Measuring overeducation with earnings frontiers and multiply imputed censored income data *

© UWE JENSEN†, HERMANN GARTNER‡ & SUSANNE RÄSSLER§

April 2006

Abstract

In this paper, we remove one serious drawback of the IAB employment sample impeding its applicability to the estimation of earnings frontiers: the censoring of the income data, by multiple imputation. Then, we estimate individual potential income with stochastic earnings frontiers, and we measure overeducation as the ratio between actual income and potential income. It is shown that the measurement of overeducation by this income ratio is a valuable addition to the overeducation literature because the well-established objective or subjective overeducation measures focus on some ordinal matching aspects and ignore the metric income and efficiency aspects of overeducation.

Keywords: Censored data, earnings frontiers, IABS employment sample, multiple imputation, overeducation

JEL codes: C15, C24, C81, J31

*We are grateful to Donald B. Rubin for fruitful discussions about our imputation algorithm and to Norbert Schanne for helpful comments on the paper.

†Institut für Statistik und Ökonometrie, Christian-Albrechts-Universität, Olshausenstr. 40, 24118 Kiel, Germany, Tel.: +49-431-880-1627, e-mail: Jensen@stat-econ.uni-kiel.de

‡Institut für Arbeitsmarkt- und Berufsforschung (IAB), Weddigenstr. 20-22, 90478 Nürnberg, Germany, Tel.: +49-911-179-3386, email: Hermann.Gartner@iab.de

§Institut für Arbeitsmarkt- und Berufsforschung (IAB), Weddigenstr. 20-22, 90478 Nürnberg, Germany, Tel.: +49-911-179-3084, email: Susanne.Raessler@iab.de

Contents

1 Introduction: Measuring overeducation

Employees are called overeducated if the knowledge they have acquired in the education process cannot be fully applied in their present jobs. The overeducation problem is important because it stands for inefficient use of individual and social resources, i.e., for superfluous individual and social costs. It is insufficient to restrict the attention to the question if individuals are employed or unemployed. Many studies have shown that the amount of overeducation is considerable in many OECD countries. See Büchel et al. (2003) for a good introduction to this field of research.

In the analysis of overeducation, many papers concentrate on the economical understanding of the problem and on its descriptive analysis. But another problem is how to measure overeducation. Borghans and de Grip (2000, p. 13ff) and Jensen (2003) give introductions to the measures that dominate the literature:

1. For the objective measure, professional job analysts try to specify the required level and type of education in particular occupations and, by this means, provide a classification of occupations into a requirement ranking of, say, 6 or 7 classes. These classifications are then converted to years of requested schooling or schooling dummies which can be compared with the acquired schooling of the individuals in the study.

2. The subjective measure is based on worker's self-assessment. Workers are asked how much formal education is required to get a job like theirs. Once again, this (classified) information is compared with the actual schooling of the individual.

3. The empirical measure uses the distribution of schooling years in a given occupation or a group of occupations. Typically, individuals are defined to be overeducated if their schooling level is more than 1 standard deviation above the mean of all individuals in that occupation.

Of course, different overeducation measures can lead to different results for the same individual – and all measures can be criticized:

1. The objective measure, for example, ignores within-occupation variation in job-specific schooling requirements. Furthermore, providing objective measures with a sufficiently detailed partition into different occupations and accounting for occupational requirement changes over time is extremely expensive (Borghans and de Grip, 2000, p. 15). From a statistical view, one should add that converting metric information (on schooling requirements) into few classes may destroy much information.

2. Employees may be inclined to over- or understate the educational requirements of their job. This leads to biased subjective measures (Borghans and de Grip, 2000, p. 16). Furthermore – see the previous item – classifying metric information into few classes usually leads to a loss of information.

3. The empirical measure produces results which depend heavily and arbitrarily on the choice of the cut-off point (Borghans and de Grip, 2000, p. 16). Furthermore, this method will always identify overeducation even if it does not exist, a well-known problem of corresponding poverty measures.

