A Continuous-Time Model of Income Dynamics

Thorsten Heimann and Mark Trede
Institute for Econometrics
University of Münster
Am Stadtgraben 9
48143 Münster, Germany
email: mark.trede@uni-muenster.de

January 2005

Abstract

Most models of income dynamics are set in a discrete framework with arbitrary choice of the accounting period. This paper introduces a continuous-time stochastic model of income flows, avoiding the problem of choosing the accounting period. Our model can be estimated using unbalanced panel data with arbitrarily spaced observations. Although our model describes the stochastic properties of income flows, estimation is based on observed incomes accrueing during time intervals (of possibly varying length). Our model of income dynamics is close in spirit to the discrete-time two-stage models. We impose a parsimoniously parametrised continuous-time stochastic process (possibly containing a unit-root) to model the deviation from a traditional earnings function. The model is estimated using micro-data from the German social security agency from 1975 to 1995.

Keywords: Earnings, Diffusion process, Ornstein-Uhlenbeck process, Panel data, Estimation

JEL classifications: C13, J31, J62

1 Introduction

Knowing the dynamics of the income or earnings process is important for various purposes, e.g., it might allow us to distinguish between different theories about income distributions; models of income dynamics can be used to base macro economic theories of saving behaviour on a micro economic foundation; they are a tool for measuring income risk; simulations of possible future income paths can be generated and might be used for pricing private income insurance contracts as proposed by Shiller and Schneider (1998).

The existing dynamic income models assume that income is a discrete-time stochastic process. This approach has two disadvantages. First, the choice of the accounting period is completely arbitrary and often driven simply by data availability. Second, even if one succeeds in finding an appropriate time series model of, say, weekly earnings, the model is bound to fail if we aggregate into annual earnings, as aggregation over time alters many time series properties. Further, although wages are usually paid weekly or monthly, the flow variable earnings is nevertheless generated in continuous time. Although discretely observed in the form of cumulated (or rather integrated) earnings over certain time intervals, the income process itself is essentially a continuous phenomenon and ought to be modelled as such.

This paper introduces a continuous-time stochastic model of income flows, avoiding the problem of choosing the accounting period. Our model can be estimated using unbalanced panel data with arbitrarily spaced observations. Estimation is not based on observations of the income flow at certain points of time (since flow variables are obviously unobservable at single points of time), but rather on observed incomes accruing during time intervals (of possibly varying length).

The literature on dynamic income models is large but mostly restricted to the discrete-time case (see Atkinson, Bourguignon and Morrisson (1992) and Alvarez, Browning and Ejrnaes (2002) for overviews). Most studies use a two-stage estimation procedure: earnings are regressed on individual characteristics such as age, education or occupation, then the residuals are modelled as an autoregressive and/or moving average process. There is no consensus on whether the income or earnings process is to be modelled as stationary or rather with a stochastic trend. Influential studies on income dynamics are Lillard and Willis (1978), MaCurdy (1982), Abowd and Card (1989), Gottschalk and Moffitt (1994), Baker (1997), Baker and Solon (2003) and Geweke and Keane (2000).

A continuous-time approach to income dynamics is developed in Geweke, Marshall and Zarkin (1986*b*) and Geweke, Marshall and Zarkin (1986*a*): Their continuous-time Markov chain model allows income to jump at any point in time between income classes. Discretising income into classes and inspecting movements between them is a standard methodology in the strand of literature dealing with income mobility (see Shorrocks (1976) and Shorrocks (1978)). Obviously, a discretisation of the income space rubs out a huge amount of valuable information. Our model, in contrast, preserves the continuous nature of both time and income itself.

The lack of continuous-time models of the income process conflicts with the long established and wide-spread use of continuous-time models in econometrics (see e.g. Phillips (1972), Bergstrom (1983) and Bergstrom (1990)) and related fields such as finance (see any textbook on finance).

Our model of income dynamics is close in spirit to the discrete-time two-stage models. We impose a parsimoniously parametrized continuous-time stochastic process (possibly containing a unit-root) to model the deviation from a traditional earnings function. To illustrate our approach, the model is estimated using micro-data from the German social security agency from 1975 to 1995. Data provider is the Institut für Arbeitsmarkt und Berufsforschung (IAB). This administrative data set forms a one percent random sample of the German labour force. The number of individuals covered is large (roughly 560,000 individuals). For each individual the dataset contains daily information on earnings and other occupational and personal characteristics, that are observed for spells of different lengths.

We estimate the model by a two-stage maximum likelihood method. First, we estimate an earnings function with individual random effects and use the resulting estimates to predict the mean income (as a function of time) for each person in the dataset. Second, we estimate an integrated Ornstein-Uhlenbeck process for the residuals. Although the Ornstein-Uhlenbeck process is the continuous equivalent to an AR(1) process, the resulting integrated process has a much richer time series structure.

The paper is organised as follows. Section 2 presents our continuous-time model of income dynamics. Section 3 describes the two-stage estimation procedure. Section 4 describes the data and the estimation results. Section 5 concludes.

2 The continuous-time model of income dynamics

The income of an individual person is assumed to follow a stochastic process in continuous time. We model income flow Y_t at time point t as

$$Y_t = \tilde{y}_t + u_t \tag{1}$$

where u_t is a (not necessarily stationary) Ornstein-Uhlenbeck process with the stochastic differential equation

$$du_t = -\eta u_t dt + \sigma dW_t, \tag{2}$$

$(W_t)_{t\geq 0}$ denotes a Wiener process, $\tilde{y}_t$ the mean income to be defined in more detail below, and $\eta \in \mathbb{R}$ and $\sigma > 0$ are the parameters of the process. For $\eta > 0$ the stochastic process $(Y_t)_{t\geq 0}$ is trend stationary (around mean income $\tilde{y}_t$). For $\eta = 0$, income is a continuous-time random walk around $\tilde{y}_t$, and for $\eta < 0$ the income process is explosive. The parameter σ determines the strength of the stochastic income component.

