The impact of sample attrition on employment participation and wages: Evidence from the British Household Panel Survey

R. Crouchley[1], S. Bradley[2] and R Oskrochi[3]

Lancaster University, Lancaster, LA1 4YF, England

April 23, 2007

[1]Department of Economics and Centre for e-Science (email: r.crouchley@lancaster.ac.uk, corresponding author)

[2]Department of Economics (email: s.bradley@lancaster.ac.uk)

[3]Now at The School of Computing and Mathematical Sciences, Oxford Brookes University, Oxford, OX3 0BP (email: roskrochi@brookes.ac.uk)

Abstract

In this paper we construct and estimate a double selection model and a truncated double selection model for evaluating the potential impact of non-ignorable missing (NIM) data in panel data on wages. Our substantive focus is on the analysis of employment participation and earnings using the British Household Panel Survey (BHPS). Simulations show that the missing data mechanism cannot be ignored, however, the estimates based on the BHPS show little effect of NIM. We conclude that researchers need to investigate the missing data mechanism in their data as a validity check on the results of their models.

1 Introduction

The main objective of this paper is to construct and evaluate several selection models that allow for the bias in a wage equation which may arise from several different sources of informative subject attrition in survey data, hereafter referred to as non-ignorable missing (NIM) data. These models are evaluated using simulated data, and by their application to the analysis of the determinants of log hourly wages using the British Household Panel Survey (BHPS) data. Specifically, we construct parametric models for a bivariate selection mechanism, where the model has one component for missing subjects and another related component for selection into the state of interest (employment). We study two situations. In the first situation it is assumed that we observe the covariates that determine which subjects are present (or missing) in the survey, and in the second situation it is assumed that the data are truncated so that we know nothing beyond the fact that missing subjects exist. In this latter situation the covariates that determine presence in the sample only exist for the observed subjects. A related objective of the paper is to assess whether the selection models developed in this paper lead to substantively different inference when compared to traditional models, such as OLS and single selection models for determing the returns to education in the the analysis of log wages. Obtaining a precise estimate of the returns to a degree, and education in general, is very important in view of the fact that this kind of evidence can shape government policy towards the expansion of higher education.

Most previous research on the determinants of hourly wages acknowledges the presence of selection effects but the potential bias created by missing subjects is either ignored, or the authors claim that the data are representative of the population (see, for example, Waldfogel, 1995; Kiernan, 1997; Gregg and Machin, 1998; Fronstin, Greenberg and Robins, 2001; Hildreth, 1999; Harmon, Walker and Westergaard-Nielsen, 2001 and Chevalier and Walker, 2001). Their argument may be that it is better to avoid reliance on untestable (however reasonable) assumptions, than adopt a more complex model. This assumption may be unjustified, and we explore it via simulation. We conclude that it is better to assume that attrition bias is present in wage equations and that in the interests of building more robust inference about the determinants of wages researchers should test for it. Our models provide researchers with some tools to do so.

The rest of this paper is structured as follows. In Section 2 we specify the double selection and the truncated double selection models, and in section 3 we evaluate these models by analysing their residuals and by conducting a sensitivity analysis. In Section 4 we explore some of the properties of the models in a small simulation study. Section 5 briefly describes the BHPS data and section 6 discusses the results of the empirical analysis. This is followed in Section 7 by our conclusions.

2 Selection and truncation models for the analysis of hourly earnings

2.1 The single selection model

Two of the most troublesome aspects of empirical research in economics are the problems of sample selection bias and (multiple) sample truncation bias (Abowd and Farber, 1980; Poirer, 1980). Both problems refer to situations where the subjects of study are selected in a way that is not independent of the response of interest, which means that biased inferences may be made. Models for a single selection mechanism are well developed (Madalla, 1983; Amemiya, 1985), and the most popular of these is Heckman's two-stage estimator, in spite of the sensitivity of this approach to violations of its assumptions. Sample truncation bias occurs if the model fails to recognise that our population has been truncated, and this is also well recognised in the literature (Bloom and Killingsworth, 1985; Muthen and Joreskog, 1983). We build on these two strands of the literature to construct double selection models to allow for the possible bias which may arise from subject attrition.

Our substantive focus is on the determinants of log hourly wage, (W_i), which for each individual is only observable if they are employed $(E_i = 1)$. This observation scheme can arise from a Tobit type II model, (Amemiya, 1985, p385) which assumes

$$\begin{aligned} E_i^* &= \eta_{ei} + e_i \\ W_i &= \eta_{wi} + w_i, \end{aligned}$$

where $\eta_{ei} = \beta_e X_{ei}$, $\eta_{wi} = \beta_w X_{wi}$, $e_i \sim N(0,1)$ and $w_i \sim N(0, \sigma_w^2)$ with $cor(e,w) =$

ρ_{ew}. E^* is not observed, but we know whether the individual is employed or not, that is, $\Pr(E_i = 1) = \Pr(E_i^* > 0)$. Identifiabilty of the model occurs if there are covariates in X_{ei} that do not appear in X_{wi}, for example, marital status is likely to affect employment status (E_i^*) but it is unlikely to affect the wage recieved (W_i). This selection model can be thought of as arising from an individual's comparison of their reservation wage (W_i^r) with their market wage (W_i^o) (Gronau, 1973). If we assume that both W_i^r and W_i^o can be written as linear combinations of independent variables plus error terms, then if $W_i^o > W_i^r$, the individual is employed and $W_i = W_i^o$, while if $W_i^o \leq W_i^r$ they are not employed and W_i^o is not observed, i.e. $W_i^o - W_i^r = E_i^*$. All the parameters of the model are identifiable except the variance of $W_i^o - W_i^r$, which can be set equal to 1 without loss of generality.

Clearly, ordinary least squares (OLS) may not be the most appropriate estimation procedure for β_w and σ_w^2 over the subsample for which $E_i = 1$. The observation plan implies that individuals with small values of e_i are more likely to have $E_i = 0$, when compared with individuals with large values of e_i. If $\rho_{ew} \neq 0$, the expected value of w_i over the subset of individuals for which $E_i = 1$ will not be zero and OLS will yield biased estimates. Maximum likelihood or a two step procedure can be used to obtain estimates for this model.

The likelihood for an individual is

$$L_i = \Pr(E_i^* \leq 0)^{(1-E_i)} \times \Pr(E_i^* > 0, W_i)^{E_i},$$

where we follow the notation of Amemiya (1985, p383) and let Pr denote a probabilty

or a density or a combination of both, as appropriate. Under the assumption of a bivariate normal distribution for e and w, the distribution of e conditional on w is also normal with mean $\eta_e + \sigma_{ew}\sigma_w^{-1}(W_i - \eta_w)$ and variance $1 - \sigma_{ew}^2\sigma_w^{-2}$ (see for example, Amemiya, 1985, pp 384-387). If $\phi(.)$ is the standard normal density function and $\Phi(.)$ the standard normal distribution function, then

$$\begin{aligned}
\Pr(E_i^* > 0, W_i) &= \Pr(E_i^* > 0 \mid W_i) \times \Pr(W_i) \\
&= \int_{-\frac{\eta_e + \sigma_{ew}\sigma_w^{-1}(W_i - \eta_w)}{\sqrt{1-\sigma_{ew}^2\sigma_w^{-2}}}}^{\infty} \phi(u)du \times \sigma_w^{-1}\phi(\frac{W_i - \eta_{wi}}{\sigma_w}) \\
&= \Phi\left[\frac{\eta_e + \sigma_{ew}\sigma_w^{-1}(W_i - \eta_w)}{\sqrt{1-\sigma_{ew}^2\sigma_w^{-2}}}\right] \times \sigma_w^{-1}\phi(\frac{W_i - \eta_{wi}}{\sigma_w}),
\end{aligned}$$

also

$$\begin{aligned}
\Pr(E_i^* \leq 0) &= \int_{-\infty}^{-\eta_e} \phi(u)du \\
&= \Phi(-\eta_e) = [1 - \Phi(\eta_e)].
\end{aligned}$$

2.2 A double selection model

Now suppose that an additional sample selection mechanism exists where the wage for each individual is only observable if they are present (retained) in the survey ($R_i = 1$) and employed ($E = 1$). We use a latent variable R_i^* for presence in the survey so that $\Pr(R_i = 1) = \Pr(R_i^* > 0)$.[1] The three sub-models are linked by allowing for a

[1] Presence in the survey implies that they also respond by answering the questions in the survey.

correlation in their errors, so that

$$\begin{aligned} R_i^* &= \eta_{ri} + r_i, \\ E_i^* &= \eta_{ei} + e_i, \\ W_i &= \eta_{wi} + w_i, \end{aligned}$$

where $\eta_{ri} = \beta_r' X_{ri}$, $\eta_{ei} = \beta_e X_{ei}$, $\eta_{wi} = \beta_w X_{wi}$. We assume a trivariate normal distribution with mean zero for (r, e, w), and variance-covariance matrix Σ, where

$$\Sigma = \begin{bmatrix} 1 & \rho_{re} & \rho_{rw}\sigma_w \\ \rho_{re} & 1 & \rho_{ew}\sigma_w \\ \rho_{rw}\sigma_w & \rho_{ew}\sigma_w & \sigma_w^2 \end{bmatrix}.$$

What does this model imply about what we observe? Suppose that an individual has a high value of r, then they have a high probability of being present in the survey $(R_i = 1)$, and if $\rho_{re} > 0$ then the individual is also likely to have a large value of e, which increases the probability that $(E_i = 1)$. To establish what a large value of r implies for W_i we need to use the first order partial correlation coefficient

$$\rho_{ew.r} = \frac{\rho_{ew} - \rho_{rw}\rho_{re}}{\sqrt{1-\rho_{rw}^2}\sqrt{1-\rho_{re}^2}}.$$

So, for example, if $\rho_{re} = 0.7$, $\rho_{rw} = 0.2$ and $\rho_{ew} = -0.4$, then $\rho_{ew.r} = -0.743$, implying that the high value of e, given by ρ_{re} is associated with a low value of w.