Finally, all overeducation measures ignore the income aspect of overeducation. That is why Jensen (2001a) presented a 4th way of measuring overeducation, the income ratio measure: individual potential income is estimated with a stochastic earnings frontier. Then, overeducation (= 'income inefficiency') is measured as the income ratio between actual income and potential income.

Applying individual income differentials as overeducation criterion is sometimes criticized because, for example, a bachelor of science working as car dealer with a considerable income will not be identified as overeducated (Büchel, 2001). The counter-argument is: from an individual and from a social perspective, it is at least questionable if a bachelor of science is more adequately employed if he works at a university in a temporary part-time job with a minor income. Following human capital theory, income maximization is an important goal of investments in schooling. And the income ratio measure allows the inclusion of the (metric) income and efficiency aspects of overeducation ignored by the well-established objective or subjective measures focusing on some (ordinal) matching aspects.

But the 2-step approach used in Jensen (2001a) has been criticized in the production frontier literature where it is originated. Therefore, Jensen (2003) improved the income ratio approach by introducing the inefficiency effects specification by Reifschneider and Stevenson (1991) to the overeducation literature. Here, overeducation can be modeled as a function of several overeducation determinants while the main model describes the dependence of potential income on determinants (like human capital) of individual income. But the results of that paper are unsatisfactory due to the small sample size of the data set (German Socioeconomic Panel, GSOEP) used in that study.

The IAB employment sample is a much larger data set for modeling individual income and overeducation (IAB stands for 'Institute for Employment Research of the Federal Employment Agency' (Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit)). But the income data in this data set are censored in the upper tails. When estimating average earnings functions, this problem is typically solved by applying

Tobit models. But the estimation of an earnings frontier – trying to estimate individual overeducation (= 'income inefficiency') with respect to maximum potential income – is immensely impeded.

In this paper, we will therefore multiply impute the censored income data in order to be able to estimate individual overeducation with the large IAB employment sample. The contribution of this paper to the literature is two-fold: We show that the income ratio approach for measuring overeducation works excellently when applied to a sufficiently large data set. And we show how the censored income data of the IAB employment sample can be imputed to estimate earnings (frontier) functions.

The reminder of this article is as follows: The following section summarizes the necessary details on earnings frontiers and their econometric origin, namely production frontiers. After an overview on the IAB employment sample in section 3, a short introduction to multiple imputation is provided in section 4. The subsequent section presents and discusses the estimation results in detail. Some conclusions close the paper.

2 Stochastic earnings frontiers

This section summarizes the theory on stochastic earnings frontiers necessary in the following.

2.1 Production frontiers

In microeconomic theory, production functions provide maximum possible output for given inputs of, say, n firms in the sample. In reality, inefficient input use may lead to lower outputs for many firms. That is why frontier functions (lying on top of the data cloud) have been developed for estimating potential output and inefficiency. See the surveys in Coelli et al. (1998), Greene (1997) or Jensen (2001a) for more details on frontiers.

Based on the seminal work of Aigner and Chu (1968), Aigner et al. (1977) and Meeusen and van den Broeck (1977) introduced the stochastic production frontier

$$Y_i = e^{\beta_0} \cdot \prod_{j=1}^{k} X_{ij}^{\beta_j} \cdot e^{v_i} \cdot TE_i, \quad i = 1, \ldots, n, \tag{1}$$

or in logs

$$y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i = v_i - u_i, \quad u_i \geq 0. \tag{2}$$

