Health and Wages
Panel data estimates considering selection and endogeneity

Robert Jäckle

Ifo Working Paper No. 43

March 2007

Ifo Working Paper No. 43

Health and Wages
Panel data estimates considering selection and endogeneity*

Abstract

This paper investigates the effects of health on wages by controlling for a number of problems: first, the unobservable genetic endowment may cause an omitted variable bias; second, using a self-reported health variable could induce measurement error; third, the issue of reverse causality arises; and fourth, panel attrition driven by the endogenous decision to participate in the labour market may result in inconsistent estimation. By using recently developed methods, I control for all of the above issues in one framework. The results show that good health raises wages for both women and men. I find the health variable to suffer from measurement error. In the male sample, applying OLS or 2SLS, instead of methods accounting for selection and individual heterogeneity, causes an upward bias in the health coefficient.

JEL Code: I1, J4, C33, C34.
Keywords: Health, wages, fixed effects, sample selection, instrumental variables.

Robert Jäckle
Ifo Institute for Economic Research
at the University of Munich
Poschingerstr. 5
81679 Munich, Germany
Phone: +49(0)89/9224-1603
jaeckle@ifo.de

* I thank Christian Holzner, Georg Wamser, and Joachim Winter for helpful discussions. Matthew Scogin is thanked for thoroughly reviewing the paper. I would like to gratefully notice Joachim Wolff who kindly shared his GSOEP preparation files. All remaining deficiencies are my responsibility.

1 Introduction

Whether there exists a measurable interrelation between health and wages is an important question in both labour and health economics. There are two important reasons which establish a link between the state of health and wages. First, health as part of one's human capital may affect labour market productivity and hence wages. Second, as Grossman (2001) points out, if marginal benefits of investment in health increase with the salary, health should rise with wages and the issue of reverse causality comes up. However, a number of further challenges arise. To start with, as self-reported health satisfaction is used for estimation, it is not possible to assess one's actual health status accurately and measurement error could be a source of bias. Another shortcoming that is unappreciated in most earlier studies of this kind is sample selection. Since labour market participation is endogenous – with one reason for selection being the health status – applying methods without selection corrections may result in inconsistent estimation. Finally, an issue particulary relevant in the health context is individual heterogeneity. The reasonable presumption that genetic endowment is correlated with health calls for panel data techniques to account for the well known omitted variable bias.

In an attempt to control for all of these problems in one framework, I utilise recently developed estimation methods proposed by Wooldridge (1995) and Semykina and Wooldridge (2005). In the first paper, Wooldridge develops new straightforward techniques to test and correct for sample selection in fixed effects models. The method relies on standard probit estimates for each year to calculate T inverse Mills ratios (IMRs) and explicitly models the conditional mean of the error terms in the main equation. It is easier to implement and more flexible than other models in the literature as it does not demand any known distribution of the errors in the equation of interest, and allows them to be time heteroscedastic and serially correlated in an unspecified way. In an application to female labour supply, Dustmann and Rochina-Barrachina (2000) compare Wooldridge's (1995) estimator to the methods proposed by Kyriazidou (1997) and Rochina-Barrachina (1999). Kyriazidou's (1997) estimator is semi-parametric and matches observations with the same selection effect in two periods. By taking the difference between any two years one gets rid of both individual heterogeneity and sample selection. A crucial point is the "conditional exchangeability" assumption, implying that the idiosyncratic errors are homoscedastic over time conditional on the covariates and unobserved effects in both equations. While Kyriazidou (1997) does not impose distributional

assumptions on the selection term, Rochina-Barrachina (1999) parameterises this effect and assumes joint normality of the error terms in the probit and main equation. Her method does not rely on the "conditional exchangeability" assumption. Dustmann and Rochina-Barrachina (2000) show how to expand the three estimators to account for the problems of non-strict exogeneity and measurement error. Similarly, Semykina and Wooldridge (2005) enhance Wooldridge's (1995) estimator and demonstrate how to test and control for sample selection in a fixed effects model with endogeneity. Again, their approach allows for time heteroscedasticity and autocorrelation in the error terms in both equations.

Turning to the literature concerned with the impact of health on wages, Lee (1982) suggests an econometric model that accounts for the simultaneous effects of health and wages in a structural multi-equation system, based on a generalised version of the Heckman (1978) treatment model. Using a male sample of US citizens, he finds that health and wages are strongly interrelated; that is the wage rate positively affects health and vice versa. Haveman, Wolfe, Kreider and Stone (1994) estimate a multiple equation system for working time, wages, and health, employing generalised methods of moments techniques. In their male sample for the US they show that poor health affects wages negatively. Contoyannis and Rice (2001) study the impact of self-assessed general and psychological health on wages using the British Household Income Survey. They apply fixed effects and random effects instrumental variable estimators and conclude that reduced psychological health decreases male wages, while positive self-assessed health increases hourly wages for women. In a recent study, Gambin (2005) investigates the relationship between health and wages for 14 European countries and finds that for men, self-reported health has a greater effect than for females, while in the case of chronical diseases the opposite holds true.

This paper uses data from the German Socio-Economic Panel (GSOEP) to estimate reduced-form wage equations for women and men augmented by a variable measuring health satisfaction. I follow Wooldridge (1995) and Semykina and Wooldridge (2005) in an attempt to account for the problems of unobserved heterogeneity, sample selection, and endogeneity. A number of tests provide evidence that for the male sample selection corrections are indicated, while for women no selection problems occur. The results show that good health raises wages. For females an increase in health satisfaction by 10% enhances (hourly) wages approximately by 0.14 to 0.47 percent. In the male sample the increase of the wage rate ranges from about 0.09 to 0.88 per-

cent. The health variable is found to suffer from measurement error. For men, employing pooled OLS or 2SLS, instead of methods accounting for selection and individual heterogeneity, is accompanied by an upward bias in the health coefficient.

The remainder of this paper is organised as follows: the starting point is a discussion of specification issues and resulting problems; that is followed by a detailed overview of the different estimation methods in section 3; the next part provides summary statistics of the data; then, in section 5, I look at estimation and test results; and, finally, section 6 concludes the paper.

2 Model Specification and Resulting Problems

In order to improve our understanding of how health affects wages, a simple model is presented. In this model, the only input factor is the quantity of effective labour L_t a firm uses to produce Y_t at time t. The production function of a firm is determined by the function $Y_t = F(L_t)$, and the amount of effective labour can be written as

$$L_t = \sum_{i=1}^{n} p_i(s_i, a_{i,t}, h_{i,t}) \times l_{i,t}, \tag{1}$$

where $l_{i,t}$ is the actual labour supply per employee i, and $p_i(\cdot)$ is a unknown function that determines the effectiveness of $l_{i,t}$. The efficiency of an individual's working hours depends on the (maximum) years of schooling s_i, age $a_{i,t}$, and her/his state of health $h_{i,t}$. In what follows, I refer to the first two variables as the human capital part of $p_i(\cdot)$ and to the latter part as health effect.

If workers are paid according to their marginal product the log wage of each employee can be written as

$$\log w_{i,t} = log[\frac{dF}{dL_t} \times \frac{\partial L_t}{\partial l_{i,t}}] = \log F_{L_t} + \log p_i(s_i, a_{i,t}, h_{i,t}). \tag{2}$$

This implies that log wages can be decomposed into the term $\log F_{L_t}$, which depends on supply and demand factors on the firm level, and a human capital and health effect, respectively, that varies on the level of the employee. In order to approximate the first part, I use yearly averages of job-seekers and

notified vacancies on the level of the federal states in Germany,[1] as well as four different categories for the firm size. To find a plausible functional form for the human capital part of the term $\log p_i(\cdot)$, a specification of the variables $a_{i,t}$ and s_i similar to the one proposed by Mincer (1958 and 1974) is assumed. Finally, to cover the health status, a function of a self-assessed health measure is included, which asks individuals for a description of their current satisfaction with health.[2]

The following parameterization captures the above model:

$$\log w_{i,t} = \mathbf{b}_{s,t}\boldsymbol{\alpha} + \mathbf{f}_{i,t}\boldsymbol{\beta} + \mathbf{a}_{i,t}\boldsymbol{\gamma} + \theta s_i + \delta f(h_{i,t}) + error, \tag{3}$$

where $\mathbf{b}_{s,t}$ is a vector that approximates supply and demand forces on the (federal) state level s, $\mathbf{f}_{i,t}$ are dummy variables capturing different firm sizes, $\mathbf{a}_{i,t}$ is the vector of a 3rd order polynomial of $a_{i,t}$, s_i are years of schooling or training, $f(h_{i,t})$ is a function of the health variable, and $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \theta, \delta)'$ is the corresponding parameter vector.

The Health Effect. There are a number of important links that connect the state of health and earnings. First, health as part of one's human capital affects labour market productivity and hence wages. Second, in the theoretical work of Grossman (2001), health is defined as an endogenous capital stock, which determines the amount of time one can spend in producing monetary income. Since average hours worked deviate substantially among individuals – with one reason for the difference being the health status – (the log of real) hourly wages rather than monthly earnings are analyzed.[3] Third, in Grossman's (2001) model the rate of return to (gross) investment in health equals the additional availability of healthy time, evaluated at the hourly wage rate. This means that health should rise with wages as the marginal benefits of health investment increase with the wage rate, implying that $h_{i,t}$ is *simultaneously* determined along with $w_{i,t}$.

When estimating the $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \theta, \delta)$ in equation (3) a number of further problems arise. To start with, *measurement error* can be an important source

[1] The corresponding figures are extracted from "Arbeitsstatistik 2005 - Jahreszahlen", provided by the Federal Employment Agency, Nuremberg.

[2] The health variable is categorial, ranging form zero to ten. It is transformed using the following function: $f(h_{i,t}) = \log(h_{i,t} + \sqrt{(h_{i,t}^2 + 1)})$.

[3] This specification also suits equation (2) well since the derivative of $F(L_t)$ with respect to the actual working time, $l_{i,t}$, suggests to utilising hourly wages as dependent variable in equation (3).

of bias when trying to explain wages by employing self-reported questions about health satisfaction. An example of an objective health measure would be a physician's diagnosis of a persons's biological state of health. However, in the absence of such a variable it is likely that δ will be biased towards zero. Another problem arises due to the non-availability of a random sample from the population. In this study, I am interested in the effect of health on the labour market productivity of *all* persons. So, taking into account only the working population induces a sample *selection problem*. In this context, a bias results from the fact that individuals endogenously decide to participate in the labour market. Since it is likely that some of the factors determining participation also affect health, the selection process might lead to inconsistent estimation. A further problem is the possible appearance of an *omitted variable bias*. In this respect one could think of the genetic endowment of a person. If somebody is genetical 'well' equipped she/he might at the same time be healthier and draw a higher salary, so that the health coefficient is upward biased. Finally, as has been noted by Contoyannis, Jones and Rice (2004) and Halliday and Burns (2005) it is likely that the state of health follows a persistent stochastic process. The literature describes two sources of persistence: individual heterogeneity and state dependence. The first one exists due to the (unobserved) degree to which a person is able to cope with individual health shocks (such as hard attacks, accidents, etc.). State dependence, as the second source of persistence, means that an individual's ability to deal with health shocks depends on her/his (former) health status.

The major focus of this paper is to control for simultaneity, measurement error, omitted variables, and selection in one common framework. Unfortunately, the methods proposed in section 3 do not allow to fully cover the dynamics in the state of health. Persistence working through the (unobserved) individual ability to cope with health problems can be controlled for by including unobserved effects. Dynamic effects due to the state dependence of the health status, on the other hand, necessitate to include an (unknown) number of lagged health variables. Yet, the estimation of a 'complete' model identifying the above sources of endogeneity plus the full dynamics of health is beyond the scope of this paper. Therefore, a parsimonious specification including only contemporaneous values of health satisfaction is employed. Non-inclusion of lagged health variables, however, leaves a source of endogeneity in the model which is controlled by applying an instrumental variable approach that uses lagged values of variables related to former health shocks (number of doctor visits in the last three months, number of days off from work due to illness last year).

The Human Capital Part. As mentioned before, the human capital part of $p_i(\cdot)$ is approximated using a Mincer-like specification. He suggests using a model, where log wages are linear in the years of schooling, and linear and quadratic in the years of labour market experience. In an empirical application using the GSOEP, Romeu Gordo (2006) finds evidence for the existence of a positive relationship between unemployment and health satisfaction. On this account, I decided to employ a specification that includes unemployment experience rather than working experience. The combination of the variables age and unemployment experience, however, (implicitly) controls for the corresponding work experience as well. Finally, human capital theory suggest using the time persons spent with their current employer (firm tenure) as a proxy for firm-specific investment in human capital. Since firm tenure (and its square) is more closely related to labour productivity than the general working experience it should cause an extra increase in wages.

To account for the potential correlation between the kind of job an individual holds and her/his health status seven dummies covering the occupational status are included.[4] In order to further control for other structural factors that may affect wages, I control for sector and time fixed effects as well as other binary variables distinguishing between the eastern and western part of Germany, full-time and part-time employment, and German versus non-German nationality.

Thus, enhancing equation (3) according to the previous discussion yields:

$$\log w_{i,t} = \mathbf{b}_{s,t}\boldsymbol{\alpha} + \mathbf{f}_{i,t}\boldsymbol{\beta} + \mathbf{a_{i,t}}\boldsymbol{\gamma} + \mathbf{ue}_{i,t}\boldsymbol{\upsilon} + \mathbf{ft}_{i,t}\boldsymbol{\tau} + \delta f(h_{i,t}) + \mathbf{du}_{i,t}\boldsymbol{\pi} + error, \quad (4)$$

where $\mathbf{b}_{s,t}$, $\mathbf{f}_{i,t}$, $\mathbf{a_{i,t}}$, s_i, and $f(h_{i,t})$ are defined as above; the vector $\mathbf{ue}_{i,t}$ stands for unemployment experience and its square, $\mathbf{ft}_{i,t}$ is the length of time (and its square) a person spent with her/his current employer, and the $\mathbf{du}_{i,t}$ are sector, occupation, part-time work, nationality, and time dummies.