At this point the question of identifiability of the correlations arises. We illustrate the conditions for identifiability by looking at the responses in pairs. For E_i^* and W_i we have indentifiabilty of ρ_{ew}, β_e, and β_w as in the single selection model providing

X_{ei} contains an exogenous covariate that is not present in X_{wi}, for instance, marital status. For R_i^* and W_i identifiability of ρ_{rw}, β_r, and β_w requires that X_{ri} contains an exogenous covariate that is not present in X_{wi}. For R_i^* and E_i^* identifiability of ρ_{re}, β_r, and β_e requires that X_{ri} contains an exogenous covariate that is not present in X_{wi}. For the last two situations we argue that in Survey data such a covariate could be the interviewer's assessment of the quality of the interview, because it is difficult to justify its inclusion, based on economic theory, in the linear predictors for W_i or E_i^*; whereas it is entirely appropriate to establish if it is significant in the linear predictor for R_i^*[2].

The likelihood for an individual is

$$L_i = \left[\Pr(R_i^* > 0, E_i^* > 0, W_i)^{E_i} \times \Pr(R_i^* > 0, E_i^* \leq 0)^{(1-E_i)}\right]^{R_i} \times \Pr(R_i^* < 0)^{(1-R_i)}.$$

From the properties of the multivariate normal we can write,

$$\begin{aligned} \Pr(R_i^* &> 0, E_i^* > 0, W_i) = \Pr(R_i^* > 0, E_i^* > 0 \mid W_i) \times \Pr(W_i) \\ &= \int_0^\infty \int_0^\infty \phi(u_1, u_2)\, du_1 du_2 \times \sigma_w^{-1} \phi\left(\frac{W_i - \eta_{wi}}{\sigma_w}\right), \end{aligned}$$

where the conditional bivariate normal random variables (u_1, u_2) have means

$$\mu_1 = \eta_{r_i} - \left(W_i - \eta_{w_i}\right)\left[a_{12}(1)(1-\rho_{rw}^2) + a_{12}(2)\left(\rho_{re} - \rho_{rw}\rho_{we}\right)\right],$$

$$\mu_2 = \eta_{e_i} - \left(W_i - \eta_{w_i}\right)\left[a_{12}(1)\left(\rho_{re} - \rho_{rw}\rho_{we}\right) + a_{12}(2)(1-\rho_{rw}^2)\right],$$

[2] We observe the quality of the interview for non-respondents because in the illustrative example we use data from sweeps of the survey before the individual dropped out.

where

$$a_{12}(1) = \frac{\rho_{re}\rho_{we} - \rho_{rw}}{\sigma_w D},$$

$$a_{12}(2) = \frac{-\rho_{we} + \rho_{re}\rho_{rw}}{\sigma_w D},$$

and

$$D = 1 - \rho_{we}^2 - \rho_{re}^2 + 2\rho_{re}\rho_{rw}\rho_{we} - \rho_{rw}^2.$$

The (u_1, u_2) also have variance-covariance matrix

$$\Sigma_{u_1,u_2} = \begin{bmatrix} 1 - \rho_{rw}^2 & \rho_{re} - \rho_{rw}\rho_{we} \\ \rho_{re} - \rho_{rw}\rho_{we} & 1 - \rho_{we}^2 \end{bmatrix},$$

for Σ_{u_1,u_2} to be positive definite, i.e. for $\phi(u_1, u_2)$ to be a proper bivariate probabilty density, we require $D > 0$.

For the non-employed respondents we have

$$\Pr(R_i^* > 0, E_i^* \leq 0) = \int_{-\eta_r}^{\infty} \int_{-\infty}^{-\eta_e} \phi(u_1, u_2)\, du_1 du_2,$$

where (u_1, u_2) have variance-covariance matrix

$$\begin{bmatrix} 1 & \rho_{re} \\ \rho_{re} & 1 \end{bmatrix}.$$

Finally, for those that did not respond we have

$$\begin{aligned} \Pr(R_i^* < 0) &= \int_{-\infty}^{-\eta_r} \phi(u)\, du \\ &= 1 - \Phi(\eta_r). \end{aligned}$$

2.3 A truncated double selection model

As before the wage for each individual is only observable if an individual is present in the survey ($R_i = 1$) and employed ($E = 1$). Therefore, if $R_i = 1$ we observe either $E_i = 1$ and $W_i > 0$ or $E_i = 0$ (the individual is not employed). We do not observe anything if $R_i = 0$. There are three sub-models as before:

$$\begin{aligned} R_i^* &= \eta_{ri} + r_i, \\ E_i^* &= \eta_{ei} + e_i, \\ W_i &= \eta_{wi} + w_i, \end{aligned}$$

However, we now want a likelihood that is conditional on being present in the survey, i.e.

$$L_i = \left[\Pr(R_i^* > 0, E_i^* > 0, W_i \mid R_i^* > 0)^{E_i} \times \Pr(R_i^* > 0, E_i^* \leq 0 \mid R_i^* > 0)^{(1-E_i)}\right]^{R_i},$$

where

$$\Pr(R_i^* > 0, E_i^* > 0, W_i \mid R_i^* > 0) = \frac{\Pr(R_i^* > 0, E_i^* > 0, W_i)}{\Pr(R_i^* > 0)},$$

and $\Pr(R_i^* > 0, E_i^* > 0, W_i)$ was obtained earlier. Also

$$\Pr(R_i^* > 0, E_i^* \leq 0 \mid R_i^* > 0) = \frac{\Pr(R_i^* > 0, E_i^* \leq 0)}{\Pr(R_i^* > 0)},$$

where $\Pr(R_i^* > 0, E_i^* \leq 0)$ was also obtained earlier.

3 Model criticism

3.1 Residuals

The validity of conclusions drawn from the selection and truncated models described above depend on the nature of the residuals. In fact, one way of assessing the correspondence between the data and the model is by analysing the residuals, however, there is very little literature on residuals for the models in the previous section because selection creates a problem for the usual diagnostic tests (Hirano *et al.*, 1998).

There are many different ways of writing $\Pr(R_i^* > 0, E_i^* > 0, W_i)$ using Bayes' formula. Each form has several conditional and/or truncated distributions for which residuals can be produced. We use the obvious candidate for the continuous response, that is, the doubly truncated distribution $\Pr(W_i \mid R_i^* > 0, E_i^* > 0)$ where

$$\begin{aligned}\Pr(W_i \quad | \quad R_i^* > 0, E_i^* > 0) &= \frac{\Pr(W_i, R_i^* > 0, E_i^* > 0)}{\Pr(R_i^* > 0, E_i^* > 0)} \\ &= \frac{\Pr(R_i^* > 0, E_i^* > 0 \mid W_i)\Pr(W_i)}{\Pr(R_i^* > 0, E_i^* > 0)}.\end{aligned}$$

The moments of this distribution are

$$E(W_i^j \mid R_i^* > 0, E_i^* > 0) = \int_{-\infty}^{\infty} W_i^j \frac{\Pr(R_i^* > 0, E_i^* > 0 \mid W_i)\Pr(W_i)}{\Pr(R_i^* > 0, E_i^* > 0)} dw_i.$$

Recall $W_i = \eta_{wi} + w_i$. Note that the denominator can come outside the integral. The variance of this distribution is given by

$$Var(W_i \mid R_i^* > 0, E_i^* > 0) = E(W_i^2 \mid R_i^* > 0, E_i^* > 0) - E^2(W_i \mid R_i^* > 0, E_i^* > 0).$$

The Pearson residual for this doubly truncated model of W_i will be conditional on

being present ($R_i = 1$) and on being employed ($E_i = 1$). It takes the form

$$r(W|R = 1, E = 1) = \frac{W_i - E(W_i \mid R_i^* > 0, E_i^* > 0)}{\sqrt{Var(W_i \mid R_i^* > 0, E_i^* > 0)}},$$

where $E(W_i \mid R_i^* > 0, E_i^* > 0)$ and $Var(W_i \mid R_i^* > 0, E_i^* > 0)$ are found analytically or numerically for each individual as they are conditional on the covariates. A plot of $r(W|R = 1, E = 1)$ against $E(W_i \mid R_i = 1, E_i = 1)$ will provide a simple check for aberrant observations. This residual can be used with either the double selection or the truncated double selection model.

3.2 Sensitivity analysis

Identification of the selection and truncation models is based on untestable assumptions about the distribution of the missing data (Horowitz and Manski, 1998). Furthermore, error structures can be particularly sensitive to changes in the systematic parts of the model, consequently sensitivity analysis has been proposed as a means of verifying the results and model diagnostics, e.g. residual plots

We could allow for changes to the distribution of the stochastic components of the selection/truncation process by assuming that G_1, G_2 are specified distribution functions for the errors (r_i, e_i) of the selection model, e.g. $G_1(r_i) = 1 - (1 + \lambda r_i)^{-1/\lambda}$ (Aranda-Ordaz, 1981) for a given value of λ. We would then assume that

$$\left(w_i, \Phi^{-1}[G_1(r_i)], \Phi^{-1}[G_2(e_i)]\right)^T,$$

has a multivariate normal distribution with an unstructured covariance matrix, see Lee (1983), for example.

Perhaps one of the most appealing ways of performing a sensitivity analysis is the Verbeke and Molenberghs (2000) treatment of the Diggle and Kenward (1994) model. This approach follows Cook (1986), who suggests that more confidence can be placed in a model which is relatively stable under small modifications. This would involve allowing σ_{we} and σ_{wr} to vary by subject. Further avenues of investigation could include the semiparametric estimation of the various models (Rotnitzky *et al*, 1998; Scharfstein *et al*, 1999). Even so, one has to accept that finding the presence of NIM could actually be more informative about the inadequacies of the underlying assumptions than any causal mechanism.

Unfortunately, analysing the data can be unhelpful for exploring these kinds of specification issues. Not only are the true values of the structural parameters unknown, but also any comparison between models can be complicated by other specification errors. Therefore, we resorted to simulation to investigate some of the properties of the models developed in section 2. We assume the presence of a sample selection mechanism, an employment selection mechanism and a wage equation. We compare the results from the double selection model, the truncated double selection model, the single selection model ignoring selection for presence in the sample and the classical OLS, generating 5000 cases in each sample. We also use 100 samples or sets of simulations; in all models the covariates are assumed to be independent of each other and independent of the error terms.