Here, y_i is the output (in logs), x_{ij} are k inputs (all in logs) of firm no. i, and β_j are unknown parameters. Then, with $TE_i = 1$,

$$Y_i^* = e^{\beta_0} \cdot \prod_{j=1}^{k} X_{ij}^{\beta_j} \cdot e^{v_i} \tag{3}$$

or in logs with $u_i = 0$

$$y_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + v_i \tag{4}$$

is maximum possible output (in logs) for given inputs. The output ratio

$$0 \leq TE_i = e^{-u_i} = \frac{Y_i}{Y_i^*} \leq 1 \tag{5}$$

is interpreted as technical inefficiency of firm no. i. Finally, the composed error term ε_i consists of the one-sided inefficiency term u_i and the symmetric part v_i representing statistical noise. x_{ij}, v_i, and u_i are assumed to be independent with the distributional assumptions

$$v_i \sim N(0, \sigma_v^2) \quad \text{and} \quad u_i \sim \text{trunc}_0 N(\mu, \sigma_u^2) \tag{6}$$

where $\text{trunc}_0 N(\cdot, \cdot)$ stands for a normal distribution truncated at $u = 0$ (see Stevenson, 1980).

The choice of the truncated normal distribution in (6) contains the half-normal distribution as (testable) special case. Other tractable alternatives for the inefficiency distribution are the exponential and the gamma distribution. Ritter and Simar (1997) have shown the bad performance of the gamma distribution. Jensen (2005) has presented some risks of working with exponentially distributed inefficiency. Therefore, we prefer the truncated normal distribution.

The log-likelihood function is given by $l(\beta, \sigma, \lambda, \mu) =$

$$-n \left[\ln(\sigma) + const + \ln\left(\Phi\left(\frac{-\mu}{\sigma\lambda} \right) \right) \right] - \sum_{i=1}^{n} \left[\frac{1}{2} \left(\frac{\varepsilon_i}{\sigma} \right)^2 - \ln\left(\Phi\left(\frac{-\mu}{\sigma\lambda} - \frac{-\varepsilon_i \lambda}{\sigma} \right) \right) \right] \tag{7}$$

with

$$\lambda = \frac{\sigma_u}{\sigma_v} \quad \text{and} \quad \sigma^2 = \sigma_v^2 + \sigma_u^2 \tag{8}$$

and the cumulative standard normal distribution function $\Phi(\cdot)$. Iterative maximization leads to consistent and asymptotically efficient maximum likelihood (ML) estimators $\hat{\beta}_j$, $\hat{\sigma}$, $\hat{\lambda}$, and $\hat{\mu}$.

How can the inefficiency terms be estimated? Since, in a stochastic frontier model, the estimation residuals only estimate the composed error ε and not u, the inefficiencies

must be estimated indirectly, for example with the help of the minimum mean-squared error predictor

$$E[u_i|\varepsilon_i] = \frac{\sigma\lambda}{1+\lambda^2}\left(\frac{\phi\left(\frac{\varepsilon_i\lambda}{\sigma}\right)}{\Phi\left(-\frac{\varepsilon_i\lambda}{\sigma}\right)} - \frac{\varepsilon_i\lambda}{\sigma}\right) \tag{9}$$

with the standard normal density function $\phi(\cdot)$.

This basic approach might be too restrictive. Independence of x_{ij} and u_i may be a hard assumption as heteroscedasticity might occur. Despite the iid assumption (6), u might still contain some structure. Until the beginning of this decade, the so-called '2-step approach' has been employed quite often: the inefficiency estimates from (9) – the first step – were used to find some inefficiency determinants in a second step. But Wang and Schmidt (2002) have shown that this procedure – claiming that u_i are iid in the first step and finding structure in them in the second step – can lead to severely biased results.

Therefore, it is advisable to use the procedure already presented by Reifschneider and Stevenson (1991). They allow the inefficiency terms u_i to depend on some explanatory variables z_{ij} (interpreted as sources of inefficiency) which may be partly identical with variables x_{ij}:

$$u_i = \delta_0 + \sum_{j=1}^{l} \delta_j z_{ij} + w_i = d_i + w_i, \quad i = 1, \ldots, n, \tag{10}$$

δ_j are unknown parameters. The distributional assumptions are

$$v_i \sim N(0, \sigma_v^2), \quad u_i \sim \text{trunc}_0 N(d_i, \sigma_u^2), \quad \text{and} \quad w_i \sim \text{trunc}_{-d_i} N(0, \sigma_u^2). \tag{11}$$

The ML estimators $\hat{\beta}_j$, $\hat{\delta}_j$, $\hat{\sigma}$ and $\hat{\lambda}$ are derived simultaneously using iterative ML techniques. See the given references for the likelihood function of the full model.