The expected income flow $\tilde{y}_t$ is a function that predicts a person's income at time point t given a list of explanatory variables. To keep the model simple we restricted the list of explanatory variables to just age and time, though in principle other variables could be included, e.g., occupational characteristics, education, etc. We include both linear and quadratic terms in age to allow mean income to exhibit an inverse u-shape. We further include a time trend in order to account for differences between people generating their incomes at different times. Let B denote a person's date of birth, then his or her age at time t can be written as $A_t = t - B$.

Our simple model for the expected income flow at time point t for an individual aged A_t is

$$\tilde{y}_t = \mu + \beta_1 t + \beta_2 A_t + \beta_3 A_t^2. \tag{3}$$

Of course, it is impossible to separate time and cohort effects if there are just observations on a single person (or a single cohort). However, since our model will be estimated with panel data from many cohorts, the parameters in (3) are identified.

The solution of the stochastic differential equation (2) is

$$u_t = u_0 e^{-\eta t} + \sigma \int_0^t e^{\eta(s-t)} dW_s$$

where u_0 is the, possibly random, start value with $Var(u_0) =: V_0$. The stochastic processes $(u_t)_{t\geq 0}$ and $(Y_t)_{t\geq 0}$ are Gaussian with mean and covariance functions

$$\begin{aligned} E(Y_t) &= \tilde{y}_t && (4)\\ E(u_t) &= 0 && (5)\\ Cov(Y_s, Y_t) &= Cov(u_s, u_t) \\ &= \left[V_0 + \frac{\sigma^2}{2\eta}\left(e^{2\eta \min(s,t)} - 1\right)\right] e^{-\eta(t+s)}. && (6) \end{aligned}$$

Since we cannot observe the income flow directly at single points of time, but rather the income generated during time intervals $[t_0, t_1], \ldots, [t_{T-1}, t_T]$, we integrate (1) to obtain the observable integrated Ornstein-Uhlenbeck process

$$\begin{aligned} S_k &= \int_{t_{k-1}}^{t_k} Y_t dt \\ &= \int_{t_{k-1}}^{t_i} \tilde{y}_t dt + \int_{t_{k-1}}^{t_k} u_t dt \end{aligned}$$

for non-overlapping intervals $k = 1, \ldots, T$. Note that the (deterministic) integral $\int \tilde{y}_t dt$ captures the first moments of the process while the (stochastic) integral $\int u_t dt$ describes its second moments. Since $(Y_t)_{t \geq 0}$ is Gaussian, so is $(S_k)_{k=1,2,\ldots}$. Notice that S_k does not inherit the Markov property of the Ornstein-Uhlenbeck process u_t. Its variance and covariance functions have been derived by Gloter (2001) for the case of equidistant intervals. The generalisation for intervals of arbitrary lengths is (see appendix for the derivations)

$$\begin{aligned} E(S_k) &= \mu \left(t_k - t_{k-1}\right) \\ &\quad + \beta_1 \frac{1}{2} \left(t_k^2 - t_{k-1}^2\right) \\ &\quad + \beta_2 \left(\frac{1}{2} \left(t_k^2 - t_{k-1}^2\right) - B \left(t_k - t_{k-1}\right) \right) \\ &\quad + \beta_3 \left(\frac{1}{3} \left(t_{ik}^3 - t_{ik-1}^3\right) + B^2 \left(t_k - t_{k-1}\right) - B \left(t_{k-1}^2 - t_k^2\right) \right) \end{aligned} \tag{7}$$

$$\begin{aligned} Var(S_k) &= \frac{\sigma^2}{\eta^3} \left(e^{\eta (t_{k-1} - t_k)} - 1 - \eta \left(t_{k-1} - t_k\right) \right) \\ &\quad + \left(\frac{V_0}{\eta^2} - \frac{\sigma^2}{2\eta^3} \right) \left(e^{t_k \eta} - e^{t_{k-1} \eta} \right)^2 \left(e^{-2\eta (t_k + t_{k-1})} \right). \end{aligned} \tag{8}$$

$$\begin{aligned} Cov\left(S_k, S_l\right) &= \frac{\sigma^2}{2\eta^3} \left(e^{(t_{k-1} - t_l)\eta} - e^{(t_{k-1} - t_{l-1})\eta} + e^{(t_k - t_{l-1})\eta} - e^{(t_k - t_l)\eta} \right) \\ &\quad + \left(\frac{V_0}{\eta^2} - \frac{\sigma^2}{2\eta^3} \right) \left(e^{t_k \eta} - e^{t_{k-1} \eta} \right) \\ &\quad \times \left(e^{t_l \eta} - e^{t_{l-1} \eta} \right) e^{-\eta (t_{k-1} + t_k + t_{l-1} + t_l)} \end{aligned} \tag{9}$$

where $1 \leq l < k \leq T$. These stochastic properties form the basis of the maximum likelihood estimation procedure to be presented in the next section.

3 Estimation Method

Estimation of the parameters of the income model in section 2 proceeds in two steps: In the first step, the parameters of the mean income function (3) are estimated from the data and then used to predict the expected income. In the second step, we use the residuals to estimate the parameters of the Ornstein-Uhlenbeck process (2). This method is in line with most of the literature in the discrete case, although a one-step maximum likelihood estimation of all model parameters simultaneously would be more efficient. The development of such an estimation procedure is subject to current research. However, we reckon that the loss in efficiency of the two-stage estimation is tolerable.