[4] Since it is likely that the state of health depends on the kind of job one holds, interaction terms between the occupational status and the health variables were included. However, the interaction terms turned out to be statistically insignificant and were, therefore, excluded from the final model. In another specification, I interacted age and health since it seems obvious that the later changes in the course of life time. However, again I did not find any significant results with respect to the interaction terms.

3 Econometric Approach

To simplify the notation in this section, the explanatory variables in (4) are approximated by the vector $\mathbf{x}_{i,t}$. The basic framework for the discussion is a linear unobserved regression model of the form:

$$w_{i,t} = \beta_0 + \mathbf{x}_{i,t}\boldsymbol{\beta} + c_i + u_{i,t}, \quad t = 1, 2, ..., T; \quad i = 1, 2, ..., N, \tag{5}$$

where $\mathbf{x}_{i,t}$ is $1 \times K$, $\boldsymbol{\beta}$ is the $K \times 1$ parameter vector of interest, c_i contains unobserved individual characteristics (genetic endowment, ability to deal with health problem, talents, etc.), and $u_{i,t}$ is an unobserved error term. Correlation between the individual effect c_i and $\mathbf{x}_{i,t}$ causes the well known omitted variable bias problem. A common way to get rid of this problem is the so called *within* or *fixed effects* estimator. It is the pooled OLS estimator from the regression of the time-demeaned $w_{i,t}$ on the equally transformed $\mathbf{x}_{i,t}$. If a balanced panel is available, and for N relatively large compared to T, the conditional mean independence assumption,

A. 1 $E(u_{i,t} \mid \mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,T}, c_i) = 0, \qquad t = 1, 2, ..., T,$

is a sufficient condition for the within-estimator to be consistent as T is constant and $N \to \infty$. Assumption A. 1 also states that the $\mathbf{x}_{i,t}$ are *strictly exogenous* conditional on c_i, which is another way of expressing that the disturbance term $u_{i,t}$ is uncorrelated with the explanatory variables in each time period ($E(\mathbf{x}'_{i,s}u_{i,t}) = \mathbf{0}$, $s \neq t$, and $s, t = 1, 2, ..., T$). Under the standard rank condition that rank $(E(\tilde{\mathbf{X}}'_i\tilde{\mathbf{X}}_i)) = K$ the within estimator is defined as:

$$\hat{\boldsymbol{\beta}}_{within} = (\sum_{i=1}^{N} \tilde{\mathbf{X}}'_i\tilde{\mathbf{X}}_i)^{-1}(\sum_{i=1}^{N} \tilde{\mathbf{X}}'_i\tilde{\mathbf{w}}_i) = (\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{\mathbf{x}}'_{i,t}\tilde{\mathbf{x}}_{i,t})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T} \tilde{\mathbf{x}}'_{i,t}\tilde{w}_{i,t}), \tag{6}$$

where $\tilde{\mathbf{X}}_i = \mathbf{J}\mathbf{X}_i$, $\tilde{\mathbf{w}}_i = \mathbf{J}\mathbf{w}_i$ ($\mathbf{w}_i$ is $T \times 1$, $\mathbf{X}_i$ is $T \times K$, and $\mathbf{J} = \mathbf{I}_T - \mathbf{i}_T(\mathbf{i}'_T\mathbf{i}_T)^{-1}\mathbf{i}'_T$), and $\tilde{\mathbf{x}}_{i,t} = \mathbf{x}_{i,t} - T^{-1}\sum_{z=1}^{T}\mathbf{x}_{i,z}$, $\tilde{w}_{i,t} = w_{i,t} - T^{-1}\sum_{z=1}^{T} w_{i,z}$.

3.1 Panel attrition under conditional mean independence assumption

If a complete panel is available, estimation of equation (6) is straightforward. However, in the GSOEP the number of observations differ over years, i.e. not

all relevant variables are observed for each person and each time period under consideration. In the study at hand, two causes for missing observations can be distinguished: 1) individuals are not willing to report information with respect to one of the explanatory variables or the dependent variable (item non-response); 2) individuals endogenously decide to participate in the labour market (self-selection). Under these circumstances the conditional mean independence assumption A. 1 becomes:

A. 2 $E(u_{i,t} \mid \mathbf{x}_i, \mathbf{s}_i, \mathbf{d}_i, c_i) = 0, \qquad t = 1, 2, ..., T,$

where $\mathbf{x}_i = (\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, ..., \mathbf{x}_{i,T})$; $\mathbf{s}_i = (s_{i,1}, s_{i,2}, ..., s_{i,T})$ are selection dummies denoting whether an individual i is participating in the labour market at time t, and $\mathbf{d}_i = (d_{i,1}, d_{i,2}, ..., d_{i,T})$ are binary variables indicating item non-response. A. 2 is valid if the $(\mathbf{s}_i, \mathbf{d}_i)$ are strictly exogenous conditional on c_i and $\mathbf{x}_i$. Assumption A. 2 allows $(\mathbf{s}_i, \mathbf{d}_i)$ to be correlated with c_i or $\mathbf{x}_i$. That is, for the within-estimator to be consistent, it is not necessary that selection into or out of the data set is completely random.

Under the further condition that $\sum_{t=1}^{T} E(s_{i,t} d_{i,t} \tilde{\mathbf{x}}'_{i,t} \tilde{\mathbf{x}}_{i,t})$ is non-singular, pooled OLS on the unbalanced panel yields the following parameter vector:

$$\hat{\boldsymbol{\beta}}_{within} = (\sum_{i=1}^{N} \sum_{t=1}^{T} s_{i,t} d_{i,t} \tilde{\mathbf{x}}'_{i,t} \tilde{\mathbf{x}}_{i,t})^{-1} (\sum_{i=1}^{N} \sum_{t=1}^{T} s_{i,t} d_{i,t} \tilde{\mathbf{x}}'_{i,t} \tilde{w}_{i,t}), \tag{7}$$

where $\tilde{\mathbf{x}}_{i,t} = \mathbf{x}_{i,t} - T_i^{-1} \sum_{z=1}^{T} s_{i,z} d_{i,z} \mathbf{x}_{i,z}$, $\tilde{\mathbf{w}}_{i,t} = \mathbf{w}_{i,t} - T_i^{-1} \sum_{z=1}^{T} s_{i,z} d_{i,z} \mathbf{w}_{i,z}$, and $T_i = \sum_{z=1}^{T} s_{i,t} d_{i,t}$.

3.2 Selection correction in unobserved effects models

The within estimator of section 3.1 is a reasonable approach when we can be sure that condition A. 2 holds. If the decision to participate in the labour market $\mathbf{s}_i$ is, however, correlated with $u_{i,t}$, the estimator in (7) is inconsistent. That means, the participation decision is neither randomly determined nor fully covered by some of the observable variables.

In the paper at hand, I consider health as an determinant of wages and labour supply, and I am interested in making statements about the impact of health on wages for *all* individuals. Sample selection arises if some unobservable components of the working decision also affect wages. In this respect, one could think of the genetic endowment and the life situation of an individual (e.g. alcohol and nicotine consume, (un)healthy lifestyle, sport activities,

etc.). It is a natural assumption that genetic conditions are time-invariant, whereas the personal life situation is likely to change in the course of time. Consequently, for the former, the relationship between the selection process and wages can be completely described by an individual specific fixed effect. The later, on the other hand, is time-variant and for this reason not covered by c_i. As a result, the selection effect of an individual's life situation is influencing wages through the error term $u_{i,t}$. Since these factors are also correlated with health as an explanatory variable in the wage equation, the failure to control for the selection process may lead to inconsistent estimation.

To overcome the selection problem, the following model is estimated:

$$w_{i,t} = \beta_0 + \mathbf{x}_{i,t}\boldsymbol{\beta} + c_i + u_{i,t}, \quad t = 1, 2, ..., T; \quad i = 1, 2, ..., N, \tag{8}$$

$$s^*_{i,t} = \gamma_0 + k_i + \mathbf{z}_{i,t}\boldsymbol{\gamma} + e_{i,t}, \tag{9}$$

$$s_{i,t} = \begin{cases} 1 & \text{if } e_{i,t} > -\gamma_0 - \mathbf{z}_{i,t}\boldsymbol{\gamma} - k_i \\ 0 & \text{otherwise,} \end{cases} \tag{10}$$

where (8) equals (5), (9) and (10) describe a person's decision to participate in the labour market, $s^*_{i,t}$ is the latent propensity to work, $\mathbf{z}_{i,t}$ is a $1 \times G$ vector of covariates, and $\boldsymbol{\gamma}$ is the corresponding parameter vector ($G \times 1$). The variable $w_{i,t}$ is only observed when $s_{i,t} = d_{i,t} = 1$, and the $(\mathbf{z}_{i,t}, s_{i,t})$ are observable for $d_{i,t} = 1$.[5] It is usually assumed that $G > K$, meaning that $\mathbf{z}_{i,t}$ includes at least one exogenous variable that identifies selection. The individual effect k_i contains unobserved characteristics and exhibits no variation over time. Furthermore, $e_{i,t}$, which is normally distributed with standard deviation σ^e_t, is uncorrelated with k_i, $\mathbf{z}_i = (\mathbf{z}_{i,1}, ..., \mathbf{z}_{i,T})$, and $\mathbf{d}_i = (\mathbf{d}_{i,1}, ..., \mathbf{d}_{i,T})$. Following Mundlak (1978), Chamberlain (1984), and Wooldridge (1995) the time-invariant effects are assumed to be linked with $\mathbf{z}_{i,t}$ through a linear function of k_i on the time averages of $\mathbf{z}_{i,t}$ (denoted as $\bar{\mathbf{z}}_i$) and an error term a_i, that is independent of $(\mathbf{z}_i, \mathbf{d}_i)$ and $e_{i,t}$. Equation (9) therefore becomes:

$$s^*_{i,t} = \gamma_0 + \psi_0 + \bar{\mathbf{z}}_i\boldsymbol{\psi} + \mathbf{z}_{i,t}\boldsymbol{\gamma} + a_i + e_{i,t} = \theta_0 + \bar{\mathbf{z}}_i\boldsymbol{\theta} + \mathbf{z}_{i,t}\boldsymbol{\gamma} + v_{i,t}, \tag{11}$$

where $\theta_0 = \gamma_0 + \psi_0$; $\boldsymbol{\theta} = \boldsymbol{\psi}$, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ are $G \times 1$ parameter vectors, and a_i is zero mean normally distributed. The distribution of the composite error term

[5] In the case of item non-response ($d_{i,t} = 0$), the corresponding observation is missing in both the selection and the main equation.

$v_{i,t} = a_i + e_{i,t}$ is normal with standard deviation $\sigma_t^v = \sigma^a + \sigma_t^e$. It is allowed to be heterogeneously distributed over time and there are no restrictions imposed on the correlation between $v_{i,t}$ and $v_{i,s}$, i.e. $Cov(v_{i,t}, v_{i,s}) \neq 0$ for $s \neq t$.

Implicitly, assumptions on the selection equations (9) and (11) were already mentioned in the above, but I summarise them in the following (see also Wooldridge (1995), p.126, and Dustmann and Rochina-Barrachina (2000), p.6):

A. 3 *The unobserved effect in the selection equation can be described as a linear projection of* k_i *on* $\bar{\mathbf{z}}_i$*, where* $\bar{\mathbf{z}}_i = P_i^{-1} \sum_{s=1}^{T} d_{i,s}\mathbf{z}_{i,s}$*, and* $P_i^{-1} = \sum_{s=1}^{T} d_{i,s}$*.*

A. 4 *The errors* $v_{i,t} = a_i + e_{i,t}$ *are independent of* $(\mathbf{z}_i, \mathbf{d}_i)$ *and they are normally distributed,* $N(0, \sigma_t^v)$*.*

The next step is to estimate equation (11) using standard probit for each t and obtain the inverse Mills ratios (IMRs) for $s_{i,t} = d_{i,t} = 1$ as $\hat{\lambda}_{i,t} = \phi(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t)/\Phi(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t)$, where $\mathbf{h}_{i,t} = (1, \bar{z}_{i,1}, ..., \bar{z}_{i,G}, z_{i,1,t}, ..., z_{i,G,t})$ and $\hat{\boldsymbol{\delta}}_t = (\hat{\theta_{0,t}}, \hat{\theta_{1,t}}, ..., \hat{\theta_{G,t}}, \hat{\gamma_{1,t}}, ..., \hat{\gamma_{G,t}})'$. At this point, it seems tempting to include the IMRs as additional regressors and to estimate equation (8) using the within-estimator described in (7). However, as Wooldridge (2002) points out, this is (usually) not a valid strategy to arrive at consistent estimates.[6] Instead, he suggests a method that allows the selection term $\hat{\lambda}_{i,t}$ to be *not* strictly exogenous in (8) (i.e there are no restrictions on how $u_{i,t}$ relates to $v_{i,s}$, $s \neq t$).[7] This strategy necessitates to specifically model the unobserved effect such that correlation between c_i and $(\mathbf{x}_i, v_{i,t})$ is possible. Explicitly, the assumptions are:

A. 5 $E(u_{i,t} \mid \mathbf{z}_i, \mathbf{d}_i, v_{i,t}) = E(u_{i,t} \mid v_{i,t}) = L(u_{i,t} \mid v_{i,t}) = \rho_t v_{i,t}$,

i.e. $u_{i,t}$ is mean independent of $(\mathbf{z}_i, \mathbf{d}_i)$ conditional on $v_{i,t}$ and the conditional mean of $u_{i,t}$ is a linear function of $v_{i,t}$.

