Health and Wages - Panel data estimates considering selection and endogeneity

Jäckle, Robert and Himmler, Oliver
TNS Infratest Social Research, Goettingen University

February 2007

Health and Wages
Panel data estimates considering selection and endogeneity

Robert Jäckle[*] Oliver Himmler

TNS Infratest Social Research *Goettingen University*

November 14, 2008

Abstract

This paper complements previous studies on the effects of health on wages by addressing the problems of unobserved heterogeneity, sample selection, and endogeneity in one comprehensive framework. Using data from the German Socio-Economic Panel (GSOEP) we find the health variable to suffer from measurement error and a number of tests provide evidence that selection corrections are necessary. Good health leads to higher wages for men, while there appears to be no significant effect for women. Contingent on the method of estimation, healthy males are estimated to earn between 1.3% and 7.8% more than those in poor health.

Keywords:
health; wages; fixed effects; sample selection; instrumental variables

JEL Classification:
I1, J4, C33, C34

[*] TNS Infratest Sozialforschung
Arnulfstr. 205
D-80687 Munich

Phone: +49 89 5600 1290
E-mail: robert.jaeckle@tns−infratest.com

We thank Christian Holzner, Georg Wamser, Joachim Winter, and participants at various seminars and conferences for helpful discussions. Three anonymous referees are thanked for insightful suggestions. Special thanks goes to Joachim Wolff who kindly shared his GSOEP preparation files. All remaining deficiencies are the authors' responsibility.

1 Introduction

Does superior health enable individuals to command higher wages? This question has spurred research in both labor and health economics and consequently led to the identification of two major channels of interaction. First, health as part of human capital may positively affect labor 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. Thus, reverse causality may lead to biased estimates of the health effect. A number of further challenges need to be dealt with: while inaccuracies in assessing health status may introduce bias due to measurement error whenever self-reported health satisfaction is used in the estimations, another problem that remains unappreciated in most earlier studies is non-random sample selection. Since labor market participation is endogenous and health status is one of the influences driving selection, failing to apply selection correction methods may result in inconsistent estimation. Finally, an issue particularly relevant in the health context is unobserved heterogeneity. Whenever unobserved factors such as genetic endowment are correlated with health, the use of panel data techniques to account for omitted variable bias is called for.

The impact of health on wages has been studied using a variety of econometric approaches, accounting for the above problems to different extents: Gambin (2005) investigates the relationship between health and wages for 14 European countries employing fixed (FE) and random effects (RE) estimation. She proposes that for men, self-reported health has a greater effect than for females, while in the case of chronic diseases the opposite holds true. An econometric model that accounts for the simultaneous effects of health and wages in a structural multi-equation system has been suggested by Lee (1982). His approach is based on a generalized version of Heckman's (1978) treatment model. Using a cross-sectional sample of male US citizens, he finds that health and wages are strongly interrelated, that is, wages positively affect health and vice versa. In a similar vein, Cai (2007) estimates a multi-equation system using cross-sectional Australian data and finds health to have a positive effect on wages once endogeneity is accounted for. He also finds that there is no endogenous selection present in his data. Haveman, Wolfe, Kreider, and Stone (1994) estimate a multiple equation system for working time, wages, and health, employing generalized methods of moments techniques on panel data. They find that in the male US population poor health affects wages negatively. The effect of self-assessed general and psychological health on wages is at the core of Contoyannis and Rice's (2001) study using the British Household Panel Survey. They apply FE and RE instrumental variable estimators and conclude that reduced psychological health decreases male wages, while positive self-assessed health increases hourly wages for women. While each of these papers tackles at least one of the mentioned econometric issues, to our knowledge there is no study that accounts for unobserved heterogeneity, non-random sample selection and endogeneity in one framework.

In order to fill this gap, we utilize a recently developed estimation method proposed by Semykina and Wooldridge (2006), which extends Wooldridge's (1995) method of testing and correcting for sample selection in fixed effects models. The latter estimator has been contrasted with alternative methods proposed by Kyriazidou (1997) and Rochina-Barrachina (1999) in an application to female wage equations by Dustmann and Rochina-Barrachina (2007). While Kyriazidou's (1997) estimator implies homoscedastic idiosyncratic errors over time, Rochina-Barrachina (1999) does not rely on this assumption. The drawback of their method, however, is that it assumes joint normality of the error terms in the probit and the main equation. Wooldridge's (1995) method relies on standard probit estimates for each year in order to calculate annual inverse Mills ratios (IMRs) and explicitly models the conditional mean of the error terms in the main equation. Its advantage over the other models that have been suggested is that it does not rely on 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. One approach to expanding these three estimators to account for non-strict exogeneity and measurement error is presented in Dustmann and Rochina-Barrachina (2007). Similarly, Semykina and Wooldridge (2006) enhance Wooldridge's (1995) estimator and demonstrate how to test and control for sample selection in a fixed effects model with endogeneity. The reason we choose to adopt the Semykina and Wooldridge (2006) approach in this paper is that, other than the alternative methods, it allows for time heteroscedasticity and autocorrelation in the error terms in both equations.

The estimator is applied to male and female samples taken from the German Socio-Economic Panel (GSOEP). We find the health variable to be reported with error and a number of tests provide evidence that corrections for non-random selection into the workforce are indicated in both the female and male sample. We show that the impact of health on wages is statistically different from zero for men only. For them, a highly significant effect of health on wages, associated with up to 7.8% of a health premium is found and cannot be eradicated by applying selection correction. Considering non-random selection into the work force is, however, associated with lower wages on each health level for both genders.

The remainder of this paper is structured as follows: the starting point is a discussion of specification issues and resulting problems, followed by a detailed overview of the estimation methods in section 3. The ensuing section 4 provides data descriptions and discusses various specifications of the health variable. In section 5 we report estimation and test results. Section 6 concludes.

2 Model Specification and Resulting Problems

To fix ideas, a simple model of how health affects wages is presented. A firm produces Y_t at time $t = (1, 2, ..., T)$, using effective labor L_t as the single input in producing Y_t. The firms's production function is given by $Y_t = F(L_t)$, and the amount of effective labor can be written as

$$L_t = \sum_{i=1}^{n} p_i(E_i, a_{i,t}, h_{i,t}) \cdot \ell_{i,t}, \tag{1}$$

where $\ell_{i,t}$ is labor supply of employee i, and $p_i(\cdot)$ is an unknown function that determines the effectiveness of an individual's working hours $\ell_{i,t}$. This function takes as arguments the years of education E_i, age $a_{i,t}$, and state of health $h_{i,t}$. In what follows, we refer to the first two variables as the *human capital part* of $p_i(\cdot)$ and to the latter part as *health effect*.

Workers are paid according to their marginal productivity, and accordingly the log wage of each employee can be written as

$$\log w_{i,t} = log[\frac{dF}{dL_t} \cdot \frac{\partial L_t}{\partial \ell_{i,t}}] = \log F_{L_t} + \log p_i(E_i, a_{i,t}, h_{i,t}), \tag{2}$$

such that wages are determined by the firm-level supply and demand factors $\log F_{L_t}$ as well as by the employee-level human capital and health effects.

In what follows, we describe the operationalization of the latter two effects and derive the baseline econometric model.

The Human Capital Part. The human capital part of $p_i(\cdot)$ is approximated using a specification similar to Mincer (1958 and 1974). He suggested that log wages are linear in the years of schooling, and linear and quadratic in the years of labor market experience. Romeu Gordo (2006), however, finds evidence for the existence of a positive relationship between unemployment and health satisfaction using GSOEP data. On this account, we include unemployment rather than working experience. Adding an age variable then implicitly controls for work experience as well. Furthermore, human capital theory suggests using firm tenure as a proxy for the firm-specific investment in human capital. Since firm tenure (and its square) is more closely related to labor productivity than the general working experience it should cause an extra increase in wages.

The Health Effect. As stated earlier, health is an essential part of human capital and will thus affect labor market productivity which in turn determines wages. We use self-assessed health satisfaction as our key explanatory variable, the definition and functional form of which is discussed in detail in section 4.2.

Dependent variable and baseline specification. While health as a part of human capital directly affects productivity, it can also be considered an endogenous capital stock, which according to Grossman (2001) determines the amount of time an individual can spend participating in the labor market. One reason that the number of hours worked diverges somewhat across individuals may therefore lie in differences in health status and so we will use hourly wages rather than monthly earnings as the dependent variable.

The above model can then be parameterised as follows:

$$\begin{aligned} \log(w_{i,t}) &= \mathbf{b}_{B,t}\boldsymbol{\alpha} + \mathbf{f}_{i,t}\boldsymbol{\beta} + \mathbf{a}_{\mathbf{i},\mathbf{t}}\boldsymbol{\gamma} + \theta E_i + \mathbf{ue}_{i,t}\boldsymbol{\upsilon} + \mathbf{ft}_{i,t}\boldsymbol{\tau} \\ &+ \mathbf{ch}_{i,t}\boldsymbol{\rho} + \delta g(h_{i,t}) + \mathbf{du}_{i,t}\boldsymbol{\pi} + error, \end{aligned} \tag{3}$$

where $w_{i,t}$ are hourly wages, $\mathbf{b}_{B,t}$ is a vector that approximates firm level supply and demand forces ($\log F_{L_t)}$ by using the average number of job-seekers, notified vacancies, and (un)employment figures at the state (*Bundesstaat*) level B.[1] The vector of dummy variables $\mathbf{f}_{i,t}$ captures four different categories of firm size, $\mathbf{a}_{i,t}$ is the vector of a 3rd order polynomial of age $a_{i,t}$ and E_i denotes years of schooling or training. Second order polynomials of uneemployment experience and firm tenure are captured in $\mathbf{ue}_{i,t}$ and $\mathbf{ft}_{i,t}$, respectively. $\mathbf{ch}_{i,t}$ are the number of children in three age categories and $g(h_{i,t})$ is a yet to be determined function of the health variable. Finally, $\mathbf{du}_{i,t}$ are indicator variables for firm sector, occupational status,[2] East Germany, part-time work, nationality, children, and time periods.[3]

In the estimation of the parameter vector $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}', \theta, \boldsymbol{\upsilon}', \boldsymbol{\tau}', \boldsymbol{\rho}', \delta, \boldsymbol{\pi}')'$ in equation (3) a number of problems arise. To start with, Grossman (2001) suggests that 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 $w_{i,t}$. As we employ self-reported health satisfaction, *measurement error* can also be an important source of bias. In the absence of an "objective" measure, such as a physician's evaluation of overall health, δ will likely be biased towards zero. Another problem arises if a random sample drawn from the overall population is not available. In this study, we aim to identify the effect of health on the labor market productivity for *all* individuals, thus a bias may result from the fact that individuals endogenously decide to participate in the labor market. If some of the factors determining participation also affect health and wages, *selection correction* methods are in order. *Omitted variable bias* is also a cause of concern. Disregarding, e.g. genetic endowment of a person could lead to biased estimates as it may at the same time impact health status and hourly wages.[4] The following section explains how we deal with these issues econometrically.