2.2 Earnings frontiers

The ideas and models presented in the previous subsection now are transferred to the estimation of an earnings frontier in a rather straightforward way: transforming schooling investments (input) to earnings (output) is also a production process which should be performed as efficient as possible.

The earnings frontier model is based on the popular human capital model by Mincer (1974) for explaining individual income. However, the rigid assumptions of the basic human capital model have often been criticized for good reasons. In contrast to these assumptions,

- labor markets and education markets are far from perfect.

- higher education is not necessarily identical to higher productivity, and the latter does not automatically imply higher wages: much knowledge learned at school or at the university does not come into adequate use in jobs. Applicants with identical human capital stock may have substantial productivity differences in a given job. And apparently identical occupations can differ considerably regarding their productivity requirements and wages.

- employees and employers are not perfectly informed about all labor market aspects relevant to them: partly considerable information deficits exist on both sides, instead. Employers often are not sufficiently informed about the productivity of their employees. And if they are informed, they may be not able to allocate the abilities of their employees to the jobs available in the firm. For employees, an efficient job search leads to direct and indirect costs. Therefore, it is rational to stop the search if a wage offer exceeds a certain reservation wage.

These and many other restrictions to the basic human capital model have led to many extensions like screening theory, search theory, etc. The earnings frontier approach accounts for these restrictions by transferring the inefficiency aspect from production frontiers to earnings frontiers.

In Daneshvary et al. (1992), individual wages $LINC$ (in logs) are assumed to depend on personal characteristics H augmenting human capital stock, job characteristics C and information I on labor market conditions, the wage distribution, and job search methods. Individuals stop their search when a wage offer exceeds the reservation wage $LINC_r$. For any set of H and C and perfect information I^*, a potential maximum attainable wage $LINC^*$ exists. Therefore, $LINC = LINC(H, C, I)$ is estimated as stochastic earnings frontier (see the production frontier (2))

$$LINC_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i, \quad \varepsilon_i = v_i - u_i, \quad u_i \geq 0 \tag{12}$$

The explanatory variables x will be introduced in the following section.

Then,

$$LINC_i^* = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + v_i \tag{13}$$

(see (4)) is maximum possible (potential) income. With untransformed individual income INC_i, Jensen (2001a) interprets the inefficiency terms

$$0 \leq \mathrm{OVER}_i = 1 - e^{-u_i} = 1 - \frac{INC_i}{INC_i^*} \leq 1 \tag{14}$$

(instead of (5)) as cost of imperfect information measuring the overeducation of individual number i as income ratio. Note that parametrization (14) is different from (5). Jensen (2003) uses the inefficiency effects specification (10) as overeducation effects specification to model overeducation as a function of several determinants. v_i represents statistical noise as in the previous subsection.

For interpreting parameter values, the untransformed stochastic earnings frontier

$$INC_i = e^{\beta_0} \cdot \prod_{j=1}^{k} X_{ij}^{\beta_j} \cdot e^{v_i} \cdot (1 - \mathrm{OVER}_i) \tag{15}$$

(instead of (1)) is preferred. All further remarks of the previous subsection concerning modeling and estimation of u, model assumptions, etc. apply likewise.

Jensen (2001a and 2001b) discusses the advantages of the earnings frontier over the 'average earnings function'. The frontier model includes the important but hardly measurable human capital investment 'search for information'. An average human capital model only explains potential income sufficiently but individuals do not obtain their maximum possible income because of imperfect information described above. The usual OLS approach does not take into account the considerable amount of individual inefficiency in finding the suitable jobs and therefore falsely interprets inefficiency (overeducation) as misspecification.