Estimation of the mean income function

As to the mean income function, we assume the existence of an individual random effect. For individuals $i = 1, \ldots, N$ and time intervals $[t_{k-1}, t_k]$, $k = 1, \ldots, T_i$, the resulting random-effects model for average daily incomes S_{ik}^* takes the form (cf. (7))

$$\begin{aligned}
S_{ik}^* &= \left(E\left(S_{ik}\right) + \alpha_i + v_{ik}\right)/\Delta_{ik} \\
&= \mu \\
&\quad +\beta_1 \frac{t_{ik}^2 - t_{ik-1}^2}{2\Delta_{ik}} \\
&\quad +\beta_2 \frac{\frac{1}{2}\left(t_{ik}^2 - t_{ik-1}^2\right) - B_i\Delta_{ik}}{\Delta_{ik}} \\
&\quad +\beta_3 \frac{\frac{1}{3}\left(t_{ik}^3 - t_{ik-1}^3\right) - B_i\left(t_{ik}^2 - t_{ik-1}^2\right) + B_i^2\Delta_{ik}}{\Delta_{ik}} \\
&\quad + \left(\alpha_i + v_{ik}\right)/\Delta_{ik} \qquad (10)
\end{aligned}$$

where $\Delta_{ik} = t_{ik} - t_{ik-1}$

$$\begin{aligned}
E\left(\alpha_i\right) &= E\left(v_{it}\right) = 0, \\
E\left(\alpha_i v_{ik}\right) &= 0, \\
E\left(\alpha_i \alpha_j\right) &= \begin{cases} \sigma_\alpha^2 & \text{if } i = j, \\ 0 & \text{if } i \neq j, \end{cases} \\
E\left(v_{ik} v_{jl}\right) &= \begin{cases} \sigma_v^2 & \text{if } i = j,\ k = l, \\ 0 & \text{otherwise,} \end{cases} \\
E\left(\alpha_i B_i\right) &= E\left(u_{ik} B_i\right) = 0.
\end{aligned}$$

Estimating the model using average daily incomes, rather than total income during each spell, has the advantage that the individual effects influence income proportional to spell length. The estimation's residuals can easily be transformed back from average values to total values by multiplying them with spell lengths.

The parameters $\mu, \beta_1, \beta_2, \beta_3$ of (3) are the same for each person, we do not allow for random coefficients. We assume that α_i and v_{ik} are normally (and independently) distributed and then recover the model parameters by maximising the restricted likelihood function.[1] We use the estimated parameters $\hat{\mu}, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ to predict mean income according to the earnings function (3) for each individual and time period. The distance between observed and predicted incomes serve as input in the following estimation step.

Obviously, the covariance structure imposed by this error component model deviates from the one implied by the Ornstein-Uhlenbeck process u_t as given in (2). However, the estimates of μ, $\beta_1, \beta_2, \beta_3$ are consistent, hence the residuals $\hat{v}_t$ can be used as valid substitutes for the unobservable $\int u_t dt$, if the sample is sufficiently large.

Estimation of the Ornstein-Uhlenbeck Process

In our data, we do not observe realisations of the flow process Y_t itself but of the integrated process. We therefore consider the income observed in the interval $[t_{k-1}, t_k]$ to be a realisation of S_k. Similarly, the consistently estimated residuals from the first estimation step are the realisations of $\int u_t dt$. The parameters σ and η can then be estimated by maximising the likelihood function. As before, the observations are assumed to be independent across individuals $i = 1, \ldots, N$, and the parameters of the residual process are the same for each person.

We start by constructing the likelihood function for a single individual i. Required are the parameters of the multivariate normal distribution of

$$\mathbf{p}_i = \begin{bmatrix} \int_{t_0}^{t_1} u_t dt \\ \vdots \\ \int_{t_T}^{t_{T-1}} u_t dt \end{bmatrix}$$

i.e., its vector of expectations and its covariance matrix. The expected values of the residuals are, of course, zero, and the covariance matrix is constructed

[1] Estimation of the parameters was performed with the lme (linear mixed effects) command of the statistical programming language R, implementing the method of Laird and Ware (1982).

using (8) and (9). Denote the resulting covariance matrix by $\Sigma\left(\sigma^2,\eta,V_0\right)$. Reparametrising $w = V_0/\sigma^2$ we can extract the volatility parameter σ^2 from the covariance matrix,

$$
\begin{aligned}
Var\left(S_k\right) &= \sigma^2\left[\frac{1}{\eta^3}\left(e^{\eta\left(t_{k-1}-t_k\right)}-1-\left(t_{k-1}-t_k\right)\eta\right)\right. \\
&\qquad \left.+\left(\frac{w}{\eta^2}-\frac{1}{2\eta^3}\right)\left(e^{\eta t_{k-1}}-e^{\eta t_k}\right)^2\cdot e^{-2\eta\left(t_{k-1}+t_k\right)}\right] \\
Cov\left(S_k,S_l\right) &= \sigma^2\left[\frac{1}{2\eta^3}\left(e^{\eta\left(t_{k-1}-t_l\right)}-e^{\eta\left(a-t_{l-1}\right)}+e^{\eta\left(t_k-t_{l-1}\right)}-e^{\eta\left(t_k-t_l\right)}\right)\right. \\
&\qquad +\left(\frac{w}{\eta^2}-\frac{1}{2\eta^3}\right)\left(e^{\eta t_{k-1}}-e^{\eta t_k}\right)\left(e^{\eta t_l}-e^{\eta t_{l-1}}\right) \\
&\qquad \left.\times e^{-\eta\left(t_{k-1}+t_k+t_{l-1}+t_l\right)}\right],
\end{aligned}
$$