3 Econometric Approach

As indicated above, the goal of this work is to make statements about the impact of health on wages for the *entire* population. Thus, with panel data, employing a simple within estimator is a reasonable approach only when we can be sure that the decision to participate in the labor market is either randomly determined or

[1] Data provided by the German Federal Employment Agency, Nuremberg.

[2] Interaction terms between the occupational status and the health variables as well as between age and health were found to be statistically insignificant and consequently dropped from the final model.

[3] Variable descriptions are shown in tables 6 and 7

[4] Past shocks (such as heart attacks, accidents, etc.) may affect current state of health (Contoyannis, Jones, and Rice (2004) and Halliday and Burns (2005)). As far as differences in the ability to cope with such (past) health shocks aren't covered by unobserved effects, endogeneity may be introduced. Considering the full dynamics of health on top of all sources of endogeneity mentioned above is, however, beyond the scope of this paper.

fully covered by the observable variables or the fixed effect. In the context of this paper it is entirely conceivable that unobserved time varying health determinants such as the lifestyle an individual engages in (think of alcohol, nicotine, sports) or motivation affect selection and will not be covered by the fixed effect. This kind of selection will then influence wages through the error term and lead to inconsistent estimation. To overcome the **selection problem**, the following model is estimated:

$$\begin{aligned} w^*_{i,t} &= \beta_0 + \mathbf{x}_{i,t}\boldsymbol{\beta}_1 + \mathbf{y}^*_{i,t}\boldsymbol{\beta}_2 + c_i + u_{i,t}, & (4)\\ w^*_{i,t} &= w_{i,t}, \quad y^*_{i,t} = \; y_{i,t} \qquad \text{if } s_{i,t} = 1 \text{ and unobserved otherwise}, & (5)\\ S^*_{i,t} &= \gamma_0 + k_i + \mathbf{z}_{i,t}\boldsymbol{\gamma} + e_{i,t}; \qquad S_{i,t} = 1[S^*_{i,t} > 0] & (6)\end{aligned}$$

where all variables superscripted with an asterisk pertain to the entire population. In (4), $w^*_{i,t}$ are hourly wages and the $1 \times K$ vector $\mathbf{x}_{i,t}$ comprises those explanatory variables in (3) that we observe irrespective of participation, including health. Variables that can only be observed for those who work make up $\mathbf{y}^*_{i,t}$ and are imposed as exclusion restrictions on the participation equation. Unobserved individual characteristics are contained in c_i, $u_{i,t}$ is an unobserved error term and $S_{i,t}$ in (5) denotes labor market participation. Equation (6) describes a person's decision to participate in the labor market, where $S^*_{i,t}$ is the latent propensity to work, 1[.] is an indicator function which equals one if its argument is true, and the $1 \times G$ vector $\mathbf{z}_{i,t}$ is a superset of $\mathbf{x}_{i,t}$. Though not strictly necessary, it is advantageous to have $G > K$, which is why we add exclusion restrictions to $\mathbf{z}_{i,t}$ that drive selection but can at the same time be omitted from equation (4). The individual effect k_i is composed of 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 and $\mathbf{z}_i$, with. $\mathbf{z}_i = (\mathbf{z}_{i,1}, ..., \mathbf{z}_{i,T})$ and $t = (1, 2, ..., T)$.

Following Mundlak (1978), Chamberlain (1984) and Wooldridge (1995), write k_i as a linear projection onto the time averages of $\mathbf{z}_i$, denoted $\bar{\mathbf{z}}_i$, a constant as well as an error a_i. Then, (6) can be rewritten as:

$$S^*_{i,t} = \theta_0 + \bar{\mathbf{z}}_i\boldsymbol{\theta} + \mathbf{z}_{i,t}\boldsymbol{\gamma} + v_{i,t}, \tag{7}$$

where the composite error term $v_{i,t} = a_i + e_{i,t}$ is independent of $\mathbf{z}_i$ and allowed to be heterogeneously distributed over time and there are no restrictions imposed on the correlation between $v_{i,t}$ and $v_{i,s}$ for $s \neq t$.

Two assumptions concerning the wage equation (Wooldridge 1995 and 2002) ensure that no restrictions are imposed on how $u_{i,t}$ relates to $v_{i,s}$, $s \neq t$.[5] First, $u_{i,t}$ is a linear function of $v_{i,t}$ and mean independent of $\mathbf{z}_i$ conditional on $v_{i,t}$. Second, similar to the selection equation, the unobserved effect is modeled as a projection of c_i onto ($\bar{\mathbf{x}}_i$, $\bar{\mathbf{y}}^*_i$, $v_{i,t}$) and an error term b_i.[6] This method specifically models

[5] Dustmann and Rochina-Barrachina (2007) call this condition "contemporaneous exogeneity" of the selection process.

[6] It should be noted that this assumption is rather restrictive, as it allows only for time-invariant unobserved effects to be correlated with the explanatory variables in equation (4). For time-variant latent variables Wooldridge's (1995) estimator may thus be inconsistent.

the unobserved effect such that correlation between c_i and $(\mathbf{x}_i, \mathbf{y}_i^*, v_{i,t})$ is possible. Under these assumptions, equation (4) can be rewritten as:

$$w_{i,t}^* = \varphi_0 + \bar{\mathbf{x}}_i\boldsymbol{\varphi}_1 + \mathbf{x}_{i,t}\boldsymbol{\beta}_1 + \bar{\mathbf{y}}_i^*\boldsymbol{\varphi}_2 + \mathbf{y}^*{}_{i,t}\boldsymbol{\beta}_2 + \xi_t\lambda_{i,t} + r_{i,t}, \tag{8}$$

where $r_{i,t} = b_i + l_{i,t}$ and $l_{i,t}$ is the remaining part of u_i after including the inverse Mills ratios (IMRs). The IMRs $\lambda_{i,t}$ are obtained by estimating equation (7) with standard probit methods for each t. Since $s_{i,s}$ $(s \neq t)$, does not influence $\lambda_{i,t}$, the error term $r_{i,t}$ is allowed to be correlated with $\lambda_{i,s}$. Equation (8) (with $\lambda_{i,t}$ replaced by $\hat{\lambda}_{i,t}$) can therefore be consistently estimated by pooled OLS. We follow Wooldridge (1995) and construct standard errors robust to serial correlation and heteroscedasticity which are also adjusted for the additional variation introduced by the estimation of T probit models in the first step.

While estimation of equation (8) assumes (strict) exogeneity of the explanatory variables, Semykina and Wooldridge (2006) provide an estimation method based on Wooldridge (1995) that allows for **endogeneity in the presence of unobserved heterogeneity and sample selection**: analogous to the above derivations, the starting point is the model in equations (4),(5) and (6). Presume, however, that the health variable (as part of $\mathbf{x}_{i,t}$ in equation (4)) is correlated with $u_{i,t}$. As it stands, health is part of $\mathbf{z}_{i,t}$ but at the same time $u_{i,t}$ must not be correlated with $\mathbf{z}_{i,t}$. Hence, the health variable is removed from $\mathbf{z}_{i,t}$ and replaced by a proxy for health which exhibits no correlation with $u_{i,t}$ and can thus serve as an additional exclusion restriction in the participation equation. The resulting $1 \times G$ vector is denoted $\mathbf{q}_{i,t}$ and its time averages $\bar{\mathbf{q}}_i$ and $\mathbf{q}_i$ itself also replace $\bar{\mathbf{z}}_i$ and $\mathbf{z}_{i,t}$ in (7).

An estimator that allows $v_{i,t}$ in (7) to be correlated with $u_{i,t}$ and c_i in (4) when the health variable is endogenous can be obtained by maintaining the assumptions underlying equation (8) and replacing $\bar{\mathbf{x}}_i$ with $\bar{\mathbf{q}}_i$. Thus, analogous to (8) we can write:

$$w_{i,t}^* = \varphi_0 + \bar{\mathbf{q}}_i\boldsymbol{\varphi}_1 + \mathbf{x}_{i,t}\boldsymbol{\beta}_1 + \bar{\mathbf{y}}_i^*\boldsymbol{\varphi}_2 + \mathbf{y}^*{}_{i,t}\boldsymbol{\beta}_2 + \xi_t\lambda_{i,t} + r_{i,t}. \tag{9}$$

Again, the first step is to estimate T standard probit models, and calculate the IMRs $\hat{\lambda}_{i,t}$. Because $r_{i,t}$ is allowed to be correlated with $\lambda_{i,s}$ for $s \neq t$ (i.e. $\lambda_{i,t}$ is not strictly exogenous in (9)), a consistent way of estimating (9) is pooled 2SLS, where $1, \bar{\mathbf{q}}_i, \mathbf{q}_{i,t}, \bar{\mathbf{y}}_i^*, \mathbf{y}_{i,t}^*, \hat{\lambda}_{i,t}$ serve as (their own) instruments. Standard errors robust to serial correlation and heteroscedasticity are calculated as suggested by Semykina and Wooldridge (2006). They are adjusted for the additional variation introduced by the estimation of T probit models in the first step and they also account for the use of the pooled 2SLS estimator.

4 Data and Descriptives

The data used in this analysis is taken from twelve consecutive annual waves of the German Socio-Economic Panel Study (GSOEP), provided by the German Institute for Economic Research (DIW). The GSOEP, which is representative of the German population, started in 1984 with about 12,200 observations from the western German states. In June 1990, another 4,400 individuals living in the territory of the former German Democratic Republic were added in order to expand the GSOEP to the eastern part of Germany.

4.1 Sample Construction

For the empirical analysis, we use observations from all sub-samples between 1995 and 2006, with the exception of samples G ("Oversampling of High Income Households") and H ("Refreshment 2006").[7] We extract data on the variables described in tables 6 and 7 in the appendix. The sample is constrained to persons older than 17 and younger than 66 years. Also excluded are those who are self-employed, self-employed in the agricultural sector, work in the family business, are on maternity leave, drafted for mandatory military or civilian service as well as individuals who serve an apprenticeship, trainees, interns, volunteers, aspirants, pensioners, and those still in education. Marginally or irregularly employed persons are also removed from the estimation sample. Motivated by two arguments, we choose to exclude (severely) handicapped people from the analysis, too. First, firms may discriminate against handicapped individuals, irrespective of their productivity. Hence, their wages may be artificially low or they might even drop out of the labor market due to discrimination, which is not meant to be captured in the selection equation. Secondly, in Germany severely handicapped people often work at special "sheltered workshops" where they are not paid according to their marginal productivity.

Hourly wages are derived by dividing gross individual earnings in the month before the interview by 4.3 (the average number of weeks per month) and then dividing the resulting weekly wage by the usual working time per week.[8] Any extra salaries like Christmas or holiday bonuses, 13th monthly pay, or child benefits are not taken into account. Suspiciously high or low wage rates were manually checked 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 equivalents.[9]

Participation in the labor market is constituted by having worked for pay in the month before the interview. In the participation equations both working and non-working adults are used for estimation. Since the econometric approach includes linear probability models, which exploit within transformations, individuals who appear for only one year are removed from the estimation sample.

