 10% of the 1991 figure, 29% had a wage increase of more than 20% over the period, whilst 12% saw a decrease of more than 20%. For women the figures were roughly comparable although 33% saw 20-percent-plus wage increases. Wage growth was significantly more likely for younger workers and for the relatively highly educated, conditional on other factors.

The fact that wages can change on the job, and when moving between jobs, means that labour supply may depend not just on the immediate wage which an individual can earn on entering work, but on his/her expectation of future wage changes. This can lead to instances of observed behaviour which are extremely difficult to rationalise in the static model, but make sense when dynamics are taken into account. For example, suppose that someone takes a job at a wage which is so low that his in-work income is lower than his out of work income. In the static theory this could only be a rational decision if the job conveyed certain benefits (e.g. benefits-in-kind, psychological benefits such as increased self esteem etc.) which were enough to more than compensate for the loss of income from working. But if wages increase over time in a job, and the individual is forward-looking, this is no longer the case. If an individual takes a job where the present value of the (net) earnings stream from working is greater than the present value of remaining unemployed, then the initial observed wage may be completely uninformative. What matters is the potential for growth in the individual's wage over time. Growth could occur by staying with the same firm or through switching jobs after a period in the initial job.

There are a number of different economic theories aimed at explaining some or all of the dynamic aspects of the labour market - why wages change within jobs and between jobs, why workers enter and leave jobs, and what this does

to the aggregate distribution of wages and patterns of labour turnover in the economy. We survey the contribution of each theory below. Some of these are complementary while some are in direct opposition to each other. As we shall explain below, there is at present much room for coexistence of opposing theories as debate is still ongoing as to whether the empirical evidence justifies one theory or another.

4.1 Human capital theory

4.1.1 The human capital model of wages and skills

Neoclassical models of wage determination almost invariably take as their starting point human capital theory, stated in its canonical form by Becker (1964)[21]. The human capital approach starts with the assumption that wages are equal to (or at least closely linked to) a worker's marginal productivity, and that productivity is largely determined by a worker's skill level or 'human capital' to use the technical term. Human capital can be augmented by investment (e.g. in schooling and/or training), and an individual decision to invest in human capital is modelled in a way similar to the way economists model a firm's decision to invest in physical capital. However, it would be naïve to suppose that *only* schooling and training determine the level of wages. The human capital approach allows for other factors such as initial ability, motivation, physical fitness and so forth to play a role. The pure model assumes a perfectly competitive labour market with no mobility costs between jobs and perfect information, so that any two workers with the same human capital level are paid the same wage. These assumptions can be relaxed (for example, by postulating imperfect competition) with the result that human capital is no longer the sole determinant of wages, although it remains an important determinant.

In a human capital model, wage growth on the job is caused by the accumulation of human capital whilst in the job. Becker (1964) suggested two different types of human capital, with different implications for wage growth when *switching* jobs:

- **General** human capital: In its purest form, general human capital is valued equally in alternative jobs as well as in the job the employee is currently in. If the human capital accumulated in a job were general, we should expect to see offered wages in alternative jobs rise by the same amount as the wage in the employee's current job.
- **Specific** human capital, valued only in the current job the employee is in. If the human capital accumulated in a job were completely specific

[21] We use 'neoclassical' in a narrow sense here to mean methodologically individualist theories emphasizing individual rationality in a perfectly competitive equilibrium market environment. Of course, a wider definition of 'neoclassical' might well encompass most if not all of the alternative theories we discuss in this survey.

then we should expect to see no reflection of the accumulated capital in outside offer wages. *Ceteris paribus*, this would be expected to make staying in the same job for a long period more attractive than switching jobs, relative to a situation where all human capital was general.

By Becker's own admission, the dichotomy between general and specific human capital is a drastic simplification. In reality, many skills acquired during a job may have some specific elements, but also general elements. To use a homegrown example, the IFS's tax and benefit microsimulation model (TAXBEN) is written in Borland's Delphi programming language. Knowledge of the exact routines which comprise TAXBEN constitutes specific human capital in that no other organisation uses those routines. But knowledge of Delphi is a much more general skill. Stevens (1994) introduced the concept of 'transferability' of skills to characterise this continuum between fully general and fully specific skills. The less specific a skill is to a single firm, the more transferable it is.

The concept of generality of skills also requires elaboration. Whilst some skills (e.g. word processing) may be useful to a vast number of firms throughout the economy, other skills may be useful only within a particular industry – for example, making hamburgers is probably only of use to fast food firms. In industries which are oligopolistic (such as aircraft construction) we have a case where industry-specific skills might be described as 'general, but only within an oligopolistic market'; as the literature on industrial organisation shows, this will probably be the hardest case to analyse.

To the extent that skills are specific to a firm, it should be noted at this point that the fact that there is only one firm in which the skill will be rewarded moves us away from the competitive paradigm (because the firm is effectively a monopoly employer of the specific skill).[22] This gives rise to a scope for bargaining between the firm and the worker over to returns to specific human capital[23]. If training is necessary for human capital to be accumulated, the question of who *funds* the training then becomes important. In the case of general training, we would expect the worker to pay the cost, as the training is fully portable between firms and so a firm which paid for general training and attempted to recoup the cost from the worker by the paying the worker less would create an incentive for the worker to 'jump ship' to another firm. In the case of specific training the worker's outside wage options do not reflect the increase in the inside wage, and so there is scope for bargaining between the firm and the worker (see, for example, Hashimoto (1981) and Stevens (1994) for models of the distribution of returns between firm and worker when training is specific).

[22] It can be argued that markets are fully contestable this may not be the case, as a new entrant firm could employ the worker in the same skill. However this is unlikely in cases such as the IFS tax benefit model described earlier; for copyright reasons if nothing else, it is unlikely that a competitor firm would duplicate the TAXBEN modules exactly.

[23] There are many examples of theoretical models of the worker-firm bargain over returns to specific capital in the literature. Two good examples are Hashimoto (1981) and Harris and Felli (1996).

4.1.2 Implications of human capital theory for wage dynamics and labour supply

If skills are transferable and the labour market is competitive then the implications of human capital theory for wage dynamics are straightforward. Wages grow in line with marginal product and we would expect outside wage offers to be in line with the employee's current wage (although this may not necessarily hold in a model where there is comparative advantage to working in different sectors, e.g. Heckman and Sedlacek (1990); or where the employer has paid up front for general training but is recouping costs through the worker). If skills are specific to the job, outside offers will probably be less than the current wage and one would expect this to provide an incentive towards longer tenure in the current job compared with the case of purely transferable skills. The wage may also diverge from marginal product, although here there are no clear-cut theoretical predictions.

The main implication of human capital theory for labour supply is that it is insufficient merely to take into account the starting wage when doing a budget-constraint based analysis of the labour supply decision. If there is scope for wage growth on the job through investment in human capital, then the starting wage is likely to provide an underestimate of the incentives to work. For example, a person might be willing to enter a low-paid job which provides zero or negative returns to work in the short run if they anticipate wage growth in the medium to long term.

However, the human capital model on its own does not seem to provide a complete picture of the dynamic operation of the labour market unless it is augmented in some way. As explained so far, it fails to account for why we see *downward* changes in wages over time for some workers in the same job. This could be due to depreciation of existing human capital, changes in the returns to human capital or a divergence of wage from marginal product under imperfect competition ('exploitation') but most literature on estimation of human capital earnings functions does not consider these factors in any great detail. Nonetheless, wage movements in both directions exist in the data, and need to be accounted for. In the next section we look at an approach which considers labour market dynamics more explicitly.

4.2 The search/matching approach

4.2.1 The basic search model

A class of models, which we label here the 'search-matching approach', departs from the perfectly competitive paradigm which was used for early expositions of human capital theory. The key assumption which is dropped in the search model is that of costless information. In a search model an employee does not have immediate and costless access to the entire job offer

distribution. Instead, in the most basic search model, the environment for a jobseeker is characterised by the following: [24]

1. the worker seeks to maximise expected present value of income, discounted to the present over an infinite horizon at rate r.

2. Whilst unemployed, the income flow net of any search costs is b and it is constant over the duration of a given spell.

3. Offers are received while unemployed according to a Poisson process with parameter δ (the arrival rate of offers). The probability of receiving at least one offer within a (short) time interval h is $\delta h + o(h)$, where $o(h)$ is the probability of receiving more than one offer in the interval and $o(h)/h \to 0$ as $h \to 0$.In some models δ is viewed as the 'offer rate', determined by employers, in other models as 'search intensity' determined by the jobseeker, or as a combination of the two.

4. A job offer is summarised by an entry wage rate w and a wage profile $w_\tau = g(\tau)$ where τ is job tenure. In the simplest case $w_\tau = w$ for all τ.

5. Successive job offers received over the course of a spell of unemployment are independent realisations from a known wage offer distribution with finite mean and variance, cumulative distribution function $F(w)$ and density $f(w)$.

6. Once rejected, an offer cannot be recalled.

7. When accepted, a job lasts until retirement, or until a layoff, where layoffs follow a Poisson process with separation rate ϕ.

Most of assumptions 1 through 7 can be relaxed at the cost of greater computational complexity in empirical modelling. We will explain the effect of relaxing some of the assumptions in what follows below. The key divergences from the perfectly competitive costless information model are:

- Job offers do not occur instantaneously but only after a period of search.

- Job offers are not all the same for a person of given characteristics.

This means in turn, that we would expect to see a certain amount of search ('frictional') unemployment in labour market equilibrium.

The worker's decision whether to accept a job or continue searching depends on an *optimal stopping strategy*, which maximises the expected value of the accepted job net of search costs. In the simplest continuous-time case, the

[24] The following framework is based on section 2 of Mortensen and Pissarides (1999) and chapter 2 of Devine and Kiefer (1991), both of which are excellent surveys of the field. It also incorporates recent advances in the literature detailed in van den Berg (1999).

instantaneous probability that a job offer will arrive at time t is denoted by the hazard function $\lambda(t)$. The associated survival function of the waiting time distribution (the probability that an offer will not arrive before time T) is denoted by $e^{-\int_0^T \lambda(t)dt}$. The value of search at time t, V_t^u, can be formulated as a Bellman equation (Bellman, 1957):

$$V_t^u = \int_t^{\infty} \left((b-a)\int_t^T e^{-r(s-t)}ds + e^{-r(T-t)} \int \max\{w, V_T^u\} dF(w) \right) \times \lambda(T) e^{-\int_0^T \lambda(t)dt} dT \quad (2.1)$$

where r is the individual's discount rate, b is income flow received contingent on unemployment less any cost of searching for a job, and a represents the cost of search. The optimal strategy involves comparing the current offered wage w_t with the value of continued search V_T^u at time T, when an offer arrives. If the former exceeds the latter then the search process stops. Thus $V_T^u = w^r$, the reservation wage (the smallest wage at which the searcher will move into employment) in this model.

The path of the value of search over time depends on the distribution assumed for $\lambda(t)$. If $\lambda(t)$ is assumed constant, the waiting time distribution is exponential and the value of search is stationary, solving the equation

$$\begin{aligned} V_t^u = V^u &= \int_0^{\infty} \left[\frac{b-a}{r}(1-e^{-rT}) + e^{-rT} \int \max\{w, V^u\} dF(w) \right] \lambda e^{-\lambda T} dT \\ &= \frac{b-a}{r+\lambda} + \frac{\lambda}{r+\lambda} \int \max\{w, V^u\} dF(w) \end{aligned} . \quad (2.2)$$

where λ is the Poisson offer arrival rate. If $\lambda(t)$ varies with time the value of search will be non-stationary, and a differential equation can be obtained relating the value of search to the time t, the search cost, benefit level and the wage distribution:

$$rV_t^u = b(t) - a(t) + \lambda(t) \int [\max\{w, U(t)\} - U(t)] dF(w) + \frac{dU(t)}{dt} . \quad (2.3)$$

Models derived from (2.3) can be empirically tested using data on unemployment spells and entry wages; we go into this in more detail in Chapter 6. Some predictions from the simple model are that:

a) The reservation wage increases if benefit income increases.

b) The reservation wage decreases if the discount rate r increases (because the value of continuing search goes down).

c) The reservation wage increases if the rate of offers λ increases (because continuing search increases in value).