A. 6 $E(c_i \mid \mathbf{z}_i, \mathbf{d}_i, v_{i,t}) = L(c_i \mid 1, \bar{x}_{i,1}, ..., \bar{x}_{i,K}, v_{i,t}) = \tau_0 + \tau_1\bar{x}_{i,1} + ... + \tau_k\bar{x}_{i,K} + \varsigma_t v_{i,t}$,

[6] It is, however, possible to use the Within estimator for testing purposes. Under the null hypothesis in A. 2, the IMRs should not be significant when using the within-estimator on an augmented version of equation (4). See also section 5.

[7] To place more emphasis on this, without abandoning the strict exogeneity assumption for the IMR at this point it is not possible to allow for serial correlation in the selection equation.

i.e. the unobserved effect in the main equation can be described as a linear projection of c_i on $(\bar{\mathbf{x}}_i,\ v_{i,t})$ and an error term b_i, where $\bar{\mathbf{x}}_i = (\bar{x}_{i,1}, ..., \bar{x}_{i,K})$, and the conditional expectation of b_i is independent of $(\mathbf{z}_i, \mathbf{d}_i)$ and $v_{i,t}$ ($E(b_i \mid \mathbf{z}_i, \mathbf{d}_i, v_{i,t}) = 0$).

At this point, it seems necessary to spend a few words on the item non-response indicators $\mathbf{d}_i$ in A. 4, A. 5, and A. 6. In the case where item non-response is entirely random, $\mathbf{d}_i$ is independent of $(\mathbf{u}_i, \mathbf{s}_i, \mathbf{z}_i, c_i, k_i)$. Hence, $\mathbf{d}_i$ is independent of $\mathbf{v}_i$ in A. 4 and assumptions A. 5 and A. 6 hold under $E(u_{i,t} \mid \mathbf{z}_i, v_{i,t}) = E(u_{i,t} \mid v_{i,t}) = \rho_t v_{i,t}$ and $E(c_i \mid \mathbf{z}_i, v_{i,t}) = L(c_i \mid 1, \bar{\mathbf{x}}_i, v_{i,t}) = \tau_0 + \bar{\mathbf{x}}_i\boldsymbol{\tau} + \varsigma_t v_{i,t}$. However, the assumption of complete randomness is stronger than actually needed. If there is item non-response, the corresponding observation is missing both in the selection and in the main equation. So, one needs to assume that $\mathbf{d}_i$ is independent of the error term $v_{i,t}$ in the participation equation, and conditional mean independent of $u_{i,t}$. Nevertheless, $\mathbf{d}_i$ is still allowed to be correlated with $(\mathbf{z}_i, k_i)$. Since $v_{i,t}$ is a determinant of c_i (see A. 6) $\mathbf{d}_i$ needs to be uncorrelated with the unobserved effect in the main equation.[8]

The conditional expectation for $w_{i,t}$ can, then, be expressed as:

$$\begin{aligned} E(w_{i,t} \mid \mathbf{z}_i, \mathbf{d}_i, v_{i,t}) &= E(w_{i,t} \mid \mathbf{z}_i, v_{i,t}) \\ &= E(c_i \mid \mathbf{z}_i, v_{i,t}) + \beta_0 + \mathbf{x}_{i,t}\boldsymbol{\beta} + E(u_{i,t} \mid \mathbf{z}_i, v_{i,t}) \\ &= (\beta_0 + \tau_0) + \bar{\mathbf{x}}_i\boldsymbol{\tau} + \mathbf{x}_{i,t}\boldsymbol{\beta} + (\varsigma_t + \rho_t)v_{i,t} \\ &= \varphi_0 + \bar{\mathbf{x}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t v_{i,t}. \end{aligned} \tag{12}$$

Here, the first and second equality hold under the assumption that item non-response is entirely random (see above); $\varphi_0 = \beta_0 + \tau_0$, $\boldsymbol{\varphi} = \boldsymbol{\tau}$, $\boldsymbol{\varphi}$ and $\boldsymbol{\tau}$ are $K \times 1$ parameter vectors, and $\xi_t = \varsigma_t + \rho_t$.[9] Using the law of iterated expectations on

[8] One should be aware of the fact that the random item non-response assumption might be doubted if persons are not willing or able to reply to the GSOEP due to their poor health status. Unfortunately, it is not possible to control for this eventuality and the random drop-out assumption needs to be maintained at this point.

[9] With the exception of the constant term, identifying the vector $\boldsymbol{\beta}$ can easily be achieved since by the law of iterated expectations:

$$\begin{aligned} E(c_i \mid \mathbf{z}_i, \mathbf{d}_i) &= \tau_{0,t} + \bar{\mathbf{x}}_i\boldsymbol{\tau}_t + \rho_t E(v_{i,t} \mid \mathbf{z}_i, \mathbf{d}_i) \\ &= \tau_{0,t} + \bar{\mathbf{x}}_i\boldsymbol{\tau}_t = \tau_0 + \bar{\mathbf{x}}_i\boldsymbol{\tau}. \end{aligned}$$

The second equality holds because $E(v_{i,t} \mid \mathbf{z}_i, \mathbf{d}_i) = 0$ in assumption A. 4 and the third equality follows due to the fact that the coefficients describing the time constant effects are necessarily time-invariant. If variables are not changing over time it is impossible to distinguish β_k and φ_k. Furthermore, there is now way to determine how much of the

equation (12) yields:

$$\begin{aligned} E(w_{i,t} \mid \mathbf{z}_i, s_{i,t}) &= \varphi_0 + \bar{\mathbf{x}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t E(v_{i,t} \mid \mathbf{z}_i, s_{i,t}) \\ &= \varphi_0 + \bar{\mathbf{x}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t f(\mathbf{z}_i, s_{i,t}), \end{aligned} \tag{13}$$

where $f(\mathbf{z}_i, s_{i,t})$ is a function of $\mathbf{z}_i$ and $s_{i,t}$. Since in the selected sample $w_{i,t}$ is only observable for $s_{i,t} = 1$, $f(.)$ can be replaced by $f(\mathbf{z}_i, s_{i,t} = 1) = f(\mathbf{z}_i, v_{i,t} > -\mathbf{h}_{i,t}\boldsymbol{\delta}_t) = \phi(\mathbf{h}_{i,t}\boldsymbol{\delta}_t)/\Phi(\mathbf{h}_{i,t}\boldsymbol{\delta}_t) = \lambda_{i,t}$.

As mentioned before, the crucial point is that $v_{i,s}$, for $s \neq t$, is not in the conditioning set of A. 5 and so Wooldridge's estimator allows for serial correlation and heterogeneity in the error terms of the selection equation. Stated differently, $s_{i,s}$, for $s \neq t$, is not in $E(v_{i,t} \mid \mathbf{z}_i, s_{i,t})$ and so the error term $r_{i,t}$ in

$$\begin{aligned} w_{i,t} &= \varphi_0 + \bar{\mathbf{x}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t\lambda_{i,t} + (b_i + l_{i,t}) \\ &= \varphi_0 + \bar{\mathbf{x}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t\lambda_{i,t} + r_{i,t} \end{aligned} \tag{14}$$

is allowed to be correlated with $\lambda_{i,s}$, for $s \neq t$, where $l_{i,t}$ is part of the composite error term $u_{i,t} = \rho_t v_{i,t} + l_{i,t}$ and b_i is defined as above. Dustmann and Rochina-Barrachina (2000) call the condition $E(r_{i,t} \mid \mathbf{z}_i, s_{i,t}) = 0$ "contemporaneous exogeneity" of the selection term with respect to $r_{i,t}$.

The simplest way to consistently estimate (14) (with $\lambda_{i,t}$ replaced by $\hat{\lambda}_{i,t}$) if strict exogeneity (with respect to the IMRs) fails is pooled OLS. When calculating the asymptotic variance of $(\varphi_0, \boldsymbol{\varphi}', \boldsymbol{\beta}', \boldsymbol{\xi}')'$, I follow Wooldridge (1995) and construct standard errors robust to serial correlation and heteroscedasticity that are also adjusted for the additional variation introduced by the estimation of T probit models in the first step. The calculation of the asymptotic variance covariance estimator is described in the appendix.

3.3 Panel attrition with endogenous regressors

Estimation of equation (14) assumes (strict) exogeneity of the explanatory variables. However, in the paper at hand – even after controlling for individual specific heterogeneity and sample selection – the health variable is likely to be endogenous. Three cases of endogenity may appear. 1) Since health satisfaction is a self-assessed variable, measurement error might pose a problem;

selection process works through c_i and how much through the time varying unobserved factors in $u_{i,t}$.

2) the health condition may benefit from rising wages as the marginal return of health investment increases with the wage rate (reverse causality); 3) if past shocks affect current health, the health variable is not strictly exogenous in the wage equation.

Semykina and Wooldridge (2005) provide an estimation method based on Wooldridge (1995) that accounts for endogeneity in the presence of unobserved heterogeneity and sample selection. Analogous to section 3.1, it seems reasonable to start with a mean independence assumption that allows for consistent estimates in an unbalanced panel framework, when some of the explanatory variables are endogenous. Presume that the health variable (as part of $x_{i,t}$ in equation (5)) is correlated with $u_{i,t}$. Furthermore, suppose that a vector of instruments $\mathbf{q}_{i,t}$ $(1 \times E)$ is available, which consists of all exogenous variables in $\mathbf{x}_{i,t}$ and at least one instrument.[10] Then, for the Within- or FE-2SLS (two step least square) estimator in an unbalanced panel framework to be consistent, the equivalent to A. 2 is:

A. 7 $E(u_{i,t} \mid \mathbf{q}_i, \mathbf{s}_i, \mathbf{d}_i, c_i) = 0, \qquad t = 1, 2, ..., T,$

where $\mathbf{q}_i = (\mathbf{q}_{i,1}, \mathbf{q}_{i,2}, ..., \mathbf{q}_{i,T})$, $\mathbf{q}_{i,t} = (q_{i,t,1}, ..., q_{i,t,E})$, and the $(\mathbf{s}_i,\mathbf{d}_i)$ are defined as in section 3.1. A. 7 requires sample attrition ($\mathbf{s}_i$, $\mathbf{d}_i$) and the vector of instruments $\mathbf{q}_i$ to be strictly exogenous conditional on c_i. Moreover, all variables in $\mathbf{q}_i$ are assumed to vary over time, $\mathbf{q}_i$ is allowed to be correlated with c_i, and the $(\mathbf{s}_i,\mathbf{d}_i)$ are either completely random or a function of $(\mathbf{q}_i, c_i)$. If there are no linear dependencies among the demeaned $\mathbf{q}_{i,t}$ (rank $E(\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{q}}'_{i,t}\tilde{\mathbf{q}}_{i,t}) = E$, $\tilde{\mathbf{q}}_{i,t} = \mathbf{q}_{i,t} - T_i^{-1}\sum_{z=1}^{T} s_{i,z}d_{i,z}\mathbf{q}_{i,z}$, and $T_i = \sum_{z=1}^{T} s_{i,z}d_{i,z}$) and if rank $E(\sum_{t=1}^{T} s_{i,z}d_{i,t}\tilde{\mathbf{x}}'_{i,t}\tilde{\mathbf{q}}_{i,t}) = K$ (i.e. the instruments are partially correlated with the endogenous variables conditional on the exogenous part of $\mathbf{x}_{i,t}$) the FE-2SLS estimator is given by:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{FE-2SLS} \;=\; & [(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{x}}'_{i,t}\tilde{\mathbf{q}}_{i,t})'(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{q}}'_{i,t}\tilde{\mathbf{q}}_{i,t})^{-1} \\ & (\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{q}}'_{i,t}\tilde{\mathbf{x}}_{i,t})]^{-1} \times (\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{x}}'_{i,t}\tilde{\mathbf{q}}_{i,t})' \\ & (\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{q}}'_{i,t}\tilde{\mathbf{q}}_{i,t})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\tilde{\mathbf{q}}'_{i,t}\tilde{w}_{i,t}). \qquad (15)\end{aligned}$$

[10] It is assumed that unemployment, experience, and years of schooling in equation (4) are strictly exogenous conditional in c_i.

As in section 3.2, it is assumed that item non-response occurs random. That means, condition A. 7 alters to:

A. 8 $E(u_{i,t} \mid \mathbf{q}_i, \mathbf{s}_i, \mathbf{d}_i, c_i) = E(u_{i,t} \mid \mathbf{q}_i, \mathbf{s}_i, c_i) = 0, \qquad t = 1, 2, ..., T.$

3.4 Selection correction in unobserved effects models with endogeneity

The final step is to derive an estimator that allows $v_{i,t}$ in (11) to be correlated with $u_{i,t}$ and c_i in (8), when the health variable is endogenous (meaning that $E(r_{i,t} \mid \mathbf{x}_i, s_{i,t}) \neq 0$ in equation (14)).