[7] 1995 is chosen as starting point because the key variable "number of doctor visits" is not available in 1994. Sub-samples A through D constitute the "base data", sub-samples E and F are refreshment samples, which start in 1998 and 2000, respectively. The 2006 refreshment sample H is excluded, because by definition every person in this sample is observable for only one year.

[8] Usual hours are chosen due to their invariance to short term health problems. Including the effects of short term health issues on hours may bias hourly wages upwards as paid sick days are common practice in Germany. Contractual hours are used instead of usual working time whenever the former exceed the latter.

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

Table 8 shows how the stepwise exclusion of different groups leads to an estimation sample of 9,277 females and 8,847 males, resulting in 57,203 and 57,419 observations, respectively. For the estimation of the wage equations, persons who participate in the labor market for only one year are dropped from the sample. Due to this restriction and because individuals with missing wages who declare participation are defined as participating in the selection equations, the number of observations in the wage equations differs from the working population in the probit sample.

In the time period considered, about 69% of the female and around 86% of the male sample population participate in the labor market and male real hourly wages are on average about 0.22 log points higher than those of women. Tables 13 and 14 in the appendix compare variables in the participation equations for working and non-working individuals, tables 15 and 16 provide detailed summary statistics for variables used in the wage equations.

4.2 Health Variable

The GSOEP health measure asks individuals to state how satisfied they currently are with their health on a categorial scale ranging from zero to ten. As the functional form is a priori unclear, three specifications of the health variable are employed in order to gain insight into the relationship between health and participation/wages: (i) without any further transformations, implying a log-linear relationship, (ii) using a log-log model, as suggested by equation (2)[10] and (iii) splitting health satisfaction into four dummy variables, thus producing a flexible nonlinear specification.[11] Table 10 in the appendix shows results of these preliminary regressions for both the wage and participation equations.[12]

The coefficients of the health variable(s) turn out to be significantly different from zero in all specifications, for both women and men, and in the wage and participation equations. An important observation is that health satisfaction affects wages and labor market participation nonlinearly. This becomes evident in both the log-log and the dummy specification. In the latter, throughout the categories excellent, good, and medium health an increasing effect of health satisfaction at a diminishing rate is revealed. For example, in the case of female labor supply, reducing health from excellent to good has a much smaller effect (0.001) than reducing it further to medium health (0.021). Equally stated, diminishing health from excellent to good in the male sample affects wages less strongly (0.006) than reducing it from good to medium (0.03).

[10] Health satisfaction is transformed as follows: $g(h_{i,t}) = \log(h_{i,t} + \sqrt{(h_{i,t}^2 + 1)})$, which is a parallel translation of the log function, where $g(h_{i,t} = 0) = 0$.

[11] According to the frequency distribution in the appendix (table 9), we define poor (cat. $0-4$), medium (cat. $5-6$), good (cat. $7-8$), and excellent health (cat. $9-10$), where the first one serves as basis category.

[12] To make parameters directly interpretable, we employ linear probability models to estimate the participation equations in columns (4), (5), and (6). In all specifications further explanatory variables (see tables 3, 4 and 11, 12) are included but not reported.

Based on these results and given that almost 90% of the observations are allocated over the categories excellent, good, and medium health (see table 9), the health measure should exhibit some kind of nonlinear specification, where wages and the probability to work increase with health at diminishing rates. For pragmatic reasons, instead of choosing the more flexible dummy variable specification, we decide to rely on the log-log structure. First, its functional form most closely approximates the model suggested in equation (2). Second, only one instrument is needed when implementing the log-log form, which is especially important for the IV-approaches to the participation equations in section 5.1. Finally, it still allows for increasing returns to health at a decreasing rate – the relevant functional form for 90% of all observations.

The observed mean of this log-health variable for working females between 1995 and 2006 is 2.579, while the value for non-working women is smaller at 2.482 log points. For males, the working to non-working health ratio is 2.594 to 2.40. The hypothesis of the equality of means between the working and non-working group can be rejected on the basis of two standard t-tests, t = 25.22 (p-value = 0) for females and t = 39.73 (p-value = 0) for males.

5 Empirical Results

5.1 Participation Equations

Health is expected to influence the decision to participate in the labor market as well as wages. Thus, in order to gain insight on the extensive margin, tables 11 and 12 in the appendix present estimation results for the Mundlak-type specification needed for the Wooldridge (1995) and Semykina and Wooldridge (2006) estimators as well as five additional specifications. The exclusion restrictions we propose are: non-labour income, a binary variable for having a partner, partner's net wage and second degree polynomials of the partner's age, labor market experience and education as well as an indicator variable for whether the partner variables were missing though the presence of a partner is reported.

As a means of coping with the possible endogeneity of health in the participation equation we employ computationally undemanding (FE-)IV linear probability specifications in columns (3) and (4). Here, the number of doctor visits in the last three months serves as an instrument for the health variable.[13] The intuition is that "doctor visits" approximate past investment and depreciation in health and account for past shocks affecting current health satisfaction. At the same time "doctor visits" should not have an effect on wages other than through health status.[14] Columns (1) and (2) display pooled OLS and within results to allow a check

[13] For an example of how an endogenously reported health measure may affect wages see Stern (1989). In his paper he uses symptoms or diseases as instruments for endogenously reported disability and labour force participation.

[14] One issue our instrument probably doesn't resolve is that people may justify non-participation in the labor market by reporting low health, such that there is actually an omitted variable, say, "motivation". If these individuals visit physicians in order to justify their non-participation in the

of the IV specifications against naïve estimators.

The estimated coefficients of the health variable turn out to be significantly different from zero for both women and men and in all four linear specifications. Comparing the parameters in columns (3), (4) with (1), (2) shows that, as is expected in the presence of measurement error, the coefficients of health satisfaction using IV methods are larger than those in the pooled OLS or within model.[15] On the other hand, the inclusion of unobserved effects reduces the estimated parameters in columns (2) and (4) in comparison to (1) and (3), i.e. correlation between the health variable and latent individual heterogeneity is associated with an upward bias.

Column (5) provides a pooled probit model, which assumes that the explanatory variables are independent of any unobserved effect.[16] Column (6) applies the Mundlak specification – as laid out in section 3 – to the pooled sample. Based on the above mentioned hints that "health satisfaction" may be endogenous in the selection equations, in the pooled probit (5) and Mundlak-type (6) specifications the possibly endogenous "health satisfaction" variable is replaced by "number of doctor visits" which we assume to be exogenous and which reflects health satisfaction. Thus, the "doctor visits" variable effectively serves as an additional exclusion restriction, increasing their total number to eleven. This procedure follows Semykina and Wooldridge (2006) and Dustmann and Rochina-Barrachina's (2007) method. It is strictly necessary for the Semykina and Wooldridge (2006) estimator and also applied to the pooled probit estimator in order to enable comparison. In line with columns (1) through (4), a higher number of doctor visits (i.e. lower health) is associated with a significantly lower probability of participation in both probit specifications.

As coefficients in linear and nonlinear models cannot readily be compared, table 1 provides participation probabilities of "average" individuals, which differ only with respect to their state of health (actually, they differ only with respect to the mean values of health/doctor visits within each of the four health categories poor, medium, good, excellent; see section 5.1). For a healthy woman the pooled probit probability of participation (column (5)) is 13 percentage points higher than for a female of poor health. Controlling for correlated individual effects (column (6)) reduces the probability difference to a mere 1.5 percentage points. The linear specifications in columns (1) and (2) reveal the same pattern: When applying pooled OLS the probability to work is about 11 percentage points higher for healthy than for unhealthy women; the gap shrinks to 2 percentage points when implementing the within transformation. Columns (3) and (4) display the instrumental variables

same fashion, the instrument may be invalid. However, as long as physicians do not issue sick notes to people who are healthy, there is really no reason to arrange such appointments. Additionally, as long as "motivation" is time invariant, the individual effects in column (6) should take care of the problem.

[15] Heteroskedasticity robust, regression based Hausman tests in the spirit of Wooldridge (2002) confirm systematic differences between the health coefficients in columns (3), (4) and (1), (2).

[16] In columns (5) and (6), a robust variance covariance matrix accounts for the fact that observations are correlated within individuals over time. Under more restrictive assumptions, 'traditional' random effects probit estimation is possible; results for these models are available on request.

estimates. Again, the fixed effects approach reduces the probability gap; however, the magnitude of the gaps is larger than without controlling for endogeneity.

Table 1: Participation Probabilities (in %), by different Health Groups, Women and Men, 1995-2006

	Women					
	OLS	Within	2SLS	FE-2SLS	Probit	Mundlak Pr.
	(1)	(2)	(3)	(4)	(5)	(6)
Poor health	61.6	67.3	54.1	55.8	71.4	72.7
Medium health	67.4	68.5	66.0	66.3	73.2	73.6
Good health	70.3	69.1	71.8	71.5	74.0	74.0
Excellent health	72.1	69.4	75.6	74.8	74.4	74.2

	Men					
	OLS	Within	2SLS	FE-2SLS	Probit	Mundlak Pr.
	(1)	(2)	(3)	(4)	(5)	(6)
Poor health	76.9	83.3	71.1	74.9	93.9	94.6
Medium health	84.2	85.5	83.0	83.8	94.7	94.9
Good health	87.7	86.5	88.8	88.1	95.0	95.0
Excellent health	90.0	87.2	92.6	90.9	95.1	95.1

Source: GSOEP 1995-2006, own calculations. Participation probabilities are based on different binary choice models (see tables 11 and 12 in the appendix). $57,203$ observations from $9,277$ female persons and $57,419$ observations from $8,847$ male individuals. Except for health satisfaction in columns (1)-(4) and doctor visits in columns (5) and (6), probabilities are accounted at the mean values of all covariates. The state of health is defined as: poor (cat. $0-4$), medium (cat. $5-6$), and good (cat. $7-8$), and excellent (cat. $9-10$) health.

The male probit estimates show the probability difference between healthy and unhealthy individuals to vary between 1 percentage point when the pooled probit estimator is considered and 0.5 percentage points when controlling for the interaction between individual effects and the health variable. In the linear specifications, the corresponding values (columns (1) and (2)) are around 13 and 4 percentage points, respectively. Finally, allowing for the endogeneity of health satisfaction expands the probability gap to 22 and 16 percentage points, respectively.

Results for most of the other variables are as expected (see tables 11 and 12 in the appendix). For both women and men the participation probability increases with age (at a decreasing rate) and education.[17] Living in the eastern part of Germany is associated with lower participation for men, while the effect is positive for women (the female population in the eastern region also has a higher participation probability than their western counterparts, probably rooted in the socialist past). Being of non-German origin and the amount of non labor income has a negative influence on the probability of labor market participation. An increasing labor market attachment of the partner tends to reduce the probability to work for women and has a tendency to increase the participation probability in the male population, yet some of the partner and children variables exhibit the same sign for women and men, which means that the effects remain somewhat ambiguous overall. For both sexes, the number of children in different age categories mostly reduce the individuals' labor market attachment and the partner's net wage is associated with a decreasing working probability in most specifications for both females and males.