and write $\mathbf{\Sigma}\left(\sigma^2,\eta,w\right) = \sigma^2\Omega\left(\eta,w\right)$ where $\Omega\left(\eta,w\right)$ collects the terms in square brackets. Since the integrated process is multivariate normal with the $(T\times 1)$ vector $\mathbf{0}$ of expected values and the covariance matrix $\sigma^2\Omega\left(\eta,w\right)$, the joint density is

$$
\begin{aligned}
& f\left(\mathbf{p}|\sigma^2,\eta,w\right) \\
= {} & (2\pi)^{-T_i/2}\left|\mathbf{\Sigma}\left(\sigma^2,\eta,w\right)\right|^{-1/2}\exp\left[-\frac{1}{2}\mathbf{p}'\mathbf{\Sigma}\left(\sigma^2,\eta,w\right)^{-1}\mathbf{p}\right].
\end{aligned}
$$

Since the incomes according to (2) are independent across individuals, the likelihood function for all persons $i = 1,\ldots,N$ in the dataset is just the product over the individual likelihood functions, and the joint log-likelihood function is

$$
\begin{aligned}
& \sum_{i=1}^{N}\ln L\left(\sigma^2,\eta,w|\mathbf{p}_i\right) \\
= {} & \sum_{i=1}^{N}\left(-\frac{T_i}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(\left|\mathbf{\Sigma}_i\left(\sigma^2,\eta,w\right)\right|\right)-\frac{1}{2}\mathbf{p}_i'\mathbf{\Sigma}_i\left(\sigma^2,\eta,w\right)^{-1}\mathbf{p}_i\right),
\end{aligned}
$$

where subscript i indicates the person and N is the number of individuals in the dataset. Notice that the covariance matrices $\mathbf{\Sigma}_i$ differ (in general) from person to person as the time intervals need not be identical for everyone. Numerical maximisation yields the ML estimates. Since σ^2 can be factored out of the covariance matrix we can easily concentrate the likelihood to speed up the maximisation procedure. The first-order condition for σ^2 is

$$\begin{aligned}
&\frac{\partial \sum_{i=1}^{N} \ln L\left(\sigma^2, \eta, w|\mathbf{p}_i\right)}{\partial \sigma^2} \\
=& \sum_{i=1}^{N}\left(-\frac{1}{2}\frac{\partial \ln\left(\left|\sigma^2\Omega_i\left(\eta, w\right)\right|\right)}{\partial \sigma^2} - \frac{1}{2}\frac{\partial\left[\mathbf{p}_i'\sigma^{-2}\Omega_i\left(\eta, w\right)^{-1}\mathbf{p}_i\right]}{\partial \sigma^2}\right) \\
=& -\frac{1}{2\sigma^2}\sum_{i=1}^{N} T_i + \frac{1}{2\sigma^4}\sum_{i=1}^{N}\mathbf{p}_i'\Omega_i\left(\eta, w\right)^{-1}\mathbf{p}_i \\
=& \ 0,
\end{aligned}$$

or

$$\tilde{\sigma}^2 = \frac{1}{\sum_{i=1}^{N} T_i}\sum_{i=1}^{N}\mathbf{p}_i'\Omega_i\left(\eta, w\right)^{-1}\mathbf{p}_i. \qquad (11)$$

Hence the concentrated log-likelihood is

$$\begin{aligned}
&\sum_{j=1}^{N} \ln L\left(\eta, w|\mathbf{p}_j\right) \\
=& -\frac{\sum_{j=1}^{N} T_j}{2}\ln\left(2\pi\right) - \frac{\sum_{j=1}^{N} T_j}{2}\ln\left(\frac{1}{\sum_{i=1}^{N} T_i}\sum_{i=1}^{N}\mathbf{p}_i'\Omega\left(\eta, w\right)^{-1}\mathbf{p}_i\right) \\
& -\frac{1}{2}\sum_{j=1}^{N}\ln\left|\Omega_j\left(\eta, w\right)\right| - \frac{\sum_{i=1}^{N} T_i}{2} \qquad (12)
\end{aligned}$$

which is to be numerically maximised with respect to η and w. The ML estimate for σ can then be recovered from (11) by inserting $\hat{\eta}$ and $\hat{w}$ for η and w. The asymptotic distribution of the estimated parameters is, of course, normal. Note that the presence of a stochastic trend in Y_t (or even an explosive process) is innocuous.

The maximum likelihood estimation becomes rather unwieldy if the number of integrals T is large, since (12) involves the $T \times T$ matrices Ω_j, $j = 1, \ldots, N$. As the largest number of spells in our illustrative application does not exceed 40 (and is mostly around 10-15) this is not a serious limitation. An alternative estimation method, based on prediction-based estimating functions (Sørensen, 2000) and capable to handle large T, is suggested by Ditlevsen and Sørensen (2004).

4 Data and Results

Estimation of our model is based on the Institute for Employment Research (Institut für Arbeitsmarkt und Berufsforschung, IAB) employment sample, which covers a one percent random sample of all employees registered with the German social security system within the period from 1975 to 1995. It includes spell data on the employment history as recorded by the social insurance system, and information on periods of drawing benefits. The variables include information on education, part/full time employment, occupation and the average daily remuneration during the spell. In addition, some socio-economic variables such as age, gender, marital status, nationality, number of children etc. are also available.

The data have been recorded by the administrative data collection procedure of the social insurance system, introduced in West Germany in 1973. It includes a common notification procedure for health insurance, unemployment insurance, and the statutory pension scheme. All employers in Germany are legally obliged to supply the social security agencies with comprehensive information about their employees. Employers have to notify the agencies of any relevant changes in the employment status. If there are no changes, an annual control notification is required. These data are collected and stored by the Federal Employment Service (Bundesanstalt für Arbeit).