Consider a model that consists of the main equation (8) and a selection process that occurs according to the following equation:

$$s^*_{i,t} = \gamma_0 + k_i + \mathbf{q}_{i,t}\boldsymbol{\gamma} + e_{i,t}; \qquad s_{i,t} = 1[s^*_{i,t} > 0], \tag{16}$$

where 1[.] is an indicator function that equals one if its argument is true, and zero otherwise. Again, the selection equation rests on assumptions A. 3 and A. 4, except that now the $1 \times G$ vector $\bar{\mathbf{z}}_i$ and the $1 \times TG$ vector $\mathbf{z}_i$ are replaced by the $1 \times E$ vector $\bar{\mathbf{q}}_i$ and the $1 \times TE$ vector $\mathbf{q}_i$. Under these assumptions, equation (16) becomes:

$$s^*_{i,t} = \gamma_0 + \psi_0 + \bar{\mathbf{q}}_i\boldsymbol{\psi} + \mathbf{q}_{i,t}\boldsymbol{\gamma} + a_i + e_{i,t} = \theta_0 + \bar{\mathbf{q}}_i\boldsymbol{\theta} + \mathbf{q}_{i,t}\boldsymbol{\gamma} + v_{i,t}. \tag{17}$$

Likewise, A. 5 and A. 6 are imposed on the relationship between the selection process and ($u_{i,t}$, c_i), where the vector $\mathbf{z}_i$ is replaced by $\mathbf{q}_i$ and $\bar{x}_{i,j}$, $j = 1, ..., K$, is now $\bar{q}_{i,p}$, $p = 1, ..., E$. Then, the conditional expectation in (12) can be rewritten as:

$$E(w_{i,t} \mid \mathbf{q}_i, v_{i,t}) \;\; = \;\; \varphi_0 + \bar{\mathbf{q}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t v_{i,t}, \tag{18}$$

where $\xi_t = (\varsigma_t + \rho_t)$. Using the law of iterated expectations on (18) and plugging into (8) yields:

$$\begin{aligned} w_{i,t} &= \varphi_0 + \bar{\mathbf{q}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t E(v_{i,t} \mid \mathbf{q}_i, s_{i,t}) + r_{i,t}, \\ &= \varphi_0 + \bar{\mathbf{q}}_i\boldsymbol{\varphi} + \mathbf{x}_{i,t}\boldsymbol{\beta} + \xi_t f(\mathbf{q}_i, s_{i,t}) + r_{i,t}. \end{aligned} \tag{19}$$

Again, the first step is to estimate T standard probit models of equation (17), and calculate the IMRs $\hat{\lambda}_{i,t}$. Then, because the selected sample has $s_{i,t} = 1$, $f(\mathbf{q}_i, s_{i,t})$ in equation (19) can be expressed as $f(\mathbf{q}_i, s_{i,t} = 1) = f_t(\mathbf{q}_i, v_{i,t} >$

$-\mathbf{h}_{i,t}\boldsymbol{\delta}_t) = \phi(\mathbf{h}_{i,t}\boldsymbol{\delta}_t)/\Phi(\mathbf{h}_{i,t}\boldsymbol{\delta}_t) = \lambda_{i,t}$, where $\mathbf{h}_{i,t} = (1, \bar{q}_{i,1}, ..., \bar{q}_{i,E}, q_{i,t,1}, ..., q_{i,t,E})$ and $\boldsymbol{\delta}_t$ is the corresponding parameter vector. Finally, since $r_{i,t}$ is allowed to be correlated with $\lambda_{i,s}$, for $s \neq t$, (i.e. $\lambda_{i,t}$ is not strictly exogenous in (19)), a consistent way of estimating (19) – with $f(\mathbf{q}_i, s_{i,t} = 1)$ replaced by $\hat{\lambda}_{i,t}$ – is pooled 2SLS, where $1, \bar{\mathbf{q}}_i, \mathbf{q}_{i,t}, \hat{\lambda}_{i,t}$ serve as instruments ($1, \hat{\lambda}_{i,t}$, and the exogenous variables in $\mathbf{x}_{i,t}$ are used as their "own" instruments).

To calculate the asymptotic variance of $(\hat{\varphi}_0, ..., \hat{\varphi}_E, \hat{\beta}_1, ..., \hat{\beta}_K, \hat{\xi}_1, ..., \hat{\xi}_T)'$, I follow Semykina and Wooldridge (2005) and construct standard errors robust to serial correlation and heteroscedasticity that are adjusted for the additional variation introduced by the estimation of T probit models in the first step. They also account for the use of the pooled 2SLS estimator. The estimation of the asymptotic variance covariance matrix is described in the appendix.

At last, it is important to describe, how many instruments are needed in the above procedure. As usual, the vector of instruments consists of all exogenous variables in $\mathbf{x}_{i,t}$ and at least as many instruments as there are endogenous variables. Moreover, for the purpose of clearly identifying the parameter vector in the main equation, at least one additional instrument is required. Thus, in the paper at hand a minimum of two instruments should be available.

4 Data and Descriptives

The data used in this analysis are made available by the German Socio-Economic Panel Study (GSOEP, see SOEP Group 2001) at the German Institute for Economic Research (DIW), Berlin. It is a representative panel data set of the German population that is drawn on a yearly basis. For the western German states, the GSOEP started with about 12,200 observations in 1984. In June 1990 another 4,400 persons living in the former territory of the German Democratic Republic were interviewed in order to expand the GSOEP to the eastern part of Germany.

For the empirical analysis, observations from all sub-samples, with the exception of sample G ("Oversampling of High Income") between 1995 and 2005, are selected.[11] I extract data on the variables described in tables 4 and 5 in the appendix and exclude (individual-year) observations from both the selection and the wage equation if there is missing data on any of these

[11] 1995 is chosen as starting point because the variable 'number of doctor visits' is not available in 1994.

variables except wages. The sample is constrained to persons older than 17 and younger than 66 years. I exclude those who are self-employed, self-employed in the agricultural sector, work in family business, are on maternity leave, as well as persons attending military/civilian service, and marginally or irregularly employed persons in any of the years under consideration. Individuals who serve an apprenticeship, trainees, interns, volunteers, aspirants, pensioners, and persons still in education are also removed from the estimation sample.

In this kind of study, it is important to discuss whether to include (severely) handicapped persons in the analysis. Motivated by two arguments, I decided to leave them out of the estimation sample. First, firms might discriminate against handicapped persons, irrespective of their productivity. Therefore, their wages might be artificially low or they might drop out of labour market due to discrimination – something that is not meant to be captured in the selection equation. Second, in Germany severely handicapped persons mainly work at special locations (Behindertenwerkstätten), where they are not paid according to their marginal productivity.

The dependent variable used in the main equation is hourly wages derived from individual gross earnings in the month before the interview divided by 4.3 and information on the actual working time per week. In case actual hours worked fall below contractual hours worked, hourly wages are constructed using the later. Any extra salaries like Christmas or holiday bonuses, 13th monthly pay, or child benefits are not taken into account. When calculating hourly wages suspiciously high or low values were manually overseen and dropped if necessary. Wages (as well as all other financial variables) are deflated to their year 2001 real values using the eastern and western CPIs and – if necessary – converted into euro figures at the rate of 1.95583 (the conversion rate of the Euro in 1999).[12]

Individuals are defined as participating in the labour market if they work for pay in the month before the interview. In the participation equations both working and non-working persons are used for estimation.[13] After the stepwise exclusion of different groups, I arrive at an estimation sample of $10,081$ female and $9,540$ male persons, resulting in $48,763$ and $48,536$ observations, respectively. For estimating earnings equations, persons who work for only one year are dropped from the sample. Due to this restriction and because observations

[12]For this purpose, Consumer Price Indices included in the $pequiv files of the GSOEP were used.

[13]I follow Dustmann and Rochina-Barrachina (2000) and include persons as working if they declare participation, but not wages (and if all explanatory variables are available).

with missing wages are included in the participation equation, the number of observations in the wage equations (29,304 and 39,048; see tables 11 and 10 in the appendix) differs slightly from the working population in the probit sample (30,689 and 40,399; see tables 9 and 8).

Tables 8 and 9 in the appendix compare variables in the participation equation for working and non-working individuals, and tables 10 and 11 depict summary statistics of the variables used in the earnings equations. The health variable, which reports current health satisfaction of individuals, is categorial, ranging from zero to ten. It is transformed using the following log-function: $f(h_{i,t}) = \log(h_{i,t} + \sqrt{(h_{i,t}^2 + 1)})$. Health satisfaction differs between the working and non-working population. On average, the transformed value for working females between 1995 and 2005 is around 2.583, while the value for non-working women is smaller at about 2.49 log points. For males, the working non-working health ratio is about 2.598 to 2.406. The hypothesis of the equality of means between the working and non-working group can be rejected on the basis of two standard t-test, t = 23.38 (p-value = 0) for females and t = 38.73 (p-value = 0) for males. In the time period considered, about 63% of the female sample and around 83% of the male sample population participate in the labour market. Men active in the labour market are on average 2.2 years younger, their school attendance was 1.1 years higher, their non-labour income is lower and they have more children than their non-working counterparts. At the same time, a lower portion of male labour market participants is single (21% vs. 32%), has a foreign nationality (12% vs. 19%), and less male workers live in the eastern part of Germany (24% vs. 37%). For women, in this respect the opposite holds true: A larger part of working females live in eastern Germany (28% vs 21%), a smaller portion is married/has a partner (75% vs. 85%), and they have less children than working females. Just like their male colleagues, female workers are slightly younger (2.75 years), spent more time in education or training (1.16 years), and have a lower non-labour income compared to the sample population of female non-workers. Finally, when comparing women and men it becomes clear that in the sample period male real hourly wages were on average about 0.22 log points higher than those of women.

5 Empirical Results

Equation (4) is estimated using six different estimation methods and tables 2 and 3 report the results. Pooled OLS in column (1) assumes that the explana-

tory variables are uncorrelated with individual heterogeneity. If an individual's genetic endowment affects health positively and if it is at the same time more likely to be in the labour market, then OLS estimates should be upward biased. The Within estimator in column (2) helps to overcome this problem as it allows for correlation between health satisfaction and unobserved heterogeneity. The upward bias should be even larger if (positively correlated) time-variant unobservables, determining wages and participation, cause a sample selection problem. An example in this respect is a person's individual life situation characterised by her/his alcohol and nicotine consume, sport activities, healthy lifestyle, etc.. Consequently, in column (3) Wooldridge's (1995) estimator is presented. It allows controlling for heterogeneity and selection in one common framework. However, as argued before, it is unlikely that the health variable is strictly exogenous in the wage equation. A solution to this problem is to use instrumental variable techniques. The pooled 2SLS estimator in column (4) assumes all exogenous variables and the instruments to be uncorrelated with the unobserved effects, whereas the FE-2SLS in column (5) additionally allows for correlated fixed effects. Finally, Semykina's and Wooldridge's (2005) estimator (column (6)) deals with heterogeneity, selection and endogeneity in one estimation approach.

The set of instruments consists of nine variables which also serve as exclusion restrictions in the participation equations: non-labour income, a binary variable for having a partner/being married, age of the partner/spouse and its square, labour market experience of the partner/spouse and its square, and education (in years) of the partner/spouse and its square.[14] Furthermore, I use two extra instruments that are not included in the selection equation. 1) the number of doctor visits in the last three months; 2) the number of days absent from work due to illness in the last year. At this juncture, the argument is that both variables approximate past investment (and depreciation) in health and account for past shocks affecting current health satisfaction. To check the rank conditions on the 2SLS estimators, F-tests on the joint-significance of the instruments in the first step regressions are conducted. For both women and men and for all econometric models the null hypotheses are rejected at any sensible level.

[14] Wooldridge (2002) suggests to add all exogenous variables, which appear in the selection equation, to the list of instruments. He argues that it can be 'dangerous' to introduce any exclusion restrictions up on reduced form equations. However, based on the prior information that some authors find a direct relationship between the number of children and wages, I do not use these kind of variables as instruments, though they are excluded from the earnings equations and included in the participation equations.

The presumption that a selection bias exists is testable. In table 1 a number of Wald tests on the joint significance of 11 Inverse Mills Ratios, each one constructed using a separate probit, are provided. In columns (1) and (2) I follow Wooldridge (1995) and conduct so called 'variable addition' tests, that were first proposed by Verbeek and Nijman (1992). It is assumed that no further endogeneity problems occur. Under the null hypothesis the Within estimator in section 3.1 is valid. In columns (3) and (4) tests in the spirit of Semykina and Wooldridge (2005) are accomplished. The null hypothesis suggests to use the FE-2SLS estimator of section 3.3. Both procedures (as well as all further estimates) are done separately for females and males to account for expected gender differences in wage determination.

Table 1: IMR Tests, Women and Men, 1995-2005

	male FE[a)]	female FE[a)]	male FE-2SLS[b)]	female FE-2SLS[b)]
	(1)	(2)	(3)	(4)
IMR 1995	.046	.011	.051	.019
	(.030)	(.022)	(.031)*	(.023)
IMR 1996	-.007	.011	-.005	.017
	(.021)	(.019)	(.021)	(.019)
IMR 1997	-.014	-.002	-.011	.004
	(.031)	(.019)	(.030)	(.020)
IMR 1998	.009	-.002	.012	.003
	(.024)	(.023)	(.023)	(.023)
IMR 1999	-.031	-.0003	-.027	.003
	(.018)*	(.020)	(.018)	(.020)
IMR 2000	-.049	-.020	-.045	-.016
	(.016)***	(.017)	(.016)***	(.017)
IMR 2001	-.058	-.034	-.052	-.031
	(.016)***	(.016)**	(.017)***	(.016)*
IMR 2002	-.020	-.023	-.014	-.020
	(.019)	(.017)	(.019)	(.017)
IMR 2003	-.038	-.029	-.032	-.024
	(.016)**	(.019)	(.016)**	(.019)
IMR 2004	-.047	-.046	-.042	-.041
	(.017)***	(.020)**	(.017)**	(.020)**
IMR 2005	-.043	-.052	-.035	-.048
	(.018)**	(.021)**	(.019)*	(.021)**
Wald-test, χ^2_{11} =	31.64	12.65	26.96	12.42
p-values	.001	.317	.005	.333
N	39,048	29,304	39,048	29,304

Source: GSOEP 1995-2005, own calculations. Within and FE-2SLS estimation. Robust standard errors are in parenthesis: * significance at ten, ** at five, and *** at one percent. Robust p-values are reported under the test statistics. *a)* Wald tests on the joint significance of the IMRs are provided. It is assumed that there are no further endogeneity problems. Under the null hypothesis the Within estimator in section 3.1 is valid. *b)* Wald test on the joint significance of the IMRs are provided. Under the null hypothesis the FE-2SLS estimator in section 3.3 is valid.