3 Data and the censoring

Many studies about wage structure in Germany use the IAB employment sample (IABS). But up to now, it was not possible to estimate earnings frontiers. The IABS is a 2 percent random sample of all employees covered by the social security system. Accordingly self-employed, family workers, and civil servants (Beamte) are not included. The data set represents about 80 percent of the employees in Germany. A detailed description of the data can be found in Bender et al. (2000) or Hamann et al. (2004).

The data are drawn from the Employment Register of the Federal Employment Agency. Employers are obliged to report at least once a year the earnings and other information about their employees. Therefore, the data are highly reliable. The IABS includes, among others, information about age, sex, education, wages, and the occupational group.

Since contributions to the social security system must be payed only up to a contribution limit, the wages are reported only up to this limit. Thus, the wages are right-censored. For employees payed above the limit, only the limit is reported. Our solution for this data problem is discussed below.

The data set includes about one million employment spells for each year. For our study, we impose several restrictions on the data: Only employment spells valid at June

30, 2001 are included. We exclude part-time workers and apprentices. Part-time workers are excluded because the wages are given as daily wages and the hours worked per day are unknown. Therefore, we compare only fulltime workers. Furthermore, we have dropped all cases where earnings are below twice the limit of minor employment because these wages are implausible for fulltime workers. We restrict our data to West German male residents at the age of 18 to 55 years. With these restrictions, the data set contains 165,425 observations with 27,104 (16.4%) censored wages.

The variables used for the earnings frontier model and the imputation procedure are defined as follows: Potential experience (linear, quadratic and cubic) is calculated as age minus six minus years of schooling. Tenure (linear, quadratic and cubic) are the cumulated years a person works in the firm. The educational levels are: no qualification, vocational training, Abitur (higher secondary school), Abitur with vocational training, technical college, and university. The job-status is separated to four categories: unskilled worker, craftsman, master craftsman, and white collar worker. Additionally, we use an occupational classification scheme to define 11 occupational groups according to Blossfeld (1985). The occupations are grouped by hierarchical levels in the firm and by main economic sectors. To control for the economic environment we include further dummy variables for 9 categories of firm size, for 15 industrial categories (aggregated from the wz73 code), and for 8 regional types. The definition of the regional types can be found in the appendix in table 4.

4 Multiple imputation

To allow the estimation of earnings frontier functions with these data, we treat the censoring of wages as a missing data problem. In this special case, we regard the missingness mechanism as 'not missing at random' (NMAR, according to Little and Rubin, 1987, 2002) as well as 'missing by design', because the missingness depends on the value itself, i.e., if the wage exceeds the limit, the true value will not be reported but the limit, say a. A common approach to handle missing data is multiple imputation which means that every missing value is randomly imputed for m times (see Rubin 1978, 1987, 1996). In our case, the missing value is the wage whenever the limit is reported.

To start with, let $Y = (Y_{obs}, Y_{mis})$ denote the random variables concerning the data with observed and missing parts. In our specific situation this means that for all units with wages below the limit a each data record is complete, i.e., $Y = (Y_{obs}) = (X, \text{wages})$. For every unit with a value of the limit a for its wage information we treat the data record as partly missing, i.e., $Y = (Y_{obs}, Y_{mis}) = (X, ?)$. Thus, we have to multiply impute the missing data $Y_{mis} = \text{wage if wage} = a$.

4.1 The basic principle

The theory and principle of multiple imputation (MI) originates from Rubin (1978). The theoretical motivation for multiple imputation is Bayesian, although the resulting multiple imputation inference is usually also valid from a frequentist viewpoint. Basically, MI requires independent random draws from the posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$ of the missing data given the observed data. Since it is often difficult to draw from $f_{Y_{mis}|Y_{obs}}$ directly, a two-step procedure for each of the m draws is useful:

(a) First, we make random draws of the parameters Ξ according to their observed-data posterior distribution $f_{\Xi|Y_{obs}}$,

(b) then, we perform random draws of Y_{mis} according to their conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$.