It is useful to compare these predictions with those of the static labour supply model featured in Chapters 1 and 2. Taking (a) first, an increase in (out-of-work) benefit income corresponds to a shift upwards on the vertical axis of the budget constraint. If the benefit is withdrawn when in work (as this basic search model assumes), individuals are unambiguously less likely to work. The static single-period model occurs in instantaneous time, and so tells us nothing regarding (b). Life-cycle optimisation models have a role for the discount rate r but the impact of an increase in the discount rate is not straightforward in most models. As for the effect of the offer rate in (c), the static model normally assumes an infinite offer rate (or, at least, a certain offer in the instantaneous time period) at a given wage. So even this simple search model introduces concepts which are glossed over or simply not mentioned in the basic static model. As we show below, this process continues as we make the model more complex.

4.2.2 Extending the basic search framework

The search model described in detail above is merely a starting point for most current empirical work. In particular, many of the assumptions have been relaxed as follows:

Varying the discount rate

The discount rate r can vary over time or across individuals as a function of other observable characteristics of the jobseeker. This complicates the equations (2.2) and (2.3) shown above but the model is still empirically tractable.

Variance in benefit levels over time

In many benefit systems the benefit level b is not constant over time. For example, in the UK, contributory job-seekers allowance (C-JSA) lasts for only six months from the date of first claim, after which a jobseeker would suffer a drop in net income in most cases as he or she moved onto non-contributory JSA (if eligible). b can be treated as time varying at the cost of a slight increase in computational complexity (see for example van den Berg, 1990).

Variance in offer rates

One might expect the offer rate to vary, either for reasons external to the individual jobseeker (e.g. macroeconomic variation) or reasons to do with changes in the jobseeker's own characteristics (e.g. skill depreciation during extended periods of unemployment). This can be incorporated by making the parameter of the Poisson arrival process a function of time or other observable characteristics, which again increases computational complexity in estimating search models. Van Den Berg (1990) and Garcia-Perez (1998) both examine this possibility.

Allowing for wage progression

The function w_τ can be amended to include possible wage growth after entering the job, resulting from returns to tenure and/or experience from human capital or some other model of the labour market. Later in this chapter we examine models which incorporate search with wage progression (e.g. Keane and Wolpin, 1997).

Incorporating multiple transitions

The basic search model considered above is merely a starting point for an analysis of labour market dynamics as it only deals with a single transition per worker, and in one direction – out-of-work to in-work – at that. The model has been extended in a number of directions to deal with multiple transitions and multiple states. We give more details on how these models work later in this chapter in section 6..1, when we look at the specifications of 'hazard' or transition models. The main types of models arising from these extensions are:

i. **Repeated search models.** In this class of models, the economic agent is allowed to search further for better jobs after a job has been found. The most straightforward implementation of this model is the *on-the-job* search model (Mortensen, 1986; Arellano and Meghir, 1991) where workers in a job maximise the present value of search as given by the Bellman equation

$$V_t^{w_c} = \int_t^\infty \left((w_c - a) \int_t^T e^{-r(s-t)} ds + e^{-r(T-t)} \int \max\{w_o, V_T^{w_1}\} dF(w_o) \right) \times \lambda(T) e^{-\int_0^T \lambda(t)dt} dT \tag{2.4}$$

where w_c refers to the worker's current wage and w_o is the offer wage . In this simple model the optimal strategy is that one accepts a job only if the offered wage exceeds the current wage. As it stands the offer rate and search costs are identical for workers compared with the unemployed, but the model can be easily altered to allow for differential search costs and offer rates for people in each labour market state.

ii. **competing-risks models.** In these models there is a possibility of exit to more than one state for people starting off in a given state at time $t = 0$. A natural extension of the search model to more than two final states would be to consider becoming inactive as another possible final state. Thus, an unemployed work seeker who fails to find work in a given period can carry on searching in the next period, or can give up. This might be reasonable as a model of the decision to take early retirement, or of job search amongst women deciding whether to return to the labour force after having children. As this model is only partially a search model we do not discuss it further here, but we do return to it in Section 2.3.1.

iii. **Successive durations models.** These models consider events occurring *after* the initial transition which might return the employee to his or her initial unemployed state, or to some other state (such as inactivity or full-time education). Obviously, to generate a situation where employees are returned to unemployment, there has to be some 'job-termination' process going on in the model (either something causing the worker to quit, or the firm to lay the worker off). We say more about possible job termination processes below. Successive durations models are of varying complexity. At the simpler end there are models such as Mealli and Pudney (1996) and Ham and Rea (1987) which look at multiple transitions between unemployment and employment. The most complex approach currently taken is that of papers such as Keane and Wolpin (1997) who model multiple labour market transitions as the outcome of far-sighted and dynamically optimising individuals in a life cycle framework. This approach draws in concepts from the intertemporal labour supply literature studied in Chapter 1 and we discuss it in detail in Section 6.4 below. Of course, the greater the number of transitions covered in an empirical model, the more stringent are the data requirements for that model to be estimable. Thus many researchers are unable to exploit successive-duration modelling to its full extent because the data will not support it in each individual case. We return to this later in this chapter when we look at different types of empirical models.

4.2.3 Two-sided search: matching firms and workers

The basic search model gives a characterisation of how workers respond to job offers. Matching models, first developed in the 1980s, extend this framework to include search by employers on the demand side of the labour market. An analogue of equation (2.2) can be derived for the expected value of the firm's future profit:

$$V^{\Pi} = \frac{-c}{r+\eta} + \frac{\eta}{r+\eta}\int \max\{V, J\}dG(J) \qquad (2.5)$$

where V^{Π} represents the value to the firm of holding a vacancy open, c is the cost of recruiting a worker to fill a vacancy, η is the frequency with which an employer encounters workers seeking employment, J is the value of filling the job (i.e. productivity), and G represents the cumulative distribution of J over the workforce. Once again r is the discount rate.

In matching models, a successful match commands quasi-rents *ex post* because it is costly in time and resources for either the firm or the worker to seek another match. The implication of this is that the 'market wage' is not unique and therefore some mechanism needs to be specified for the division of the gains from the match between worker and firm. To satisfy individual

rationality, the share of the quasi-rents distributed to each partner must be at least as large as the forgone option of continued search.

In search equilibrium models, a matching function $M(u,v)$ is specified which relates the aggregate rate at which new matches are formed to the number of unemployed workers u and the number of vacancies v. A full definition of search equilibrium specifies the measures of search participants u and v as well as V^u and V^Π. Much of the literature assumes a constant supply of (potential) workers with an unlimited number of (potential) vacancies. This seems to ignore any general equilibrium considerations, in that in a general equilibrium framework, we might expect feedback from changes in labour supply and demand to the overall distribution of wages (as modelled by, e.g., Creedy and Duncan, 2001). On a deeper philosophical level, the search/matching model itself means that the coherency of a single 'aggregate labour supply schedule' (or indeed a labour demand schedule) is called into question. We examine attempts to incorporate general equilibrium effects into empirical search matching models in Section 6.3 below.

4.2.4 Exits from employment in the matching model

Obviously, unless job matches are terminated at some point, then with positive inflows into work unemployment would decrease to zero in long-run equilibrium. Pissarides (1990) examines an 'equilibrium unemployment' model where on the job creation side, $m(u,v)$ is increasing, concave and homogenous of degree 1. All jobs have the same productivity. Employer and worker negotiate a wage by generalised Nash bargaining when they meet and subsequently produce until an idiosyncratic shock destroys the job match. Thus job destruction in this case is assumed completely exogenous. The model generates a negative relationship between vacancies and unemployment, known as the 'Beveridge curve'. As firms are behaving as competitive profit-maximisers, aggregate employment is determined by the condition that the value of a new vacancy should be zero in equilibrium, i.e. there should be no unexploited rents arising from vacancy creation. It should be noted, however, that whilst this framework provides a means of allowing for exits from employment, it doesn't really *explain* those exits as the process is treated as completely exogenous (a situation analogous to the treatment of the growth mechanism in growth models before the invention of endogenous growth theory, for example).

Some basic predictions for the relationship between economic variables from this simple matching model are:

- Increases in benefit income increase equilibrium unemployment. The average wage for those in work also increases.

- Increases in the worker's share of the quasi-rents from job matching relative to the firm increase unemployment and wages.

- Increases in the productivity from successful matches reduce unemployment and increase wages.

- An increase in the rate of job destruction reduces wages and increases unemployment.

The model can be increased in complexity in several ways, such as introducing heterogeneity of workers (e.g. by differential human capital endowments), allowing for more complex job destruction processes (such as the model of Mortensen and Pissarides (1994) where match productivity evolves according to a Markov process) and introducing specific rules for the division of the match surplus between workers and firms (e.g. bargaining models such as insider-outsider or efficiency wage models, or competitive pricing models where third-party 'market makers' ensure that wage is equal to marginal product by exploiting all opportunities for gains from trade). However in many of the extended models the straightforward predictions given above do not necessarily hold, because there are often different effects of changes in parameters such as out-of-work income, entry wages and wage growth on the job creation and job destruction rates. In general, the comparative static results from different theoretical models in this vein can vary widely according to the specific assumptions made in each model. What this shows is that making policy predictions from the search-matching literature taken as a whole can be a somewhat hit-and-miss affair. Thus in what follows we confine ourselves to some fairly basic predictions about what may happen to wages and job entry and exit in this framework.

4.2.5 Implications of the search model for wage growth

For the moment, let us assume that there is no human capital accumulation on the job, and that instead, the labour market is described by a search/matching model[25]. The implication of such a model is *not* that everyone of a certain schooling level gets the same wage ad infinitum. Because of the assumption of a non-degenerate offer distribution, some workers receive better job offers than others, despite having the same characteristics. The better a job match is, the higher the wage, and so the greater the incentive to stay in work. Even with a Poisson layoff process present in the model, it can be shown that the attractiveness of well-matched job offers generates a positive relationship between the wage and tenure in the job in cross-sectional data. This is due to selectivity bias – the set of jobs with high tenure is a selected sample of well-matched jobs. A simple regression of wages on tenure would find a positive relationship even if wage growth on the job were zero! Furthermore, if we assume that job offers can arrive even when a person is already in work (thus allowing 'on the job search' in the model)[26] then we would expect to see a positive relationship between

[25] This is the framework described by Manning (2001b).

[26] an assumption is often made that on-the-job search is less effective (e.g. produces fewer offers per time period) than off-the-job search because an unemployed individual can engage in search as a full-time activity whereas an employed person has to search part-time. But this is not strictly necessary; for example, it might be easier for a working person to find contacts for new job offers.

age and earnings purely because an older person is more likely to have received a well-matched job offer at some point. In this case there *is* a positive relationship between age and earnings but no growth in earnings on the job – wage growth is secured by changing jobs. This is the explanation of job changes offered by Topel and Ward (1992) among others.

Thus the search model indicates that for an individual worker, we would expect to see returns to experience which are positive, but returns to tenure which are zero. Furthermore, the returns to experience should come from switching jobs (to a better matched job). We would also expect that a person who is laid off from a well-matched job would be likely to suffer a substantial pay reduction in his or next job as it is unlikely that he or she would be able to find another job as well-matched straight off. For a person in a badly matched job this kind of loss to changing jobs should not occur.

4.2.6 Implications of the search/matching model for labour supply

The fact that the search/matching class of models makes the dynamics of the labour market much more explicit is all well and good, but how much do these models tell us about labour supply responses which the models examined in Chapters 1 and 2 do not? The most obvious innovation is that the search model provides a possible explanation for unemployment in a way which the static labour supply model does not. Unemployment and non-participation are clearly behaviourally distinct states in the search/matching model, whereas in the analysis of chapter 1, there was no functional distinction to be made between them.

The concept of 'match quality' of a job also provides a rationale for wage dispersion amongst observationally identical workers in jobs; if correct it suggests that the focus of the static model on a fixed wage w for each worker may be misplaced. Instead workers are offered a wage from a distribution $f(w \mid X)$, where X are observable characteristics. Changes in match productivity over time may also be able to account for why workers exit jobs and move back to unemployment. Later on we will examine the reasons for job exit more closely.