Since the purpose of the data collection is to set up a social insurance account for each employee, and since substantial legal sanctions are imposed for incorrect or missing notifications, the data are much more reliable than survey income data collected on a voluntary basis. Furthermore, the dataset does not suffer from panel mortality or attrition.

However, there are some limitations to the data quality:

First, the social insurance agency records a person's wage only up to the contribution assessment ceiling of the social security system. For wages exceeding this threshold, the data are censored (from above). The threshold increases over time roughly in line with the increase in the general wage level. For the 1980s, the fraction of censored observations lies between 8 and 11 percent, but it is substantially higher for subgroups such as highly qualified employees (see Steiner and Wagner (1997, p. 639)).

Second, the German social insurance system does not include civil servants, self-employed persons, nor employees with an income below a certain threshold and thus not subject to social insurance contributions. In 1995, the employees registered with the social insurance system in West Germany accounted for roughly 80 percent of the total workforce, varying across occupations and industries (see Bender, Haas and Klose (2000, p. 651)).

Third, there is a structural break: From 1984 onwards, all establishments were required to report wages including various forms of extra allowances and bonuses. Before 1984, they were free to decide whether or not to include extra payments in reported wages, and there is no information on how these payments have been treated by each establishment (see Steiner and Wagner (1997, p. 639) and references cited there). It is likely that some establishments did not report extra payments while others did so.

For each employee, the dataset contains socio-economic variables on the person as well as information on their establishment. However, for the estimation of our model we only need: total income during each spell, the spell itself, and the person's year of birth. Since each change in employment status triggers a notification the information is constantly updated, and each time a new spell is created. The spell lengths vary according to the frequency of the notifications submitted by the employer. Earnings during each spell are reported as average daily earnings (in Deutsche Mark); hence, total earnings are simply the product of average daily earnings times spell length.

We chose the day as time unit and set 1 January 1900 as day 1. Since only the year of birth is given, but not the exact day, we set each person's day of birth to 30 June. The complete IAB dataset contains 7 847 553 observations (i.e., spells) on 559 540 individuals. In order to reduce the computational requirements and to facilitate the statistical analysis, we eliminate individuals from the original dataset in the following way.

We eliminate persons holding more than one job at a time (multiple employment). Women are excluded from the sample. Further, employees in East Germany are not considered since their data do not cover the period before 1992. Concerning occupational status, we exclude apprentices, trainees, home workers, part-time workers, and people with unspecified or unknown status. We also exclude people born before 1925 or after 1970, people with only one spell, and employees reaching the contribution assessment ceiling of the social security system in at least one spell. In addition, we do not take into account interrupted employment histories, that is observations with non-adjacent spells.

Finally, we eliminate all individuals with obviously implausible data, such as persons with non-constant identification number, changing dates of birth, spell lengths of more than 366 days,[2] or daily incomes of less than one or more than 300 DM (far above the assessment ceiling). Drawing this

[2] Since an annual control notification is mandatory, spell lengths cannot exceed 366 days.

subsample from the original IAB data reduces our dataset to 292 913 observations on 23 150 individuals. This final dataset is used in both estimation steps. For each person, it contains an identification number, the beginning and end of each spell, the date of birth, and total earnings in each spell. For all individuals and spells, the lengths of the spells vary from one to 366 days, with a mean of 323.5 days. Mean earnings across spells (of possibly different lengths) are 36 920 DM, and average daily earnings range from 10 to 255 DM with a mean of 113.6 DM.

Estimation of the mean earnings model (3) using (10) yields the estimates given in the upper panel of table 1. All coefficients are significant at the 1 percent level. Mean earnings flow as a function of time is depicted in figure 1 for selected cohorts (born 1925, 1940, 1955 and 1970). The curves exhibit the typical concave pattern of earnings functions: Earnings are increasing over lifetime but at a decreasing rate; and younger cohorts' earnings are higher at a given age (due to the positive time effects). The line for the 1925 cohort stops in 1990 at the normal retirement age of 65 years. The line for the 1970 cohort starts only in 1988 at the age of 18 years.

Figure 2 shows the mean earnings flow in a cross sectional perspective for the year 1995. Apparently, for a given year, earnings are increasing with age reaching a maximum at the age of about 54 and declining thereafter. This is in line with other studies of earnings functions. Remember that we estimate the earnings function only for the sub-sample of men never reaching the assessment ceiling, i.e., many persons in our sample are likely to be blue-collar workers where the (cross sectional) earnings function is known to have a maximum, while mean earnings for white-collar workers are usually increasing over the entire age range (although the rate is decreasing in age).

First step

fixed effects	estimate	std.err.
intercept	-366.7	0.9712
time trend	1.2521×10^{-2}	4.4125×10^{-5}
age	9.5233×10^{-3}	7.6253×10^{-5}
age^2	-2.4331×10^{-7}	2.1387×10^{-9}
random effects		
σ^2_α	27.97	
σ^2_v	12.60	

Second step

parameter	estimate	std.err.
η	1.181	6.3333×10^{-3}
w	0.299	8.5198×10^{-3}
$\hat{\sigma}$	8182.456	

Table 1: Estimation results for the earnings function (upper panel) and residual process (lower panel)