We generate data from a double selection model of the form

$$R_i^* = \eta_r + r_i,$$

$$E_i^* = \eta_e + e_i,$$

$$W_i = \eta_w + w_i.$$

The linear predictor η_r takes the form

$$\eta_{r_i} = -1.75 + 1.0\frac{Age_i}{100} + 0.6Q_i + 1.0\frac{Ed_i}{10},$$

where Age_i is obtained from a uniform random number between 16 and 65, Q_i is a binary indicator which represents the quality of the interview, also obtained from a uniform random number so that $Q_i = 1$ for 70% of the sample, and 0 otherwise. Ed_i is obtained from a uniform random number, and has three values to represent years of education, that is, 70% have 12 years of education, 20% have 14 and 10% have 17.

The linear predictor η_e takes the form

$$\eta_{e_i} = -0.75 + 0.5\frac{Age_i}{100} + 0.2Mar_i + 1.0\frac{Ed_i}{10},$$

where the variable, Mar_i, is a binary indicator to represent marital status, also obtained form a uniform random number so that $Mar_i = 1$ for 60% of the sample, and 0 otherwise.

The linear predictor η_w takes the form

$$\eta_{w_i} = 0.1 + 1.0\frac{Age_i}{100} + 2.0\frac{Ed_i}{10} - 0.2Race,$$

The variable, $Race_i$, is a binary indicator to represent ethnic background, obtained from a uniform random number, so that $Race_i = 1$ for 20% of the sample, and 0 otherwise.

The stochastic errors (r_i, e_i, w_i) are from a multivariate normal distribution with mean zero and variance-covariance structure Σ, where

$$\Sigma = \begin{bmatrix} 1 & \rho_{re} & \rho_{rw}\sigma_w \\ \rho_{re} & 1 & \rho_{ew}\sigma_w \\ \rho_{rw}\sigma_w & \rho_{ew}\sigma_w & \sigma_w^2 \end{bmatrix}.$$

We use a range of values for ρ_{re}, ρ_{rw} and ρ_{ew}, but we assume that $\sigma_w^2 = 1$. Marginally, this model gives approximately 62% of individuals with $R = 1$, and approximately 80% of individuals with $E = 1$. W_i has log mean of 3.05, i.e. $Wage = £21.1$ per hour. For $\sigma_w^2 = 1$ the probability of a negative W_i is very low.

We used NAG (1996) routine G05DAF for the uniform random variables and routines G05EAF and G05EZF to generate the error terms (r, e, w). We simulated data for a range of situations, but only three are needed to provide a picture of what is happening: (1) $\rho_{re} = 0.7$, $\rho_{rw} = 0.2$ and $\rho_{ew} = -0.4$; (2) $\rho_{re} = -0.7$, $\rho_{rw} = -0.2$ and $\rho_{ew} = 0.4$; and (3) $\rho_{re} = -0.7$, $\rho_{rw} = 0.2$ and $\rho_{ew} = 0.4$. To obtain a data set of 5000 observations, we compute (R_i^*, E_i^*, W_i) given the error terms (r, e, w). If $R_i^* > 0$, we set $R_i = 1$ and zero otherwise. If $R_i = 0$, the values of E_i^*, W_i are set to missing. If $R_i = 1$, we test the value E_i^*, if $E_i^* \leq 0$ we set $E_i = 0$ and the value of W_i is set to missing. If $E_i^* > 0$ we set $E_i = 1$ and the value of W_i is retained.

To minimise the number of constraints that need to be imposed during the es-

timation of the various models we have parameterised σ_w^2 as $\sigma_w^2 = \exp(\alpha_w)$ and $\rho_{jk} = 2\left(1/\left(1+\exp\left(-\alpha_{jk}\right)\right) - 1/2\right)$, so that α_w and the α_{jk} are free to take on any values on the real line, i.e. $\alpha_w = \log(\sigma_w^2)$ and

$$\alpha_{jk} = log(1+\rho_{jk}) - log(1-\rho_{jk}).$$

Consequently $\rho_{re} = 0.7$, implies $\alpha_{jk} = 1.734601$, $\rho_{rw} = 0.2$, implies α_{jk} =0.405 and $\rho_{ew} = -0.4$, implies $\alpha_{jk} = -0.847$.

We used NAG (1996) routine E04UCF, a quasi Newton algorithm, to maximize the log-likelihood subject to the single constraint $D > 0$, which ensures that Σ is positive definite. Various problems with starting values were encountered. The presence of the constraint seemed to make it difficult to search over all of the parameter space and the algorithm sometimes converged to a local maxima. To overcome this problem we adopted the following four step procedure: (1) apply OLS to the W_i conditional on $R_i = 1$ and $E_i = 1$; (2) obtain the Probit model results for $E_i = 1$ and $E_i = 0$, conditional on $R_i = 1$, and then use these in the single selection model to estimate α_w, α_{ew}; (3) obtain the Probit model results for $R_i = 1$ and $R_i = 0$ and use these results in the double selection model to estimate α_{rw}, α_{re} and re-estimate α_{ew}, α_w; and the covariate parameters and (4) use the double selection model results as the starting values for the truncated double selection model. Even this procedure failed on some data sets, and when this happened the model was re-estimated with some elements of Σ fixed. If this model behaved properly then this solution would be taken as the starting values for the model with Σ free.

The results for the three sets of simulations are presented in Tables 1 to 3, which contain the means and standard deviations of the estimated parameters $\widehat{\boldsymbol{\theta}} = \left(\widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\beta}}\right)$. To give some idea of the magnitude of the bias obtained for the parameters in the linear predictor we performed a t-test

$$t = \frac{\text{mean}(\widehat{\theta}) - true\,(\theta)}{St.D./\sqrt{100-1}},$$

and these values are also included in Tables 1 to 3. The values that are significant at 5 percent level are highlighted. We could also test the β parameters estimated by OLS against the β parameters estimated by the other models following Hausman (1978), but eschew this approach here.

The results of the simulations suggest that the double selection model is superior to the other models insofar as the estimated parameters are almost identical to their true values. By far the worst model is the single selection model, which fails to recover the true parameters in any of the simulations. The OLS model is not much better. Consequently, there are differences in the parameter estimates of the OLS and single selection model on the one hand and the double selection model on the other, which implies that if NIM is present a misleading picture can be obtained. Where NIM is present in the data it is therefore important to control for this kind of sample selection effect. The double selection model is also superior to the truncated double selection model. The main difference between the two arises in the model for selection into the survey, where the truncated double selection model fails to recover the true parameters. In view of these findings, and in the interests of parsimony, our

analysis of the BHPS data focuses on a comparison of the OLS, single selection and the double selection models.

4 The data

The BHPS was set up in 1989 to further understanding of social and economic change in Britain at the individual and household level. It was designed as a nationally representative sample of more than 5000 households, which were selected by a two stage stratified systematic sample of postcode sectors. A total of 250 postcode sectors were selected from an implicitly stratified listing of all sectors (8980) and a systematic procedure was used to select a number of households from each sector. The same individuals are interviewed in successive years (waves) and if individuals split off from the original household all adult members of the new households are also interviewed. Although the characteristics of the population of Britain changed in the 1990s, it is claimed that by weighting the BHPS sample it should remain representative of that population. In this analysis we use wave 1, 2 and 7 data on employment and earnings, in addition to a host of individual, family and local labour market data. The rationale for using selected waves of the BHPS data is that our focus is on illustrating the importance, or otherwise, of sample attrition. Furthermore, one might expect that any bias induced by NIM would be larger the greater the gap between sweeps of the survey.

Wave 1 data was collected in 1991 and wave 7 in 1997, and we to conduct our analysis using 'paired' waves 1, 2 and waves 1 and 7, which enables us to use the

covariate data at the initial wave (e.g. wave 1) to explain participation in the sample in the subsequent wave (i.e. wave 2 and wave 7). This gave a sample of 3,998 economically active males aged 16-65 who were present at wave 1 (1991) with relevant covariate data, but only 3316 of these remain in the survey at wave 2. Thus, the BHPS survey had lost 682 (17%) of its original set of economically active males between just two waves. The equivalent figures for wave 7 are 2097 and 1901, which means that between waves 1 and 7 48% of the original sample had been lost to the survey. We treat those individuals that become 65 before wave 2 and wave 7 as independently right censored and as such they do not contribute to the analysis of employment participation and earnings in 1992 or 1997. Also, given our substantive focus on employment and wage determination, respondents who were in full time education in 1991 are dropped from this analysis.

The BHPS contains a wealth of information relating to the worker and to the firm if they are employed. We also map labour market data relating to the unemployment rate, the vacancy rate and the industrial structure in the travel-to-work area in which the individual lives (see Tables A and B, Appendix).

Using the BHPS data we illustrate the potential effect of ignoring attrition by estimating the following models:

1. Classical OLS model of log earnings, W,for those subjects that were employed and present in the survey, hence ignoring the condition on $E = 1$ and $R = 1$.

2. Single selection model for data in which we have (a) log wage and employment,

$(W, E = 1)$ and (b) the non-employed $(E = 0)$, but ignoring the condition on $R = 1$.

3. Double selection model for data in which we have (a) log wage, employment and presence in the survey, $(W,\ E = 1,\ R = 1)$, (b) non-employment and presence in the survey and $(E = 0$ and, $R = 1)$ (c) Not present in the survey $(R = 0)$.

   Throughout, we restrict the analysis to males, since there is a lot of evidence that the labour market behaviour of females is quite different to that of males. It is also likely that female dropout mechanisms are quite different to those of males.