[17] For women, in some of the linear specifications the probability to work decreases with age.

5.2 Wage Equations

Since the core interest of this study is the estimation of the wage equation (eq. 3), results for six different estimation methods are given in tables 3 and 4. Columns (1) through (3) in each table display results for OLS, FE and Wooldridge's (1995) estimator, all of which assume health to be exogenously determined. In both tables, endogeneity of health is allowed for in the pooled 2SLS (4) and FE-2SLS (5) specifications as well as in Semykina and Wooldridge's (2006) estimator (6).

The Instruments. For specifications (4) through (6) the set of instruments consists of all eleven variables which serve as exclusion restrictions in the participation equations (including "doctor visits", see section 5.1). 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. Overidentification tests strongly reject the null hypotheses of no correlation between the instruments and the error of the wage equation for both sexes in the pooled IV and FE-2SLS estimations (columns (4) and (6)). When testing for overidentifying restrictions in Semykina's and Wooldridge's (2006) framework, however, *no correlation* between the instruments and the error in the wage equation is detected. This is in line with Semykina (2007), who shows that if instruments enter the selection equation, "[...] they will be inevitably correlated with [...]," the error term of the selected sample. Consequently, if a selection bias exists – which is the case here (see table 2) – overidentification tests will detect endogeneity of the exclusion restrictions. Thus, rejecting the null hypothesis in the pooled IV and FE-2SLS approach is just another way of stating that selection bias is present.

The Selection Effects. A preliminary check for the presence of selection bias can be carried out by Wald tests on the joint significance of the Inverse Mills Ratios (table 2). In columns (1) and (2) we follow Wooldridge (1995) and conduct "variable addition" tests, as first proposed by Verbeek and Nijman (1992). It is assumed that no further endogeneity problems occur and under the null the standard within estimator is valid. In columns (3) and (4) tests in the spirit of Semykina and Wooldridge (2006) are carried out, where the null hypothesis suggests to use the FE-2SLS estimator. For women and men alike the null hypothesis is strongly rejected and this evidence of selection bias in both the FE and the FE-2SLS framework indicates that use of the methods introduced in section 3 is in order.[18]

Health and Wages. While good health significantly increases participation for both men and women, the impact of health on the wage rate differs quite a bit

[18] For both women and men, the inverse Mills ratios are negatively correlated with wages in most years (coefficients not reported). Since the IMRs are inversely related to the estimated probabilities of being employed, the negative coefficients indicate that a higher participation probability is associated with an above average salary.

Table 2: IMR Tests, Women and Men, 1995-2006

	Within[a)]		FE-2SLS[b)]	
	Male	Female	Male	Female
Wald-test, χ^2_{12} =	139.80	44.52	131.68	44.11
P-values	0.000	0.000	0.000	0.000
N	47,746	37,670	47,746	37,670

Source: GSOEP 1995-2006, own calculations. Within and FE-2SLS estimation. 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 estimators are valid. *b)* Wald tests on the joint significance of the IMRs are provided. Under the null hypothesis the FE-2SLS estimators are valid.

across genders. For males (table 3), the parameter of the health variable using pooled OLS (0.041) is higher than the coefficient in the fixed effects model (0.013). Both effects are significantly different from zero at the 1% level. Controlling for selection lowers the significance level to 5% and reduces the coefficient even further (0.011), but the differences between the FE and the Wooldridge (1995) estimator are practically small. This suggests that using the FE estimator already accounts for most of the bias introduced by the correlation between the health variable and unobserved individual heterogeneity. 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 their 2SLS counterparts in columns (4), (5), and (6) which is to be expected if self-assessed health is error-ridden. Within the instrumental variable framework, the (significantly estimated) parameters again exhibit substantial differences. Using pooled 2SLS is associated with a coefficient of 0.046, whereas implementing FE-2SLS yields the highest parameter of 0.062. Though less precisely estimated, controlling for selection scales the health coefficient down to 0.041. For the Mundlak-type estimators in columns (3) and (6), a Wald of the joint significance of the unobserved individual effects is carried out and in both cases indicates correlated individual effects. Selection tests, where now the assumptions under the null hypothesis are more restrictive than those underlying the tests in table 2 again reject the null of no selection effects in columns (3) and (6). Finally, endogeneity tests show systematic differences between the health coefficients in columns (2) and (5).

The same six econometric models using the female sample are presented in table 4. The results, however, are less intuitive than in the male sample. As with men, selection corrections are indicated by Wald tests on the joint significance of the IMRs for the models in columns (3) and (6). In these specifications Wald tests confirm the presence of correlated individual effects just as in the male sample, whereas endogeneity tests suggest that the health variable is exogenous in columns (1), (2), and (3). Throughout all specifications, only pooled OLS points to a significant effect of health for females. Therefore, summarizing the above, it seems that for women health has only a negligible effect on wages (intensive margin), though there exists a significant effect on labour market participation (extensive level).

In an attempt to give an idea of the economic significance of the above results and in order to facilitate comparison of the various estimators, table 5 provides

Table 3: Wage equations, Men, 1995-2006

	OLS$^{a)}$	Within$^{a)}$	Wooldr95$^{c)}$	2SLS$^{b)}$	FE-2SLS$^{b)}$	SemWool06$^{d)}$
	(1)	(2)	(3)	(4)	(5)	(6)
Log. health sat.	0.041 (0.004)***	0.013 (0.004)***	0.011 (0.005)**	0.046 (0.014)***	0.062 (0.02)***	0.041 (0.024)*
Age	0.083 (0.007)***	.	.	0.083 (0.007)***	.	.
Age sqare	-.002 (0.0002)***	-.002 (0.0002)***	-.002 (0.0003)***	-.002 (0.0002)***	-.002 (0.0002)***	-.002 (0.0003)***
Age triple	1.00e-05 (1.29e-06)***	1.00e-05 (1.72e-06)***	1.00e-05 (2.21e-06)***	1.00e-05 (1.29e-06)***	1.00e-05 (1.73e-06)***	1.00e-05 (2.22e-06)***
Unempl. exp.	-.047 (0.003)***	-.105 (0.012)***	-.096 (0.017)***	-.047 (0.003)***	-.106 (0.012)***	-.097 (0.017)***
Unempl. exp. sq.	0.003 (0.0004)***	0.004 (0.002)*	0.004 (0.003)	0.003 (0.0004)***	0.004 (0.002)*	0.004 (0.003)
Firm tenure	0.014 (0.0005)***	0.004 (0.0008)***	0.005 (0.001)***	0.014 (0.0005)***	0.005 (0.0008)***	0.005 (0.001)***
Firm tenure sq.	-.0002 (1.00e-05)***	-.0001 (0.00002)***	-.0001 (0.00003)***	-.0002 (1.00e-05)***	-.0001 (0.00002)***	-.0001 (0.00003)***
Education	0.034 (0.0008)***	.	.	0.034 (0.0008)***	.	.
Du. education	-.023 (0.004)***	-.006 (0.004)	-.003 (0.005)	-.023 (0.004)***	-.006 (0.004)	-.003 (0.005)
Part-time	-.115 (0.016)***	-.038 (0.019)**	-.025 (0.022)	-.115 (0.016)***	-.036 (0.019)*	-.025 (0.022)
Foreigner	-.002 (0.005)	.	.	-.002 (0.005)	.	.
State level variables						
Log. unempl. (fed. st.)	-.115 (0.006)***	-.017 (0.017)	-.005 (0.021)	-.115 (0.006)***	-.015 (0.017)	-.005 (0.021)
Log. vac. (fed. st.)	-.040 (0.008)***	-.003 (0.007)	-.004 (0.009)	-.040 (0.008)***	-.002 (0.007)	-.004 (0.009)
Log. empl. (fed. st.)	0.171 (0.01)***	0.025 (0.018)	0.017 (0.023)	0.171 (0.01)***	0.023 (0.018)	0.018 (0.023)
East Germany	-.221 (0.006)***	-.041 (0.01)***	-.037 (0.012)***	-.221 (0.006)***	-.041 (0.01)***	-.037 (0.012)***
Number of children						
≤ 2 years of age	0.036 (0.005)***	0.013 (0.005)**	0.01 (0.006)*	0.036 (0.005)***	0.013 (0.005)***	0.01 (0.006)*
$3-5$ years of age	0.036 (0.004)***	0.018 (0.005)***	0.015 (0.006)***	0.036 (0.004)***	0.019 (0.005)***	0.015 (0.006)***
$6-16$ years of age	0.013 (0.003)***	0.002 (0.003)	0.0006 (0.004)	0.013 (0.003)***	0.002 (0.003)	0.0009 (0.004)
Du. num. child.	-.018 (0.005)***	-.008 (0.006)	-.008 (0.007)	-.017 (0.005)***	-.007 (0.006)	-.008 (0.007)
Firm size (base cat.: <20 employees						
$20-199$	0.087 (0.005)***	0.046 (0.006)***	0.036 (0.008)***	0.087 (0.005)***	0.045 (0.006)***	0.036 (0.008)***
$200-1,999$	0.153 (0.005)***	0.059 (0.008)***	0.047 (0.009)***	0.153 (0.005)***	0.058 (0.008)***	0.047 (0.009)***
≥ 2000	0.192 (0.005)***	0.067 (0.008)***	0.055 (0.01)***	0.192 (0.005)***	0.066 (0.008)***	0.054 (0.01)***
Firm size miss.	0.08 (0.017)***	0.022 (0.017)	0.037 (0.019)*	0.08 (0.017)***	0.021 (0.017)	0.038 (0.019)**
Constant	-.146 (0.106)	.	.	-.159 (0.113)	.	.
N	47,746	47,746	47,746	47,746	47,746	47,746
D.f.	47,695	40,020	47,651	47,695	40,020	47,641
Wald tests on the joint significance of						
12 IMRs	.	.	83.04***	.	.	64.21***
11 time dummies	349.62***	266.15***	138.27***	349.71***	265.53***	137.37***
6 occ. dummies	2709.41***	17.80***	728.22***	2685.21***	17.97***	687.12***
9 sector dummies	1369.84***	65.81***	398.07***	1372.60***	66.01***	405.00***
Unobs. effects^{e}	.	.	1080.87***	.	.	1146.49***

Source: GSOEP 1995-2006, 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)* Standard errors are are robust to serial correlation and heteroscedasticity; *b)* robust standard errors as in *a)*, but the 2SLS estimator is used and accounted for; *c)* robust standard errors as in *a)*, but the variation introduced by the probit first-stage estimation is accounted for; *d)* robust standard errors as in *c)*, but the 2SLS estimator is used and accounted for; *e)* χ^2 test statistics for the joint significance of 35 variables (vector $\bar{\mathbf{x}}_i$) or 45 variables (vector $\bar{\mathbf{q}}_i$) are reported.