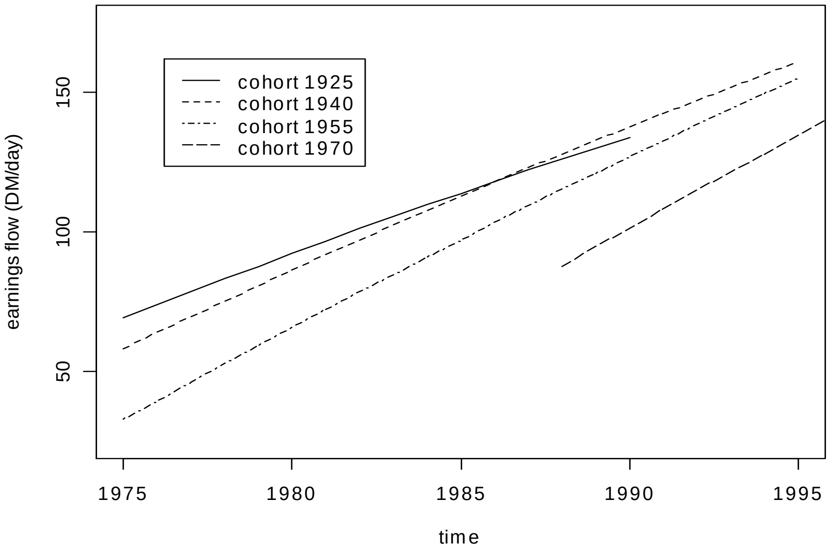

Figure 1: Mean earnings flow (in DM/day) as a function of time for selected cohorts

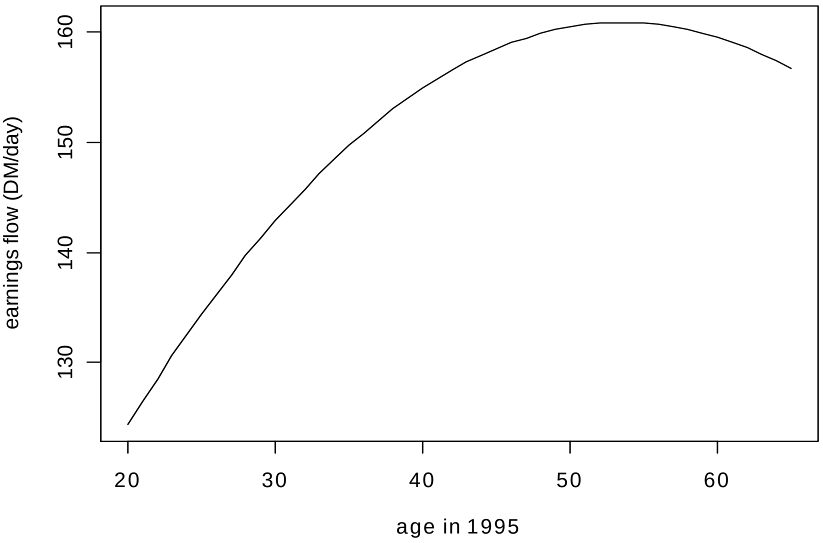

Figure 2: Mean earnings flow (in DM/day) as a function of age in 1995

The results of the second estimation step are reported in the lower panel of table 1. The estimates $\hat{\eta}$ and $\hat{w}$ are significant at the 1 percent level, we have not yet derived the standard error of $\hat{\sigma}$. Using (8) and (9) with $V_0 = w\sigma^2$ we can compute the (estimated) variance-covariance matrix for incomes accruing during intervals of arbitrary lengths.

5 Conclusions

We suggest a model of labour income in which income evolves continuously over time. The income flow of an individual follows a (not necessarily stationary) Ornstein-Uhlenbeck process around a traditional earnings function. The model is capable of handling observations on income accruing during time intervals of different lengths. Our framework avoids arbitrarily defined income periods. We derive a maximum-likelihood estimation technique for the model parameters and use a dataset that contains spell earnings information.

Further research is under work in various directions: Our dataset is subsampled from the IAB employment sample in a non-random manner, we systematically dropped a large number of individuals to avoid estimation problems due to non-adjacent time intervals or censored data. In order to ensure a more representative sample, it may be worthwhile to modify the model in a way that it can handle these problems. A Tobit-like extension to retain observations censored at the assessment ceiling will be incorporated into the estimation procedure.

Furthermore, our model presumably suffers from missing exogenous variables in the earnings function. Rather than including age only, further individual attributes (e.g. occupation) or business cycle variables will we added.

From estimations for single individuals it seems that parameters assumed to be homogenous for all individuals appear to be heterogenous in fact. It may be worthwhile to examine this further, and where appropriate to allow for a certain degree of heterogeneity either by assuming common coefficients for more homogenous subgroups, or by assuming random coefficients.

We have used nominal income rather than real income and therefore we did not account for decreasing buying power of the income over time. It may be suggestive to deflate wages by a consumer price index before using the data for our analysis.

References

Abowd, J. and Card, D. (1989), 'On the covariance structure of earnings and hours changes', *Econometrica* **57**, 411–445.

Alvarez, J., Browning, M. and Ejrnaes, M. (2002), 'Modelling income processes with lots of heterogeneity', CAM Working Paper 2002-01.

Atkinson, A. B., Bourguignon, F. and Morrisson, C. (1992), *Empirical Studies of Earnings Mobility*, Harwood, Chur.

Baker, M. (1997), 'Growth-rate heterogeneity and the covariance structure of life-cycle earnings', *Journal of Labor Economics* **15**(2), 338–375.

Baker, M. and Solon, G. (2003), 'Earnings dynamics and inequality among Canadian men, 1976-1992: Evidence from longitudinal income tax records', *Journal of Labor Economics* **21**(2), 289–321.

Bender, S., Haas, A. and Klose, C. (2000), 'The IAB employment subsample 1975–1995', *Schmollers Jahrbuch* **120**, 649–662.

Bergstrom, A. (1983), 'Gaussian estimation of structural parameters in higher order continuous time dynamic models', *Econometrica* **51**(1), 117–152.