Because

$$f_{Y_{mis}|Y_{obs}}(y_{mis}|y_{obs}) = \int f_{Y_{mis}|Y_{obs},\Xi}(y_{mis}|y_{obs},\xi) f_{\Xi|Y_{obs}}(\xi|y_{obs}) d\xi \tag{16}$$

holds, with (a) and (b) we achieve imputations of Y_{mis} from their posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$. Due to the data generating model used, for many models the conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$ is rather straightforward.

In contrast, the corresponding observed-data posteriors $f_{\Xi|Y_{obs}}$ usually are difficult to derive for those units with missing data, especially when the data have a multivariate structure and different missing data patterns. The observed-data posteriors are often no standard distributions from which random numbers can easily be generated. However, simpler methods have been developed to enable multiple imputation based on Markov chain Monte Carlo (MCMC) techniques (extensively discussed by Schafer, 1997). In MCMC, the desired distributions $f_{Y_{mis}|Y_{obs}}$ and $f_{\Xi|Y_{obs}}$ are achieved as stationary distributions of Markov chains which are based on the complete-data distributions which are easier to compute.

To proceed further, let θ denote a scalar quantity of interest that is to be estimated, such as a mean, variance, or correlation coefficient. Notice that now θ can be completely different from the data model used before to create the imputations. Although θ could be an explicit function of the parameters ξ of the imputation model one of the strengths of the multiple imputation approach is that this need not be the case. In fact, $\theta = \xi$ is called congenial (Meng, 1994). However, multiple imputation is designed for situations where the analyst and the imputer are different, thus, the analyst's model could be quite different from the imputer's model. As long as the two models are not overly incompatible or the fraction of missing information is not high, inferences based on the multiply imputed data

should still be approximately valid. Moreover, if the analyst's model is a sub-model of the imputer's model, i.e., the imputer uses a larger set of covariates than the analyst and the covariates are good predictors of the missing values, then MI inference is superior to the best inference possible using only the variables in the analyst's model. This property is called superefficiency by Rubin (1996). On the other hand, if the imputer ignores some important correlates of variables with missing data, but these variables are used in the analyst's model, then the results will be biased. To account for uncongeniality we use the same variables in the imputation model as in the earnings frontiers model.

Now let $\widehat{\theta} = \widehat{\theta}(Y)$ denote the statistic that would be used to estimate θ if the data were complete. Furthermore, let $\widehat{var}(\widehat{\theta}) = \widehat{var}(\widehat{\theta}(Y))$ be the variance estimate of $\widehat{\theta}(Y)$ based on the complete data set.

We also assume that with complete data, usually tests and interval estimates which are based on the normal approximation

$$\frac{\widehat{\theta} - \theta}{\sqrt{\widehat{var}(\widehat{\theta})}} \sim N(0,1) \tag{17}$$

should work well.

Notice that the usual maximum-likelihood estimates and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions.

Suppose now that the data are missing and we make $m > 1$ independent simulated imputations $(Y_{obs}, Y_{mis}^{(1)})$, $(Y_{obs}, Y_{mis}^{(2)})$, $\ldots$, $(Y_{obs}, Y_{mis}^{(m)})$ enabling us to calculate the imputed data estimate $\widehat{\theta}^{(t)} = \widehat{\theta}(Y_{obs}, Y_{mis}^{(t)})$ along with its estimated variance $\widehat{var}(\widehat{\theta}^{(t)}) = \widehat{var}(\widehat{\theta}(Y_{obs}, Y_{mis}^{(t)}))$, $t = 1, 2, \ldots, m$. From these m imputed data sets the multiple imputation estimates are computed.