Recent work by Alan Manning (2001a, 2003) has argued that the implications of search theory for labour supply are more fundamental than anything we have alluded to so far, and indeed that they may make some of the standard results of static labour supply models invalid. Manning notes firstly that 'there is a curious dichotomy in which analysis of unemployment insurance generally uses a search framework while analysis of tax changes uses a labour supply framework even though both sets of papers are about the impact of changing incentives to work'. He analyses a case where individuals have no flexibility in choosing hours levels within a job, so in effect, the choice is a dichotomous one between working at the designated hours point or not working at all. Within this framework he analyses the impact of a change in marginal and average tax rates.

In the standard framework as outlined in chapter 1, in a single-period analysis an increase in the average tax rate leaving marginal rates unchanged, through (for example) a poll tax, can only induce an individual who was previously not working to start working; nobody who is already working will become inactive if the utility function is increasing in both consumption and leisure. In the Manning framework whether this holds depends on the effectiveness of off-the-job search compared with on-the-job search. If both are equally effective, or off-the-job search is more effective, then the standard result (or something close to it) holds. However, if on-the-job search is more effective, and workers are risk-averse, it is possible that an increase in average tax rates can *increase* the probability of employment. This is because the extra effectiveness of on-the-job search increases the attractiveness of working relative to not working even though the financial returns to work have decreased. The Manning (2001a) paper also analyses the impact of a change in marginal rates although here it is difficult to draw a straight comparison with the standard model as, because of the fact that there is a distribution of offer wages $F(w)$ rather than a single market wage for each individual, there is no single in-work income level to which a currently non-working individual can be assigned. His results are that an increase in marginal rates in the standard model holding total tax revenue constant (i.e. increasing marginal rates but reducing average tax rates, perhaps through a lump sum subsidy) has an ambiguous effect on labour supply in the standard model but unambiguously increases incentives to work in the search model.

Although Manning discusses a very specific class of model (where hours are completely fixed within a job) which is probably unrealistic, the standard neoclassical model where hours are perfectly flexible is also probably unrealistic and so his model can be viewed as an attempt to test the sensitivity of the standard model outlined in Chapter 1 to varying the assumptions slightly. His results seem to suggest that if assumptions over flexibility of hours choices do not hold then some of the predictions of the standard (static) labour supply model do not go through. This suggests that there may be serious problems in relying on the assumptions of simple labour supply models when doing structural labour supply modelling as outlined in Chapter 2.

4.3 Deferred compensation models

4.3.1 The Lazear model

Another type of theory which predicts a rising wage profile over time within the job is the 'deferred compensation' model, exemplified by Lazear (1981). This model departs from the perfectly competitive paradigm in that the worker's wage diverges from his or her marginal productivity for much of his or her employment duration, although in the Lazear version, the sum of *lifetime* wages is still equal to the sum of productivity over the duration of the job. In

the simplest version of the model, the worker's marginal product is constant over the duration of the job, as in the search/matching models described earlier. However, the firm faces a principal-agent problem in inducing the worker not to 'shirk' because it is assumed that monitoring the worker's activity so that the worker puts in the required effort is a costly process. The optimal solution is for the firm to 'backload' wage payments towards later in the worker's career. In the early part of the worker's tenure with the firm, he/she is paid less than marginal product; later on, wages rise above marginal product. The promise of higher wages later on in the duration of the job is an incentive mechanism to encourage greater effort earlier in the job than would be the case if the wage tracked marginal productivity throughout the duration of the job.

The Lazear model assumes that the labour market is competitive, which raises the question of how the firms are able to offer a wage less than marginal product to workers early on in the job – why do the workers not move to other firms paying them the value of marginal product? Lazear's view is that, because the deferred compensation contract is the profit-maximising solution to the problem of imperfect monitoring, only firms which offer deferred compensation contracts will be able to survive in equilibrium.[27] Other writers have taken more of an 'internal labour market' view of the wage-tenure relationship along the lines of Doeringer and Piore (1971). In these closely related models, firms are able to offer wages which diverge from marginal product because of imperfect competition in the labour market – in particular, the cost to the worker of moving from job to job.

4.3.2 Deferred compensation, the returns to tenure and experience and labour supply

In the pure deferred compensation model, we would expect positive returns to job tenure, but no returns to experience. People leaving jobs unexpectedly in the latter phase of their tenure with the company, when wage is above marginal product (e.g. being laid off, or fired due to shirking) would expect to see a reduction in wages on return to work. The model as it stands is rather incomplete, as it does not specify what the wage structure should be when returning to work after being laid off from a previous job. The analysis tends to be couched in terms of a 'job for life' and so has little to say regarding job to job moves. It is thus unlikely to be appropriate for workers who expect to move from job to job during their working life. Nonetheless the deferred compensation model is often offered as a possible explanation of returns to tenure in 'stable' employment, (e.g. in the professions, or many public sector jobs). If the model is correct, the implcations for labour supply modelling would appear to be:

- as with human capital theory, we can expect to see wage growth on the job, although for different reasons. Thus, it may be worthwhile for a rational farsighted individual to take a job which offers zero or even negative net

[27] An apparently unaddressed issue is what happens to workers who get laid off unexpectedly (e.g. due to firm level negative demand shocks). A layoff would appear to generate large costs for the employee as on resuming work, he or she would lose any payoff to seniority accumulated in the previous job.

gain to work in the short run, because wages are likely to rise later in the job.

- in the later phases of the worker's tenure with a single employer, wage levels are likely to outstrip outside offers because of the backloading of pay over the contract period. Thus the model makes voluntary separations less likely than a human capital model where the spot wage is equalised with marginal product *and* skills are transferable. (If skills are specific to the firm then the implications of both models are similar).

4.4 Summary: the implications of different wage growth models for returns to tenure and experience

The following table summarises the discussion above by showing what different theories of the evolution of wages within and across jobs would predict for the returns to tenure and experience, as well as for whether there is a wage penalty after being laid off from a job.

Table 4.1

Theory	**Return to:**		**Wage**
	tenure	**Experience**	**Penalty after layoff?**
Human capital (general)	0	+	No
Human capital (specific)	+	0	Yes
Search/matching	0	+	Possibly
Deferred compensation	+	0	Yes

Table 4.1 shows that the implications of models of wage growth through *specific* human capital and deferred compensation appear to be broadly similar, implying a positive return to tenure but no separate return to experience. On the other hand, the search/matching approach and *general* human capital models deliver the opposite result: zero returns to tenure, but positive returns to experience. It should be made clear however that the theories are not necessarily mutually exclusive. It is quite possible to imagine a world where there were returns to both general and specific human capital, but where wages were also affected by labour marker search and the quality of the employee-employer match, and where deferred compensation played a role in wage setting over the employee's career. If this were the case we would expect to see positive returns to both experience *and* tenure.

4.5 Theories of labour turnover and job separation

4.5.1 Introducing job exit

The search/matching theory explained above gives a lot of attention to the process by which people move *into* work, but doesn't by itself explain why people move *out* of work, i.e. the process of job exit. Of course in an economy with finitely lived agents, there are bound to be at least a certain number of job exits due to retirement (just as there are bound to be at least a certain number of entries into work due to new cohorts entering the labour market). However in this section we are more concerned with job exits over and above this (i.e. prior to compulsory retirement age). In this section we analyse theories of what causes job separations and how the process of job exit might interact with job entry. This will be important for Stage 1B of the project when we develop models of job entry and job exit.

In some ways the analysis of transitions from work to unemployment (or inactivity) is more complicated than the analysis of work entry because whereas the decision by a formerly unemployed worker to enter work is presumably always a voluntary decision initiated by the worker, an exit from work may or may not be voluntary in the same sense. Defining a job exit as a move from work into non-work (unemployment or inactivity), rather than a move from one job to another, we can distinguish three exit scenarios, from the most voluntary to the least voluntary:

1. **quits:** a 'quit' occurs when an employee decides to leave his or her job despite having the option of continuing in the job.

2. **layoffs:** a layoff occurs when the firm decides to terminate the employee's contract.

3. **plant closures:** a closure occurs when a firm has to shut down completely and hence all workers in the plant (or the firm) are laid off.

As we explain in Section 5.4 below, distinguishing between quits and layoffs in an empirical context can be very difficult. Here, we confine ourselves to a purely theoretical discussion.

4.5.2 Economic explanations of why job separations occur

Deterioration in match-specific productivity

This was discussed to an extent in section 4.2.3 earlier. The simplest search-matching model assumes that match productivity is constant within a particular job match. Given a constant external environment, the job continues

indefinitely. However, in more complex models such as Mortensen and Pissarides (1994) match productivity can decline over time. This means that other match options for the firm and/or worker become more attractive and means that there is a positive probability of job separation. The question *why* match productivity might decline over time is not usually explicitly addressed by microeconomic matching models, but of course this is the really crucial question – particularly as declining wages in a job (i.e. negative returns to job tenure) seem to go against the thrust of the human capital model shown earlier, where workers can enhance their wages on the job by on-the-job learning or training. Macroeconomic models have more to say about why productivity in jobs might change over time, a subject we return to later in this section.

Search-related explanations

Even in the absence of changes in match-specific productivity, job separations can be optimal in a search model. There are two particular situations where we might expect to see job exit:

(i) if a person moves from one job to another job with higher wages via on-the-job search. This is a job-to-job move with no intervening period of unemployment.

(ii) If off-the-job search is more effective than on-the-job search a person might need to leave his or her current job before searching for a higher paid job. This is a very different scenario from that considered in single period static labour supply models as the initial return to leaving the first job is negative; it is only after finding a new job at higher wages that any gain is realised.

If other features of the labour market change (for example, if the wage offer distribution is shifted due to changes in the returns to skill, or tax and benefit changes) then we may see job exit in a variety of scenarios. Some of these are considered by Manning (2001b), discussed earlier.

Changes in individual attributes

There are a number of situations where an individual might choose to leave work, either temporarily or permanently, due to a change in some aspect of his or her observable characteristics. Examples include:

- Ageing (e.g. retirement);
- Poor health;
- Women leaving the labour force to have children;
- Geographical relocation.

These kind of changes are often modelled as exogenous to the individual in empirical work although in some cases they may be factors which the household has some choice over (e.g. location, or childbirth); this is mainly

due to the difficulties of endogenising these choices in empirical models. The life-cycle framework of Keane and Wolpin (1997) among others, discussed in Section 2.3.4, seems probably the most promising model in which to endogenise decisions such as whether to have children or not, and whether to retire, but as discussed later, at present this can only be done under strong assumptions about the structure of individual preferences, individual rationality, and the way the labour market works.

Intertemporal substitution of labour supply

The intertemporal labour supply literature, discussed to some extent in chapter 1, places labour supply decisions within a life-cycle context. In a world with perfect foresight and rational optimising agents, we would expect individual decisions about the trade-off between work and leisure to be made on a lifetime basis. Individuals would decide the percentages of their time to allocate to human capital investment (schooling/training), work and non-market activities (leisure, child-rearing, etc) during their lives. We would be bound to see pre-planned job exits at various points during the life-cycle in this framework. For example, early retirement might be built into an individual's lifetime labour supply 'plan', if he or she knew that his or her earnings capacity would diminish even before compulsory retirement age.

To stand any change of being relevant to the real world, the intertemporal model needs to be amended to introduce uncertainty about productivity at different points in the life-cycle. Introducing uncertainty gives rise to the possibility that we might see unplanned exits from employment (or entry to employment) as the returns to work change and work becomes more or less worthwhile at a given time. Essentially agents update their optimal life-time labour supply plan in each time period as new information on wage rates becomes available. The possibility that unemployment in certain time periods might reflect such intertemporal substitution in labour supply was first formalised by Lucas and Rapping (1969). One of the first syntheses of the intertemporal model in a full life cycle setting was MaCurdy (1981, 1983). Recent empirical applications include Low (1999) and Keane and Wolpin (1997). There is a lot of debate as to how realistic this branch of the literature is in a real-world context, given that to be empirically tractable the models normally have to make sweeping assumptions (e.g. perfect competition in the labour market) – for a sceptical view see Card (1994). However if valid then it does provide an explanation for job exit at some point(s) during the life-cycle.