Bergstrom, A. (1990), *Continuous-Time Econometric Modelling*, Recent Advances in Econometrics, Oxford University Press, Oxford.

Ditlevsen, S. and Sørensen, M. (2004), 'Inference for observations of integrated diffusion processes', *Scandinavian Journal of Statistics* **31**(3), 417–.

Geweke, A. and Keane, M. (2000), 'An empirical analysis of income dynamics among men in the PSID: 1968-1989', *Journal of Econometrics* **96**(2), 293–356.

Geweke, J., Marshall, R. and Zarkin, G. (1986*a*), 'Exact inference for continuous time Markov chain models', *Review of Economic Studies* **53**, 653–669.

Geweke, J., Marshall, R. and Zarkin, G. (1986*b*), 'Mobility indices in continuous time Markov chains', *Econometrica* **54**, 1407–1423.

Gloter, A. (2001), 'Parameter estimation for a discrete sampling of an integrated Ornstein-Uhlenbeck process', *Statistics* **35**, 225–243.

Gottschalk, P. and Moffitt, R. (1994), 'The growth of earnings instability in the U.S. labor market', *Brookings Papers on Economic Activity* **2**, 217–272.

Karatzas, I. and Shreve, S. E. (1991), *Brownian Motion and Stochastic Calculus*, 2nd edn, Springer, New York.

Laird, N. and Ware, J. (1982), 'Random-effects models for longitudinal data', *Biometrics* **38**, 963–974.

Lillard, L. and Willis, R. (1978), 'Dynamic aspects of earnings mobility', *Econometrica* **46**, 985–1012.

MaCurdy, T. (1982), 'The use of time series processes to model the error structure of earnings in a longitudinal data analysis', *Journal of Econometrics* **18**, 83–114.

Phillips, P. (1972), 'The structural estimation of a stochastic differential equation system', *Econometrica* **40**(6), 1021–1041.

Shiller, R. J. and Schneider, R. (1998), 'Labor income indices designed for use in contracts promoting income risk management', *Review of Income and Wealth* **44**(2), 163–182.

Shiryaev, A. N. (1999), *Essentials of Stochastic Finance*, World Scientific, Singapore.

Shorrocks, A. (1976), 'Income mobility and the markov assumption', *Economic Journal* **86**, 566–578.

Shorrocks, A. (1978), 'The measurement of mobility', *Econometrica* **46**, 1013–1024.

Sørensen, M. (2000), 'Prediction-based estimating functions', *Econometrics Journal* **3**, 123–147.

Steiner, V. and Wagner, K. (1997), 'Entwicklung der Ungleichheit der Erwerbseinkommen in Westdeutschland', Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, Vol. 30, Institut für Arbeitsmarkt- und Berufsforschung, 638–641.

Appendix

We first consider the variance of the integrated process for a certain time interval,

$$Var(S_k) = Var\left(\int_{t_{k-1}}^{t_k} u_t dt\right),$$

for $0 \leq t_{k-1} < t_k$. Let $I_{t_{k-1}}^{t_k} = \frac{1}{n}\sum_{i=1}^{n} u_{i\Delta_k + t_{k-1}}$ with $\Delta_k = (t_k - t_{k-1})/n$ denote the corresponding Riemann sum of the process with mesh Δ_k. Generally, the variance of $I_{t_{k-1}}^{t_k}$ can be written as

$$Var\left(I_{t_{k-1}}^{t_k}\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Cov\left(u_{t_{k-1}+i\Delta_k}, u_{t_{k-1}+j\Delta_k}\right). \qquad (13)$$

For the Ornstein-Uhlenbeck process $(u_t)_{t\geq 0}$ (Karatzas and Shreve (1991, p. 358) or Shiryaev (1999, p. 239))

$$\begin{aligned} & Cov\left(u_{i\Delta_k+t_{k-1}}, u_{j\Delta_k+t_{k-1}}\right) \\ = & \left[Var\left(u_0\right) + \frac{\sigma^2}{2\eta}\left(e^{2\eta \min(i\Delta_k+t_{k-1}, j\Delta_k+t_{k-1})} - 1\right)\right] \\ & \times e^{-\eta(\Delta_k(i+j)+2t_{k-1})}. \end{aligned} \qquad (14)$$

Since we do not know the value of the process at time $t = 0$ we consider u_0 to be stochastic and set $Var\left(u_0\right) =: V_0$. Equation (13) then becomes

$$\begin{aligned} Var\left(I_{t_{k-1}}^{t_k}\right) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[V_0 + \frac{\sigma^2}{2\eta}\left(e^{2\eta \min(i\Delta_k+t_{k-1}, j\Delta_k+t_{k-1})} - 1\right)\right] \\ & \quad \times e^{-\eta(\Delta_k(i+j)+2t_{k-1})} \\ &= \frac{\sigma^2}{2\eta}\cdot\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(e^{-\eta\Delta_k|i-j|}\right) \\ & \quad + \left(V_0 - \frac{\sigma^2}{2\eta}\right)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} e^{-\eta(\Delta_k(i+j)+2t_{k-1})}, \end{aligned}$$

where we have exploited the fact that $|i - j| = -2\min(i, j) + i + j$. The first double sum converges to

$$\begin{aligned} & \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(e^{-\eta\Delta_k|i-j|}\right) \\ = & \frac{2}{\eta^2}\left(e^{t_{k-1}\eta - t_k\eta} - 1 - t_{k-1}\eta + t_k\eta\right) \end{aligned}$$

and the second double sum converges to

$$\begin{aligned}
&\lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\eta(\Delta_k(i+j)+2t_{k-1})}\\
=&\left(\frac{V_0}{\eta^2}-\frac{\sigma^2}{2\eta^3}\right)\left(e^{t_k\eta}-e^{t_{k-1}\eta}\right)^2\left(e^{-2\eta(t_{k-1}+t_k)}\right),
\end{aligned}$$

so that we finally obtain

$$\begin{aligned}
Var\int_{t_{k-1}}^{t_k}u_t dt &= \lim_{n\to\infty}\left(Var\left(I_{t_{k-1}}^{t_k}\right)\right)\\
&= \frac{\sigma^2}{\eta^3}\left(e^{t_{k-1}\eta-t_k\eta}-1-t_{k-1}\eta+t_k\eta\right)\\
&\quad+\left(\frac{V_0}{\eta^2}-\frac{\sigma^2}{2\eta^3}\right)\left(e^{t_k\eta}-e^{t_{k-1}\eta}\right)^2\left(e^{-2\eta(t_{k-1}+t_k)}\right).
\end{aligned}$$