5 The determinants of hourly earnings: Evidence from the BHPS

As mentioned in the Introduction we prefer to assume NIM and hope to show either that it is not present, or that it has little effect on the parameters of interest. There are two specifications of the wage equation, one of which is a simple human capital model, reflecting the impact of supply-side (individual) factors on wage determination (Specification A), and the other model which incorporates demand side influences on wage determination by including employer characteristics (Specification B). The rationale for estimating two models for the wage is to enable us to investigate if, and how, the impact of NIM changes as the specification of the model changes. The covariates included in Specification A are the highest level of educational qualification attained rather than years of education to enable us to examine the returns to a

degree. Experience and its square is the only other covariate included in this model. Turning to Specification B, a set of covariates are included in the model to capture the type of contract that the worker is employed under. These include part-time, fixed term and seasonal or temporary contracts and it is expected that workers on these types of contract will have lower hourly earnings than their counterparts on permanent contracts, either because of shorter tenure with the firm or because such workers are also less skilled. A dummy variable for self-employment is included and insofar as this is an alternative to unemployment, it is expected that those workers who are self-employed will have lower hourly earnings than permanently employed workers. A variable to capture whether the firm recognises a union in pay bargaining is included in the model and it is expected that this will have a positive effect on hourly earnings. Firm size is also included since it might be expected that larger firms pay higher wages. We control for the industry in which the firm operates to capture product market effects that might feed through to influence workers wages.

The specification of the employment participation and retention models are the same for models A and B but clearly parameter estimates can vary because of the difference in the specification of the wage equation. A crucial issue is that of the identifiability of the interdependent sub-models, which we discussed in theoretical terms in a previous section. In practice, identification comes down to the need to include at least one covariate in each sub-model that does not influence the wage, and so in the employment participation model we include marital status and the number of dependent children, since they are expected to have their primary effect on

the probability of being in employment rather than on the wage received for those in work. These variables are also included in the retention model, however, to achieve identification in this model two other variables are included – the number of contacts attempted by the interviewer with the respondent, including the contact that led to the collection of data, and the interviewer's assessment at wave 1 of the quality of the interview. It is expected that the greater the number of contacts made and the lower the quality of the interview, the less likely the respondent is to remain in the Survey. It is unlikely that these variables would influence the probability of an individual being in employment or the wage they receive. We return to a discussion of the magnitude and statistical significance of the covariates below, and start by asking whether NIM is present in the BHPS.

5.1 Testing the significance of NIM for log Wages in the BHPS

Tables 4 and 5 report the correlations between the correlations of the random (omitted) effects for each sub-model (ρ) and their re-parameterisation (α) for 1992 and 1997, respectively. All of the correlations are large and statistically significant, especially in the case of the correlation between the omitted effects for the retention model and that for employment participation. This correlation ranges from 0.574 for Specification A in 1992 to 0.788 for Specification B in 1997. In contrast, the correlations between the omitted effects for employment participation and wage determination are negative and statistically significant suggesting that there is evidence of non-random assignment into employment, which is a common finding.

We can check to see if we can remove the retention model (R) from our specification, by testing to see if $cov\ (re) = cov(rw) = 0$, which is effectively a test of the double selection model versus the single selection model. The values of the chi-square test (2df) range from 120.1 for Specification A in 1992 to 231.5 for Specification B in 1997, Tables 4 and 5. It is also interesting to note that when more covariates are included in the model (Specification B) the value of the correlations between the omitted effects in the employment participation and retention models rise considerably whereas the correlations between the omitted effects for the employment and wage models fall. For instance, Table 4 shows that ρ_{re} changes from 0.574 in Specification A to 0.732 in Specification B, whereas ρ_{ew} changes from -0.580 to -0.341. These changes imply that adding more covariates reduces the importance of NIM, which is consistent with the findings of Rubin (1996), however, even with a relatively large number of covariates in Specification B of the wage model and in the sub-models the importance of NIM remains. Furthermore, as the gap between the initial wave (1991) and each successive wave (1992 and 1997) increases, then NIM becomes more important as one might expect due to the larger number of dropouts from the Survey. To see this compare the values of the correlations and chi-square tests in Table 4 with their equivalent in Table 5, which suggests that NIM may also be a non-stationary process. An alternative explanation is that the change in the correlations could be influenced by changes in the state of the local economy over the period 1991-97, however, this is unlikely because we control for this by including the local unemployment and vacancy rates in the employment and retention models.

5.2 The impact of NIM on parameter estimates - returns to education

In this section we investigate how the parameter estimates change as we increase the level of complexity of the model, or more specifically as we move from the OLS to the Single Selection model (SS) and then on to the Double Selection model (DS). We illustrate the impact of NIM by focusing upon the wage model and in particular the effect of having a degree versus having pre-University qualifications known as A Levels. An enormous literature exists on the returns to a degree and much of this literature estimates either OLS or SS models. However, obtaining a precise estimate of the returns to a degree, and education in general, is very important in view of the fact that this kind of evidence can shape government policy towards the expansion of higher education and the introduction of student fees. Tables 6 and 7 report the estimates for all of the covariates in the wage models for 1992 and 1997, respectively, whereas Table 8 summarises our findings on the variables of interest.

For the simple human capital model the magnitude of the parameter estimates fall for degree and rise for A Level as we move from the OLS to the DS specification. This picture is replicated for Specification B where demand side determinants of the wage are also included, but only in 1997. In 1992 there is actually an increase in the returns to a degree and to A level, although the magnitude of the estimates from all Specification B models are lower as one would expect given the larger number of covariates that are included in these models. Since three out of four models show a consistent pattern, we regard the 1992 results for Specification B as an anomaly. In

general we also find that the differential between degree and A Level falls by more for DS-SS comparison than for the SS-OLS comparison, especially in 1997 when a much larger fraction of the sample has attrited. However, the differences between the parameter estimates for each of the models are not substantially, or statistically, different. The inference for policy makers in particular is unlikely to change whether one uses a double selection model or not. We therefore conclude that, although NIM is present in the BHPS, it does not introduce substantial bias into the analysis of log wages. There is, however, some improvement in the precision of the estimates, reflected by the smaller standard errors in the DS versus the OLS models (see Tables 6 and 7).

5.3 Identification and dropouts from the BHPS

The importance of our results rests in part on whether the DS model is actually identified. Tables 9 to 12 show the estimates of the employment participation models (Tables 9 and 10) and the retention model (Tables 11 and 12). Recall that in the employment participation model we sought identification through the inclusion of marital status and the number of children. The marital status variables perform better than the variables for the number of children, especially in Specification B, and are more highly significant in 1997. There is some inconsistency in the sign of these variables between Specifications A and B, and hence it could be claimed that the employment participation models are only weakly identified, though no formal test is conducted. However, there are other variables in this model which will aid identifica-

tion, such as cumulative employment and cumulative unemployment experience. We therefore argue that the employment participation models are likely to be identified.

Turning to the retention models two variables are included that are excluded from both the wage model and the employment participation model, that is, the number of contacts and the cooperativeness of the interviewee. The estimates on these variables are correctly signed and statistically significant throughout, suggesting that the retention models are identified. Not surprisingly, more cooperative interviewees are more likely to remain in the BHPS, whereas individuals requiring a greater number of contacts are less likely to remain in the Survey. In addition, Tables 11 and 12 reveal that the respondents who are more likely to remain in the survey are those individuals with higher qualifications, a disability or health problem and those workers with greater employment experience. Interestingly, workers who live in local labour markets with more job vacancies are more likely to remain in the survey, whereas those workers in areas with a higher unemployment rate are more likely to attrit. Part of the attrition from the BHPS must therefore be related to the migration of workers in search of employment.

Finally, Figure 1 plots the residuals $r(W|R = 1, E = 1)$ against $E(W_i \mid R_i = 1, E_i = 1)$ for each of the models, which suggests that there are no seriously outlying observations. However, further confirmation of model adequacy requires a sensitivity analysis along the various lines suggested in section 3.2.

6 Conclusion

In this paper we have presented a double selection model and a truncated double selection model as a means of combating the widely made assumption that dropouts from survey data are ignorable for estimating wage equations. We do acknowledge that approach we have presented in this paper could be extended by using polynomials or splines of the covariates in the various linear predictors and by allowing some of the covariates (e.g. education) in the wage equation to be endogenous.

The problem of non-ignorable missing data (NIM) has received insufficient attention in the literature on sample selection in economics, yet non ignorable drop out is likely to be a generic feature of social survey data, a feature that wont go away by ignoring it. To illustrate the potential effects of NIM we focused on a substantive issue that has received considerable attention amongst economists, namely the determinants of hourly earnings. A comparison was made between the results from the classical OLS model without selection into employment and the survey, a model with selection into employment following Heckman and the more complex double selection and truncated double selection models that do allow for selection into employment and selection into the survey. Simulations suggest that in the presence of NIM the double selection model performs best insofar as its parameters are closest to the true parameters. We also applied the double selection models to data from the BHPS in an analysis of the determinants of employment participation and hourly wages, and find that the presence NIM does not substantially bias the estimates. Consequently,

we conclude that for this particular analysis and for this particular data, analysts can be reasonably confident that SS or OLS estimates are not substantially affected by NIM so that the inference for policy makers of the returns to education in this data set, in particular, is unlikely to change whether one uses a double selection model or not. However, this does not mean that NIM should be ignored in other analyses of wage equations. On the contrary, we argue that researchers should test for the presence and impact of NIM, and the models we present offer one way of doing this.

References

Abowd, J.M., and Farber, H.S., (1982), Job queues and the Union status of workers, *Industrial and Labour Relations Review*, 35(3), pp 354-367.

Amemiya, T., (1985), *Advanced Econometrics,* Basil Blackwell. Oxford, England.

Aranda-Ordaz, F. J., (1981), On two families of transformations to additivity for binary response data, *Biometrika*, 68, 2, 357-63.

Blalock, H.M. (JR), (1979), *Social Statistics*, McGraw Hill, London.

Bloom, D.E., and Killingsworth, M.R., (1985), Correcting for truncation bias caused by a latent truncation variable, *Journal of Econometrics,* 27, pp131-135.

Chevalier A., and Walker, I., (2001), *Ch. 16, United Kingdom*, pp 302-330, in Harmon, C., Walker, I. and Westergaard-Nielsen, N., (eds), Education and Earning in Europe, A cross country analysis of the returns to education, Edward Elgar, Cheltenham, UK.

Cook R. D., (1986), Assessment of local influence, *Journal of the Royal Statistical Society*, Series B, 48, 133-169.

Diggle, P. J. and Kenward, M. G., (1994), Informative Drop-out in longitudinal data

analysis, *Journal of Applied Statistics,* 43, 49-93.