Business cycle explanations

The intertemporal labour supply literature has been influential in contemporary macroeconomics, particularly as a part of theories which attempt to explain why economies exhibit business cycles. The business cycle literature – both 'real' and 'monetary' is vast, and a full discussion is outside the scope of this paper; but our main focus here is on the fact that macroeconomic business

cycle theories can provide an explanation for the changes in match-specific productivity which microeconomic matching models normally take as exogenous.

The exact mechanism by which jobs are created and destroyed in the business cycle varies according to the source of economic fluctuations. In the real business cycle approach of Kydland and Prescott (1982) and Long and Plosser (1983), fluctuations in the rate of technological change are 'propogated' throughout the economy, leading to different levels of optimal employment in different periods in an economy with perfect foresight and far-sighted rational agents. Andolfatto (1996) combines the real business cycle framework with a search model of the labour market to generate a model which appears to correspond better to observed data on output and employment correlations than do the traditional RBC models. Other macroeconomic models of employment reallocation and job entry and exit include the 'sectoral shift' hypothesis of Lilien (1982) and Hosios (1994), where unemployment fluctuates due to the costly reallocation of labour between sectors, and the 'creative destruction' model of Caballero and Hammour (1994), where production units in the economy are of different vintages, with different productivity levels, and the timing of the gradual replacement of older plants by newer plants gives rise to employment fluctuations.

The key message from these macroeconomic models for our purposes is that there are a number of reasons why the productivity of a job match might vary over time, and these provide a theoretical underpinning for the job exit process in matching models, and bring an extra realism into our discussion of labour market dynamics. In section 5.4 below we examine the empirical evidence on the distribution of job changes and employment fluctuations over time.

4.5.3 Assymetries in the treatment of work exit and work entry in labour market models

There is an interesting dichotomy in the relationship between labour market transitions and the theoretical explanations for those transitions. In the case of work entry, this tends to be much more related to the budget constraint and to labour *supply* theory than does work exit, which is much more associated with labour *demand* theory. This has to do with the fact that it makes a lot more sense to treat layoffs as being initiated by the *firm* in the framework of labour demand adjustment than it does to treat them as labour supply responses. In the case of quits however, work incentives may still have a role to play. Clearly, in most circumstances quitting a job and moving into unemployment involves a loss of net income. The role of the benefit system here is sometimes overstated, as under current UK rules, former employees who leave their job 'voluntarily' are not entitled to Jobseekers Allowance or other out-of-work benefits for a period of six months from the quit point.[28] Nonetheless factors such as the rate of tax on earned income affect the net

[28] This of course creates an incentive on the employee's part for quits to be disguised as layoffs.

income loss associated with quitting and so one would expect changes in the tax system to influence the quit rate. Changes in the generosity of the tax and benefit system for those in work vis-à-vis those out of work also affect the net costs of on-the-job search versus off-the-job search and so one would expect an additional influence of the system on exit rates via this second channel.

Chapter 5. Empirical evidence on labour market dynamics

Introduction

Having covered the appropriate theoretical ground in chapter 4, this section discusses what empirical evidence there is in favour of, or against, the various theories of dynamics which we have examined. We begin by looking at evidence on the returns to experience and tenure.

5.1 Evidence on the returns to experience and tenure

A large amount of empirical work on the returns to experience and tenure has been undertaken in the last fifteen years or so. This remains a contentious area in which a firm consensus has not yet been reached. In this section we look at three recent papers in the field, each of which uses state-of-the-art empirical techniques, yet which achieve very different estimates of the returns to experience and tenure.

5.1.1 Topel (1991)

During the 1980s many studies reported that there were positive and substantial returns to experience, but that the returns to tenure were minor; for example Abraham and Farber (1987), Altonji and Shakotko (1987), and Marshall and Zarkin (1987). Topel uses data from the US Panel Survey of Income Dynamics for the period 1968 to 1983. He estimates the returns to tenure and experience in two stages. The first stage is a wage growth equation which just looks at wage growth in all jobs, not distinguishing between returns to experience and returns to tenure. In stage 2, the returns to experience are estimated by a comparison of workers who started new jobs at different points in their careers (and hence who have zero tenure to start off with, but differing levels of experience). The returns to tenure can then be identified as total wage growth minus the returns to experience.

Topel finds that the returns to tenure are large and statistically significant – “of the order that one would obtain from a simple OLS regression”. The first year of tenure in a job gives a 5% increase in the wage on average conditional on other factors. Over 10 years the returns to tenure are estimated at 2.8% per year. Meanwhile the return to experience is estimated at around 7% per year on average. These results lead Topel to conclude that theories that emphasise the role of job-specific human capital, and/or deferred compensation mechanisms in wage setting, are empirically important.

5.1.2 Altonji and Williams (1997)

Altonji and Williams conduct a very detailed examination of the methodologies used in the studies by Abraham and Farber (1997), Altonji and Shakotko

(1987) and Topel (1991), all of whom estimated returns to tenure from the US PSID data. They identify some issues concerning Topel's methodology which may have caused him to overestimate the returns to tenure:

- Discrepancies between the wage index measure which Topel uses to control for aggregate wage growth in the economy and the growth in wages in the PSID data.
- Topel uses wages measured over the year prior to interview in PSID and tenure at interview whereas Altonji and Williams recommend using wages measured over the year *after* interview.
- A number of other minor discrepancies in the way Topel handles the data.

Altonji and Williams find that after correcting for these problems, the estimate of the returns to tenure over a ten-year period is around 11% - larger than the Altonji and Shakotko (1987) estimate of 7%, but below Topel's estimate of 28%. Positive and significant returns to experience are also found, although the estimation methodology means that no single percentage number can be quoted here.

5.1.3 Dustmann and Meghir (2001)

Dustmann and Meghir use data from Germany to look at the returns to experience and tenure. They argue that the research from the PSID suffered from inaccurate measurements of labour market experience and job tenure prior to the start of the surveys, and additionally that the data are substantially prone to measurement error in experience, tenure and wages. They use administrative data from Germany, with full labour market histories since leaving full time education for each individual, to address the measurement error problem. A first stage random coefficients wage equation is estimated, with the control variables including the number of periods the individual has worked (overall labour market experience), the time spent in the sector (sector tenure) and the time spent in the firm (job tenure). This methodology takes account of the findings of Neal (1995) that returns to time spent in a specific *industry* may be important[29]. Dustmann and Meghir criticise Topel's method of identifying returns to tenure using the entire sample of workers in new jobs at a given wave of the panel, because this will be a mixture of workers who are improving on the old offer, workers fired from an ongoing firm and victims of plant closure. By contrast, here a sample of people made redundant due to the closure of their plant is used, the argument being that this represents an exogenous job termination and so the return to new jobs started following plant closure has a much cleaner interpretation.

The main identification assumption of the model is that conditional on experience, potential experience (years since leaving full time education) and educational attainment, age does not affect wages for young workers. This

[29] Neal (1995) examines industry stayers and industry changers using the PSID data. He finds that the value of tenure before and after a dislocation is about the same, consistent with a small return to firm seniority. In contrast, industry specific returns may be important.

means that age is not featured as a regressor in the wage equation. In the first stage, reduced form equations for overall experience and labour market participation are estimated. In each case the variable is regressed on age dummies, potential experience, the interaction of age and potential experience, year indicators and the education level. In the second stage, the wage is regressed on experience, potential experience, sector-specific tenure, education and the interaction of the first stage residuals with the experience terms. This allows estimates of the returns to experience, sector-specific tenure and job tenure.

Dustmann and Meghir find that the returns to experience have a different pattern for skilled workers and unskilled workers. The returns are non-linear for both groups; for the unskilled, the average return is 9% to the first year of experience, 7% to the second year, 1% to the third year and insignificantly different from zero thereafter. For skilled workers the pattern is closer to linear, with average returns of 6% per year for the first two years and around 4% per year thereafter. Hence there appear to be substantially greater benefits to experience over the medium to long term for skilled workers than for unskilled workers. The returns to tenure also seem to differ between groups; they are estimated at around 1% per year for unskilled workers – close to the Altonji/Williams estimate – but at around 2% per year for skilled workers, around midway between the Altonji/Williams and Topel results. It should be noted in addition that these are 'pure' returns to tenure after exogenous plant closure, whereas the results from other studies come from a more heterogenous set of job starts. A return to sector tenure of around 0.5 to 1.0% per year is also found (this is not estimated as a separate component in the earlier studies and so would presumably be subsumed into the 'experience' term).

5.1.4 Myck and Paull (2001)

Myck and Paull construct a grouped panel data set from 20 years of the Family Expenditure Survey cross-sections to generate average values of experience for groups of individuals born after 1961. The groups are defined by sex, age, year-of-birth, education and the number and age of children. They calculate average employment rates for the groups in the cross sections and then match "past" employment data with current observations on wages to find how labour market history influences the observed wage levels. Because of the grouped panel method it is impossible to distinguish between experience and tenure and the measured returns combine both of them. The "return to experience" measure they use also captures the effect of job matching which goes on while on the job.

The authors use the artificial panel structure of the data to estimate a fixed effects model of the returns to labour market experience. This approach allows them to control for group specific unobserved characteristics which could influence the estimates of returns, such as ability or motivation. Ignoring these is shown to lead to overestimates of the returns to experience.

Labour market experience is demonstrated to have an important effect on wages in the first six years of its accumulation. Men see their wages rise by over 16% in the first two years of accumulating labour market experience, and returns then fall to around 5-6% for the following four years. After that there is no statistically significant effect of experience on wages. Returns are shown to be lower for women in the first two years, but are still positive at 13% and 10%.

Regressions are then run separately for people in different education groups and returns to labour market experience are shown to be highest among the least educated people. There is no statistically significant effect of accumulating labour market experience for the highest educated groups which is surprising given the findings in other studies.

5.1.5 Conclusions: the returns to experience and tenure

The three studies we have examined in detail reach very different conclusions on the relative size of the returns to experience and tenure. All three studies find that there are significant and substantial returns to labour market experience, although the Dustmann-Meghir results suggest that this only applies to the first few years of experience for unskilled workers. Returns to tenure are estimated at anything between 1.1% and 2.8% per year for the overall sample. The Dustmann and Meghir results give 'clean' returns to tenure for new jobs started after an exogenous job separation. This has the benefit of being easy to interpret but on the other hand these jobs are only a small proportion of new jobs. Can we expect large returns to tenure in new jobs started in other circumstances (e.g. after a voluntary move to a new job)? One thing that we should also remember about the deferred compensation literature in particular is that it is probably not a model which is applicable to the whole of the economy. It is more likely to be applicable in stable industries where company lifetimes and job tenures are relatively long than in volatile industries where companies come and go, where there is substantial use of short-term and contract-based working. Myck and Paull show that first years of labour market experience are crucial in terms of increases in wages. However, their approach fails to distinguish between experience accumulated in different periods of working life. It also fails to differentiate between general experience and tenure. It has to be said that the magnitude of experience and tenure effects on wages is far from a settled matter empirically, and further work is needed in this area.

5.2 Evidence on entry wages

How do 'entry wages' compare with the overall distribution of wages in the economy? Gregg and Wadsworth (2000) use data from the UK Labour Force Survey and General Household Survey to analyse the distribution of entry wages compared with three other wage measures:

1. wages in all jobs

2. wages in jobs newly entered from other jobs ('job-to-job moves')
3. wages in 'continuing' jobs (all jobs which were not entry jobs or job-to-job moves).