As it turns out, for both women and men, the inverse Mills ratios are negatively correlated with wages for most years. Since the IMRs are inversely related to the estimated probabilities of being employed, derived from the first step probit equations, the negative coefficients indicate that a higher participation probability is associated with an above average salary. For men, the test procedures provide evidence on a selection bias in both the Within and the FE-2SLS framework. The χ^2 statistics, with 11 degrees of freedom, are 31.64 and 26.96, respectively, which gives p-values of about 0.001 and 0.005. Interestingly, for women a selection correction is not indicated. The χ^2 statistics for females are 12.65 and 12.42, resulting in p-values of 0.317 and 0.312. Thus, for women the null hypothesis of no selection bias can not be rejected, a result also found in Dustmann and Rochina-Barrachina (2000). This suggests that in the female sample the selection process is already accounted for by the observable variables and the latent effect c_i. No further evidence is found that any unobservable characteristics in the participation equation affect wages through the error term of the main equation.

For males (table 2), the parameter of the health variable using pooled OLS (0.043) is higher than the coefficient in the fixed effects model (0.012). Wooldridge's (1995) estimator, in turn, exhibits the lowest coefficient (0.009) under the assumption of no further endogeneity. All estimates are significant at the 1%-5% confidence level. These results suggest that using the FE estimator already accounts for most of the upward bias introduced by the correlation between the health variable and unobserved individual heterogeneity. Controlling for selection reduces the coefficient even further, but differences between the FE and the Wooldridge (1995) estimator are small. Turning to the 2SLS models, a comparison of the parameters shows that the coefficients of health satisfaction in columns (1), (2), and (3) are smaller than those in columns (4), (5), and (6), which is expected if there exists a measurement error problem. Yet, within this framework, the parameter ranking follows the same pattern as in the specifications without instruments. The pooled 2SLS parameter exhibits the highest (significant) parameter (0.088). Using the FE-2SLS estimator reduces the coefficient to a value of 0.071. Though insignificantly different from zero, controlling for selection scales the coefficient even further down to 0.013. For the estimators in columns (3) and (6) a Wald test on the joint significance of the $\boldsymbol{\varphi}$ was accomplished. In both cases the resulting values of the test statistics are larger than the critical value, indicating correlated individual effects. Selection tests, where now the assumptions under the null hypothesis are more restrictive than those underlying the tests in table 1, exhibit χ^2 statistics

Table 2: Wage equations, Men, 1995-2005

	OLS[a)]	Within[a)]	Wooldr95[b)]	2SLS[a)]	FE-2SLS[a)]	Wooldr05[c)]
	(1)	(2)	(3)	(4)	(5)	(6)
Health sat.	0.043 (0.004)***	0.012 (0.004)***	0.009 (0.004)**	0.088 (0.012)***	0.071 (0.017)***	0.013 (0.023)
Age	0.097 (0.006)***	.	.	0.097 (0.006)***	.	.
Age square	-.002 (0.0001)***	-.002 (0.0002)***	0.0005 (0.00005)***	-.002 (0.0001)***	-.002 (0.0002)***	0.0005 (0.00006)***
Age triple	1.00e-05 (1.18e-06)***	1.00e-05 (1.62e-06)***	-5.83e-06 (5.77e-07)***	1.00e-05 (1.11e-06)***	1.00e-05 (1.63e-06)***	-5.82e-06 (7.00e-07)***
Unempl. exp.	-.048 (0.003)***	-.097 (0.011)***	-.074 (0.014)***	-.047 (0.003)***	-.098 (0.011)***	-.075 (0.014)***
Unempl. exp. sq.	0.003 (0.0003)***	0.004 (0.002)***	0.003 (0.002)*	0.003 (0.0003)***	0.005 (0.002)***	0.003 (0.002)*
Firm tenure	0.013 (0.0005)***	0.004 (0.0007)***	0.007 (0.0009)***	0.013 (0.0005)***	0.005 (0.0007)***	0.007 (0.0009)***
Firm tenure sq.	-.0002 (1.00e-05)***	-.0001 (0.00002)***	-.0002 (0.00002)***	-.0002 (1.00e-05)***	-.0001 (0.00002)***	-.0002 (0.00003)***
Education	0.032 (0.0008)***	.	.	0.032 (0.0007)***	.	.
Dummy Educ.	-.020 (0.004)***	.	.	-.019 (0.004)***	.	.
Part Time	-.103 (0.015)***	-.042 (0.017)**	-.035 (0.021)*	-.100 (0.011)***	-.041 (0.017)**	-.036 (0.021)*
Foreigner	0.01 (0.005)**	.	.	0.009 (0.005)*	.	.
Lg. unempl. (fed. st.)	-.046 (0.004)***	0.021 (0.013)	0.029 (0.018)	-.045 (0.004)***	0.022 (0.013)*	0.03 (0.02)
Lg. vac. (fed. st.)	0.058 (0.004)***	0.01 (0.007)	-.001 (0.01)	0.057 (0.004)***	0.009 (0.007)	-.002 (0.011)
Firm size (<20 employees)[d)]						
20 - 199	0.082 (0.005)***	0.046 (0.006)***	0.034 (0.007)***	0.081 (0.004)***	0.046 (0.006)***	0.034 (0.007)***
200 - 1999	0.147 (0.005)***	0.058 (0.007)***	0.044 (0.009)***	0.146 (0.005)***	0.058 (0.007)***	0.045 (0.009)***
$\geq$ 2000 employees	0.191 (0.005)***	0.067 (0.008)***	0.051 (0.01)***	0.191 (0.005)***	0.067 (0.008)***	0.051 (0.01)***
Firm size missing	0.085 (0.018)***	0.022 (0.016)	0.024 (0.019)	0.083 (0.015)***	0.021 (0.016)	0.026 (0.019)
Region, where person works (Western Germany)						
East Germany	-.262 (0.006)***	-.032 (0.01)***	-.242 (0.011)***	-.262 (0.005)***	-.033 (0.01)***	-.238 (0.011)***
constant	0.361 (0.089)***	.	.	0.235 (0.089)***	.	.
N	39,048	39,048	39,048	39,048	39,048	39,048
d.f.	39,003	32,035	38,980	39,003	32,035	38,970
Wald tests on the joint significance of						
11 IMRs[e)]	.	.	29.62***	.	.	22.04**
10 time dummies	308.26***	265.76***	42.80***	325.46***	262.63***	40.82***
6 occup. dummies	2802.72***	13.26**	750.72***	3063.53***	13,50**	703.23***
9 sector dummies	1301.85***	64.62***	380.13***	1310.33***	64,86***	381.91***
unobs. effects [f)]	.	.	719.17***	.	.	437,86***

Source: GSOEP 1995-2005, own calculations. Standard errors in parenthesis: * significance at ten, ** at five, and *** at one percent. Year, sector, and occupation dummies are included but not reported. *a)* Robust standard errors are provided using the Huber/White/sandwich estimator; *b)* standard errors are robust to serial correlation and heteroscedasticity. They are also adjusted for the first-stage estimation; *c)* robust standard errors as in *b)*, but the 2SLS estimator is used and accounted for; *d)* for dummy variables, the basis categories are given in parenthesis; *e)* a Wald test on the joint significance of the IMRs is conducted; *f)* the χ^2 test statistics for joint significance of $\bar{\mathbf{x}}_i$ or $\bar{\mathbf{q}}_i$ are reported.

of 29.62 and 22.04. Thus, the null hypothesis of no selection can again be rejected.

For women too (table 3), six different econometric models are presented, but results are less intuitive than in the case of the male sample. As mentioned before, selection corrections are not indicated; Wald tests on the joint significance of the $(\xi_1, ..., \xi_T)$ for the models in columns (3) and (6) confirm this finding. In the specifications without instrumental variables, only pooled OLS brings about a significant result. When considering the different 2SLS estimators, only the fixed effects approach provides a coefficient which is significantly different from zero. The fact that the coefficients in columns (4), (5), and (6) are all larger than those in (1), (2) and (3) may again indicate measurement error problems in the self assessed health variable. The (significant) parameter of health satisfaction in column (5) exhibits a value of 0.047, whereas the pooled OLS coefficient in column (1) lies at 0.014.

Interpreting the results is straightforward: Since both the dependent and the health variable are given in logs, interrelations between the two can be approximated employing elasticities.[15] For males, raising health satisfaction by 10% increases (hourly) wages approximately by 0.09 to 0.88 percent. In the case of females, the increase of the wage rate ranges from about 0.14 to 0.47 percent.

Turning to the other factors affecting earnings, concave wage profiles are found with respect to the time a person spent at the same firm in all specifications and for women and men. Starting, for example, at a value of two years on the job experience, an additional year at the same firm increases female (male) wages by 0.45% (0.6%), when controlling for selection. Given the high unemployment rates in Germany, it is interesting to see that in all models past unemployment periods significantly decrease wages (at an increasing rate). If the coefficient of education is identified, the returns to an additional year of schooling are almost 4% for women and approximately 3.2% for men.

Results for most of the other variables are as expected. For both women and men wages increase at an decreasing rate with age, and working in the eastern part of Germany or being in part-time employment reduces salaries. In the pooled specifications in columns (1) and (4) a larger average number of job seekers per federal sate negatively influences wages, whereas an increasing amount of notified vacancies raises the wage rate. Finally, as for the structural factors effecting wages, I find industry and occupational wage differentials.[16]

[15]It is implicitly assumed that health satisfaction is a continuous variable. Assessing health as an categorial variable, a 10% rise in health satisfaction equals roughly the increase by one category.

[16]In all models and both for females and males Wald tests confirm the joint significance of six occupational and nine sector dummies at any sensible level.

Table 3: Wage equations, Women, 1995-2005

	OLS[a)]	Within[a)]	Wooldr95[b)]	2SLS[a)]	FE2SLS[a)]	Wooldr05[c)]
	(1)	(2)	(3)	(4)	(5)	(6)
Health sat.	0.014	0.005	0.002	0.018	0.047	0.021
	(0.005)***	(0.005)	(0.005)	(0.013)	(0.018)***	(0.024)
Age	0.071	.	.	0.071	.	.
	(0.007)***			(0.007)***		
Age sq.	-.001	-.001	0.0008	-.001	-.001	0.0008
	(0.0002)***	(0.0003)***	(0.00006)***	(0.0002)***	(0.0002)***	(0.00007)***
Age tr.	7.70e-06	5.06e-06	-9.06e-06	7.69e-06	5.19e-06	-9.02e-06
	(1.50e-06)***	(2.02e-06)**	(6.72e-07)***	(1.43e-06)***	(2.01e-06)**	(8.77e-07)***
Unempl. exp.	-.034	-.116	-.100	-.033	-.116	-.101
	(0.003)***	(0.015)***	(0.018)***	(0.003)***	(0.015)***	(0.018)***
Unempl. exp. sq.	0.002	0.008	0.007	0.002	0.008	0.007
	(0.0002)***	(0.002)***	(0.003)**	(0.0003)***	(0.002)***	(0.003)**
Firm tenure	0.015	0.002	0.005	0.015	0.002	0.005
	(0.0007)***	(0.001)**	(0.001)***	(0.0007)***	(0.001)**	(0.001)***
Firm tenure sq.	-.0002	-.00005	-.0001	-.0002	-.00005	-.0001
	(0.00002)***	(0.00003)	(0.00004)***	(0.00002)***	(0.00003)*	(0.00004)***
Education	0.039	.	.	0.039	.	.
	(0.0009)***			(0.0009)***		
du. educ.	-.028	.	.	-.028	.	.
	(0.005)***			(0.005)***		
Part time	-.048	-.004	0.004	-.048	-.004	0.002
	(0.004)***	(0.007)	(0.008)	(0.004)***	(0.007)	(0.008)
Foreigner	0.006	.	.	0.006	.	.
	(0.006)			(0.007)		
Lg. unempl. (fed. st.)	-.033	0.006	0.015	-.033	0.007	0.014
	(0.005)***	(0.015)	(0.02)	(0.005)***	(0.015)	(0.023)
Lg. vac. (fed. st.)	0.026	-.003	-.019	0.026	-.003	-.018
	(0.005)***	(0.008)	(0.011)*	(0.005)***	(0.008)	(0.013)
Firm size (<20 employees)[d)]						
20 - 199	0.087	0.04	0.036	0.087	0.04	0.036
	(0.005)***	(0.007)***	(0.009)***	(0.005)***	(0.007)***	(0.009)***
200 - 1999	0.133	0.06	0.052	0.133	0.06	0.052
	(0.006)***	(0.009)***	(0.011)***	(0.005)***	(0.009)***	(0.011)***
$\geq$ 2000 employees	0.17	0.061	0.049	0.17	0.061	0.049
	(0.006)***	(0.009)***	(0.011)***	(0.006)***	(0.009)***	(0.011)***
firm size missing	0.128	0.057	0.094	0.128	0.056	0.093
	(0.02)***	(0.017)***	(0.021)***	(0.017)***	(0.017)***	(0.021)***
Region, where person works (Western Germany)						
East Germany	-.224	-.035	-.216	-.224	-.033	-.213
	(0.006)***	(0.013)***	(0.012)***	(0.006)***	(0.013)***	(0.012)***
constant	0.707			0.697	.	
	(0.104)***			(0.108)***		
N	29,304	29,304	29,304	29,304	29,304	29,304
d.f.	29,259	23,544	29,236	29,259	23,544	29,226
Wald tests on the joint significance of						
11 IMRs[e)]	.	.	9.00	.	.	12.58
10 time dummies	79.49***	102.12***	17.97*	82.79***	103.78***	18.22**
6 occup. dummies	2259.16***	29.57***	674.32***	2270.60***	28.87***	633.81***
9 sector dummies	563.88***	30.05***	156.18***	593.48***	30.06***	158.48***
unobs. effects [f)]	.	.	811.42***	.	.	769.38***

Source: GSOEP 1995-2005, own calculations. Standard errors in parenthesis: * significance at ten, ** at five, and *** at one percent. Year, sector, and occupation dummies are included but not reported. *a)* Robust standard errors are provided using the Huber/White/sandwich estimator; *b)* standard errors are robust to serial correlation and heteroscedasticity. They are also adjusted for the first-stage estimation; *c)* robust standard errors as in *b)*, but the 2SLS estimator is used and accounted for; *d)* for dummy variables, the basis categories are given in parenthesis; *e)* a Wald test on the joint significance of the IMRs is conducted; *f)* the χ^2 test statistics for joint significance of $\bar{\mathbf{x}}_i$ or $\bar{\mathbf{q}}_i$ are reported.