Table 4: WAGE EQUATIONS, WOMEN, 1995-2006

	OLS$^{a)}$	Within$^{a)}$	Wooldr95$^{c)}$	2SLS$^{b)}$	FE-2SLS$^{b)}$	SemWool06$^{d)}$
	(1)	(2)	(3)	(4)	(5)	(6)
Log. health sat.	0.023 (0.005)***	0.007 (0.005)	0.005 (0.005)	0.008 (0.014)	0.03 (0.022)	0.007 (0.024)
Age	0.08 (0.008)***	.	.	0.08 (0.008)***	.	.
Age sqare	-.002 (0.0002)***	-.002 (0.0003)***	-.002 (0.0003)***	-.002 (0.0002)***	-.002 (0.0003)***	-.002 (0.0004)***
Age triple	9.57e-06 (1.60e-06)***	1.00e-05 (2.25e-06)***	1.00e-05 (2.69e-06)***	9.59e-06 (1.60e-06)***	1.00e-05 (2.25e-06)***	1.00e-05 (2.85e-06)***
Unempl. exp.	-.029 (0.002)***	-.091 (0.017)***	-.088 (0.021)***	-.030 (0.002)***	-.091 (0.017)***	-.088 (0.021)***
Unempl. exp. sq.	0.001 (0.0002)***	0.002 (0.003)	0.002 (0.004)	0.001 (0.0002)***	0.002 (0.003)	0.002 (0.004)
Firm tenure	0.016 (0.0007)***	0.003 (0.001)***	0.003 (0.001)**	0.016 (0.0007)***	0.003 (0.001)***	0.003 (0.001)**
Firm tenure sq.	-.0002 (0.00002)***	-.00007 (0.00003)**	-.00009 (0.00004)**	-.0002 (0.00002)***	-.00007 (0.00003)**	-.00009 (0.00004)**
Education	0.042 (0.0009)***	.	.	0.042 (0.0009)***	.	.
Du. education	-.030 (0.005)***	-.013 (0.006)**	-.010 (0.008)	-.030 (0.005)***	-.012 (0.006)**	-.010 (0.008)
Part-time	-.047 (0.004)***	0.018 (0.007)**	0.025 (0.008)***	-.047 (0.004)***	0.018 (0.007)**	0.025 (0.008)***
Foreigner	0.014 (0.006)**	.	.	0.014 (0.006)**	.	.
State level variables						
Log. unempl. (fed. st.)	-.101 (0.007)***	0.0008 (0.018)	0.018 (0.023)	-.101 (0.007)***	0.002 (0.018)	0.017 (0.024)
Log. vac. (fed. st.)	-.041 (0.008)***	-.003 (0.009)	0.002 (0.01)	-.040 (0.008)***	-.003 (0.009)	0.002 (0.012)
Log. empl. (fed. st.)	0.133 (0.012)***	0.031 (0.022)	0.005 (0.029)	0.132 (0.012)***	0.031 (0.022)	0.006 (0.03)
East Germany	-.190 (0.007)***	-.039 (0.013)***	-.037 (0.014)**	-.191 (0.007)***	-.039 (0.013)***	-.037 (0.014)**
Number of children						
≤ 2 years of age	0.047 (0.017)***	-.013 (0.017)	0.031 (0.021)	0.047 (0.017)***	-.013 (0.017)	0.031 (0.022)
$3-5$ years of age	0.025 (0.009)***	-.009 (0.01)	0.015 (0.013)	0.024 (0.009)***	-.008 (0.01)	0.014 (0.014)
$6-16$ years of age	-.006 (0.005)	-.012 (0.007)*	-.004 (0.009)	-.006 (0.005)	-.012 (0.007)*	-.004 (0.009)
Du. num. child.	-.011 (0.008)	0.008 (0.01)	0.007 (0.012)	-.011 (0.008)	0.008 (0.01)	0.007 (0.013)
Firm size (base cat.: <20 employees						
$20-199$	0.088 (0.005)***	0.031 (0.007)***	0.026 (0.009)***	0.088 (0.005)***	0.031 (0.007)***	0.026 (0.009)***
$200-1,999$	0.139 (0.005)***	0.052 (0.008)***	0.044 (0.01)***	0.138 (0.005)***	0.052 (0.008)***	0.044 (0.01)***
≥ 2000	0.177 (0.006)***	0.055 (0.009)***	0.04 (0.011)***	0.177 (0.006)***	0.054 (0.009)***	0.04 (0.011)***
Firm size miss.	0.102 (0.02)***	0.04 (0.017)**	0.071 (0.021)***	0.102 (0.02)***	0.039 (0.017)**	0.072 (0.021)***
Constant	0.08 (0.12)	.	.	0.122 (0.125)	.	.
N	37,670	37,670	37,670	37,670	37,670	37,670
D.f.	37,619	31,063	37,575	37,619	31,063	37,565
Wald tests on the joint significance of						
12 IMRs	.	.	53.52***	.	.	44.50***
11 time dummies	142.69***	203.37***	110.07***	143.24***	203.97***	96.64***
6 occ. dummies	2293.65***	23.41***	705.71***	2292.66***	23.37***	664.39***
9 sector dummies	537.381***	25.40***	162.58***	538.35***	25.45***	162.13***
Unobs. effects^{e}	.	.	985.57***	.	.	1024.72***

Source: GSOEP 1995-2006, 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)* Standard errors are are robust to serial correlation and heteroscedasticity; *b)* robust standard errors as in *a)*, but the 2SLS estimator is used and accounted for; *c)* robust standard errors as in *a)*, but the variation introduced by the probit first-stage estimation is accounted for; *d)* robust standard errors as in *c)*, but the 2SLS estimator is used and accounted for; *e)* χ^2 test statistics for the joint significance of 35 variables (vector $\bar{\mathbf{x}}_i$) or 45 variables (vector $\bar{\mathbf{q}}_i$) are reported.

predicted wages of four "average" individuals, who differ only in their state of health.[19] For a male in excellent health, pooled OLS predicts real wages to be about 5% (0.62 Euro) higher than for a male person suffering from poor health.[20] Accounting for individual heterogeneity in column (2) reduces the wage gap to about 0.19 Euro or 1.5%. Whenever non-random selection into the work force (*Wooldr95*) is additionally considered, hourly wages decline on each health level and the wage differential in column (3) shrinks to 1.3 percentage points (0.15 Euro) when compared to the within estimator. Predictions are slightly different when implementing instrumental variable techniques. The wage gap between individuals who are highly satisfied with their health status and those suffering from poor health is about 0.69 Euro (5.7%) in column (4) and 0.90 Euro (7.8%) in column (5). Again, wages are reduced in each health group whenever sample selection is accounted for. Here, the wage premium for being in excellent health is estimated to be around 0.56 Euro (5.1%).

Table 5: Wage Predictions (per hour), by different Health Groups, Women and Men, 1995-2006

	Men					
	OLS	Within	Wooldr95	TSLS	FETSLS	Wooldr05
	(1)	(2)	(3)	(4)	(5)	(6)
Poor health	12.137	12.115	11.162	12.086	11.612	10.967
Medium health	12.475	12.218	11.246	12.463	12.104	11.274
Good health	12.641	12.268	11.286	12.650	12.350	11.425
Excellent health	12.752	12.301	11.313	12.774	12.514	11.526

	Women					
	OLS	Within	Wooldr95	TSLS	FETSLS	Wooldr05
	(1)	(2)	(3)	(4)	(5)	(6)
Poor health	9.870	8.882	5.779	9.993	8.712	5.782
Medium health	10.020	8.925	5.799	10.045	8.891	5.811
Good health	10.095	8.947	5.809	10.071	8.980	5.826
Excellent health	10.144	8.960	5.815	10.088	9.038	5.835

Source: GSOEP 1995-2006, own calculations. Predicted hourly wages (in Euros, deflated to unity at year end 2001) based on different wage equations (see tables 3 and 4). 47,746 observations from 7,679 male persons and 37,670 observations from 6,560 female individuals. Except for health satisfaction and the IMRs, wages are accounted at the mean values of all explanatory variables. The state of health is defined as: poor (cat. $0-4$), medium (cat. $5-6$), and good (cat. $7-8$), and excellent (cat. $9-10$) health.

The differences in real hourly wages between healthy and unhealthy females range from 0.6% to 3.7%, conditional on the method of estimation. Since all estimators with the exception of pooled OLS lack sufficient precision, simulation results in the female sample must be interpreted with caution. Keeping this in

[19] Since the individuals' health status also affects the probability of participating, we calculate "mean" IMRs (over time and person), which differ with respect to the corresponding health groups. Hence, the simulations in table 5 are based on mean values of all explanatory variables in the wage equations except for health satisfaction and doctor visits, respectively, as well as the inverse Mills ratios.

[20] In order to obtain hourly wages we exponentiate predicted log wages. This procedure differs somewhat from the one proposed by Kennedy (1984). However, we conduct sensitivity tests based on the pooled OLS estimates and find the differences between Kennedy's method and our predictions negligible.

mind, the decline in wages induced by accounting for non-random selection into work (columns (3) and (6)) is fairly strong in the female sample. This finding is rooted in the fact that integrating the large share of non-participating females into the workforce would drastically reduce average wages.

Other Results. Aside from our main interest in the effect of health on wages, concave wage profiles are found with respect to firm tenure in all specifications and for women and men. Given the high unemployment rates in Germany, it is of practical relevance to see that in all models past unemployment periods go with significantly lower wages (at an increasing rate), whereas education positively affects participation and comes with a rate of return per additional year of schooling of 4.2% for women and approximately 3.4% for men.

Results for most of the other variables are as expected. For both women and men wages increase at a decreasing rate with age. 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 in the federal sate negatively influences wages, whereas an increase in the number of employed raises the wage rate. Women and men working in large firms ($\geq$ 2000 employees) earn significantly more than in medium-sized firms, which in turn earn more than males and females employed in small firms. These effects persist when controlling for individual heterogeneity and selection, albeit smaller in magnitude. The number of children (especially 0 to 5 year olds) is associated with higher wages in the male sample, whereas results in the female sample are insignificant in most specifications. Finally, as for the structural factors affecting wages, we find industry and occupational wage differentials as in all models and irrespective of gender. Wald tests confirm the joint significance of six occupational and nine sector dummies at any sensible level.

6 Conclusions

In this article, we employ recently developed estimation methods by Semykina and Wooldridge (2006) in order to control for selection, individual heterogeneity and endogeneity in one comprehensive framework and apply them to the question of whether health has a causal effect on wages. A number of tests provide evidence that corrections for non-random selection into the workforce are necessary in both the female and male sample and the health variable is found to suffer from measurement error. Our results show that good health raises wages for men, while for women there appears to be no significant effect. The fact that predicted participation probabilities in the probit models are more contingent on health for the average woman than they are for the average man is in line with this finding, providing tentative evidence that for females health may mainly affect participation, while for males the effect is essentially to be found on the intensive margin.