Using LFS data from 1997-98, Gregg and Wadsworth find that median hourly entry wage pay, at around £4.40 per hour, was around 65 percent of the median hourly wage for all jobs. This put median entry jobs at the 20th percentile of the overall wage distribution. A regression of hourly wages on dummies for entry jobs and job-to-job movers shows that the raw hourly wage gap between continuing jobs and entry jobs was 41 log points. Around a quarter of this gap could be explained by including controls for age, education, gender and region. Adding current job tenure controls to the regression showed that the gap between entry jobs and jobs with 3-12 months' tenure was around 20 log points. Adding further controls for industry, firm size and the public sector shrank the hourly wage gap to 15 points. This suggests that around 35% of the wage gap between entrants and other workers was unaccountable for by any *observable* characteristic in the data. This interpretation should be viewed as suggestive only however, as the controls for job characteristics and tenure are potentially individual choice variables and hence may be endogenous to the wage. Adding dummies to the model for the length of time spent out of work before entering a job gives the result that people out of work for 7 to 12 months face an average wage gap of 27 log points, rising to 39 points for a spell of two to three years out of work. A possible human capital explanation of this is that the long-term unemployed face a depreciation of work-related skills which leads to a wage penalty through lower human capital stocks on the return to work. An alternative explanation is that the results are picking up unobserved heterogeneity – individuals who spend long times of periods out of the labour market may simply be poorer workers to begin with, but the correlation between time out of work and lower wages is not causal. However this would not explain why wages *before* separation were so high. Gregg and Wadsworth argue that the results show that 'simply using individual characteristics observable in the cross-section for both the working and non-working populations will be misleading.'

Another result of the analysis was that workers often leave short tenured job matches. Half of job-to-job and exit moves came from workers with less than one year's job tenure. This suggests that the entry wage distribution reflects a lot of 'cycling' between employment and unemployment.

Gregg and Wadsworth also present some evidence on the evolution of entry wages over the last twenty years vis-à-vis the wage distribution as a whole. Data from the General Household Survey between 1980 and 1990 suggest that median overall wages grew by about 25% (adjusted for price inflation) whereas median entry wages grew not at all. However, the regression-adjusted entry wage gap shows no clear trend over the period 1980-1997 in GHS and LFS data; the gap is estimated at anything between 17 and 31 percentage points. A decomposition of the rise in the raw wage gap between entry jobs and continuing jobs indicates that about one-third of the rise is due to job entrants being increasingly less well educated and having less

experience relative to other workers. Much of the rest of the rise appears to be down to changes in unobservables.

5.3 Wage mobility

The studies reported in section 5.1 suggested that the most important increases in wages take place in the first few years of employment and experience accumulation. Given that working lives span about 40 to 50 years for most people, this means that over most of this period wages change little in real terms. This is to some extent surprising. Yet, explanation of what we observe in the data may be two-fold. We might observe little effect of experience on changes in wage levels because:

- productivity (human capital) - and wages as a result - does not increase after the first few years of employment
- people experience both increases as well as falls in wages over their working lives so that “on average” wages seem to stagnate

Wages may fall as a result of depreciation of human capital or as a consequence of changes in technology, which make some skills obsolete. A spell of unemployment cause an individual to suffer more severely from these effects. Also, when beginning a new job, employees may have to finance job-specific training if it is funded ‘upfront’ by their employer; this might be reflected in lower wages during or after the training. Temporary falls in wages may also be a result of agreements between employers and employees aimed at minimising the level of job losses at the time of a recession.

Cross-sectional data on experience and wages will not be able to distinguish between the two hypotheses and will assign an average value of the wage to an average value of labour market experience. To be able to make a distinction between the two explanations we need to follow individuals over time and see what happens to their wages as their labour market history develops. Two studies reported below have done precisely this using two different data-sets, the British Household Panel Survey (BHPS) and the National Earnings Survey.

Gosling, Johnson, McCrae and Paull (1997) use the first four waves of the BHPS (1991-1994) and look at wage changes within and between jobs. Wage fluctuations occur both in terms of absolute wage levels as well as on a relative scale (there are movements between different points in the wage distribution). The authors report that even looking at changes in wages over a period as long as three years, a substantial number of people experience wage reductions.

Absolute wage reductions are more pronounced for people who experience a spell of unemployment over this period, but are relatively common also among those who don’t. 18% of men who did not spend any time out of work between

1991 and 1994 experienced wage reductions of over 10%. For men who did experience spells of unemployment in this period wages fell by more than 10% in 44% of cases. At the same time around 29% of men (both among those who did and did not become unemployed) saw their wage increase by over 20%.

The BHPS data suggests that there is a large degree of relative wage mobility as well. Gosling *et al.* look at relative position of people in the wage distribution by quartile. They find that by 1994 about 52% of men and women were in the same wage quartile as in 1991, though during the three years some people moved in and out of these quartiles as well experiencing both increases and falls in wages. People from the higher end of the income distribution were less likely to change their relative wage position and least likely to find themselves out of work after the three years.

Dickens (1999) confirms the above findings using the New Earnings Survey. The degree of reported wage mobility is slightly higher than that found for the BHPS. Dickens finds that 48% of men and 44% of women in the lowest wage decile (the bottom 10% of wage earners) were still there after a year. Relative wage mobility seemed to be highest among those in the middle of the wage distribution and as found in Gosling *et al.* it was lowest among those at the top of the distribution. Over 70% of men and 65% of women from the top wage decile were still there after twelve months.

5.4 Patterns of Job Displacement

In this section we examine the empirical evidence on job separations and job exits. Empirical labour market data often contain information on quits and layoffs but there are often problems distinguishing between these cases. One can think of several examples where a quit might appear as a layoff in empirical data or vice versa:

- an employee might quit in preference to being sacked to avoid the negative labour market signal attached to dismissal. This would show up as a quit in empirical data, but is really a pre-empted layoff.

- In countries with substantial employment protection legislation (such as France and Germany as analysed by Bender *et al* (1999)) layoffs are time-consuming and require substantial expenditure on the part of the firm to meet all the legal requirements for consultation with the workforce, a statutory minimum notice period, and so on. A firm may agree to compensate employees who quit rather than being laid off if by doing so they bypass the costs imposed by the legislation.

- A firm might terminate the contract of a worker who would have left of his or her own accord anyway within a short time.

Even if the data are reliable in distinguishing quits from layoffs, the reason for layoffs may vary widely. Some might be because productivity on a job has declined for exogenous reasons (such as macroeconomic factors); this is the scenario which causes job separations in the model of Mortensen and Pissarides (1994), for example. On the other hand, the layoff might be for reasons endogenous to the worker or firm (for example, poor effort on the job). A plant closure is the scenario most likely to be exogenous to the worker and can probably be regarded as an extreme example of a layoff for exogenous reasons.

The upshot of all this is that any model of work exit which is flexible enough to accommodate the various reasons why somebody might leave a job has to take account of the fact that the data may misclassify the various scenarios if it is to be empirically estimable. We return to the appropriate specification for an exit equation in Report 2. In the remainder of this section, we examine empirical work which has looked at the rate of job exit and worker displacement in various countries, the relationship between exit wages and overall wages, and the wages earned on *re*-entry into the labour market.

Empirical studies tend to agree that worker displacement (defined by van den Berg *et al* (1999) as 'permanent job separations[30] initiated by an employer because of adverse economic conditions') are counter-cyclical (i.e. displacement is more likely in a recession). This is found by van den Berg (1999) for the US and the Netherlands, Kletzer (1998) for the US, and Bender *et al* (1999) for France and Germany. By contrast, quits appear to be pro-cyclical in empirical work (e.g. McLaughlin, 1991). This makes sense if individuals are more confident about the prospects for securing a better job if the economy is doing well.

Do the observed patterns of worker turnover support any particular macroeconomic theory of employment dynamics over the business cycle? Most empirical studies find that sectoral shifts in employment over time are a much *less* important component of job reallocation than simultaneous job destruction and creation *within* each sector, perhaps corresponding to the 'creative destruction' model outlined in Cabellero and Hammour (1994, 1996).[31]

Comparisons between different countries appear to indicate that more loosely regulated labour markets such as the US and the UK have higher job exit rates but also higher job entry rates than more heavily regulated labour markets such as France and Germany. The obvious explanation for this is that the extra 'hiring and firing' costs associated with employment protection legislation make employment adjustment more costly in the more regulated

[30] The stress on 'permanent' here allows us to distinguish between permanent job separations and temporary layoffs, which are an important feature of certain sections of the US labour market.
[31] See Davis and Haltiwanger (1992) and Caballero, Engel and Haltiwanger (1997) for more details.

countries, and this manifests itself in smaller inflows and outflows of workers in these countries.[32]

5.5 Exit wages and overall wages

What is the relationship between 'exit wages' and overall wages in the economy? Empirical work on this question by Jacobson, LaLonde and Sullivan (1993), using detailed US panel data, found that it was important to make a distinction between workers who were laid off for reasons exogenous to the individual worker and workers who quit their jobs. Jacobson *et al* constructed a 'mass-layoff' sample of individuals who were laid off by firms who suffered aggregate employment reductions of 30 per cent or more in the 5 years previous to the survey. For these workers, average quarterly earnings began to diverge from the sample-wide average for workers of similar characteristics up to three years prior to separation. Immediately before separation, the earnings of workers in the separating sample were around 10% lower on average than the sample mean as a whole. By contrast, there appeared to be little or no wage 'dip' prior to separation for workers who claimed to have left their jobs voluntarily.

These empirical findings, which are largely confirmed by more recent work for the US and the Netherlands by Abbring et al (1999), show that the reason for job exit appears to be vital. In the case of layoffs, the finding of a wage 'dip' prior to separation appears to tally with the prediction of the matching models shown earlier, which imply that exit wages will tend to be lower than wages in ongoing jobs conditional on the characteristics of the worker, as it is the poorly matched (and hence low-productivity) jobs which terminate. The business cycle models considered earlier, which stress the role of reallocation of labour from low-productivity jobs to high-productivity jobs (given worker attributes), would also draw us towards this conclusion.

On the other hand, where the worker initiates the separation because he or she is moving to (or searching for) a higher paid job, the relationship between the separation wage and the average level of wages in general is much less clear. In a simple search model where the wage within each job is static over time, the most we can say is that the wage in the new job should be higher than the wage in the old job; but this does not necessarily imply that the wage in the old job is low compared with *average* wages for someone of those given characteristics. In the case of the intertemporal labour supply theory considered earlier, there are (at least) two possible scenarios. If it is the case that returns to work in the labour market *as a whole* reduce at a point in time which drives temporary exit from the labour market (for example, if there is a negative shock to the returns to skill), this tells us nothing about what exit wages look like relative to overall wages at that point in time. But on the other hand, if someone is dropping out of the labour market due to a perception that

[32] Of course, the fact that inflows and outflows are smaller does not in itself mean that unemployment will be higher in the more regulated countries. Some commentators have argued to this conclusion (e.g. Siebert (1997)) but this is by no means a universally accepted hypothesis (see, e.g. Nickell (1997)).

his or her returns to work are relatively low at that point in the life-cycle (e.g. the decision to retire early), and the age-earnings profile is relatively smooth, then we might expect to find that wages decline prior to job exit.

To sum up this section, it appears that the relationship between exit wages and the overall sample of wages is not at all obvious, being the outcome of the interplay between a number of factors. This has direct relevance for the exit equations which we are planning to estimate in our model (discussed in the Report 2). The estimation of our dynamic model in Report 3 takes into account the comparison of the exit wage information in the UK Labour Force Survey with the overall wage distribution in the Family Resources Survey and examines whether exit wages are lower conditional on observed characteristics.

5.6 Re-entry wages for displaced workers

There are conflicting findings on the question of whether displaced workers suffer a wage penalty on re-employment. Ruhm (1991) finds that displacement leads to a average wage loss of 14 to 18 percent based on data from the US Displaced Workers Survey. However, van den Berg et al (2000) appear to show no significant wage penalty using more recent data for the US and a *positive* effect of displacement using data for the Netherlands. Bender et al (1999) similarly find only small earnings losses in France and Germany. For the UK, Gregory and Jukes (1997) find a re-entry penalty of around 10%, whilst Gregg, Knight and Wadsworth (1999) find a penalty of around 9% for workers who are displaced from a full-time job and re-enter work in another full-time job. The existence of a re-entry penalty would tally with the findings that entry wages are on average lower than the average of the overall wage distribution, although an alternative hypothesis would be that individuals who leave work and return quickly get wages which were similar to those in their previous job, whereas those who have been away for a longer period earn much less on average. Even if the entry wage distribution were a mixture of these two types of workers it could still have a much lower mean than the overall wage distribution on average.