Women and men working in large firms ($\geq$ 2000), ceteris paribus, earn significantly more than in medium-sized firms, which in turn earn more than males and females employed in small firms. These effects are still observed when controlling for individual heterogeneity and selection effects, however, the magnitude of the parameters declines.

6 Conclusions

In this article, I employ recently developed estimation methods, which control for selection, individual heterogeneity, and endogeneity in one common framework, and apply them to the question whether health has an effect upon wages. There are a number of important links that connect the state of health and earnings. First, health as part of one's human capital affects labour market productivity and hence wages. Second if the rewards to health investment increase in the wage rate health should rise with wages, implying that there exists the problem of reverse causality. Furthermore, as self-reported health satisfaction is used for estimation, it is not possible to assess one's actual health status accurately and measurement error could be a source of bias. Another shortcoming may arise due to the fact that labour market participation is endogenous, where one reason for the endogeneity is an individual's health status. If panel attrition is not a random phenomena but driven by the individual participation decision employing standard methods may result in inconsistent estimation. Finally, since it is likely that unobserved effects (e.g. genetic endowment) are correlated with health the use of panel data techniques is necessary in order to control for a potential omitted variable bias.

In this paper reduced form wage equations for women and men augmented by a variable measuring health satisfaction are estimated. In an attempt to control for unobserved heterogeneity, sample selection, and endogeneity the estimators proposed by Wooldridge (1995) and Semykina and Wooldridge (2005) are applied. Due to the panel structure of the data it is possible to control for unobserved effects. A number of tests provide evidence that for the male sample selection corrections are indicated, while this issue does not cause any problems in the female population. The results show that good health raises wages. For females an increase in health satisfaction by 10% enhances (hourly) wages approximately by 0.14 to 0.47 percent. In the male sample the increase of the wage rate ranges from about 0.09 to 0.88 percent. The health variable is found to suffer from measurement error. For men, applying pooled OLS or

pooled 2SLS is accompanied by an upward bias in the health coefficient.

The estimated effects of health on wages work only for contemporaneous changes in health and wages. It is, however, likely that the state of health follows a persistent stochastic process, where the first source of persistence can easily be controlled for by including fixed effects. Non-inclusion of lagged health variables, to account for state dependency as the second reason of persistency, leaves a source of endogeneity in the model, and I try to compensate for it by utilising instrumental variables. Yet, it seems to be a task for the future to estimate a 'complete' model that allows for identifying all potential sources of endogeneity plus the dynamics of health in one common framework.

Appendix

Asymptotic variance-covariance matrices for the estimators in section 3.[17] Given the estimated parameter vector $\hat{\boldsymbol{\varrho}}^{OLS} = (\hat{\varphi}_0, \hat{\boldsymbol{\varphi}}', \hat{\boldsymbol{\beta}}', \hat{\boldsymbol{\xi}}')'$ in section 3.2, the asymptotic variance is $\text{Avar}(\hat{\boldsymbol{\varrho}}^{OLS}) = \hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{-1}$. Consistent estimators of $\hat{\mathbf{A}}$ and $\hat{\mathbf{B}}$ are:

$$\hat{\mathbf{A}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\hat{\mathbf{m}}'_{i,t}\hat{\mathbf{m}}_{i,t}, \tag{20}$$

where $\hat{\mathbf{m}}_{i,t}$ is $(1, \overline{\mathbf{x}}_i, \mathbf{x}_{i,t}, 0, ..., 0, \hat{\lambda_{i,t}}, 0, ..., 0)$, a $1 \times (1 + 2K + T)$ vector; and

$$\hat{\mathbf{B}} = N^{-1}\sum_{i=1}^{N}\hat{\mathbf{p}}_i\hat{\mathbf{p}}'_i. \tag{21}$$

The $(1 + 2K + T) \times 1$ vector $\hat{\mathbf{p}}_i$ is defined as

$$\hat{\mathbf{p}}_i = \hat{\mathbf{j}}_i - \hat{\mathbf{D}}\hat{\mathbf{k}}_i, \quad i = 1, ..., N, \tag{22}$$

and $\hat{\mathbf{j}}_i = \sum_{t=1}^{T} s_{i,t}d_{i,t}\hat{\mathbf{m}}'_{i,t}\hat{r}_{i,t}$, where $\hat{r}_{i,t}$ is the OLS residual from equation (14). Next, construct the $(1 + 2G)T \times 1$ vector $\hat{\mathbf{k}}_i$ as $(k'_{i,1}, ..., k'_{i,T})'$ and obtain each $k_{i,t}$ by multiplying the estimated information matrix, $I_t(\hat{\boldsymbol{\delta}}_t)$, for each t with the score, $sc_{i,t}(\hat{\boldsymbol{\delta}}_t)$, of the log-likelihood function for person i at time t.[18]The formulas are given in Maddala (1983) or Semykina and Wooldridge (2005) and need to be calculated using $\mathbf{h}_{i,t}$ and $\hat{\boldsymbol{\delta}}_t$, defined in section 3.2. Using e.g. the statistical software Stata© allows to easily derive the two terms. First, extract the variance-covariance matrix for the T probit models, calculate the inverse and divide it by the number of observations in each participation equation. Second, use the score option for each probit and multiply it with the corresponding $(1 + 2G) \times 1$ covariate-vector. Third, multiply the two to obtain T $k_{i,t}$ vectors and stack them as described above.

Finally, a consistent estimator for $\hat{D}$ is

$$\hat{\mathbf{D}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\hat{\mathbf{m}}'_{i,t}\hat{\boldsymbol{\varrho}'}^{OLS}\hat{\mathbf{F}}_{i,t}. \tag{23}$$

Here, $\hat{\mathbf{F}}_{i,t}$ is the $(1 + 2K + T) \times T(1 + 2G)$ matrix

$$\hat{\mathbf{F}}_{i,t} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \hat{\mathbf{Z}}_{i,t} & \mathbf{0} & \dots & \mathbf{0} \end{pmatrix}, \tag{24}$$

[17]The derivations in this section are based on Wooldridge (1995) and Semykina and Wooldridge (2005).

[18]$(1 + 2G)$ is the number of covariates in each participation equation, see section 3.2.

where each zero in the first row block is a $(1 + 2K) \times (1 + 2G)$ matrix and each zero in the second row block is a $T \times (1 + 2G)$ matrix. At last, the $T \times (1 + 2G)$ matrix $\hat{\mathbf{Z}}_{i,t}$, which is in the tth column block of $\hat{\mathbf{F}}_{i,t}$, is defined as $\hat{\mathbf{Z}}_{i,t} = (\mathbf{0}', \mathbf{0}', ..., (\hat{v}_{i,t}\mathbf{h}_{i,t})', \mathbf{0}', ..., \mathbf{0}')'$, where each zero is a $1 \times (1 + 2G)$ vector, and

$$\hat{v}_{i,t} = -\frac{\phi(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t)[\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t\Phi(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t) + \phi(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t)]}{\Phi(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t)^2}. \tag{25}$$

To calculate the asymptotic variance of the coefficient vector $\hat{\boldsymbol{\varrho}}^{2SLS} = (\hat{\varphi}_0, ..., \hat{\varphi}_E, \hat{\beta}_1, ..., \hat{\beta}_K, \hat{\xi_1}, ..., \hat{\xi}_T)'$ in section 3.4, define

$$\text{Avar}(\hat{\boldsymbol{\varrho}}^{2SLS}) = N^{-1}(\hat{\mathbf{C}}'\hat{\mathbf{O}}^{-1}\hat{\mathbf{C}})^{-1}\hat{\mathbf{C}}'\hat{\mathbf{O}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{O}}^{-1}\hat{\mathbf{C}}(\hat{\mathbf{C}}'\hat{\mathbf{O}}^{-1}\hat{\mathbf{C}})^{-1}. \tag{26}$$

First, use the $1 \times (1 + E + K + T)$ vector of regressors $\hat{\mathbf{y}}_{i,t} = (1, \overline{\mathbf{q}}_i, \mathbf{x}_{i,t}, 0, ..., 0, \hat{\lambda_{i,t}}, 0, ..., 0)$ and the $1 \times (1 + 2E + T)$ vector of instruments $\hat{\mathbf{n}}_{i,t} = (1, \overline{\mathbf{q}}_i, \mathbf{q}_{i,t}, 0, ..., 0, \hat{\lambda_{i,t}}, 0, ..., 0)$ to calculate

$$\hat{\mathbf{C}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\hat{\mathbf{n}}'_{i,t}\hat{\mathbf{y}}_{i,t} \quad \text{and} \quad \hat{\mathbf{O}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\hat{\mathbf{n}}'_{i,t}\hat{\mathbf{n}}_{i,t}. \tag{27}$$

The formula for $\hat{\mathbf{B}}$ is given in (21), but its dimension is now $(1+2E+T)\times(1+2E+T)$, and the $(1 + 2E + T) \times 1$ vector $\hat{\mathbf{p}}_i$ has the form

$$\hat{\mathbf{p}}_i = \sum_{t=1}^{T}(s_{i,t}d_{i,t}\hat{\mathbf{n}}'_{i,t}\hat{r}_{i,t} - \hat{\mathbf{M}}\hat{\mathbf{k}}_i), \tag{28}$$

where $\hat{r}_{i,t}$ is the 2SLS residual from equation (19).[19] The $(1 + 2G)T \times 1$ vector $\hat{\mathbf{k}}_i$ is constructed as described above. $\hat{\mathbf{M}}$, a $(1+2E+T)\times(1+2G)T$ matrix, has the form $\hat{\mathbf{M}} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} s_{i,t}d_{i,t}\hat{\mathbf{n}}'_{i,t}\hat{\boldsymbol{\varrho}}'^{2SLS}\nabla_\delta\hat{\mathbf{y}}'_{i,t}$. Finally, define $\nabla_\delta\hat{\mathbf{y}}'_{i,t}$ like $\hat{\mathbf{F}}_{i,t}$ in (24), except that now each of the T zeros in the first row block is $(1 + K + E) \times (1 + 2G)$ and each zero in the second row block is $T \times (1 + 2G)$. The $T \times (1 + 2G)$ matrix $\hat{\mathbf{Z}}_{i,t}$, which is in the tth column block of $\nabla_\delta\hat{\mathbf{y}}'_{i,t}$, has the form

$$\hat{\mathbf{Z}}_{i,t} = \begin{pmatrix} -\mathbf{h}_{i,t}\hat{\lambda}_{i,t}(\mathbf{h}_{i,t}\hat{\boldsymbol{\delta}}_t + \hat{\lambda}_{i,t}) \\ \mathbf{0} \\ \cdots \\ \mathbf{0} \end{pmatrix}. \tag{29}$$

[19] Note that $r_{i,t}$ is not the residual from the second stage OLS regression. Instead, it is defined as $\hat{r}_{i,t} = w_{i,t} - \hat{\mathbf{y}}_{i,t}\hat{\boldsymbol{\varrho}}^{2SLS}$.

Table 4: Description of Variables (part i)

Variable	Description
Probit	dummy variable indicating participation in the labour market (probit = 1) or no participation (probit = 0)
Log hourly wage	log gross hourly real wage (deflated to 2001 Euros)
Health satisfaction	variable indicating current health satisfaction of an individual; categories range from $0-10$; transformation: $f(h_{i,t}) = \log(h_{i,t} + \sqrt{(h_{i,t}^2+1)})$
Age	age in years
Unemployment experience	length of unemployment in a person's career; in years, with months in decimal form
Firm tenure	length of time with firm; in years, with months in decimal form
Education	amount of education or training in years
Dummy education	after intensive checks, wrong values of the education variable are changed to their maximum (Du. educ. = 1)
Part-time	dummy variable indicating part-time work
Foreigner	dummy variable indicating non-German nationality
Log unemployment[a]	(log) yearly averages of job seekers (per federal state)
Log vacancies	(log) notified vacancies (per federal state)
Firm size	four dummy variables indicating different firm sizes; categories: up to 20 employees ; $20-199$ employees; $200-1999$ employees; larger than 2000 employees
Region live/work	dummy variables indicating where a person lives (probit equ.) or works (wage equ.); Region = 0 if Western Germany
Occupation	seven occupation dummies, constructed using the Erikson, Goldthorpe Class Category IS88 (basis: high serv.)
Sector	ten aggregated sector dummies, based on the NACE classification (basis: agric., forestry, fishing)

(continued)

[a]Both unemployment and vacancy figures are extracted from Arbeitsstatistik 2005 - Jahreszahlen, provided by the Federal Employment Agency, Nuremberg.

Table 5: Description of Variables (part ii)

Variable	Description
Time	eleven time dummies (1995 - 2005) (basis: 1995)
Number of children	no. of children in three categories; 1) up to 2 years old; 2) between 3 - 5 years old; 3) between 6 - 16 years old
Non labour income	household income minus net wage income (in 2001 Euros)
No. of visits doctor	number of doctor visits last three months
No. days off	number of days absent from work due to illness last year; the variable is set to zero if a persons was not working last year
Partner or Spouse variables	
Single	dummy variable indicating whether a person has a partner/is married (single = 0)
Net wage[a]	net wage of partner or spouse
Age	age in years of partner or spouse
Experience	labour market experience of partner/spouse
Education	amount of education or training in years of partner/spouse

[a]All partner/spouse variables equal zero, if single = 1.