A question for further research that immediately comes to mind is whether investment in improved health status is worthwhile at the micro or macro level. While monetary gains from being in good health are calculated in table 5, these results are not very informative from a welfare point of view, which must take into account individual utility gains and adequate cost measures. Another task for future research could be the estimation of an "all-encompassing" model which takes into account all sources of endogeneity mentioned in this article and additionally tries to address dynamic effects in the state of health.

Appendix

Table 6: 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. earnings per hour (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
Education	amount of education or training in years
Dummy education	whenever years of education or training *decrease* over time, the lower values are changed to the former maximum and the dummy education variable is set to 1
Unemployment experience	duration of unemployment in a person's career; in years, with months in decimal form
Firm tenure	duration of time with firm; in years, with months in decimal form
Log. non labour income	log. household income minus net wage income (in 2001 Euros)
No. of doctor visits	number of doctor visits in the last three months
Part-time	dummy variable indicating part-time work
Foreigner	dummy variable indicating non-German nationality
Firm size	four dummy variables indicating different firm sizes; categories: up to 20 employees ; $20-199$ employees; $200-1999$ employees; larger than 2000 employees
Occupation	seven occupation dummies, constructed using the Erikson, Goldthorpe Class Category IS88 (basis: high serv.)

(continued)

Table 7: Description of Variables (part ii)

Variable	Description
Sector	ten aggregated sector dummies, based on the NACE classification (basis: agric., forestry, fishing)
Time	eleven time dummies (1996 - 2006) (basis: 1995)
State level variables	
Log unemployment[a]	(log) yearly averages of job seekers in the individual state of residence
Log vacancies	(log) yearly average of notified vacancies (per state)
Log employed	(log) yearly average of employed persons (per state)
Dummy East Germany	dummy variables indicating where a person lives (probit equ.) or works (wage equ.); Region = 0 if Western Germany
Parent variables	
Number of children[b]	no. of children in three categories; 1) up to 2 years old; 2) between 3 - 5 years old; 3) between 6 - 16 years old
Dummy no. of children	dummy variable indicating the presence of children under the age of 17 (du. num. child. = 1 if no children present)
Partner or Spouse variables[c]	
Single	dummy variable indicating whether a person has a partner/is married (single = 1 if person has no partner)
Flag missing	dummy variable indicating missing data on partner/spouse variables (du. flag miss. = 1 if partner present but data missing)
Net wage	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] Unemployment, vacancy, and employment figures are provided by the Federal Employment Agency, Nuremberg.

[b] All children variables equal zero, if dummy no. child. = 1.

[c] All partner/spouse variables equal zero, if single = 1 or flag miss. = 1.

Table 8: Stepwise adjustment of samples (in %) and average years of individuals in sample

	Men				Women			
	Obs.	%	Inidv.	Avrg. years in sample	Obs.	%	Indiv.	Avrg. years in sample
(1) complete sample	104540	.	16037	6.52	113468	.	16997	6.68
(2) between 18 and 65 years	88194	15.64	14181	6.22	92445	18.53	14467	6.39
(3) not pensioners	81108	8.03	13378	6.06	84654	8.43	13677	6.19
(4) not in education	77163	4.86	12969	5.95	80081	5.40	13141	6.09
(5) not self-employed	69974	9.32	12249	5.71	76378	4.62	12855	5.94
(6) not on maternity leave	69922	.07	12246	5.71	72461	5.13	12726	5.69
(7) not military/civilian service	69651	.39	12219	5.70	72455	.008	12723	5.69
(8) no apprenticeship, etc.	66203	4.95	11765	5.63	69428	4.18	12316	5.64
(9) not marginally or irregularly part-time employed	64983	1.84	11601	5.60	65028	6.34	12008	5.42
(10) not in 'sheltered workshops'	64866	.18	11589	5.60	64901	.20	11997	5.41
(11) with valid information on all probit variables	57419	11.48	8847	6.49	57203	11.86	9277	6.17
(12) labour market participants	49397	.	8239	6.00	39336	.	7305	5.38
(13) with valid information on all wage equation variables$^{a)}$	47746	3.34	7679	6.22	37670	4.24	6560	5.74

Data: GSOEP, samples A-F, 1995-2006. The sample is pooled on the individual-year level.

a) For estimating earnings equations, individuals who work for only one year are dropped from the sample. Observations with missing data on wages are included in the participation equation if they report to have worked for pay in the month before the interview.

Table 9: Frequency Distribution, Health Satisfaction

	Overall Sample		Male Sample		Female Sample	
	Abs.	in %	Abs.	in %	Abs.	in %
Cat. 0	814	.71	371	.65	443	.77
Cat. 1	769	.67	371	.65	398	.70
Cat. 2	2,351	2.05	1,155	2.01	1,196	2.09
Cat. 3	4,828	4.21	2,372	4.13	2,456	4.29
Cat. 4	6,004	5.24	2,898	5.05	3,106	5.43
Cat. 5	14,718	12.84	6,840	11.91	7,878	13.77
Cat. 6	12,053	10.52	6,093	10.61	5,960	10.42
Cat. 7	21,008	18.33	10,776	18.77	10,232	17.89
Cat. 8	29,600	25.82	15,004	26.13	14,596	25.52
Cat. 9	13,767	12.01	7,038	12.26	6729	11.76
Cat. 10	8,710	7.60	4,501	7.84	4,209	7.36
All	114,622	100.00	57,419	100.00	57,203	100.00

Data: GSOEP, samples A-F, 1995-2006. Observations are on individual-year level.

Table 10: Different Functional Forms of the Health Variable

	Male Sample					
	Wage Equ.			Part. Equ.		
	Spec. (1)	Spec. (2)	Spec. (3)	Spec. (4)	Spec. (5)	Spec. (6)
Lin. health sat.	.008	.	.	.019	.	.
	(.0008)***			(.0008)***		
Log. health sat.	.	.041	.	.	.109	.
		(.004)***			(.004)***	
Medium health	.	.	.003	.	.	.080
			(.005)			(.005)***
Good health	.	.	.033	.	.	.115
			(.005)***			(.005)***
Excell. health	.	.	.039	.	.	.113
			(.006)***			(.005)***
N	47,746	47,746	47,746	57,419	57,419	57,419

	Female Sample					
	Wage Equ.			Part. Equ.		
	Spec. (1)	Spec. (2)	Spec. (3)	Spec. (4)	Spec. (5)	Spec. (6)
Lin. health sat.	.005	.	.	.015	.	.
	(.0009)***			(.0009)***		
Log. health sat.	.	.023	.	.	.087	.
		(.005)***			(.004)***	
Medium health	.	.	-.012	.	.	.073
			(.006)*			(.006)***
Good health	.	.	.013	.	.	.094
			(.006)**			(.006)***
Excell. health	.	.	.017	.	.	.095
			(.007)***			(.006)***
N	37,670	37,670	37,670	57,203	57,203	57,203

Data: GSOEP, samples A-F, 1995-2006. All specifications (including the linear probability models in columns (4), (5), and (6)) are estimated using pooled OLS. Robust standard errors are in parenthesis: * significance at ten, ** at five, and *** at one percent. The state of health in columns (3) and (6) is defined as: poor (cat. $0-4$), medium (cat. $5-6$), good (cat. $7-8$), and excellent (cat. $9-10$). Further explanatory variables are included (see tables 3, 4 and 11, 12), but not reported.

Table 11: Participation Equation, Men, 1995-2006

	OLS[a)]	Within[a)]	2SLS[a),b)]	FE-2SLS[a),b)]	Probit[c)]	Mundlak Pr.[c),d)]
	(1)	(2)	(3)	(4)	(5)	(6)
Age	.029	.	.030	.	.166	.
	(.006)***		(.006)***		(.045)***	
Age square	-.0005	.0005	-.0005	.0005	-.004	-.001
	(.0002)***	(.0002)**	(.0002)***	(.0002)**	(.001)***	(.001)
Age triple	1.36e-06	-7.46e-06	1.23e-06	-7.77e-06	.00002	-3.36e-06
	(1.20e-06)	(1.72e-06)***	(1.20e-06)	(1.73e-06)***	(9.30e-06)**	(1.00e-05)
Education	.015	.	.014	.	.109	.
	(.0006)***		(.0006)***		(.009)***	
Dummy Education	-.005	-.012	-.004	-.012	-.066	-.049
	(.004)	(.004)***	(.004)	(.004)***	(.034)*	(.034)
Foreigner	-.042	.	-.043	.	-.224	.
	(.005)***		(.005)***		(.052)***	
Lg. health sat.	.109	.032	.178	.133	.	.
	(.004)***	(.005)***	(.012)***	(.020)***		
Doctor visits	.	.	.	.	-.028	-.013
					(.002)***	(.002)***
Lg. non-lab.-inc.	-.036	-.033	-.036	-.033	-.362	-.367
	(.0004)***	(.0006)***	(.0004)***	(.0006)***	(.025)***	(.025)***
State level variables						
Lg. unempl. (fed. st.) a	-.069	-.095	-.068	-.092	-.438	-.447
	(.005)***	(.016)***	(.005)***	(.016)***	(.067)***	(.117)***
Lg. vac. (fed. st.)	.008	-.021	.006	-.021	.078	-.064
	(.007)	(.008)***	(.007)	(.008)***	(.052)	(.051)
Lg. empl. (fed. st.)	.060	.101	.060	.096	.353	.353
	(.009)***	(.019)***	(.009)***	(.019)***	(.088)***	(.144)**
East Germany	-.028	-.013	-.027	-.016	-.178	-.151
	(.006)***	(.022)	(.006)***	(.022)	(.065)***	(.161)
Number of children						
≤ 2 years of age	.003	.012	.001	.011	.062	.061
	(.005)	(.006)**	(.005)	(.006)*	(.045)	(.044)
$3-5$ years of age	-.010	.005	-.010	.005	-.053	.005
	(.004)**	(.005)	(.004)**	(.005)	(.038)	(.039)
$6-16$ years of age	-.016	-.009	-.017	-.008	-.075	-.049
	(.003)***	(.004)**	(.003)***	(.004)**	(.026)***	(.030)*
Du. num. child.	-.066	-.011	-.064	-.010	-.216	.010
	(.005)***	(.006)*	(.005)***	(.006)	(.055)***	(.048)
Partner/Spouse variables						
Single	.708	.040	.689	.026	4.420	1.196
	(.052)***	(.061)	(.052)***	(.062)	(.489)***	(1.038)
Net wage partner/spouse	-.00002	-.00005	-.00002	-.00005	-.0002	-.0004
	(3.17e-06)***	(4.63e-06)***	(3.18e-06)***	(4.69e-06)***	(.00003)***	(.00005)***
Age partner/spouse	.019	.013	.019	.013	.135	.093
	(.002)***	(.003)***	(.002)***	(.003)***	(.016)***	(.024)***
Age sq. partner/spouse	-.0002	-.0001	-.0002	-.0001	-.001	-.0009
	(.00002)***	(.00004)***	(.00002)***	(.00004)***	(.0002)***	(.0003)***
Exp. partner/spouse	-.0007	-.004	-.0009	-.004	-.002	-.012
	(.0006)	(.002)**	(.0006)	(.002)**	(.007)	(.014)
Exp. sq. partner/spouse	-.00003	.00009	-.00002	.00009	-.0002	.0002
	(.00002)	(.00005)*	(.00002)	(.00005)*	(.0002)	(.0004)
Educ. partner/spouse	.060	-.021	.058	-.023	.279	-.044
	(.006)***	(.011)*	(.006)***	(.011)**	(.060)***	(.152)
Educ. sq. partner/spouse	-.002	.0005	-.002	.0006	-.010	-.0003
	(.0002)***	(.0004)	(.0002)***	(.0004)	(.002)***	(.006)
Du. flag miss.	.847	.136	.827	.122	5.196	1.633
	(.054)***	(.060)**	(.053)***	(.061)**	(.504)***	(1.039)
Constant	-.633	.	-.819	.	-4.305	.
	(.111)***		(.114)***		(.912)***	
Time dummies χ^2_{11} =	40.7***	25.74***	41.704***	24.954***	30.269***	17.273*
Unobs. effects χ^2_{36} =	.	.	.	.	.	451.32***
LL	.	.	.	.	-16958.95	-16718.69