Chapter 6. Empirical estimation of dynamic models

In this section we examine empirical models which explicitly address the dynamics of the labour market discussed in chapter 5. These models fall into four main groups:

i. models of the transition into work (the 'hazard modelling' literature).
ii. Structural models of labour market search, entry wages and the reservation wage.
iii. General equilibrium search/matching models.
iv. Life cycle models of work and schooling choice in an intertemporal context under uncertainty.

We also devote a section to analysing the model of work entry and the tax benefit system devised by Gregg, Johnson and Reed (1999), as the model which we develop in the Report 2 of this project is a direct descendent of that approach in many ways.

In each case the data requirements of the approach are discussed, as are some basic results from the recent literature. The sensitivity of the results to identifying assumptions and the robustness of the results will also be placed under the spotlight.

6.1 Hazard modelling

6.1.1 The basic hazard model

In hazard, or duration, analysis, the dependent variable is the time that an individual takes to transit from an initial state to an end state. Hazard modelling was first used in labour economics for analysing the time taken to exit from unemployment into work. Early papers using this methodology include Nickell (1979), Lancaster (1979) and Atkinson et al (1984) for the UK, and Meyer (1990) for the US.

The basic idea of the hazard approach is to specify a function for the hazard rate out of unemployment in terms of explanatory variables that produce variation in the reservation wage (w^r), the offer distribution $F(w)$ and the arrival rate (δ). A straightforward specification often used in the empirical literature is the *proportional hazards* model,

$$\omega(x(t),t) = \lambda_0(t)\Phi(x(t)) \qquad (2.6)$$

where λ_0 is referred to as the *baseline hazard*. $x(t)$ contains a vector of exogenous controls. In the proportional hazards specification, conditional on certain values of the regressors x any variation in the hazard *over time* is

captured entirely in the baseline hazard; regressors just shift the level of the hazard around according to some functional form. In many cases, the regressors in the x vector are time-invariant factors such as age at start of sample period, region of residence, and educational attainment. The specification can be made more flexible however by including time varying regressors in x (for example, to take account of macroeconomic conditions, or changes in the local labour market). Common choices for the regressor function Φ include the exponential specification $\Phi(x) = e^{x'\beta}$ and the more flexible Weibull specification $\Phi(x) = e^{\alpha t^{\alpha-1}}$, $\alpha > 0$. The model can be made more flexible still by estimating the baseline hazard λ_0 using semiparametric or nonparametric techniques as shown in Lancaster (1990).

6.1.2 Multiple end states: the competing risks framework

Although early hazard models in labour economics concentrated on the transition from unemployment into work, more recently the approach has been extended to deal with more than one initial state and end state (the 'competing-risks' approach). In the competing risks framework, there are (at least) two hazards:

$$\omega_1(x(t),t) = \lambda_{1,0}(t)\Phi_1(x(t))$$

$$\omega_2(x(t),t) = \lambda_{2,0}(t)\Phi_2(x(t)) \qquad (2.7)$$

$\omega_1(t) + \omega_2(t) \leq 1$ for all t.

The hazard functions describe the conditional probability of exit to the mutually exclusive final states 1 and 2 (for example, the initial state could be unemployment, final state 1 employment, and state 2 inactivity). The model is called a competing-risk model as the individual has two options or 'risks' to leave the current state, and the realisation of one option is necessary and sufficient for leaving the state. The particular formulation in equations (2.7) above is what Van Den Berg (2001) labels the 'Multiple Mixed Proportional Hazards' model (MMPH), where each individual hazard is proportional. Examples of labour market transition analysis in the competing-risks framework include Narendrenathan and Stewart (1990) for the UK. Also, the hazard framework shown above only accommodates a single transition; but the framework can be extended to allow multiple transitions. Van den Berg (2001) surveys recent empirical developments in this direction.

6.1.3 Models with multiple start and end states and multiple transitions

The competing-risks model shown above only accommodates a single start state and single transition to an end state. However, many researchers have estimated models which allow for multiple initial states and multiple transitions. Honoré (1993) gives details of the MMPH model with multiple

transitions. Examples of models which use data on multiple transitions from unemployment to employment and back again, and multiple unemployment durations, are Nielsen et al (1992), Bonnal et al (1997) and Gonul and Srinivasan (1993).

6.1.4 Data requirements and identification of the hazard model

The data requirements of the hazard approach are quite stringent. In a single-transition model, for each individual in a sample, estimation ideally requires

(i) the time at which the individual entered the initial state,
(ii) the time at which the individual made the transition to the end state, and
(iii) the other covariates included in x.

Right-censored data, where (i) is known but (ii) is unknown in some cases, can be accommodated in the estimation procedure without great difficulty. *Left-censored* data, where (i) is unknown, can also be controlled for, but is more problematic.

A major problem in hazard modelling is *unobserved heterogeneity* – differences between individuals in the sample which cannot be controlled for by observed variables. Because the composition of the sample effectively alters over time in the duration analysis (as people transit from the start state to the end state), without further assumptions about how the unobservables are distributed it is impossible to say whether a hazard is declining over time because being in the initial state for a longer period diminishes the probability of exit (state-dependence) or because the people who are more likely to exit *conditional on observable factors* do so early on, leaving behind other individuals who have *always* been less likely to exit (unobserved heterogeneity). Elbers and Ridder (1982) and Heckman and Singer (1984) discuss estimation of hazard models in the presence of unobserved heterogeneity. It turns out that if a model is specified as follows:

$$\omega(t \mid x, v) = \lambda_0(t)\Phi(x(t)).v \qquad (2.8)$$

where ω, t, x, λ_0 and Φ are defined as in equation (2.7) and v denotes an unobserved heterogeneity term (assumed *uncorrelated* with x)[33], then it can be shown that failing to take account of unobserved heterogeneity biases the results from estimating the model.[34] Heckman and Singer (1984) show that it is possible to estimate the model consistently if unobserved heterogeneity $G(v)$ follows a Gamma distribution. However, there has been some sensitivity analysis done on the effects of misspecifying the functional form of unobserved heterogeneity which indicates that the single-spell model is quite susceptible to estimation biases arising from misspecification. As van den

[33] This model is known as the Mixed Proportional Hazards model (see van den Berg, 2001).
[34] Specifically, if the elements of x are positively correlated with the hazard then the neglect of unobserved heterogeneity leads to results that are biased in favour of negative duration dependence (i.e it looks as if the hazard decreases for each person with time). See Lancaster (1979) for proof.

Berg (2001) puts it, 'in the absence of strong prior information on the determinants of the MPH model, single-spell data do not enable a robust assessment of the relative importance of these determinants as explanations of random variation in the observed durations… Estimation results from single spell data are sensitive to misspecification of the functional forms associated with these determinants. Therefore, the interpretation of such results in terms of the shape of the individual hazard are often unstable and should be viewed with extreme caution.'

Using multiple-spell models mitigates some of the above problems. If data are available on multiple spells then duration analysis becomes more akin to standard panel data analysis where the data can be used to separate 'fixed effects' (including individual unobserved heterogeneity, if this is assumed to be fixed over time) from other factors. Also the assumption on the distribution of the unobserved heterogeneity can be done away with (see Lancaster, 1998, for example). However, the data requirements for multiple spell analysis are obviously far more stringent. Panel data are required, with reasonably accurate timing information on state transitions. The number of available data sets with accurate information on multiple spells in labour market states is much smaller than the number of available data sets with single spell information (indeed many surveys are designed as single spell surveys). And of course, using panel data gives rise to additional sample selection problems (due to attrition and reduced response rates in general).

6.1.5 Choice and construction of regressors in hazard modelling

The regressors included in the control variable set x in a hazard model usually comprise age, family status, some measures of health status, and a financial incentive variable or variables to take account of the financial gains to moving into work. In early work such as Nickell (1979) and Lancaster (1979), a replacement rate variable – the ratio of out-of-work benefits to in-work net income – was often used. Atkinson et al (1984) showed that the exact approach taken to measuring both the numerator and denominator of the replacement rate could result in widely differing estimates of the effect of replacement ratios on unemployment durations, indicating that the early hazard model studies were not very robust to measurement error in benefit or net wage levels.

6.1.6 The limitations of the hazard approach

Techniques for modelling both out-of-work income and post-tax wages have fortunately improved a lot over the 1980s and 1990s (see for example Meyer, 1990). However, the hazard approach still seems lacking in its treatment of the hours decision element of labour supply in particular (as opposed to the participation decision, which is modelled very exactly). Labour market transitions in this framework tend to just look at 'the move into work' as opposed to 'the move into work at a given hours level and wage'. Thus, for the purposes of modelling work incentives in a dynamic labour market framework, the hazard methodology as it stands only seems to be useful up to a point.

In the next section, we consider more structural approaches to empirical estimation of search and matching models.

6.2 Structural models of labour market search, entry wages and the reservation wage

The data requirements for structural estimation of a search model such as that shown in section 2.2.2 are onerous. The reservation wage w^r – the lowest wage at which an unemployed worker will take a job – is not normally observable in empirical data. Although a number of studies attempt to derive reservation wage information from survey questions asking "what is the lowest wage you would be willing to accept?" or similar questions, there are very few surveys where this question is asked more than once of the same individual to enable researchers to look at the path of the reservation wage through time. Additionally, many economist are wary of relying on self-reported measures of economic variables like the reservation wage, preferring to deduce them from actual market actions and outcomes.

In a fully structural model the reservation wage can be determined from other data through the search equation, but the data requirements are quite stringent. Even in the simplest case outlined in section 4.2.1, for direct estimation the following variables would be needed;

- Out of work income b
- Search costs a
- The discount rate r
- The conditional offer wage distribution $F(w)|x$, where x is a variable of characteristics that we would normally put in a wage equation.
- The rate of arrival of offers δ.

Given these, the reservation wage can be derived as:

$$w^r = (b-a) + (\delta/r)\int_{w^r}^{\infty}(w-w^r)dF(w) \qquad (2.9)$$

Data on out-of-work income b are often available, either directly from survey data or by imputation using simulation models of the benefit system. The cost of search a is problematic as although some data sets contain information on methods used to look for work (e.g. the Labour Force Survey) the information on search intensity necessary to translate this into monetary-equivalent costs is lacking. The discount rate r is unobservable directly although a sensible guess on its magnitude can be made in many cases. The rate of offer arrivals δ is also problematic, as few data sets contain information on job *offers* (as opposed to accepted jobs). Given these limitations, assumptions have to be made about the distribution of δ and a in the same way that assumptions were made about the distribution of (for example) fixed costs in estimation of the static labour supply model in Chapter 1. A parametric (or semi-parametric) assumption on the shape of the offer distribution $F(w)$ can be made, but even

so, it is often hard to derive a closed form expression for the reservation wage as a function of *observed* data. Some approximation to equation (2.9) through expansion of the integrand term is often made, before estimation proceeds by maximum likelihood. Examples of papers which follow this approach include Narendrenathan and Nickell (1985), Devine (1988) and Neumann (1997).

This approach to estimation of the dynamics of the labour market has powerful predictive power if we believe the model which is being used as a starting point. It can be used to give precise estimates of the response of work entry rates to changes in out of work income b or a reduction in search costs a resulting from government support to jobseekers through institutions such as Jobcentre Plus (for example). However the tightness of the structural specification means that the model is vulnerable to misspecification. Advocates of the budget-constraint approach presented in chapters 1 and 2 can point to a number of omissions from the search model; for example, fixed costs of work are not accounted for in equation (2.9), and once again there is no treatment of hours of work. Refinements and extensions exist that take account of both of these, but there is an additional criticism; these partial equilibrium models are only looking at a transition from unemployment into employment (saying nothing about job exit or destruction), and also the employer side of the labour market is not covered. For an empirical estimation strategy which estimates a full set of search equilibrium conditions taking employee/employer matching and job entry *and* exit into account, we have to move to recent empirical studies of search/matching models under general equilibrium.

6.3 General equilibrium search/matching models

We focus here on studies which attempt to estimate empirical versions of search/matching models relying at least to some extent on microeconomic data rather than aggregate time-series evidence or cross-country comparisons[35].