Participation equations. Tables 6 and 7 present estimation results for the participation equations (see equations (9), (10), and (11) in section 3.2) between 1995 and 2005 using pooled and 'traditional' random effects probit models and two Mundlak (1978) versions of Chamberlain's (1980) random effects probit model. Columns (3) and (4) depict results where the unobserved effects, k_i, are written as linear predictions on the means of all regressors and an error term a_i, which is assumed to be independent of $\mathbf{h}_{i,t}$ with (constant) variance σ_a^2. This explicitly allows some of the regressors to be correlated with the individual effects (k_i), but means that coefficients of time-invariant regressors, like education, are not identified. Under the further assumption that the participation indicators $(s_{i,1}, ..., s_{i,T})$ are independent conditional on $(\mathbf{h}_i, a_i)$, a random effects probit model is estimated; results are depicted in column (4).[20] The pooled probit model in column (3) (where again the unobserved effects are parameterised using the (within) means of the regressors) offers an estimation approach under less restrictive assumptions. Here, the

[20]For a detailed description of the different estimators and corresponding assumptions see Wooldridge (2002), chapter 15.8.

independence assumption with respect to $(s_{i,1}, ..., s_{i,T})$ can be relaxed. However, a robust variance covariance matrix estimator is required to account for the fact that observations are correlated within individuals over time.[21] Equivalently, in columns (1) – pooled probit – and (2) – random effects probit – the same specifications are considered, but here it is assumed that the unobserved effects, k_i, are uncorrelated with any of the regressors.

The estimated coefficients of the health variable show that for both women and men good health significantly increases the probability to work. To compare the different results, I calculate probability differences of being in (very) good health versus suffering from poor health. On this account, participation probabilities of 'average' individuals are predicted, where persons differ only with respect to their state of health. For a healthy man, using pooled probit (column (1)), the probability to work is $P(s = 1|health = 10, \overline{\mathbf{h}}) - P(s = 1|health = 0, \overline{\mathbf{h}}) = 43$ percentage points higher than for a unhealthy male person. Estimating the same model, but controlling for correlated individual effects (column (3)), strongly reduces the probability difference to 8.5 percentage points. In the random effects specification of column (2) the probability to work is 3.6 percentage points higher for healthy than for unhealthy men. Here, controlling for correlated fixed effects results in a probability difference of a single percentage point (column (4)). For women the impact of health satisfaction on labour market participation is also positive and significant in all econometric models. A comparison of healthy and unhealthy females results in probability differences of about 36 percentage points, when the pooled probit estimator without correlated individual effects is considered, and 5.6 percentage points, when controlling for the interaction between individual effects and the health variable. In the random effects models the corresponding values (columns (2) and (4)) are around 9.5 and 3.6 percentage points, respectively.

Results for most of the other variables are as expected. For both women and men, the participation probability increases with age (at an decreasing rate) and education. Living in the eastern part of Germany, being of non-German origin, and the amount of non labour income has a negative effect on the probability to work. Interestingly, many of the partner and children variables exhibit the same sign for women and men. For both sexes, the number of children in different age categories mostly reduce the participation probability. The partner's net wage and her/his labour market experience is associated with a decreasing working probability in most specification for both females and males.

[21] Wald tests for the joint significance of the $\boldsymbol{\theta}$ coefficients confirm the presence of correlated unobserved effects. The resulting values of the test statistic in columns (3) and (4) are for both women and man larger than the critical value of the χ^2 at the one percent level.

Table 6: Participation Equation, Men, 1995-2005

	Pooled[a)]	Random Effects[b)]	Mundlak, pooled[a)c)]	Mundlak, R.E.[b)c)]
	(1)	(2)	(3)	(4)
Age	.117	.013	.	.
	(.041)***	(.059)		
Age square	-.002	.002	-.0001	.0008
	(.001)**	(.001)	(.0002)	(.0004)**
Age triple	1.00e-05	-.00004	-7.05e-06	-.00003
	(8.45e-06)	(1.00e-05)***	(2.73e-06)***	(4.09e-06)***
Education	.102	.211	.	.
	(.007)***	(.012)***	.	.
Dummy Education	-.038	-.095	.	.
	(.032)	(.041)**		
Foreigner	-.221	-.431	.	.
	(.045)***	(.072)***		
Health Sat.	.472	.445	.123	.222
	(.024)***	(.032)***	(.022)***	(.036)***
Non labour inc.	-.0004	-.0009	-.0005	-.001
	(.00002)***	(.00002)***	(.00003)***	(.00002)***
Number of children				
up to 2 years old	-.026	-.072	-.119	-.158
	(.037)	(.056)	(.037)***	(.062)**
between 3 - 5	-.036	-.027	-.078	-.084
	(.034)	(.049)	(.034)**	(.055)
between 6 - 16	-.060	-.094	-.081	-.112
	(.020)***	(.028)***	(.024)***	(.035)***
Partner/Spouse variables				
Single	3.502	3.609	1.804	1.932
	(.403)***	(.605)***	(.666)***	(.831)**
Net wage partner/spouse	-.00009	-.0003	-.0003	-.0005
	(.00003)***	(.00004)***	(.00003)***	(.00005)***
Age partner/spouse	.110	.117	.108	.114
	(.015)***	(.021)***	(.021)***	(.028)***
Age sq. partner/spouse	-.001	-.001	-.001	-.001
	(.0002)***	(.0003)***	(.0003)***	(.0004)***
Exp. partner/spouse	-.002	-.010	-.023	-.050
	(.006)	(.010)	(.014)*	(.018)***
Exp. sq. partner/spouse	-.00009	.00005	.0008	.002
	(.0002)	(.0003)	(.0004)**	(.0005)***
Educ. partner/spouse	.225	.206	-.007	-.008
	(.047)***	(.074)***	(.088)	(.107)
Educ. sq. partner/spouse	-.009	-.008	-.0008	-.001
	(.002)***	(.003)**	(.004)	(.004)
Region, where person lives (Western Germany)				
East-Germany	-.516	-1.032	-.515	-1.003
	(.031)***	(.060)***	(.032)***	(.060)***
constant	-5.829	-4.687	.	.
	(.680)***	(.990)***		
time dummies, χ^2_{10} =	65.115***	115.718***	52.737***	96.926***
unobs. effects, χ^2_{17} =	.	.	573.04***	730.26***
LL	-17572.54	-13532.97	-17303.95	-13332.33
scale parameter ρ_a	.	.778	.	.769
		(.007)		(0.007)

Source: GSOEP 1995-2005, own calculations. Different Probit specifications. 48,536 observations from 9,540 individuals. Standard errors in parenthesis: * significance at ten, ** at five, and *** at one percent. Year dummies are included in each procedure but not reported. a) Standard errors are robust to serial correlation in the individual scores across t; b) 24 points of quadrature; c) unobserved effects are specified as a linear projection on the (within) means of the regressors.

Table 7: Participation Equation, Women, 1995-2005

	Pooled[a)]	Random Effects[b)]	Mundlak, pooled[a)c)]	Mundlak, R.E.[b)c)]
	(1)	(2)	(3)	(4)
Age	.078	.030	.	.
	(.040)**	(.061)		
Age square	-.0005	.002	.001	.003
	(.001)	(.002)	(.0002)***	(.0004)***
Age triple	-1.00e-05	-.00005	-.00003	-.00005
	(7.81e-06)	(1.00e-05)***	(2.53e-06)***	(4.23e-06)***
Education	.102	.239	.	.
	(.007)***	(.012)***		
Dummy Education	.018	.023	.	.
	(.032)	(.041)		
Foreigner	-.255	-.517	.	.
	(.043)***	(.074)***		
Health Sat.	.313	.276	.050	.120
	(.024)***	(.032)***	(.017)***	(.035)***
Non labour inc.	-.0003	-.0006	-.0002	-.0006
	(.00002)***	(.00002)***	(.00002)***	(.00002)***
Number of children				
up to 2 years old	-1.476	-2.557	-1.168	-2.268
	(.043)***	(.063)***	(.044)***	(.065)***
between 3 - 5	-.827	-1.392	-.563	-1.154
	(.029)***	(.043)***	(.029)***	(.046)***
between 6 - 16	-.361	-.566	-.187	-.385
	(.017)***	(.025)***	(.018)***	(.030)***
Partner/Spouse variables				
Single	1.008	1.827	-.026	.435
	(.439)**	(.688)***	(.584)	(.953)
Net wage partner/spouse	-.0002	-.0002	-.0001	-.0002
	(.00002)***	(.00002)***	(.00002)***	(.00002)***
Age partner/spouse	.023	.052	.002	.036
	(.016)	(.025)**	(.022)	(.037)
Age sq. partner/spouse	-.0003	-.0005	.0003	.0002
	(.0002)*	(.0003)*	(.0002)	(.0004)
Exp. partner/spouse	.008	-.002	-.038	-.063
	(.008)	(.013)	(.013)***	(.021)***
Exp. sq. partner/spouse	-.0002	-.0001	.0004	.0006
	(.0002)	(.0003)	(.0002)	(.0004)
Educ. partner/spouse	.073	.114	-.002	-.027
	(.049)	(.073)	(.059)	(.101)
Educ. sq. partner/spouse	-.002	-.005	-.0005	-.0009
	(.002)	(.003)*	(.002)	(.004)
Region, where person lives (Western Germany)				
East-Germany	-.062	-.200	-.076	-.285
	(.032)*	(.060)***	(.033)**	(.064)***
constant	-3.249	-4.355	.	.
	(.665)***	(1.034)***		
time dummies, χ^2_{10} =	51.25***	62.198***	46.025***	55.414***
unobs. effects, χ^2_{17} =	.	.	624.86***	760.78***
LL	-25488.93	-17217.55	-25226.81	-17027.29
scale parameter ρ_a	.	0.818	.	.824
		(.005)		(.005)

Source: GSOEP 1995-2005, own calculations. Different Probit specifications. 48,763 observations from 10,081 persons. Standard errors in parenthesis: * significance at ten, ** at five, and *** at one percent. Year dummies are included in each procedure but not reported. a) Standard errors are robust to serial correlation in the individual scores across t; b) 24 points of quadrature; c) unobserved effects are specified as a linear projection on the (within) means of the regressors.

Table 8: Summary, Participation Equation, Men, 1995-2005

	Entire Sample	Probit = 0	Probit = 1
Probit	.832	0	1
	(.374)	(0)	(0)
Age	41.286	43.127	40.915
	(10.997)	(13.310)	(10.431)
Age sq.	1825.462	2037.099	1782.835
	(929.872)	(1125.620)	(879.096)
Age tr.	85498.640	102698.600	82034.300
	(62658.360)	(76567.650)	(58860.570)
Education	12.206	11.253	12.398
	(2.613)	(2.238)	(2.641)
Dummy educ.	.141	.150	.139
	(.348)	(.358)	(.346)
Foreigner	.133	.194	.120
	(.339)	(.395)	(.325)
Health sat.	2.566	2.406	2.598
	(.414)	(.581)	(.363)
Non labour inc.	774.283	1438.180	640.564
	(970.113)	(1054.049)	(894.576)
Number of children			
up to 2 years old	.082	.056	.087
	(.288)	(.240)	(.297)
between 3 - 5	.118	.074	.127
	(.353)	(.289)	(.364)
between 6 - 16	.480	.351	.506
	(.816)	(.750)	(.827)
Partner/Spouse variables[a)]			
Single	.226	.318	.208
	(.418)	(.466)	(.406)
Net wage partner/spouse	586.735	501.900	601.453
	(637.155)	(649.150)	(633.907)
Age partner/spouse	40.648	44.220	40.028
	(10.185)	(11.762)	(9.754)
Age sq. partner/spouse	1756.003	2093.697	1697.415
	(856.123)	(1017.939)	(810.646)
Exp. partner/spouse	10.507	12.887	10.094
	(9.176)	(11.168)	(8.719)
Exp. sq. partner/spouse	194.591	290.777	177.903
	(299.486)	(394.703)	(276.306)
Educ. partner/spouse	11.732	11.024	11.855
	(2.452)	(2.365)	(2.446)
Educ. partner/spouse	143.648	127.112	146.517
	(63.607)	(57.693)	(64.147)
Region, where person lives			
East-/West-Germany	.261	.374	.238
	(.439)	(.484)	(.426)
N	48,536	8,137	40,399

Source: GSOEP 1995-2005, own calculations. All summary statistics are on individual-year level. Standard errors are in parenthesis.
a) The reported sample statistics for these variables are conditional on having a partner/ being married (Single = 0);

Table 9: Summary, Participation Equation, Women, 1995-2005

	Entire Sample	Probit = 0	Probit = 1
Probit	.629	0	1
	(.483)	(0)	(0)
Age	41.474	43.207	40.453
	(11.194)	(12.232)	(10.400)
Age sq.	1845.396	2016.503	1744.624
	(945.628)	(1063.793)	(852.646)
Age tr.	87019.930	100022.100	79362.450
	(63840.140)	(73899.680)	(55690.880)
Education	11.911	11.182	12.340
	(2.472)	(2.265)	(2.489)
Dummy Educ.	.127	.126	.127
	(.333)	(.332)	(.333)
Foreigner	.130	.192	.093
	(.336)	(.394)	(.291)
Health sat.	2.549	2.490	2.583
	(.428)	(.497)	(.378)
Non labour inc.	802.090	1078.961	639.030
	(981.320)	(1045.618)	(902.513)
Number of children			
up to 2 years old	.055	.119	.017
	(.236)	(.340)	(.129)
between 3 - 5	.109	.189	.062
	(.340)	(.438)	(.255)
between 6 - 16	.505	.617	.440
	(.822)	(.933)	(.742)
Partner/Spouse variables[a)]			
Single	.214	.153	.250
	(.410)	(.360)	(.433)
Net wage partner/spouse	1451.518	1422.424	1470.860
	(1119.752)	(1219.612)	(1047.693)
Age partner/spouse	45.412	46.961	44.382
	(11.295)	(12.267)	(10.474)
Exp. partner/spouse	2189.811	2355.833	2079.437
	(1054.607)	(1173.289)	(951.816)
Age tr. partner/spouse	22.225	23.679	21.258
	(11.327)	(11.961)	(10.777)
Exp. sq. partner/spouse	622.232	703.759	568.032
	(528.356)	(583.731)	(480.496)
Educ. partner/spouse	12.039	11.673	12.283
	(2.663)	(2.610)	(2.670)
Educ. sq. partner/spouse	152.038	143.073	157.997
	(71.345)	(68.426)	(72.613)
Region, where person lives			
East-/West-Germany	.255	.209	.282
	(.436)	(.407)	(.450)
N	48,763	18,074	30,689