Source: GSOEP 1995-2006, own calculations. Different binary choice specifications. $57,419$ observations from $8,847$ 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) robust standard errors are provided; b) t-tests on the significance of the instrument in the 1st step regressions confirm that the rank condition for identification of the IV estimators is fulfilled; heteroskedasticity robust, regression based Hausman tests provide evidence for the endogeneity of the health variable on the 1% significance level.c) Standard errors are robust to serial correlation in the individual scores across t; d) unobserved effects are specified as a linear projection on the (within) means of the regressors.

Table 12: Participation Equation, Women, 1995-2006

	OLS$^{a)}$	Within$^{a)}$	2SLS$^{a),b)}$	FE-2SLS$^{a),b)}$	Probit$^{c)}$	Mundlak Pr.$^{c),d)}$
	(1)	(2)	(3)	(4)	(5)	(6)
Age	-.010 (.006)	.	-.009 (.006)	.	.097 (.041)**	.
Age square	.0007 (.0002)***	.0001 (.0002)	.0007 (.0002)***	.0002 (.0002)	-.0009 (.001)	-.002 (.001)
Age triple	-1.00e-05 (1.26e-06)***	-4.98e-06 (1.67e-06)***	-1.00e-05 (1.26e-06)***	-5.33e-06 (1.69e-06)***	-8.06e-06 (8.00e-06)	-6.01e-07 (7.87e-06)
Education	.026 (.0008)***	.	.024 (.0008)***	.	.103 (.007)***	.
Dummy Education	.002 (.005)	.006 (.005)	.005 (.005)	.006 (.005)	-.015 (.032)	.020 (.025)
Foreigner	-.094 (.006)***	.	-.093 (.006)***	.	-.289 (.044)***	.
Lg. health sat.	.087 (.004)***	.017 (.004)***	.176 (.012)***	.156 (.019)***	.	.
Doctor visits	.	.	.	.	-.021 (.002)***	-.011 (.001)***
Lg. non-lab.-inc.	-.032 (.0006)***	-.024 (.0006)***	-.032 (.0006)***	-.023 (.0006)***	-.145 (.006)***	-.112 (.004)***
State level variables						
Lg. unempl. (fed. st.) a	-.088 (.007)***	-.085 (.018)***	-.088 (.007)***	-.084 (.018)***	-.322 (.057)***	-.323 (.092)***
Lg. vac. (fed. st.)	-.048 (.008)***	-.034 (.008)***	-.051 (.008)***	-.034 (.008)***	-.173 (.051)***	-.137 (.038)***
Lg. empl. (fed. st.)	.120 (.012)***	.153 (.023)***	.123 (.012)***	.157 (.023)***	.439 (.080)***	.604 (.117)***
East Germany	.048 (.008)***	.023 (.030)	.051 (.008)***	.025 (.030)	.125 (.058)**	.064 (.165)
Number of children						
≤ 2 years of age	-.317 (.010)***	-.203 (.010)***	-.319 (.010)***	-.204 (.010)***	-.981 (.044)***	-.641 (.040)***
$3-5$ years of age	-.221 (.007)***	-.116 (.007)***	-.221 (.007)***	-.114 (.007)***	-.677 (.032)***	-.359 (.029)***
$6-16$ years of age	-.105 (.004)***	-.030 (.005)***	-.106 (.004)***	-.030 (.005)***	-.315 (.023)***	-.100 (.021)***
Du. num. child.	-.050 (.007)***	-.001 (.008)	-.050 (.007)***	-.001 (.008)	-.051 (.044)	.031 (.034)
Partner/Spouse variables						
Single	.261 (.075)***	-.038 (.136)	.270 (.077)***	-.038 (.137)	.466 (.437)	-.101 (.402)
Net wage partner/spouse	-.00005 (2.79e-06)***	-.00003 (2.85e-06)***	-.00005 (2.85e-06)***	-.00003 (2.92e-06)***	-.0002 (.00002)***	-.0001 (1.00e-05)***
Age partner/spouse	.006 (.002)**	.005 (.004)	.007 (.002)***	.007 (.004)	.009 (.015)	.014 (.017)
Age sq. partner/spouse	-.0001 (.00002)***	8.58e-06 (.00004)	-.0001 (.00003)***	-9.99e-06 (.00004)	-.0002 (.0002)	.00009 (.0002)
Exp. partner/spouse	.005 (.001)***	-.003 (.002)	.005 (.001)***	-.003 (.002)	.014 (.009)	-.012 (.011)
Exp. sq. partner/spouse	-.0001 (.00002)***	-.00003 (.00004)	-.0001 (.00002)***	-.00003 (.00004)	-.0003 (.0002)*	-.0001 (.0002)
Educ. partner/spouse	.021 (.008)***	-.024 (.017)	.019 (.008)**	-.029 (.017)*	.037 (.054)	-.071 (.061)
Educ. sq. partner/spouse	-.0006 (.0003)**	.0006 (.0006)	-.0006 (.0003)**	.0009 (.0006)	-.001 (.002)	.002 (.002)
Du. flag miss.	.291 (.076)***	-.015 (.137)	.298 (.077)***	-.016 (.138)	.519 (.441)	-.063 (.404)
Constant	.010 (.128)	.	-.253 (.133)*	. .	-2.051 (.794)***	.
Time dummies $\chi^2_{11} =$	169.007***	25.96***	59.867***	59.595***	99.211***	43.765***
Unobs. effects $\chi^2_{36} =$	.	.	.	.	.	691.46***
LL	.	.	.	.	-28309.25	-27964.73

Source: GSOEP 1995-2006, own calculations. Different binary choice specifications. 57,203 observations from 9,277 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) robust standard errors are provided; b) t-tests on the significance of the instrument in the 1st step regressions confirm that the rank condition for identification of the IV estimators is fulfilled; heteroskedasticity robust, regression based Hausman tests provide evidence for the endogeneity of the health variable on the 1% significance level.c) Standard errors are robust to serial correlation in the individual scores across t; d) unobserved effects are specified as a linear projection on the (within) means of the regressors.

Table 13: Summary, Participation Equation, Men, 1995-2006

	Entire Sample	Probit = 0	Probit = 1
Probit	.860	0	1
	(.347)	(0)	(0)
Age	41.481	43.022	41.231
	(10.939)	(13.452)	(10.453)
Age square	1840.336	2031.813	1809.240
	(927.205)	(1132.517)	(885.515)
Age triple	86389.600	102508.200	83771.970
	(62600.990)	(76750.590)	(59579.680)
Education	12.192	11.192	12.354
	(2.606)	(2.158)	(2.636)
Dummy Education	.145	.162	.143
	(.353)	(.369)	(.350)
Foreigner	.128	.189	.118
	(.334)	(.392)	(.323)
Doctor visits	1.892	2.732	1.755
	(3.694)	(5.115)	(3.388)
Lg. health sat.	2.567	2.400	2.594
	(.411)	(.581)	(.370)
Lg. non-lab.-inc.	5.713	7.712	5.388
	(2.881)	(1.140)	(2.946)
State level variables			
Lg. unempl. (fed. state)	12.797	12.755	12.804
	(.550)	(.547)	(.550)
Lg. vac. (fed. state)	10.459	10.258	10.491
	(.822)	(.865)	(.810)
Lg. empl. (fed. state)	14.689	14.512	14.718
	(.740)	(.784)	(.729)
Du. East-Germany	.257	.386	.237
	(.437)	(.487)	(.425)
Number of children			
up to 2 years old	.081	.052	.086
	(.287)	(.233)	(.295)
between $3-5$	.117	.073	.124
	(.350)	(.287)	(.359)
between $6-16$	.475	.338	.497
	(.813)	(.738)	(.822)
Du. no. child.	.601	.737	.579
	(.490)	(.440)	(.494)
Partner/Spouse variables[a)]			
Single	.225	.334	.207
	(.417)	(.472)	(.405)
Net wage partner/spouse	587.805	499.016	600.032
	(638.363)	(653.367)	(635.307)
Age partner/spouse	40.912	44.535	40.413
	(10.068)	(11.591)	(9.735)
Age sq. partner/spouse	1775.179	2117.684	1728.013
	(850.881)	(1007.411)	(815.836)
Exp. partner/spouse	10.565	13.017	10.227
	(9.156)	(11.176)	(8.788)
Exp. sq. partner/spouse	195.440	294.326	181.823
	(298.963)	(393.187)	(280.838)
Educ. partner/spouse	11.902	11.181	12.001
	(2.418)	(2.330)	(2.413)
Educ sq. partner/spouse	147.499	130.442	149.848
	(63.827)	(57.959)	(64.240)
Du. flag miss.	.025	.015	.026
	(.156)	(.123)	(.160)
N	57,419	8,022	49,397

Source: GSOEP 1995-2006, 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 non-missing data (Du. flag miss. = 0) and having a partner/being married (Single = 0).