Once again the models considered here are highly structural. Kiefer and Neumann (1993) and Ridder and van den Berg (1998) both estimate a simple version of the Burdett-Mortensen model described in section 5.2, with the simplification that workers and employers are homogenous. Even this simple version is consistent with a number of stylised facts: wage offers are in general higher than reservation wages, there are positive returns to experience and tenure, larger firms offer higher wage rates, and exit rates are negatively correlated with wage rates in the cross section. However, the predicted wage distribution differs markedly from any known real-world wage distribution unless worker and employer heterogeneity are built into the model. To get around this, both of the studies mentioned above assume a labour market which is completely segmented by education, experience, occupation

[35] recent empirical papers analyzing the effects of labour market policy in a general equilibrium search framework using aggregate data include Coe and Snower (1996), Mortensen and Pissarides (1997) and Ljungqvist and Sargent (1998)

and industry subgroups, so that the simplified Burdett-Mortensen model holds within subgroups but not across them. The estimated models do provide an accurate fit to unemployment duration and cross-sectional wage data, but they underestimate the returns to experience observed in the panel.

Bontemps et al (1999) assume that there is heterogeneity in employer productivity over a continuous distribution of employer types. In their model, profit maximisation implies that there will be a one-to-one correspondence between the distribution of wage offers and the distribution of employer productivity. They obtain joint estimates of the offer arrival rate and separation rate parameters and a non-parametric estimate of the distribution of employer labour productivity using unemployment and job duration data and earnings data drawn from a French panel survey on individual worker histories. After stratifying the data by industry, the model appears to fit well, even though workers in each industry are assumed equally productive. Interestingly, the empirical results suggest that the most productive employers have monopoly power which they use to pay wages below marginal product.

These models seem to do fairly well in fitting observed data on wages, even though they incorporate assumptions which are unlikely to hold in the real world (e.g. that workers within each industry are equally productive). However, whilst it has made large strides in the last few years this literature is still in its early stages in terms of producing truly realistic models of the entire labour market. Further progress may be retarded by the available data: there are few data sets in existence in either Europe or the US which match employer and employee data, and none currently available for Britain. Certainly, the Labour Force Survey and Family Resources Survey, which we are using for our labour supply model, would not allow one to estimate a structural model of this type. However, in Report 2 we discuss more fully the implications from search and matching models that we *can* use in our model.

6.4 Combining human capital and transition modelling: life-cycle labour supply and wage progression in a dynamic framework

6.4.1 The Keane/Wolpin model

The basic search model surveyed in the previous section deals only with a single transition into work, while transitions out of work in the Mortensen/Pissarides framework are treated as exogenous shocks, after which the job search process starts again. Can we extend the search model to treat the range of transitions from non-participation to unemployment to work (perhaps moving from job to job while employed), and indeed into and out of full-time education, as the result of life-cycle optimisation by rational and far-sighted economic agents? This has been attempted by Keane and Wolpin (1997) using an econometric framework developed by Keane and Wolpin (1994). They combine the human capital methodology discussed in Section 5.1 with the search approach of 5.2 to estimate a model of how individuals

make the decision to leave school, and thereafter, when to work and not to work. The model is of course only valid under stringent assumptions concerning individual rationality and the structure of the wage offer distribution and employment creation and reduction, but once estimated, it can be used to derive powerful predictions of how changes to wage levels (through tax and benefit reforms, for example) might affect work profiles.

The model which Keane and Wolpin use views individuals as having a finite time horizon beginning at age 16 and ending at age *A.* At each age *a*, an individual chooses among four mutually exclusive and exhaustive alternatives:

1. work in a 'blue-collar' occupation
2. work in a 'white-collar' occupation
3. attend full-time education
4. 'home production ' (i.e. non-participation in the labour market).

The payoff per period at any age *a* is given by

$$R(a) = \sum_{m=1}^{4} R_m(a) d_m(a) \qquad (2.10)$$

where $R_m(a)$ is the reward per period associated with the m^{th} alternative. In alternatives 1 and 2, the payoff (i.e. the wage) is determined by human capital; it thus depends on the number of years schooling $g(a)$ and work experience in the occupation $x_m(a)$. The model allows for work in blue-collar and white-collar occupations to attract differentiated human capital in the manner first suggested by Roy (1951)[36]. Alternative 3 (full-time education) incurs tuition and 'effort' costs. Alternative 4 (non-participation) allows the individual to consume leisure, which is assumed to have a value equal to the return to a skill endowment at age 16 plus a component which fluctuates randomly with age. This implies that the structure of rewards is given by:

$$\begin{aligned} R_m(a) &= w_m(a) = r_m \exp[e_m(16) + e_{m1} g(a) + e_{m2} x_m(a) + \varepsilon_m(a)], m = 1,2 \\ R_3(a) &= e_3(16) - tc.I[g(a) \geq 16] + \varepsilon_3(a) \\ R_4(a) &= e_4(16) + \varepsilon_4(a) \end{aligned} \qquad (2.11)$$

Each alternative m contains a productivity 'shock' term ε_m. The productivity shocks are assumed to be jointly normally distributed and uncorrelated.

The individual's choice problem is to maximise the expected present value of remaining lifetime payoffs. This is done through a dynamic programming formulation (Bellman, 1957). Defining $V(S(a), a)$, the value function, to be the maximum expected present value of lifetime rewards at age *a* given the individual's state *S(a)* and discount factor δ, the problem is formulated as:

[36] The original Keane and Wolpin model also includes the military as a separate occupation and includes a quadratic in experience in the human capital earnings function as well as splitting the tuition costs variable *tc* into costs associated with grad school and costs associated with college. Here we simplify and abstract from the particular institutional features of the US.

$$V(S(a),a) = \max_{d_m(a)} E\left[\sum_{\tau=a}^{A} \delta^{\tau-a} \sum_{m=1}^{4} R_m(a) d_m(a) \mid S(a)\right] \quad (2.12)$$

The individual's decision process is described as follows: beginning at age 16, given endowments at age 16, the individual draws four random shocks from the joint $\varepsilon(16)$ distribution, then uses them to calculate the realised current rewards and the four alternative-specific value functions, and then chooses the alternative that yields the highest value. The state space is then updated according to the alternative chosen and the process is repeated. The solution of the optimisation problem does not have a closed-form representation. The problem has to be solved numerically using backward recursion using the approximation methods developed in Keane and Wolpin (1994). Estimation involves an iterative process. Assuming that shocks are serially independent the probability of any sequence of choices and rewards can be written as:

$$\Pr(c(16),...,c(\bar{a})) \mid g(16), \mathbf{e}(16)) = \prod_{a=16}^{\bar{a}} \Pr[c(a) \mid \bar{\mathbf{S}}(a)] \quad (2.13)$$

where $c(a)$ is the set of choices available and the rewards from making each given choice at age a and $\bar{\mathbf{S}}(a)$ denotes the 'predetermined' elements of the 'state-space', i.e. endowments at age 16, plus the work experience and human capital at future ages which individuals would choose if there were zero shocks in all periods. The sample likelihood is the product of the probabilities in (the equation given above) over the N individuals in the sample used for the empirical estimation. The dynamic programming problem is solved numerically for given parameter values. The likelihood function is then computed, and new parameter values entered into the dynamic programming problem, iteratively until the likelihood is maximised. Because the likelihood function involves the calculation of multivariate integrals, estimation is conducted using simulated maximum likelihood as described in Keane and Wolpin (1994). Keane and Wolpin (1997) also extend the model to allow for unobserved heterogeneity in the endowments at age 16 due to comparative advantages for different individuals in different sectors (for example, some individuals might be better at acquiring schooling due to higher innate ability than others).

Keane and Wolpin estimate a version of the model shown above on data for the US National Longitudinal Survey of Youth for around 13,000 men who were aged 14-21 in 1979. The results indicate that although the basic model fails to fit the observed data very well, an extended version which incorporates skill depreciation, job search and mobility costs, and nonpecuniary aspects to job remuneration does a much better job of fitting to the data.

6.4.2 An assessment of the life-cycle work and schooling choice model

If we believe the assumptions underlying the life-cycle work and schooling choice model as implemented by Keane and Wolpin, then it is clearly an extremely powerful way of modelling labour market transitions. This is because the solution to the model provides a way to relate the entire observed lifetime labour market decisions of the individuals in a data set to the observed features of the labour market – wage rates and the costs of and returns to schooling – available at the time. The model can then be used for various simulation exercises which have great potential to inform policy analysis, for example:

- forward extrapolation of the model to examine how assumptions about what will happen to overall wage growth and the returns to schooling in the future might affect aggregate employment rates for people of different ages as they grow earlier, and the likely pattern of labour market transitions for individuals as they head towards retirement. This can be a valuable tool in assisting with the design of pensions policy, for example (see Phelan and Rust, 1997, and Kenc and Perraudin, 1997).
- Estimates of the impact of policy interventions which affect the exogenous parameters in the model – wages and/or schooling costs in this case. For example, Heckman, Lochner and Taber (1999) use a dynamic programming approach to study the impact of tuition subsidies on the long-run accumulation of human capital in the US labour market. Trostel (1994) looks at the effect of a proportional income tax on human capital formation in a life-cycle model and finds that the estimated elasticity of labour supply with respect to the post-tax wage appears to be substantially greater in a life-cycle framework than in a traditional static framework.

If successful, these simulation exercises would appear to be more general in their application, and powerful in their predictions, than many of the other modelling approaches we have looked at in Chapters 4-6. Are there any drawbacks to the life-cycle choice model? On the face of it there appear to be three, which we discuss below. The first two can definitely be overcome (or indeed may already have been overcome), while the third may face more problems.

1. data requirements. Obviously the estimation of a life-cycle model of labour market choices requires a different type of data than models which rely on single labour market transitions or even multiple spells over a short time period. The ideal sources of data for this kind of analysis are long-running panel data sets with labour market status in each time period (e.g. every year) for each individual, as well as individual schooling histories and data on any training or adult education undertaken. This needs to be combined with information on the wage distribution, the returns to skill, experience and tenure for the period which the data covers. These can be spliced in from

other data sets, although it is useful to have wage data for the working individuals in the panel itself to calibrate the model properly. In the US, such data sets are readily available (e.g. the National Longitudinal Survey of Labour Market Experience, or the Panel Survey of Income Dynamics). In the UK the current situation is not as promising. The British Household Panel Survey would seem to be an ideal candidate for this kind of analysis, but it is still relatively young compared with the US surveys, having been running only since 1991. In future years it will certainly possible to use the BHPS for this kind of analysis.

2. computational complexity. The dynamic programming aspect of the model solution involves the approximation of multi-dimensional integrals using a simulation method originally developed by McFadden (1989) and adapted to the current context by Keane and Wolpin (1994). Even using the approximation technique, the method involves a huge number of computations to solve the model, a problem which was sufficient to debar it from widespread use until the mid-1990s. However, with computing power increasing exponentially with each new generation of processors, this problem will become smaller and smaller over time.

3. realism of the assumptions. At present, this kind of dynamic life-cycle choice model has only been implemented under quite restrictive assumptions. For example:

- individual agents are assumed to be fully rational and far-sighted in that they are able to behave as if they are computing the expected return from a vast number of alternative patterns of labour market choices, maximising this function, and then re-computing the function in each subsequent time period as new information becomes available. This relates back to one of the criticisms of standard labour supply theory which we encountered in chapter 3, i.e. that the rational utility maximising individuals encountered in microeconomic theory are so unlike any real-world person that the theory is misconceived. We relate the reader back to the discussion in chapter 1 for more on this point.
- The labour market is assumed to be perfectly competitive. This is necessary so that the returns to work can be described as solely a function of an individual's skill attributes (plus unobserved 'ability'), and simplifies the model considerably. Under imperfect competition it might well be the case that jobs in different industries or sectors attracted different returns due to rents, for example. This would mean that the choice problem of the individual would be of increased complexity as the work decision would have to consider choice of occupation as well as choice over whether to work or not. As explained in the last section, with current computing power the choice problem in each time period has to be kept as simple as possible to make the model tractable.
- Markets are in equilibrium and clear continuously. This is necessary so that the individual can have an unconstrained choice of whether to work or not to work in each time period. Deviating from this assumption would lead to huge additional complexity.