The covariances of the integrated process,

$$Cov\left(S_k,S_l\right)=Cov\left(\int_{t_{k-1}}^{t_k}u_t dt,\int_{t_{l-1}}^{t_l}u_t dt\right),$$

for time intervals $[t_{k-1},t_k]$ and $[t_{l-1},t_l]$, $t_{k-1}<t_k<t_{l-1}<t_l$, can be obtained in a similar manner: Using (14) and defining $\Delta_k=\left(t_k-t_{k-1}\right)/n$ and $\Delta_l=\left(t_l-t_{l-1}\right)/n$, we have the covariance of the Riemann sums

$$\begin{aligned}
Cov\left(I_{t_{k-1}}^{t_k},I_{t_{l-1}}^{t_l}\right) &= \frac{1}{n^2}Cov\left(\sum_{i=1}^{n}u_{i\Delta_k+t_{k-1}},\sum_{j=1}^{n}u_{j\Delta_l+t_{l-1}}\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}Cov\left(u_{i\Delta_k+t_{k-1}},u_{j\Delta_l+t_{l-1}}\right)\\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[V_0+\frac{\sigma^2}{2\eta}\left(e^{2\eta\min(i\Delta_k+t_{k-1},j\Delta_l+t_{l-1})}-1\right)\right]\\
&\quad\times e^{-\eta(t_{k-1}+i\Delta_k+t_{l-1}+j\Delta_l)}\\
&= \frac{\sigma^2}{2\eta}\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\eta(|i\Delta_k+t_{k-1}-j\Delta_l-t_{l-1}|)}\\
&\quad+\left(V_0-\frac{\sigma^2}{2\eta}\right)\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\eta(t_{k-1}+i\Delta_k+t_{l-1}+j\Delta_l)}
\end{aligned}$$

Even though slightly more complex, rearranging the elements of the first double sum proceeds analogously to the procedure above, and it can be shown that the rearranged expression converges towards

$$\begin{aligned} & \lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\eta(|i\Delta_k+t_{k-1}-j\Delta_l-t_{l-1}|)}\right) \\ = \; & \frac{\sigma^2}{2\eta^3}\left(e^{(t_{k-1}-t_l)\eta}-e^{(t_{k-1}-t_{l-1})\eta}+e^{(t_k-t_{l-1})\eta}-e^{(t_k-t_l)\eta}\right). \end{aligned}$$

The second double sum converges towards

$$\begin{aligned} & \lim_{n\to\infty}\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}e^{-\eta(t_{k-1}+i\Delta_k+t_{l-1}+j\Delta_l)}\right) \\ = \; & \left(e^{t_k\eta}-e^{t_{k-1}\eta}\right)\cdot\left(e^{t_l\eta}-e^{t_{l-1}\eta}\right)e^{-\eta(t_{k-1}+t_k+t_{l-1}+t_l)}, \end{aligned}$$

so that we finally obtain

$$\begin{aligned} & Cov\left(\int_{t_{k-1}}^{t_k}u_t dt,\int_{t_{l-1}}^{t_l}u_t dt\right) \\ = \; & \frac{\sigma^2}{2\eta^3}\left(e^{(t_{k-1}-t_l)\eta}-e^{(t_{k-1}-t_{l-1})\eta}+e^{(t_k-t_{l-1})\eta}-e^{(t_k-t_l)\eta}\right) \\ & +\left(\frac{V_0}{\eta^2}-\frac{\sigma^2}{2\eta^3}\right)\cdot\left(e^{t_k\eta}-e^{t_{k-1}\eta}\right) \\ & \times\left(e^{t_l\eta}-e^{t_{l-1}\eta}\right)e^{-\eta(t_{k-1}+t_k+t_{l-1}+t_l)}. \end{aligned}$$

The expected income in $[t_{k-1}, t_k]$ is

$$\begin{aligned} E(S_k) &= \int_{t_{k-1}}^{t_k}\tilde{y}_t dt \\ &= \int_{t_{k-1}}^{t_k}\left(\mu+\beta_1 t+\beta_2 A_t+\beta_3 A_t^2\right)dt \\ &= \int_{t_{k-1}}^{t_k}\left(\mu+\beta_1 t+\beta_2\left(t-B\right)+\beta_3\left(t-B\right)^2\right)dt \\ &= \mu\left(t_k-t_{k-1}\right)+\beta_1\frac{1}{2}\left(t_k^2-t_{k-1}^2\right) \\ &\quad +\beta_2\left(\frac{1}{2}\left(t_k^2-t_{k-1}^2\right)-B\left(t_k-t_{k-1}\right)\right) \\ &\quad +\beta_3\left(\frac{1}{3}\left(t_{ik}^3-t_{ik-1}^3\right)+B^2\left(t_k-t_{k-1}\right)-B\left(t_{k-1}^2-t_k^2\right)\right). \end{aligned}$$