Fronstin, P., Greenberg, D.H., and Robins, P.K., (2001), Parental disruption and labour market performance of children when they reach adulthood, *Journal of Population Economics*, 14, pp 137-172.

Gregg, P., and Machin, S., (1998), *Child development and success or failure in the youth labour market*, Centre for Economic Performance Discussion Paper No. 397, London School of Economics, London UK.

Gronau, R., (1973), The effects of children on the housewife's value of time, *Journal of Political Economy*, 81, pp S168-S199.

Harmon, C., Walker, I. and Westergaard-Nielsen, N., (2001), *Ch. 1, Introduction*, pp 1- 37 in Harmon, C., Walker, I. and Westergaard-Nielsen, N., (eds), Education and Earning in Europe, A cross country analysis of the returns to education, Edward Elgar, Cheltenham, UK.

Hausman, J.A., (1978), Specification tests in Econometrics, *Econometrica*, 46, pp 1251-1271.

Heckman, J.J., (1976), The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models, *Annals of Economic and Social Measurement*, 5, 475-492.

Heckman, J. J., (1979), Sample selection bias as a specification error, *Econometrica* 47, 153-161.

Hildreth, A., (1999), What has happened to the union wage differential in Britain in the 1990s, *Oxford Bulletin of Economics and Statistics*, 61, pp 4- 31.

Hirano, K., Imbens, G., Ridder, G. and Rubin, D., (1998), *Combining panel data sets with attrition and refreshment samples*, Technical Working Paper, National Bureau of Economic Research, Cambridge, Massachusetts.

Horowitz, J. L. and Manski, C. F. (1998), Censoring of outcomes and regressors due to survey nonresponse: identification and estimation using weights and imputations, *Journal of Econometrics,* 84, 37-58.

Kenward, M. G., (1998), Selection models for repeated measurements with non-random dropout: an illustration of sensitivity, *Statistics in Medicine*, 17, 2723-2732.

Kiernan K., (1997), *The legacy of parental divorce: social,economics and demographic experiences in adulthood*, Centre for Analysis of Social Exclusion, CASE paper No. 1, London School of Economics, London UK.

Lee, L. F., (1983), Generalized econometric models with selectivity, *Econometrica,* 51, 507-512.

Maddala, G. S., (1983), *Limited dependent and qualitative variables in econometrics,* Cambridge University Press, Cambridge.

Muthen, B., and Joreskog, K.G., (1983), Selectivity problems in quasi-experimental studies, *Evaluation Review*, 7, pp139-174.

NAG (1996), *Numerical Algorithms Group Manual*, Mark 16, NAG, Oxford, UK.

Poirer, D.J., (1980), Partial Observability in Bivariate Probit Models, *Journal of Econometrics,* 12, pp209-217.

Rotnitzky, A., Robins, J. M. and Scharfstein, D. O., (1998), Semiparametric regression for repeated outcomes with nonignorable nonresponse, *Journal of the American*

Statistical Association, 93, 1321-1339.

Scharfstein D. O., Rotnitzky A. and Robins J. M., (1999), Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association*, 94, 1096-1120.

Verbeke, G. and Molenberghs, G., (2000), *Linear mixed models for longitudinal data*, Springer.

Waldfogel, J., (1995), The price of motherhood: family status and women's pay in a young British Cohort, *Oxford Economic Papers,* 47, pp 548-610.

Appendix

Table A. Descriptive statistics for covariates in the wage models, 1992

Variable	Mean	Std. Dev.	Min	Max
ln(hourly wage)	1.923	0.561	0.104	4.185
Experience	1.999	1.273	0.008	6.358
Experience squared	5.616	6.001	0.000	40.428
Higher degree	0.024	0.153	0	1
Degree	0.105	0.306	0	1
HND (equivalent)	0.069	0.254	0	1
A Level	0.211	0.408	0	1
O Level (equivalent)	0.258	0.438	0	1
Below O Level	0.056	0.231	0	1
Part-time	0.036	0.185	0	1
Self-employed	0.154	0.361	0	1
Seasonal/Temporary	0.018	0.134	0	1
Fixed term	0.031	0.175	0	1
<500 employees	0.231	0.421	0	1
501-999	0.067	0.249	0	1
>1000	0.095	0.293	0	1
Union recognition	0.088	0.284	0	1
Energy	0.037	0.188	0	1
Minerals	0.043	0.202	0	1
Engineering	0.135	0.341	0	1
Manufacturing	0.112	0.315	0	1
Construction	0.091	0.287	0	1
Distribution	0.145	0.353	0	1
Transport	0.083	0.276	0	1
Banking/Finance	0.131	0.338	0	1
Other services	0.192	0.394	0	1

Table B. Descriptive statistics for covariates in the wage models, 1997

Variable	Mean	Std. Dev.	Min	Max
ln(hourly wage)	2.148	0.550	0.158	4.725
Experience	2.360	1.179	0.025	6.858
Experience squared	6.960	6.165	0.001	47.037
Higher degree	0.034	0.182	0	1
Degree	0.128	0.335	0	1
HND (equivalent)	0.089	0.285	0	1
A Level	0.229	0.420	0	1
O Level (equivalent)	0.239	0.427	0	1
Below O Level	0.062	0.241	0	1
Part-time	0.046	0.210	0	1
Self-employed	0.105	0.307	0	1
Seasonal/Temporary	0.016	0.127	0	1
Fixed term	0.029	0.168	0	1
<500 employees	0.250	0.433	0	1
501-999	0.078	0.269	0	1
>1000	0.089	0.285	0	1
Union recognition	0.438	0.496	0	1
Energy	0.026	0.160	0	1
Minerals	0.052	0.221	0	1
Engineering	0.137	0.344	0	1
Manufacturing	0.113	0.316	0	1
Construction	0.076	0.265	0	1
Distribution	0.137	0.344	0	1
Transport	0.089	0.285	0	1
Banking/Finance	0.139	0.346	0	1
Other services	0.209	0.407	0	1

List of Tables and Figures Referred to in the Text

Table 1. Parameter estimates for data simulated with ρ_{re}=0.7, ρ_{rw}=0.2 and ρ_{ew}=-0.4
Table 2. Parameter estimates for data simulated with ρ_{re}=-0.7, ρ_{rw}=-0.2 and ρ_{ew}=0.4
Table 3. Parameter estimates for data simulated with ρ_{re}=-0.7, ρ_{rw}=0.2 and ρ_{ew}=0.4
Table 4. Covariance parameter estimates and test statistics, 1992
Table 5. Covariance parameter estimates and test statistics, 1997
Table 6. The determinants of log hourly wages, 1992
Table 7. The determinants of log hourly wages, 1997
Table 8. The impact of NIM on rates of return to education
Table 9. The determinants of employment participation 1992
Table 10. The determinants of employment participation, 1997
Table 11. The determinants of retention in the BHPS, 1992
Table 12. The determinants of retention in the BHPS, 1997

Figure 1. Residual Plots

Table 1. Parameter estimates for data simulated with ρ_{re}=0.7, ρ_{rw}=0.2 and ρ_{ew}=-0.4

	OLS			Single Selection			Double Selection			Truncated Double Selection		
Parameter	Mean	s.d	t-ratio	Mean	s.d.	t-ratio	Mean	s.d	t-ratio	Mean	s.d	t-ratio
$\beta_{0(w)}$ (0.1)	**0.150**	0.161	3.102	**0.314**	0.190	11.224	0.056	0.255	-1.727	**-0.090**	0.934	-2.028
$\beta_{Age(w)}$ (1.0)	**0.916**	0.130	-6.428	**0.915**	0.135	-6.278	1.013	0.150	0.842	1.036	0.593	0.612
$\beta_{Ed(w)}$ (2.0)	2.004	0.111	0.400	**1.934**	0.122	-5.397	2.018	0.135	1.304	2.050	0.556	0.890
$\beta_{Race(w)}$ (-0.2)	-0.193	0.040	1.724	-0.193	0.040	1.638	-0.193	0.040	1.713	-0.192	0.041	1.864
$\beta_{0(e)}$ (-0.75)				**0.418**	0.331	35.125	-0.734	0.351	0.448	-0.939	2.783	-0.677
$\beta_{Age(e)}$ (0.5)				**0.031**	0.253	-18.477	0.476	0.209	-1.132	0.498	1.791	-0.011
$\beta_{Mar(e)}$ (0.2)				**0.244**	0.071	6.115	0.195	0.057	-0.820	0.210	0.065	1.577
$\beta_{Ed(e)}$ (1.0)				**0.707**	0.245	-11.897	1.001	0.212	0.064	1.043	1.850	0.230
$\beta_{0(r)}$ (-1.75)							-1.726	0.192	1.222	**-3.432**	5.637	-2.969
$\beta_{Age(r)}$ (1.0)							1.002	0.138	0.117	1.035	4.854	0.071
$\beta_{Q(r)}$ (0.6)							0.598	0.039	-0.464	**1.696**	1.999	5.456
$\beta_{Ed(r)}$ (1.0)							0.984	0.136	-1.191	**1.816**	3.777	2.149
α_{w} (0.0)				**-0.017**	0.021	-7.835	0.008	0.042	1.872	0.003	0.037	0.782
α_{ew} (-0.8473)				**-1.147**	0.486	-6.131	-0.791	0.351	1.598	-0.867	0.461	-0.428
α_{re} (1.7346)							1.748	0.350	0.393	**1.503**	0.615	-3.750
α_{rw} (0.40547)							0.438	0.238	1.346	0.411	0.337	0.175

Table 2. Parameter estimates for data simulated with ρ_{re}=-0.7, ρ_{rw}=-0.2 and ρ_{ew}=0.4