- In many models of this form, credit markets are assumed to be perfect, so that individuals can borrow against future earnings at the market interest rate (subject to the constraint of not being in debt at time of death). This assumption is obviously questionable as many theorists have pointed out that the moral hazard problems involved in borrowing against future earnings may lead to credit rationing (e.g. Stiglitz and Weiss, 1981) However, this assumption can be relaxed at relatively minor cost.
- The models become much more complex if other 'life-choices' such as family formation (marriage and childbirth), geographical location, and choice of housing tenure are considered. Some of these choices have been considered in similar dynamic programming frameworks (e.g. Rosenzwieg and Wolpin (1995) look at the effect of teenage child-bearing on the human capital formation of children who were born to these families). But a model which incorporates *all* major choices which an individual or household could face in its working life still looks to be some way off.

In summary, the life-cycle approach to a dynamic analysis of labour market transitions and labour supply decisions is both controversial in its application and extremely demanding in the computational apparatus needed to make it viable. However, if we believe the assumptions then it provides an extremely powerful tool for policy analysis – one that will surely grow in importance in the future as the technological constraints on its implementation diminish. Hopefully there will also be future work on what the implications of the model under less restrictive assumptions about individual rationality and the state of the labour market.

6.5 Gregg, Johnson and Reed (1999)

6.5.1 The modelling strategy

In their 1999 report, Paul Gregg, Paul Johnson and Howard Reed (GJR) estimated a model which is a direct antecedent of the methodology which we use in Report 3 of the project. The GJR model used data from the UK Labour Force Survey (LFS) and Family Resources Survey (FRS) for the tax year 1994-95. The model related the information from the LFS on who moved into work to the increases in disposable income that people expected to receive from moving into work. To get an accurate assessment of how the tax and benefit system affects the financial gains from working, the IFS's tax and benefit microsimulation model TAXBEN was used to evaluate post-tax incomes in and out of work for unemployed and economically inactive people in the FRS in the same period as the LFS data. An ordered probit regression was used to estimate the probability of moving into work at the deciles of the entry wage distribution in the LFS. The ordered probit controlled for the individual's sex, age group, region, education (degree, A level or equivalent, O level or equivalent, other education, no education), age-education

interactions, length of previous unemployment or inactivity, and being made redundant within the quarter prior to the start of the LFS panel.

An hours equation was used to relate the probability of entering work at full-time hours (over 30 hours) versus part-time hours (less than 30 hours) to a vector of controls (family status, homeownership, region, education, age group, and age of youngest child). This gave a vector of probabilities for each person of entering work at various deciles of the entry wage distribution and at part-time or full-time hours level. TAXBEN was run 20 times (i.e. once at each wage and hours combination) to give the financial gains from working at these wage and hours points.

The data from each dataset were averaged into 'cells', where a cell comprised all individuals of a given sex, age, educational attainment, region and family status. This made it possible to combine the information on gains from working in the FRS and the wage information and data on who moves into work in the LFS, using cell averages. For each individual *j* in cell *g*, the vector of gains from work, Γ_j, was averaged over the cell to give the cell-level vector of gains to working:

$$\Gamma_g = \frac{\sum_j \Gamma_j}{N_g} \qquad (2.14)$$

The final 'moving-into-work' equation had the following format:

$$\Pr(M_g) = \phi(\overline{X}_g'\beta + \overline{\Gamma}_g\delta + \varepsilon_g) \qquad (2.15)$$

where M_g is the cell-based probability of moving into work in quarter 5 based on being unemployed or inactive in quarter 1, $\overline{\Gamma}_g$ is a within-cell average of gains to work derived form the TaxBen run on the vector of wage and hours points for each person; $\overline{X}_g$ is a vector of extra regressors that control for other factor affecting the probability of moving into work, and ε_g a normally distributed error term. A variety of different specifications were used for $\overline{X}_g$. Table 6.1, which is a reprint of a results summary table from the Gregg, Johnson and Reed report, shows four of the different specifications, with the predicted impacts on work probabilities of increases in out-of-work income and the expected gains from working.

Table 6.1

Specification number	A		B		C		D	
Impact of £10 increase in predicted out of work income on probability of moving into work (% pts.): men	-0.1		0.9	**	0.7	**	0.0	
: women	-0.3	**	-0.2	**	-0.1		-0.3	**
Impact of £10 increase in expected gains from working on probability of moving into work (% pts.): men	1.2	**	1.3	**	0.9	**	0.0	
: women	2.5	**	2.5	**	1.9	**	0.6	
CONTROLS:								
male/female	YES		YES		YES		YES	
age	NO		YES		YES		YES	
redundancy in last 3 months	NO		NO		YES		YES	
level of unemployment (by education/region)	NO		NO		YES		YES	
age of youngest child (women only)	NO		NO		YES		YES	
sick and disabled people	NO		NO		NO		YES	
home ownership	NO		NO		NO		YES	
family status (single, living with parents, married, working partner)	NO		NO		NO		YES	

The main results were that:

- There was a positive relationship between what the GJR model predicted was the financial return from working for men and women who were unemployed or economically inactive and their propensity to move into work, controlling for age, redundancy and the level of unemployment.
- The relationship appeared to be stronger for women than for men.
- The relationship was *not* robust to controlling for family type in the model.
- Some spline regression of the 'gains to work' variable was done to examine the effect of increases in financial incentives for men and women who gained little financially from working in the first place. When the total (initial) expected gain from working was less than £100, increases in work incentives appeared to have a greater impact on the probability of moving into work.
- The elasticity of 'work entry' (the percentage increase in the work entry rate arising from a 1% increase in the expected gain to work) was calculated at about 0.08 for men and 0.12 for women.

6.5.2 Policy simulation

The results of the moving into work equation were used to predict the effects of various policies which were then (1999) scheduled to be introduced by the government. These included the 10p starting rate of income tax, the Working

Families Tax Credit, and the reforms to employee National Insurance Contributions. Two estimates of the employment effects were provided:

- The **short-run** estimate (over one year). This was arrived at by running a modified version of the TAXBEN model on each person in the FRS data to produce a new vector of out-of-work income and gains to work. These new incentive variables were then combined with the original coefficients of the moving-into-work equation to produce estimates of the increase in labour market entry for each cell over one year. These were then grossed up to give employment changes over one year.
- The **long-run** estimate. This was arrived at by exploiting the accounting identity that if the rate of labour market exit in each period were unchanged as a result of a change to the tax and benefit system, the increase (decrease) in the entry rate arising from a tax and benefit reform would lead to an increase (decrease) in the long run equilbrium stocks of employed and unemployed people in the economy.

GJR predicted that the WFTC (minus its childcare credit component, which they did not model) would lead to a net employment increase of around 12,000 in its first year, and 33,000 in the long run. The introduction of the 10p starting rate of tax was predicted to lead to an increase of 14,000 in its first year, and 52,000 in the long run.[37]

6.5.3 Limitations of the GJR approach

Whilst the GJR report was a useful contribution to the labour supply modelling literature, and was the first British empirical study to use entry wages for modelling, there were a number of weaknesses in the approach, which we discuss below on a point-by-point basis.

Failure to model labour market exit

GJR only modelled entry into work – there was no equation for job exit. This is probably a reasonable starting point for looking at the impact of labour market reforms on people who are currently unemployed or inactive – the target group for 'welfare to work' policies. However, it is a poor methodology for evaluating the overall employment impact of labour market policy because GJR were unable to say anything about how tax and benefit reforms might affect exit from work. Hence the predictions of the employment effects of the various reforms relied on the *ad hoc* assumption that the exit rate would be unaffected by the reforms.

[37] the numbers quoted here are from Gregg, Johnson and Reed (1999b) which differed from the original IFS report in that the assumption on the exit side of the labour market was that the exit *rate* would remain unchanged rather than that the *number of people exiting each period* would remain unchanged. If the number of people exiting each period remained unchanged, the exit rate would effectively fall if an increase in the entry rate increased employment in the short run. The constant exit rate assumption was held to be more realistic after some discussion of the original results (which showed higher long run effects).

Lack of time series variation

The GJR study only used data from a single year's FRS and LFS – the FRS data for the 1994/95 financial year and the four quarterly LFS panels which ended between March 1994 and February 1995. This meant that there was little or no time-series variation in the tax and benefit system and the distribution of wages which would have been useful for identifying the model. In the end, as explained in Chapter 4 of the GJR report, there was a serious identification problem with the model: in the specification of the moving-into-work equation which controlled for age group, previous redundancy, regional unemployment, age of youngest child, owner-occupiers, sick and disabled people and family type, there was little systematic variation in work incentives across groups left with which to identify any incentive effects which might have existed. This was largely due to the absence of time-series variation in the data.

Modelling of entry wages

The wage modelling strategy used by GJR was innovative in that previous labour supply studies had not used entry wages, relying instead on predictions from the wage distribution as a whole (often adjusted for non-random selection into employment). The entry wage model controlled for differences in the mean and shape of the entry wage distribution for job entrants with different observable characteristics by using the ordered probit method discussed earlier. However, it could be argued that this method was still lacking in (at least) two respects:

1. There was no allowance for wage progression in work. If individuals are forward looking then we would expect them to take the possibility of wage progression into account when calculating a 'long run' return to work, or some implicit measure of the 'net present value' of taking a job. We have seen from the evidence presented in Chapter 5 that there do seem to be returns to tenure and experience on average even for low-skilled workers, and that there is substantial mobility in the wage distribution over time. If the expected growth in wages for people entering work is large (and if they take this into account when deciding whether to work or not) then focusing merely on the entry wage could lead to an underestimate of the financial benefits to moving into work.
2. There was no allowance for self-selection *into work entry* (as opposed to selection into employment). A priori, if we believe the Heckman selectivity model of wages whereby people who are currently not in work will tend to earn lower wages on average, conditional on observable characteristics, then people who are already in work, then an analogous argument is likely to hold for people who enter work from unemployment in the LFS sample compared with people who stay unemployed over the LFS sample period. In other words the GJR entry wage modelling process may *overstate* entry wages for the unemployed or inactive who would be predicted to move into work if work incentives increased. How serious this selectivity effect is will

depend on the characteristics of current job entrants compared with potential job entrants. If current job entrants are mainly composed of people who were 'accidentally' displaced from employment and return quickly (e.g. people made redundant from previous jobs for reasons exogenous to their own performance in the job) then they are likely to be very different from the long-term unemployed and inactive people who are the primary target of welfare-to-work policies. If, on the other hand, current entrants are mainly people who were previously long-term unemployed and inactive, then these are people who are probably from the lower end of the distribution of unobservable determinants of wages already, and the selectivity effect for these entrants vis-à-vis the people who are still unemployed is likely to be smaller than the effect for the average employed person in the entire distribution of employed people vis-à-vis the unemployed. GJR did control for previous redundancy in their moving-into-work equations, which should reduce the selectivity problem to some extent.

Modelling of couples

The GJR approach to modelling couples' labour supply was as follows: when looking at the labour supply decision of one member of a couple (the husband, for example), the wife's labour supply was treated as fixed and her net income was treated as a component of the husband's unearned income. Similarly, when modelling the wife's labour supply, the husband's income was treated as fixed. This is a common simplification of the labour supply framework, but seems in the end rather unsatisfactory. Ideally we would like a modelling strategy that allows for the fact that the labour supply decision of one member of a couple can interact with the other partner's decisions. In Chapter 4 (?) of Report 2 we examine our preferred strategy for modelling couples, which aims to model these interactions by treating the couple as the basic unit of observation and using a multi-state framework.

6.5.4 Moving forward from Gregg-Johnson-Reed

As we have seen there are serious criticisms one can make of the Gregg, Johnson and Reed report. Many of the improvements which we hope to achieve in the modelling strategy for the current project arise in response to problems with the GJR study. However, we have also sought to learn from a much wider set of theoretical and empirical research on labour supply and labour market dynamics in the build-up to this project – which is the very reason for this paper. This is explained in detail in Reports 2 and 3.