Source: GSOEP 1995-2005, own calculations. All summary statistics are on individual-year level. Standard errors are in parenthesis.
a) The reported sample statistics for these variables are conditional on having a partner/being married (Single = 0);

Table 10: Summary, Wage Equation, Men, 1995-2005

	Mean	Std. dev.	10% pctl.	90% pctl.
Log hourly wage	2.578	.407	2.071	3.093
Health sat.	2.599	.360	2.095	2.893
Age	41.020	10.298	28	56
Age sq.	1788.715	870.012	784	3136
Age tr.	82257.240	58301.720	21952	175616
Unempl. exp.	.380	1.056	0	1.100
Unempl. exp. sq.	1.258	8.746	0	1.210
Firm tenure	11.314	10.052	1.100	27
Firm tenure sq.	229.057	348.743	1.210	729
Education	12.418	2.642	10.500	18
Dummy educ.	.142	.349	0	1
Part-time	.018	.135	0	0
Foreigner	.119	.324	0	1
Lg. unempl. (fed. st.)	12.768	.569	12.150	13.630
Lg. vac. (fed. st.)	10.443	.839	9.136	11.404
Firm size (<20 employees)[a)]				
20 - 199	.301	.459	0	1
200 - 1999	.237	.425	0	1
$\geq$ 2000 employees	.262	.440	0	1
Firm size miss.	.017	.127	0	0
Region, where person works (Western Germany)				
Eastern Germany	.223	.416	0	1
Occupation Dummies (High Service)				
Low Service	.185	.388	0	1
Routine Non-Manual	.041	.198	0	0
Skilled Manual	.308	.462	0	1
Semi-unskilled Manual	.211	.408	0	1
Farm Labour	.011	.106	0	0
Missing occ.	.086	.280	0	0
Sector Dummies (Agr., forestry, fishing)				
Unknown sector	.022	.147	0	0
Energy, water, mining	.015	.123	0	0
Manufacturing	.369	.483	0	1
Construction	.111	.315	0	1
Trade	.086	.280	0	0
Transport, communication	.042	.200	0	0
Financial serv., insurance	.024	.154	0	0
Other services	.089	.285	0	0
State	.229	.420	0	1
Instruments				
Num. vis. doc. (last 3 months)	1.759	3.316	0	4
Days off due to illness $(t-1)$[b)]	8.951	21.365	0	21
Non labour inc.	629.209	879.454	0	1711.065
Single	.203	.402	0	1
Net wage partner/spouse[c)]	603.609	633.797	0	1450.677
Age partner/spouse	40.027	9.667	28	53
Age sq. partner/spouse	1695.632	802.981	784	2809
Exp. partner/spouse	10.109	8.686	.700	23.500
Exp. sq. partner/spouse	177.640	274.742	.490	552.250
Educ. partner/spouse	11.867	2.446	9	15
Educ. sq. partner/spouse	146.809	64.209	81	225

Source: GSOEP 1995-2005, own calculations. All summary statistics are on individual-year level (39,048 observations). Persons with participation in only one year and individuals with missing wages are dropped from the sample. *a)* For dummy variables, the basis categories are given in parenthesis; *b)* the reported sample statistics is conditional on whether the person was working last year. The variable is set to zero otherwise; *c)* the reported sample statistics for these variables are conditional on having a partner/ being married (Single = 0).

Table 11: Summary, Wage Equation, Women, 1995-2005

	Mean	Std. Dev.	10% pctl.	90% pctl.
Log horuly wage	2.362	.400	1.839	2.834
Health sat.	2.584	.376	2.095	2.893
Age	40.610	10.264	26	55
Age sq.	1754.498	843.538	676	3025
Age tr.	79830.600	55148	17576	166375
Unempl. exp.	.449	1.105	0	1.400
Unempl. exp. sq.	1.423	10.536	0	1.960
Firm tenure	9.405	8.647	1	23.200
Firm tenure sq.	163.230	270.144	1	538.240
Education	12.360	2.488	10	16
Dummy Educ.	.129	.336	0	1
Part-time	.367	.482	0	1
Foreigner	.092	.289	0	0
Lg. unempl. (fed. st.)	12.752	.566	12.150	13.626
Lg. vac. (fed. st.)	10.367	.861	9.060	11.404
Firm size (<20 employees)[a)]				
20 - 199	.295	.456	0	1
200 - 1999	.228	.420	0	1
$\geq$ 2000 employees	.200	.400	0	1
Firm size miss.	.018	.132	0	0
Region, where person works (Western Germany)				
Eastern Germany	.275	.447	0	1
Occupation Dummies (High Service)				
Low Service	.259	.438	0	1
Routine Non-Manual	.202	.402	0	1
Skilled Manual	.068	.252	0	0
Semi-unskilled Manual	.172	.377	0	1
Farm Labour	.009	.093	0	0
Missing occ.	.219	.414	0	1
Sector Dummies (Agr., forestry, fishing)				
Unknown sector	.023	.149	0	0
Energy, water, mining	.004	.061	0	0
Manufacturing	.171	.376	0	1
Construction	.017	.130	0	0
Trade	.154	.361	0	1
Transport, communication	.023	.149	0	0
Financial serv., insurance	.031	.174	0	0
Other services	.200	.400	0	1
State	.370	.483	0	1
Instruments				
Num. vis. doc. (last 3 months)	2.382	3.470	0	5
Days off due to illness $(t-1)$[b)]	9.567	22.253	0	21
Non labour inc.	629.226	890.755	0	1693.780
Single	.247	.431	0	1
Net wage partner/spouse[c)]	1472.655	1045.818	0	2636.535
Age partner/spouse	44.476	10.391	31	59
Age sq. partner/spouse	2086.081	945.872	961	3481
Exp. partner/spouse	21.349	10.705	6.900	36
Exp. sq. partner/spouse	570.349	477.907	47.610	1296
Educ. partner/spouse	12.297	2.673	10	18
Educ. sq. partner/spouse	158.374	72.758	100	324

Source: GSOEP 1995-2005, own calculations. All summary statistics are on individual-year level (29,304 observations). Persons with participation in only one year and individuals with missing wages are dropped from the sample. *a)* For dummy variables, the basis categories are given in parenthesis; *b)* the reported sample statistics is conditional on whether the person was working last year. The variable is set to zero otherwise; *c)* the reported sample statistics for these variables are conditional on having a partner/being married (Single = 0).

References

Chamberlain, Gary, "Analysis of Covariance with Qualitative Data," *Review of Economic Studies*, 1980, *47*, 225–238.

___ , "Panel data," *in Zvi Griliches and Michael D. Intriligator (eds), Handbook of Econometrics*, 1984, *Volume 2*, 1247–1318.

Contoyannis, Paul and Nigel Rice, "The Impact of Health on Wages: Evidence from the British Household Panel Survey," *Empirical Economics*, 2001, *26*, 599–622.

___ **, Andrew Michael Jones, and Nigel Rice**, "The dynamics of health in the British Household Panel Survey," *Journal of Applied Econometrics*, 2004, *19* (4), 473–503.

Dustmann, Christian and Maria Engracia Rochina-Barrachina, "Selection correction in panel data models: an application to labour supply and wages," *IZA Discussion Paper 162*, 2000.

Gambin, Lynn M., "The Impact of Health on Wages in Europe – Does gender matter?," *HEDG Working Paper*, June 2005, *03.*

Gordo, Laura Romeu, "Effects of short- and long-term unemployment on health satisfaction *evidence from German data," *Applied Economics*, 2006, *38* (20), 23352350.

Grossman, Michael, "The Human Capital Model," in Anthony J. Culyer and Joseph P. Newhouse, eds., *Handbook of Health Economics*, Vol. 1A, Amsterdam: Elsevier Science B.V., 2001, pp. 347–409.

Halliday, Timothy J. and John A. Burns, "Heterogeneity, State Dependence and Health," *University of Hawaii at Manao, Department of Economics Working Paper Series*, 2005, *3.*

Haveman, Robert, Barbara Wolfe, Brent Kreider, and Mark Stone, "Market Work, Wages, and Men's Health," *Journal of Health Economics*, 1994, *13*, 163–182.

Heckman, James J., "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 1978, *46* (4), 931–60.

Kyriazidou, Ekaterini, "Estimation of a panel data sample selection model," *Econometrica*, 1997, *55*, 1335–1364.

Lee, Lung-Fei, "Health and Wage: A Simultaneous Equation Model with Multiple Discrete Indicators," *International Economic Review*, February 1982, *23* (1), 199–221.

Maddala, G.S., *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge: Cambridge University Press, 1983.

Mincer, Jacob, "Investment in Human Capital and Personal Income Distribution," *Journal of Political Economy*, August 1958, *66* (4), 281–302.

___ , "Schooling, Experience and Earnings," *New York: National Bureau of Economic Research*, 1974.

Mundlak, Yair, "On the pooling of time series and cross section data," *Econometrica*, 1978, *46*, 69–85.

Rochina-Barrachina, Maria Engracia, "A new Estimator for Panel Data Sample Selection Models," *Annales d'Economie et de Statistique*, 1999, *55/56*, 153.181.

Semykina, Anastasia and Jeffrey M. Wooldridge, "Estimating Panel Data Models in the Presence of Endogeneity and Selection: Theory and Application," *mimeo*, October 2005.

SOEP Group, "The German Socio-Economic Panel (GSOEP) after more than 15 years - Overview," in Elke, Holst, Dean Lillard, and Thomas A. DiPrete, eds., *Proceedings of the 2000 Fourth International Conference of German Socio-Economic Panel Study Users (GSOEP2000)*, Vierteljahreshefte zur Wirtschaftsforschung, 2001, 70(1), pp. 7-14.

Verbeek, Marno and Theo Nijman, "Testing for Selectivity Bias in Panel Data Models," *International Economic Review*, 1992, *33*, 681–703.

Wooldridge, Jeffrey M., "Selection correction for panel data models under conditional mean independence assumption," *Journal of Econometrics*, 1995, *68*, 115–132.

___ , *Econometric analysis of cross section and panel data*, Cambridge and London: MIT Press, 2002.

Ifo Working Papers

No. 42 Mayr, J. and D. Ulbricht, Log versus Level in VAR Forecasting: 16 Million Empirical Answers – Expect the Unexpected, February 2007.

No. 41 Oberndorfer, U., D. Ulbricht and J. Ketterer, Lost in Transmission? Stock Market Impacts of the 2006 European Gas Crisis, February 2007.

No. 40 Abberger, K., Forecasting Quarter-on-Quarter Changes of German GDP with Monthly Business Tendency Survey Results, January 2007.

No. 39 Batchelor, R., Forecaster Behaviour and Bias in Macroeconomic Forecasts, January 2007.

No. 38 Sülzle, K., Innovation and Adoption of Electronic Business Technologies, December 2006.

No. 37 Overesch, M. and G. Wamser, German Inbound Investment, Corporate Tax Planning, and Thin-Capitalization Rules – A Difference-in-Differences Approach, December 2006.

No. 36 Kempkes, G. and C. Pohl, The Efficiency of German Universities – Some Evidence from Non-Parametric and Parametric Methods, October 2006.

No. 35 Kuhlmann, A., German Productivity – A Reassessment via the New Ifo Productivity Database, October 2006.

No. 34 Kuhlmann, A., What is the X-Factor in the German Electricity Industry?, September 2006.

No. 33 Temple, J. and L. Wößmann, Dualism and cross-country growth regressions, August 2006.

No. 32 Baumann, F., V. Meier and M. Werding, Transferable Provisions in Individual Health Insurance Contracts, July 2006.

No. 31 Abberger, K., Qualitative Business Surveys in Manufacturing and Industrial Production – What can be Learned from Industry Branch Results?, May 2006.

No. 30 Ruschinski, M., Investigating the Cyclical Properties of World Trade, May 2006.

No. 29 Holzner, Chr., V. Meier and M. Werding, Time Limits in a Two-tier Unemployment Benefit Scheme under Involuntary Unemployment, April 2006.

No. 28 Eggert, W. and A. Haufler, Company Tax Coordination cum Tax Rate Competition in the European Union, April 2006.

No. 27 Lachenmaier, S. and H. Rottmann, Employment Effects of Innovation at the Firm Level, April 2006.

No. 26 Radulescu, D.M. and M. Stimmelmayr, Does Incorporation Matter? Quantifying the Welfare Loss of Non-Uniform Taxation across Sectors, March 2006.

No. 25 Lessmann, Chr., Fiscal Decentralization and Regional Disparity: A Panel Data Approach for OECD Countries, March 2006.

No. 24 Fuchs, Th., Industry Structure and Productivity Growth: Panel Data Evidence for Germany from 1971–2000, December 2005.

No. 23 Holzner, Chr. and A. Launov, Search Equilibrium, Production Parameters and Social Returns to Education: Theory and Estimation, December 2005.

No. 22 Sülzle, K., Stable and Efficient Electronic Business Networks: Key Players and the Dilemma of Peripheral Firms, December 2005.

No. 21 Wohlrabe, K. and M. Fuchs, The European Union's Trade Potential after the Enlargement in 2004, November 2005.

No. 20 Radulescu, D.M. and M. Stimmelmayr, Implementing a Dual Income Tax in Germany: Effects on Investment and Welfare, November 2005.

No. 19 Osterkamp, R. and O. Röhn, Being on Sick Leave – Possible Explanations for Differences of Sick-leave Days Across Countries, November 2005.

No. 18 Kuhlmann, A., Privatization Incentives – A Wage Bargaining Approach, November 2005.

No. 17 Schütz, G. und L. Wößmann, Chancengleichheit im Schulsystem: Internationale deskriptive Evidenz und mögliche Bestimmungsfaktoren, Oktober 2005.