	OLS			Single Selection			Double Selection			Truncated Double Selection		
Parameter	Mean	s.d	t-ratio	Mean	s.d	t-ratio	Mean	s.d	t-ratio	Mean	s.d	t-ratio
$\beta_{0(w)}$ (0.1)	**0.380**	0.171	16.366	**-0.177**	0.345	-7.976	0.105	0.301	0.181	0.164	1.037	0.612
$\beta_{Age(w)}$ (1.0)	**0.923**	0.126	-6.070	**1.093**	0.172	5.401	0.992	0.166	-0.508	0.950	0.883	-0.562
$\beta_{Ed(w)}$ (2.0)	**1.864**	0.118	-11.404	**2.097**	0.166	5.821	1.998	0.153	-0.130	2.048	0.441	1.084
$\beta_{Race(w)}$ (-0.2)	-0.204	0.045	-0.993	-0.204	0.045	-0.971	-0.204	0.045	-0.988	-0.204	0.046	-0.812
$\beta_{0(e)}$ (-0.75)				**-1.789**	0.249	-41.587	-0.756	0.300	-0.183	-0.453	1.883	1.571
$\beta_{Age(e)}$ (0.5)				**0.918**	0.183	22.675	0.493	0.185	-0.388	0.641	1.728	0.812
$\beta_{Mar(e)}$ (0.2)				**0.229**	0.052	5.585	0.202	0.049	0.462	0.198	0.060	-0.387
$\beta_{Ed(e)}$ (1.0)				**1.417**	0.178	23.311	1.001	0.188	0.071	0.932	0.995	-0.683
$\beta_{0(r)}$ (-1.75)							-1.744	0.174	0.321	**-5.689**	6.803	-5.760
$\beta_{Age(r)}$ (1.0)							1.000	0.135	0.032	0.842	5.185	-0.303
$\beta_{Q(r)}$ (0.6)							0.599	0.037	-0.400	**2.799**	4.300	5.089
$\beta_{Ed(r)}$ (1.0)							0.996	0.121	-0.319	**3.159**	4.752	4.520
α_{w} (0.0)				0.001	0.037	0.393	**0.016**	0.074	2.160	**0.028**	0.061	4.546
α_{ew} (0.8473)				0.826	0.496	-0.419	0.788	0.466	-1.275	0.895	0.530	0.905
α_{re} (-1.7346)							-1.843	0.716	-1.510	-1.829	1.292	-0.725
α_{rw} (-0.40547)							-0.356	0.292	1.688	-0.346	0.451	1.318

Table 3. Parameter estimates for data simulated with ρ_{re}=-0.7, ρ_{rw}=0.2 and ρ_{ew}=0.4

	OLS			Single Selection			Double Selection			Truncated Double Selection		
Parameter	Mean	s.d.	t-ratio	Mean	s.d.	t-ratio	Mean	s.d.	t-ratio	Mean	s.d.	t-ratio
$\beta_{0(w)}$ (0.1)	**1.229**	0.146	77.160	**0.286**	0.213	8.691	0.140	0.226	1.751	**-0.182**	1.016	-2.759
$\beta_{Age(w)}$ (1.0)	**0.604**	0.138	-28.461	**0.881**	0.148	-8.019	0.984	0.150	-1.050	1.101	0.767	1.314
$\beta_{Ed(w)}$ (2.0)	**1.540**	0.102	-44.721	**1.939**	0.121	-5.020	1.983	0.125	-1.336	2.067	0.445	1.488
$\beta_{Race(w)}$ (-0.2)	-0.201	0.050	-0.112	-0.201	0.050	-0.247	-0.201	0.050	-0.271	-0.204	0.051	-0.755
$\beta_{0(e)}$ (-0.75)				**-1.823**	0.211	-50.572	-0.741	0.242	0.373	-0.420	1.780	1.845
$\beta_{Age(e)}$ (0.5)				**0.912**	0.145	28.320	0.465	0.150	-2.347	0.317	1.474	-1.237
$\beta_{Mar(e)}$ (0.2)				**0.227**	0.050	5.409	0.204	0.043	0.838	0.196	0.053	-0.739
$\beta_{Ed(e)}$ (1.0)				**1.440**	0.158	27.662	0.998	0.165	-0.148	1.003	0.829	0.038
$\beta_{0(r)}$ (-1.75)							-1.766	0.174	-0.913	-2.297	3.348	-1.627
$\beta_{Age(r)}$ (1.0)							1.023	0.142	1.633	1.291	2.774	1.045
$\beta_{Q(r)}$ (0.6)							0.597	0.035	-0.934	**1.543**	2.085	4.498
$\beta_{Ed(r)}$ (1.0)							1.009	0.125	0.723	0.895	2.298	-0.456
α_{w} (0.0)				**0.015**	0.035	4.285	0.002	0.055	0.437	**0.024**	0.051	4.557
α_{ew} (0.8473)				**1.503**	0.295	22.092	0.778	0.348	-1.982	0.794	0.671	-0.793
α_{re} (-1.7346)							-1.804	0.372	-1.871	-1.740	0.718	-0.081
α_{rw} (0.40547)							0.443	0.286	1.300	0.486	0.415	1.937

Table 4. Covariance parameter estimates and test statistics, 1992

	Specification A						Specification B					
	Single Selection Estimates		Probit R=1	Double Selection Estimates			Single Selection Estimates		Probit R=1	Double Selection Estimates		
σ_w (α_w)	0.519	-1.312		0.540	-1.234		0.493	-1.415		0.508	-1.355	
	.0087[a]				0.026		0.008				0.026	
ρ_{ew} (α_{ew})	-0.494	-1.082		-0.580	-1.326		-0.435	-0.933		-0.341	-0.710	
	0.055				0.031		0.057				0.039	
ρ_{re} (α_{re})			0	0.574	1.307				0	0.732	1.868	
					0.040						0.045	
ρ_{rw} (α_{rw})			0	0.325	0.675				0	0.378	0.796	
					0.032						0.042	
Log L	-2517.591		-1439.110	-3896.675			-2423.963		-1439.110	-3790.004		
Total cases	3316		3998	3998			3316		3998	3998	3998	
Uncensored cases	2131			2131			2131			2131		
χ^2	50.63[b]			120.05[c]			39.39[b]			146.14[c]		
df	1			2			1			2		

Note: a= Standard Error
b = test of ρ_{ew}=0
c = test of ρ_{re}=ρ_{rw}=0

Table 5. Covariance parameter estimates and test statistics, 1997

	Specification A						Specification B					
	Single Selection Estimates		Probit R=1	Double Selection Estimates			Single Selection Estimates		Probit R=1	Double Selection Estimates		
σ_w (α_w)	0.497	-1.399		0.542	-1.224		0.474	-1.492		0.521	-1.304	
	.0098[a]				0.032		0.009				0.032	
ρ_{ew} (α_{ew})	-0.274	-0.563		-0.266	-0.544		-0.224	-0.456		-0.119	-0.238	
	0.084				0.028		0.084				0.025	
ρ_{re} (α_{re})			0	0.692	1.703				0	0.788	2.130	
					0.040						0.039	
ρ_{rw} (α_{rw})			0	0.509	1.123				0	0.516	1.142	
					0.024						0.017	
Log L	-1441.649		-2341.229	-3687.591			-1383.172		-2341.229	-3608.635		
Total cases	2097		3998	3998			2097		3998	3998	3998	
Uncensored cases	1339			2131			1339			2131		
χ^2	9.31[b]			190.57[c]			6.37[b]			231.53[c]		
df	1			2			1			2		

Note: a= Standard Error
b = test of ρ_{ew}=0
c = test of ρ_{re}=ρ_{rw}=0

Table 6. The determinants of log hourly wages, 1992

	Specification A						Specification B					
	OLS (1)		Single Selection (2)		Double Selection (3)		OLS (1)		Single Selection (2)		Double Selection (3)	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Constant	1.369	0.039	1.538	0.044	1.518	0.034	1.104	0.070	1.252	0.072	1.326	0.042
Experience	0.340	0.031	0.213	0.034	0.183	0.020	0.336	0.030	0.236	0.033	0.243	0.022
Experience squared	-0.065	0.007	-0.036	0.007	-0.026	0.004	-0.063	0.006	-0.041	0.007	-0.042	0.004
Higher degree	0.803	0.075	0.772	0.076	0.786	0.067	0.707	0.074	0.675	0.074	0.674	0.063
Degree	0.653	0.042	0.624	0.042	0.640	0.041	0.564	0.042	0.538	0.043	0.588	0.040
HND (equivalent)	0.468	0.048	0.452	0.048	0.475	0.051	0.379	0.047	0.366	0.047	0.390	0.048
A level	0.298	0.033	0.282	0.034	0.314	0.032	0.242	0.033	0.230	0.033	0.255	0.031
O Level (equivalent)	0.199	0.031	0.179	0.032	0.174	0.031	0.171	0.030	0.153	0.030	0.168	0.030
Below O Level	0.106	0.053	0.088	0.053	0.110	0.058	0.113	0.051	0.098	0.051	0.109	0.057
Part-time							-0.019	0.061	0.042	0.060	0.040	0.027
Self-employed							-0.100	0.034	-0.092	0.033	-0.116	0.014
Seasonal/Temporary							-0.318	0.081	-0.300	0.079	-0.295	0.035
Fixed term							-0.009	0.062	0.013	0.061	-0.039	0.027
<500 employees							0.100	0.028	0.095	0.027	0.111	0.012
501-999							0.171	0.045	0.155	0.045	0.147	0.021
>1000							0.144	0.039	0.135	0.038	0.170	0.017
Union recognition							0.003	0.039	0.003	0.038	0.024	0.017
Industry dummies							Yes		Yes		Yes	