The MI point estimate for θ is simply the average

$$\widehat{\theta}_{MI} = \frac{1}{m} \sum_{t=1}^{m} \widehat{\theta}^{(t)}. \tag{18}$$

To obtain a standard error $\sqrt{\widehat{var}(\widehat{\theta}_{MI})}$ for the MI estimate $\widehat{\theta}_{MI}$, we first calculate the 'between-imputation' variance

$$\widehat{var}(\widehat{\theta})_{between} = B = \frac{1}{m-1} \sum_{t=1}^{m} (\widehat{\theta}^{(t)} - \widehat{\theta}_{MI})^2, \tag{19}$$

and then the 'within-imputation' variance

$$\widehat{var}(\widehat{\theta})_{within} = W = \frac{1}{m} \sum_{t=1}^{m} \widehat{var}(\widehat{\theta}^{(t)}) \,. \tag{20}$$

Finally, the estimated total variance is defined by

$$\begin{aligned} \widehat{var}(\widehat{\theta}_{MI}) &= T = \widehat{var}(\widehat{\theta})_{within} + \left(1 + \frac{1}{m}\right) \widehat{var}(\widehat{\theta})_{between} \\ &= W + \frac{m+1}{m} B. \end{aligned} \tag{21}$$

The term $((m+1)/m)B$ enlarges the total variance estimate T compared to the usual analysis of variance with $T = B + W$; $(m+1)/m$ is an adjustment for finite m. An estimate of the fraction of missing information γ about θ due to non-response is given by

$$\widehat{\gamma} = \frac{(1 + 1/m)B}{T}. \tag{22}$$

For large sample sizes, tests and two-sided $(1-\alpha)100\%$ interval estimates can be based on the Student's t-distribution

$$\frac{\widehat{\theta}_{MI} - \theta}{\sqrt{T}} \sim t_v \qquad \text{and} \qquad \widehat{\theta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \tag{23}$$

with the degrees of freedom

$$v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2. \tag{24}$$

From (23) we can see that the multiple imputation interval estimate is expected to produce a larger interval than an estimate based only on one single imputation (SI). The multiple imputation interval estimates are widened to account for the missing data uncertainty and simulation error. Using only one singly imputed data set, in general, will lead to an underestimation of uncertainty and thus produce variance estimates that are too low and p-values that are too significant.

4.2 Imputation model

To adopt the multiple imputation principle for solving the problem of censored wages we assume that for person i the wage in logs is given by

$$y_i^* = x_i'\beta + \epsilon_i \tag{25}$$

where $\epsilon \overset{iid}{\sim} N(0, \tau^{-2})$.

We observe the wage $y_{obs} = y_i^*$ only if the wage is under the threshold a. If the wage is above a, we observe a instead of y_i^*:

$$y_i = \begin{cases} y_{obs} & \text{if } y_i^* \leq a \\ a & \text{if } y_i^* > a \end{cases}, \tag{26}$$

thus y_i is right censored. We now impute for a itself estimations z of the true wages. Thus, we define $y = (y_{obs}, a)$ and $y_z = (y_{obs}, z)$. Then, z is a truncated variable in the range (a, ∞) and its conditional predictive distribution is given by

$$f(z|y, \beta, \tau^2) = \frac{f_N(z|x'\beta, \tau^{-2})}{1 - \Phi(\tau a - \tau x'\beta)} \tag{27}$$

where $a < z < \infty$. According to Chib (1992) we get a data augmentation algorithm and Gibbs sampler based on the full conditional distributions:

$$f(\beta|y, z, \tau^2) = f_N(\beta|\hat{\beta}_z, \tau^{-2}(X'X)^{-1}) \tag{28}$$

$$f(\tau^2|y, z, \beta) = f_G(\tau^2|n/2, \sum_{i=1}^{n}(y_z - x'\beta)^2/2) \tag{29}$$

where $\widehat{\beta}_z^{(t)} = (X'X)^{-1}X'y_z^{(t)}$ is the usual OLS estimate based on the complete data set.