Table 14: Summary, Participation Equation, Women, 1995-2006

	Entire Sample	Probit = 0	Probit = 1
Probit	.688	0	1
	(.463)	(0)	(0)
Age	41.714	43.930	40.708
	(11.026)	(11.973)	(10.413)
Age square	1861.637	2073.182	1765.550
	(933.362)	(1049.170)	(858.688)
Age triple	87851.120	103525.300	80731.660
	(63049.890)	(73259.050)	(56400.720)
Education	11.926	11.099	12.302
	(2.467)	(2.209)	(2.486)
Dummy Education	.131	.136	.129
	(.337)	(.342)	(.335)
Foreigner	.124	.195	.091
	(.329)	(.396)	(.288)
Doctor visits	2.575	3.013	2.375
	(4.013)	(4.844)	(3.554)
Lg. health sat.	2.549	2.482	2.579
	(.425)	(.499)	(.384)
Lg. non-lab.-inc.	5.870	6.997	5.358
	(2.842)	(1.923)	(3.038)
State level variables			
Lg. unempl. (fed. state)	12.801	12.841	12.783
	(.558)	(.569)	(.553)
Lg. vac. (fed. state)	10.459	10.541	10.422
	(.823)	(.792)	(.835)
Lg. empl. (fed. state)	14.692	14.766	14.658
	(.742)	(.724)	(.748)
Du. East-Germany	.252	.202	.275
	(.434)	(.402)	(.446)
Number of children			
up to 2 years old	.039	.082	.020
	(.200)	(.286)	(.140)
between $3-5$	.100	.175	.066
	(.326)	(.423)	(.264)
between $6-16$	.502	.648	.436
	(.818)	(.952)	(.740)
Du. no. child.	.607	.517	.648
	(.488)	(.500)	(.478)
Partner/Spouse variables$^{a)}$			
Single	.212	.145	.242
	(.409)	(.352)	(.428)
Net wage partner/spouse	1450.245	1409.635	1471.591
	(1117.829)	(1234.772)	(1050.547)
Age partner/spouse	45.789	47.774	44.745
	(11.164)	(12.125)	(10.474)
Age sq. partner/spouse	2221.217	2429.394	2111.787
	(1046.820)	(1169.002)	(958.543)
Exp. partner/spouse	22.569	24.399	21.607
	(11.186)	(11.769)	(10.743)
Exp. sq. partner/spouse	634.506	733.833	582.293
	(525.858)	(582.168)	(485.620)
Educ. partner/spouse	12.171	11.776	12.379
	(2.622)	(2.560)	(2.630)
Educ sq. partner/spouse	155.015	145.226	160.161
	(71.257)	(68.108)	(72.329)
Du. flag miss.	.041	.032	.046
	(.199)	(.175)	(.209)
N	57,203	17,867	39,336

Source: GSOEP 1995-2006, 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 non-missing data (Du. flag miss. = 0) and having a partner/being married (Single = 0).

Table 15: Summary, Wage Equation, Men, 1995-2006

	Mean	Std. dev.	10% pctl.	90% pctl.
Log. hourly wage	2.571	.427	2.051	3.093
Log. health sat.	2.596	.366	2.095	2.893
Age	41.174	10.350	28.0	56.0
Age sq.	1802.446	875.795	784.0	3136.0
Age tr.	83201.150	58815.110	21952.0	175616.0
Unempl. exp.	.403	1.099	0	1.200
Unempl. exp. sq.	1.371	9.031	0	1.440
Firm tenure	11.392	10.118	1.100	27.200
Firm tenure sq.	232.165	351.910	1.210	739.840
Education	12.371	2.634	10.5	18.0
Du. educ.	.143	.350	0	1
Part-time	.022	.147	0	0
Foreigner	.117	.321	0	1
State level variables				
Log. unempl. (fed. st.)	12.805	.550	12.160	13.660
Log. vac. (fed. st.)	10.494	.809	9.192	11.428
Log. empl. (fed. st.)	14.720	.728	13.586	15.563
East Germany	.221	.415	0	1
Number of children				
up to 2 years old	.087	.296	0	0
between $3-5$	.125	.361	0	1
between $6-16$	.501	.825	0	2
Du. no. child.	.575	.494	0	1
Firm size (< 20 employees)$^{a)}$				
$20-199$	.301	.459	0	1
$200-1999$	.235	.424	0	1
≥ 2000	.256	.437	0	1
Firm size miss.	.023	.149	0	0
Occupation Dummies (High Service)				
Low Service	.185	.388	0	1
Routine non-manual	.040	.196	0	0
Skilled manual	.305	.460	0	1
Semi-unskilled manual	.212	.408	0	1
Farm labour	.012	.109	0	0
Missing occ.	.090	.287	0	0
Sector Dummies (Agr., forestry, fishing)				
Unknown sector	.029	.169	0	0
Energy, water, mining	.016	.124	0	0
Manufacturing	.363	.481	0	1
Construction	.108	.311	0	1
Trade	.084	.278	0	0
Transport, communication	.042	.200	0	0
Financial serv., insurance	.024	.153	0	0
Other services	.090	.287	0	0
State	.230	.421	0	1
Exclusion restrictions/Instruments				
num. vis. doc. (last 3 months)	1.745	3.328	0	4.0
Log. non lab. inc.	5.382	2.939	0	8.114
Single	.205	.404	0	1
Flag miss.	.026	.161	0	0
Net wage partner/spouse$^{b)}$	602.324	633.724	0	1450.677
Age partner/spouse	40.308	9.656	28.0	54.0
Age sq. partner/spouse	1717.997	806.792	784.0	2916.0
Exp. partner/spouse	10.179	8.735	.800	23.700
Exp. sq. partner/spouse	179.920	277.926	.640	561.690
Educ. partner/spouse	12.017	2.415	9.0	16.0
Educ. sq. partner/spouse	150.249	64.354	81.0	256.0

Source: GSOEP 1995-2006, own calculations. All summary statistics are on individual-year level 47,746 observations). Individuals 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 for these variables are conditional on non-missing data (Du. flag miss. = 0) and having a partner/being married (Single = 0).

Table 16: Summary, Wage Equation, Women, 1995-2006

	Mean	Std. dev.	10% pctl.	90% pctl.
Log. hourly wage	2.350	.432	1.807	2.845
Log. health sat.	2.580	.381	2.095	2.893
Age	40.661	10.321	26.0	55.0
Age sq.	1759.862	849.240	676.0	3025.0
Age tr.	80244.280	55620.660	17576.0	166375.0
Unempl. exp.	.488	1.209	0	1.500
Unempl. exp. sq.	1.699	12.051	0	2.250
Firm tenure	9.340	8.688	.900	22.900
Firm tenure sq.	162.715	271.736	.810	524.410
Education	12.325	2.481	10.0	16.0
Du. educ.	.129	.335	0	1
Part-time	.382	.486	0	1
Foreigner	.090	.286	0	0
State level variables				
Log. unempl. (fed. st.)	12.785	.552	12.150	13.630
Log. vac. (fed. st.)	10.422	.835	9.118	11.428
Log. empl. (fed. st.)	14.658	.748	13.560	15.562
East Germany	.268	.443	0	1
Number of children				
up to 2 years old	.019	.139	0	0
between $3-5$	.065	.262	0	0
between $6-16$	.434	.736	0	2
Du. no. child.	.648	.478	0	1
Firm size (< 20 employees)$^{a)}$				
$20-199$	.295	.456	0	1
$200-1999$	.219	.414	0	1
≥ 2000	.193	.395	0	1
Firm size miss.	.026	.160	0	0
Occupation Dummies (High Service)				
Low Service	.256	.437	0	1
Routine non-manual	.197	.398	0	1
Skilled manual	.068	.251	0	0
Semi-unskilled manual	.173	.378	0	1
Farm labour	.009	.092	0	0
Missing occ.	.229	.421	0	1
Sector Dummies (Agr., forestry, fishing)				
Unknown sector	.032	.176	0	0
Energy, water, mining	.004	.060	0	0
Manufacturing	.167	.373	0	1
Construction	.016	.125	0	0
Trade	.154	.361	0	1
Transport, communication	.023	.150	0	0
Financial serv., insurance	.031	.172	0	0
Other services	.204	.403	0	1
State	.364	.481	0	1
Exclusion restrictions/Instruments				
num. vis. doc. (last 3 months)	2.365	3.518	0	5.0
Log. non lab. inc.	5.346	3.039	0	8.175
Single	.242	.428	0	1
Flag miss.	.046	.208	0	0
Net wage partner/spouse$^{b)}$	1476.293	1048.302	0	2644.976
Age partner/spouse	44.698	10.410	31.0	59.0
Age sq. partner/spouse	2106.237	951.393	961	3481
Exp. partner/spouse	21.570	10.701	7.0	36.0
Exp. sq. partner/spouse	579.760	482.760	49.0	1296.0
Educ. partner/spouse	12.395	2.632	10.50	18.0
Educ. sq. partner/spouse	160.553	72.444	110.250	324.0

Source: GSOEP 1995-2006, own calculations. All summary statistics are on individual-year level (37,670 observations). Individuals 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 for these variables are conditional on non-missing data (Du. flag miss. = 0) and having a partner/being married (Single = 0).

References

Cai, L. (2007): "Effects of Health on Wages of Australian Men," *Melbourne Institute Working Paper Series 2007/02.*

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

Contoyannis, P., A. M. Jones, and N. Rice (2004): "The dynamics of health in the British Household Panel Survey," *Journal of Applied Econometrics*, 19(4), 473–503.

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

Dustmann, C., and M. E. Rochina-Barrachina (2007): "Selection correction in panel data models: An application to the estimation of females' wage equations," *Econometrics Journal*, 10, 263–293.

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

Grossman, M. (2001): "The Human Capital Model," in *Handbook of Health Economics*, ed. by A. J. Culyer, and J. P. Newhouse, vol. 1A, pp. 347–409. Elsevier Science B.V., Amsterdam.

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

Haveman, R., B. Wolfe, B. Kreider, and M. Stone (1994): "Market Work, Wages, and Men's Health," *Journal of Health Economics*, 13, 163–182.

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

Kennedy, P. (1984): "Logarithmic Dependent Variables and Prediction Bias," *Oxford Bulletin of Economics and Statistics*, 463, 389–392.

Kyriazidou, E. (1997): "Estimation of a panel data sample selection model," *Econometrica*, 55, 1335–1364.

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

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

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

Mundlak, Y. (1978): "On the pooling of time series and cross section data," *Economatrica*, 46, 69–85.

ROCHINA-BARRACHINA, M. E. (1999): “A new Estimator for Panel Data Sample Selection Models,” *Annales d‘Economie et de Statistique*, 55/56, 153–181.

ROMEU GORDO, L. (2006): “Effects of short- and long-term unemployment on health satisfaction *evidence from German data,” *Applied Economics*, 38(20), 2335–2350.

SEMYKINA, A. (2007): “Specification Tests in Panel Data Models with Selection,” *unpublished manuscript, Michigan State University.*

SEMYKINA, A., AND J. M. WOOLDRIDGE (2006): “Estimating Panel Data Models in the Presence of Endogeneity and Selection: Theory and Application,” *unpublished manuscript, Michigan State University.*

STERN, S. (1989): “Measuring the Effect of Disability on Labor Force Participation,” *The Journal of Human Resources*, 24(3), 361–395.

VERBEEK, M., AND T. NIJMAN (1992): “Testing for Selectivity Bias in Panel Data Models,” *International Economic Review*, 33, 681–703.

WOOLDRIDGE, J. M. (1995): “Selection correction for panel data models under conditional mean independence assumption,” *Journal of Econometrics*, 68, 115–132.

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