Table 7. The determinants of log hourly wages, 1997

	Specification A						Specification B					
	OLS (1)		Single Selection (2)		Double Selection (3)		OLS (1)		Single Selection (2)		Double Selection (3)	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Constant	1.473	0.064	1.541	0.068	1.312	0.049	1.214	0.108	1.260	0.108	0.961	0.047
Experience	0.362	0.046	0.302	0.050	0.316	0.026	0.327	0.046	0.284	0.049	0.366	0.022
Experience squared	-0.065	0.009	-0.051	0.010	-0.046	0.004	-0.057	0.009	-0.047	0.010	-0.056	0.004
Higher degree	0.776	0.081	0.774	0.081	0.796	0.088	0.690	0.080	0.687	0.079	0.674	0.086
Degree	0.623	0.051	0.622	0.051	0.616	0.044	0.550	0.052	0.548	0.051	0.543	0.042
HND (equivalent)	0.541	0.055	0.543	0.055	0.597	0.051	0.475	0.055	0.475	0.054	0.485	0.048
A level	0.303	0.043	0.302	0.043	0.323	0.041	0.259	0.042	0.258	0.042	0.268	0.037
O Level (equivalent)	0.191	0.042	0.187	0.042	0.218	0.039	0.165	0.041	0.161	0.040	0.128	0.038
Below O Level	0.095	0.065	0.095	0.065	0.094	0.075	0.106	0.063	0.105	0.062	0.137	0.066
Part-time							-0.068	0.070	-0.052	0.069	-0.085	0.012
Self-employed							-0.034	0.049	-0.030	0.048	0.001	0.008
Seasonal/Temporary							-0.209	0.108	-0.190	0.106	-0.211	0.017
Fixed term							-0.039	0.080	-0.037	0.079	-0.032	0.015
<500 employees							0.072	0.034	0.071	0.033	0.072	0.006
501-999							0.103	0.052	0.100	0.052	0.154	0.009
>1000							0.135	0.050	0.138	0.049	0.115	0.008
Union recognition							0.057	0.032	0.056	0.031	0.049	0.006
Industry dummies							Yes		Yes		Yes	

Table 8. The impact of NIM on rates of return to education

Wave/Year	Specification	Qualification	OLS	SS	DS	SS-OLS	DS-SS	DS-OLS
1991	A	(1) Degree	0.653	0.624	0.640			
		(2) A Levels	0.298	0.282	0.314			
		(2)-(1)	0.355	0.342	0.326	-0.130	-0.160	-0.290
	B	(1) Degree	0.564	0.538	0.588			
		(2) A Levels	0.242	0.230	0.255			
		(2)-(1)	0.322	0.308	0.333	-0.140	0.250	0.110
1997	A	(1) Degree	0.623	0.622	0.616			
		(2) A Levels	0.303	0.302	0.323			
		(2)-(1)	0.320	0.320	0.293	0.000	-0.270	-0.270
	B	(1) Degree	0.550	0.548	0.543			
		(2) A Levels	0.259	0.258	0.268			
		(2)-(1)	0.291	0.290	0.275	-0.010	-0.150	-0.160

Table 9. The determinants of employment participation, 1992

	Specification A						Specification B					
	Independent Selection		Single Selection		Double Selection		Independent Selection		Single Selection		Double Selection	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Constant	-2.415	0.842	-3.181	0.824	-2.700	0.446	-2.415	0.842	-3.036	0.832	-3.944	0.511
Owner occupier	0.088	0.117	0.230	0.114	0.104	0.039	0.088	0.117	0.198	0.115	0.074	0.042
Council tenant	-0.371	0.135	-0.276	0.133	-0.269	0.049	-0.371	0.135	-0.304	0.133	-0.229	0.053
Disabled	-0.698	0.161	-0.715	0.154	-0.726	0.076	-0.698	0.161	-0.702	0.156	-0.381	0.076
Non-white	0.318	0.184	0.238	0.177	0.003	0.087	0.318	0.184	0.248	0.178	0.150	0.109
Grammar school	0.107	0.110	0.188	0.107	0.288	0.066	0.107	0.110	0.177	0.108	0.178	0.074
Sixth form	0.186	0.262	0.191	0.249	0.441	0.169	0.186	0.262	0.185	0.252	0.255	0.195
Independent school	0.287	0.171	0.400	0.162	0.314	0.096	0.287	0.171	0.391	0.164	0.516	0.113
Secondary modern	0.061	0.084	0.079	0.081	0.095	0.049	0.061	0.084	0.076	0.082	0.097	0.055
Other school	0.128	0.163	0.110	0.162	0.355	0.094	0.128	0.163	0.096	0.163	0.072	0.107
Health problem	-0.938	0.096	-0.911	0.093	-0.873	0.028	-0.938	0.096	-0.930	0.094	-0.824	0.034
Age	0.851	0.352	1.337	0.343	1.365	0.182	0.851	0.352	1.269	0.347	1.121	0.213
Age squared	-0.160	0.033	-0.197	0.032	-0.189	0.015	-0.160	0.033	-0.192	0.032	-0.176	0.018
ln(duration-employed)	0.521	0.041	0.406	0.040	0.424	0.022	0.521	0.041	0.424	0.040	0.483	0.023
ln(duration-unemployed	-0.267	0.024	-0.281	0.023	-0.261	0.014	-0.267	0.024	-0.281	0.024	-0.206	0.016
Higher degree	0.297	0.283	0.219	0.273	0.314	0.162	0.297	0.283	0.219	0.275	0.229	0.150
Degree	0.182	0.145	0.090	0.142	0.027	0.093	0.182	0.145	0.105	0.143	0.462	0.093
HND (equivalent)	-0.135	0.145	-0.193	0.143	-0.064	0.100	-0.135	0.145	-0.184	0.144	0.118	0.100
A Levels	-0.146	0.105	-0.146	0.103	0.112	0.068	-0.146	0.105	-0.152	0.104	0.054	0.073
O Levels (equivalent)	0.033	0.095	0.055	0.094	0.027	0.066	0.033	0.095	0.047	0.094	0.182	0.067
Below O Level	-0.097	0.163	-0.024	0.161	0.200	0.115	-0.097	0.163	-0.040	0.162	0.066	0.118
ln(vacancy)	-0.063	0.089	0.001	0.088	0.018	0.025	-0.063	0.089	-0.001	0.089	0.071	0.029
ln(unemployment rate)	0.052	0.171	-0.002	0.166	-0.255	0.078	0.052	0.171	-0.007	0.168	0.122	0.089
Children - 1-2	-0.014	0.082	-0.036	0.079	-0.026	0.027	-0.014	0.082	-0.028	0.080	0.171	0.031
Children - 3+	-0.274	0.136	-0.269	0.132	0.045	0.055	-0.274	0.136	-0.264	0.133	0.100	0.065
Married	0.003	0.114	0.131	0.113	0.104	0.046	0.003	0.114	0.104	0.113	-0.078	0.048
Widowed/Divorced	-0.277	0.151	-0.200	0.148	-0.344	0.062	-0.277	0.151	-0.235	0.149	-0.289	0.066
Working partner	0.514	0.080	0.441	0.078	0.371	0.023	0.514	0.080	0.452	0.079	0.402	0.028
Local industry mix	Yes		Yes		Yes		Yes		Yes		Yes	
Cohort dummies	Yes		Yes		Yes		Yes		Yes		Yes	

Table 10. The determinants of employment participation, 1997

	Specification A						Specification B					
	Independent Selection		Single Selection		Double Selection		Independent Selection		Single Selection		Double Selection	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Constant	0.133	1.442	0.035	1.438	-2.459	0.482	0.133	1.442	0.088	1.438	-0.226	0.275
Owner occupier	0.168	0.179	0.229	0.182	0.103	0.020	0.168	0.179	0.197	0.180	-0.004	0.014
Council tenant	-0.073	0.214	-0.085	0.216	-0.156	0.029	-0.073	0.214	-0.098	0.216	-0.171	0.021
Disabled	-0.799	0.218	-0.810	0.217	-0.528	0.047	-0.799	0.218	-0.799	0.217	-0.636	0.047
Non-white	0.855	0.344	0.973	0.339	0.828	0.112	0.855	0.344	0.976	0.343	0.582	0.110
Grammar school	-0.064	0.152	-0.076	0.150	0.014	0.069	-0.064	0.152	-0.072	0.151	-0.125	0.068
Sixth form	-0.066	0.391	-0.027	0.390	0.229	0.165	-0.066	0.391	-0.032	0.392	0.173	0.171
Independent school	0.284	0.260	0.316	0.252	0.519	0.121	0.284	0.260	0.308	0.254	0.307	0.110
Secondary modern	0.128	0.124	0.126	0.123	0.062	0.057	0.128	0.124	0.127	0.123	0.116	0.056
Other school	-0.079	0.241	-0.160	0.242	-0.154	0.104	-0.079	0.241	-0.156	0.243	0.022	0.105
Health problem	-1.092	0.120	-1.079	0.119	-0.913	0.017	-1.092	0.120	-1.080	0.120	-0.894	0.014
Age	-0.874	0.628	-0.616	0.631	-0.039	0.233	-0.874	0.628	-0.674	0.631	-1.307	0.113
Age squared	-0.011	0.053	-0.028	0.053	-0.061	0.019	-0.011	0.053	-0.024	0.053	0.062	0.012
ln(duration-employed)	1.010	0.130	0.882	0.130	0.808	0.033	1.010	0.130	0.916	0.131	0.833	0.024
ln(duration-unemployed)	-0.128	0.033	-0.155	0.034	-0.110	0.007	-0.128	0.033	-0.148	0.034	-0.129	0.005
Higher degree	0.555	0.353	0.493	0.355	0.497	0.154	0.555	0.353	0.505	0.354	0.107	0.137
Degree	0.317	0.188	0.260	0.187	0.200	0.082	0.317	0.188	0.274	0.188	0.274	0.068
HND (equivalent)	0.037	0.185	-0.001	0.185	0.107	0.087	0.037	0.185	0.010	0.185	-0.094	0.074
A Levels	0.006	0.145	0.000	0.144	0.001	0.071	0.006	0.145	0.000	0.145	0.026	0.059
O Levels (equivalent)	0.174	0.136	0.144	0.136	0.161	0.069	0.174	0.136	0.150	0.136	-0.097	0.060
Below O Level	0.068	0.276	0.074	0.274	-0.294	0.121	0.068	0.276	0.063	0.274	0.232	0.097
ln(vacancy)	0.214	0.129	0.200	0.128	0.173	0.014	0.214	0.129	0.202	0.129	0.018	0.010
ln(unemployment rate)	-0.228	0.119	-0.255	0.118	-0.124	0.016	-0.228	0.119	-0.257	0.119	-0.091	0.012
Children - 1-2	-0.156	0.145	-0.116	0.144	-0.111	0.013	-0.156	0.145	-0.126	0.145	-0.173	0.009
Children - 3+	-0.392	0.235	-0.359	0.235	-0.315	0.022	-0.392	0.235	-0.359	0.235	-0.127	0.016
Married	0.068	0.181	0.133	0.183	0.244	0.023	0.068	0.181	0.115	0.183	0.163	0.015
Widowed/Divorced	0.010	0.213	0.074	0.215	0.407	0.028	0.010	0.213	0.050	0.215	0.056	0.022
Working partner	0.506	0.110	0.493	0.109	0.194	0.012	0.506	0.110	0.504	0.109	0.363	0.009
Local industry mix	Yes		Yes		Yes		Yes		Yes		Yes	
Cohort dummies	Yes		Yes		Yes		Yes		Yes		Yes	

Table 11. The determinants of retention in the BHPS, 1992

	Independent Selection		Specification A Double Selection		Specification B Double Selection	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Constant	0.748	0.749	0.505	0.464	1.020	0.454
Owner occupied	0.131	0.090	0.032	0.040	-0.079	0.039
Council tenant	0.025	0.109	-0.121	0.053	-0.137	0.051
Disabled	0.394	0.155	0.563	0.088	0.373	0.100
Non-white	-0.122	0.126	0.035	0.099	-0.001	0.103
Grammar school	0.334	0.097	0.409	0.076	0.386	0.073
Sixth form	0.523	0.241	0.487	0.199	0.619	0.189
Independent school	0.133	0.140	0.308	0.112	0.167	0.110
Secondary Modern	0.128	0.067	0.152	0.055	0.050	0.054
Other school	0.063	0.129	-0.035	0.105	0.061	0.099
Health problem	0.096	0.089	0.102	0.035	0.156	0.033
Age	-0.813	0.237	-0.871	0.201	-0.894	0.196
Age squared	0.080	0.020	0.087	0.016	0.088	0.016
ln(duration-employed)	0.340	0.016	0.349	0.014	0.357	0.014
ln(duration-unemployed)	0.233	0.025	0.227	0.016	0.219	0.016
Higher degree	0.292	0.217	0.244	0.174	0.162	0.147
Degree	0.525	0.135	0.477	0.093	0.501	0.093
HND (equivalent)	0.316	0.132	0.421	0.099	0.313	0.096
A Level	0.315	0.089	0.207	0.067	0.248	0.067
O Level (equivalent)	0.023	0.078	0.031	0.063	0.007	0.063
Below O Level	-0.010	0.127	-0.065	0.104	-0.079	0.112
Father - Professional/manager	-0.022	0.075	-0.212	0.020	-0.101	0.020
Father-Skilled non-manual	0.160	0.113	-0.028	0.027	0.014	0.028
Father - Skilled manual	-0.082	0.065	-0.063	0.018	-0.244	0.019
ln(vacancy)	0.235	0.070	0.214	0.027	0.312	0.027
Ln(unemployment rate	-0.263	0.130	-0.181	0.076	-0.286	0.074
Children - 1-2	0.271	0.066	0.235	0.031	0.209	0.031
Children 3+	0.423	0.123	0.330	0.061	0.357	0.063
Married	0.003	0.095	0.085	0.052	-0.090	0.048
Widowed/Divorced	-0.123	0.123	-0.053	0.073	-0.208	0.067
Working partner	0.005	0.073	-0.184	0.027	0.133	0.028
Co-operative - interview	0.399	0.112	0.245	0.042	0.279	0.048
Number of contacts	-0.023	0.013	-0.042	0.004	-0.035	0.004
Local industry mix	Yes		Yes		Yes	
Regional dummies	Yes		Yes		Yes	
Cohort dummies	Yes		Yes		Yes	

Table 12. The determinants of retention in the BHPS, 1997

	Independent Selection		Specification A Double Selection		Specification B Double Selection	
	Estimate	s.e.	Estimate	s.e.	Estimate	s.e.
Constant	-1.300	0.598	-1.133	0.295	-0.695	0.175
Owner occupied	0.153	0.079	-0.060	0.017	0.166	0.012
Council tenant	0.100	0.094	-0.065	0.022	0.093	0.016
Disabled	0.214	0.117	0.126	0.049	0.106	0.037
Non-white	-0.143	0.117	-0.103	0.098	-0.036	0.094
Grammar school	0.118	0.074	0.136	0.062	0.166	0.059
Sixth form	0.402	0.182	0.445	0.147	0.378	0.145
Independent school	-0.097	0.111	-0.183	0.105	-0.038	0.093
Secondary Modern	0.017	0.056	0.032	0.050	-0.028	0.048
Other school	-0.007	0.106	0.076	0.090	-0.013	0.092
Health problem	-0.089	0.070	-0.127	0.018	-0.061	0.013
Age	0.061	0.195	-0.211	0.172	-0.080	0.080
Age squared	-0.020	0.016	-0.002	0.014	-0.009	0.009
ln(duration-employed)	0.397	0.020	0.407	0.010	0.370	0.007
ln(duration-unemployed)	0.078	0.018	0.078	0.007	0.080	0.005
Higher degree	0.524	0.175	0.356	0.064	0.421	0.054
Degree	0.467	0.100	0.338	0.048	0.211	0.036
HND (equivalent)	0.384	0.102	0.430	0.047	0.263	0.036
A Level	0.257	0.071	0.225	0.040	0.150	0.031
O Level (equivalent)	0.022	0.063	0.101	0.038	-0.002	0.030
Below O Level	0.186	0.107	0.110	0.048	0.185	0.050
Father - Professional/manager	-0.039	0.061	-0.020	0.011	-0.045	0.008
Father-Skilled non-manual	0.031	0.086	0.027	0.016	0.190	0.010
Father - Skilled manual	-0.057	0.053	0.043	0.010	-0.100	0.007
ln(vacancy)	0.070	0.060	0.000	0.013	0.056	0.010
Ln(unemployment rate	-0.072	0.103	0.010	0.022	-0.029	0.015
Children - 1-2	0.094	0.053	0.021	0.010	0.121	0.007
Children 3+	-0.024	0.092	-0.070	0.018	0.062	0.012
Married	0.030	0.078	0.104	0.020	-0.058	0.013
Widowed/Divorced	-0.030	0.104	0.042	0.028	0.011	0.020
Working partner	-0.102	0.056	-0.114	0.011	-0.036	0.007
Co-operative - interview	0.335	0.112	0.335	0.027	0.260	0.020
Number of contacts	-0.031	0.011	-0.022	0.002	-0.026	0.001
Local industry mix	Yes		Yes		Yes	
Regional dummies	Yes		Yes		Yes	
Cohort dummies	Yes		Yes		Yes	

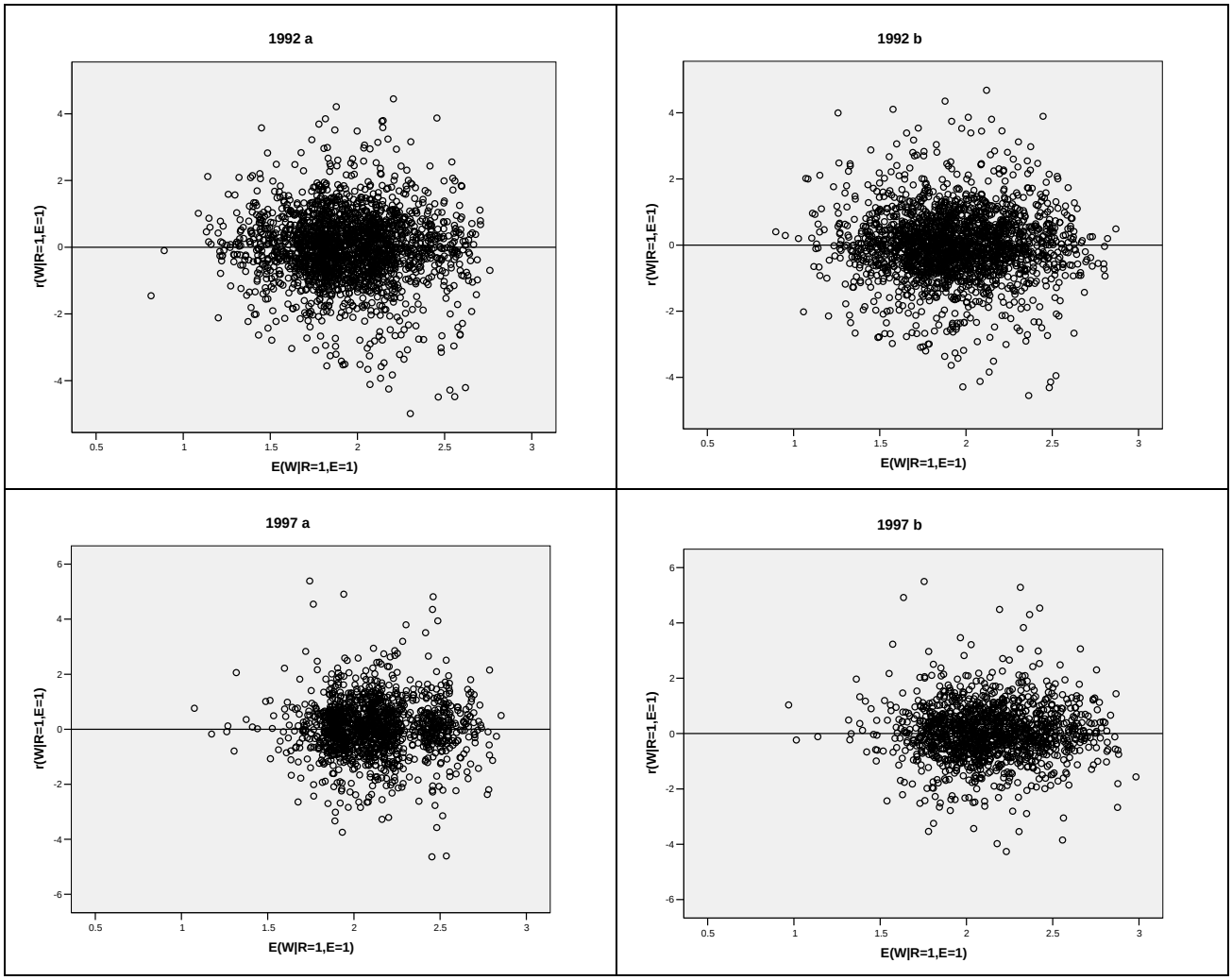

Figure 1. Residual plots