To receive valid imputations and random draws of the parameters from their observed data distribution in accordance with the rule presented in (16), we finally propose a MCMC technique as mentioned earlier. To start the chain we take the starting values $\beta^{(0)}, \tau^{2(0)}$ from a conditional ML Tobit estimation.

Imputation-step:

First, we randomly draw values for the missing variables from the truncated distribution according to (28), that is

$$z_i^{(t)} \sim \text{trunc}_a N(x_i'\beta^{(t)}, \tau^{-2(t)}). \tag{30}$$

Note that alternatively an accept-rejection algorithm could be applied instead of drawing directly from the truncated distribution. But the computational time gets too large with such an amount of missing data and these large data sets. More details on the algorithm used can be found in Gartner and Rässler (2005).

Second, we compute an OLS regression based on the imputed data sets, resulting in the estimator:

$$\widehat{\beta}_z^{(t)} = (X'X)^{-1}X'y_z^{(t)}. \tag{31}$$

Then we produce new random draws for the parameters according to their complete data posterior distribution. To run random drawings of τ with STATA we would need the inverse of a gamma distribution. Because this is complicated we use a slight modification of (29).

Posterior-step:

In the posterior-step we draw first $\tau^{2(t+1)}$ according to

$$g \sim \chi^2(n-k) \tag{32}$$

$$\tau^{2(t+1)} = \frac{g}{RSS} \tag{33}$$

where

$$RSS = \sum_{i=1}^{n} \left(y_{z_i}^{(t)} - x_i' \widehat{\beta}_z^{(t)} \right)^2 \tag{34}$$

is the residual sum of squares, and k is the number of columns of X.

Then we draw $\beta^{(t+1)}$:

$$\beta^{(t+1)} \sim N\left(\widehat{\beta}_z^{(t)}, \tau^{-2(t+1)} (X'X)^{-1} \right). \tag{35}$$

The covariates contained in X are the same as in the wage regression. We repeat the imputation-step and the posterior-step 11,000 times and storage the values $z_i^{(2,000)}$, $z_i^{(3,000)}$, ... , $z_i^{(11,000)}$ to obtain 10 different completed data sets. The imputation routine required a completion time of about six hours. Different analyses of the convergence of the chains did not exhibit any problems.

We have assumed a log-normal distribution of the wages. But the normal distribution is notoriously sensitive to outliers (see Gelman et al., 2003, p. 443). Especially by using transformations of a normal distribution this problem may be considerable (as discussed by Rubin, 1983). To examine the applicability of our distribution assumption, Gartner and Rässler (2005) compare the distribution of imputed wages with the distribution calculated with the German Socioeconomic Panel (GSOEP) and find that the imputed wages lie about in the same range as the wages in the GSOEP.

Since the distribution of the imputed wages seems plausible, the data set is used to estimate the earnings frontier model.

5 Estimation results

Descriptive statistics of the variables used to estimate the stochastic earnings frontier (12) with overeducation effects submodel (10) are given in the appendix in table 1. The estimation has been performed with LIMDEP 8.0.

Table 2 in the appendix provides the estimation results for the stochastic earnings frontier (12). Signs and size of the parameter estimates are as expected. With the untransformed earnings frontier (15), their interpretation is straightforward. For example,

studying successfully at a university means: multiply the potential income without qualification by the factor

$$e^{0.3366 \cdot 1} = 1.400, \tag{36}$$

ceteris paribus and on average. Potential income (like average income) is higher

- with more experience,
- with higher tenure,
- with higher education,
- with higher status,
- in more densely populated regions,
- in larger firms.

As motivated in subsection 2.2, overeducation (= 'income inefficiency') is measured as the income ratio between actual income and potential income. But the problematic part of the previous work on the income ratio measure has been to model overeducation. The 2-step approach in Jensen (2001a) has been criticized in the production frontier literature because it may lead to biased results. The overeducation effects specification in Jensen (2003) gives unsatisfactory results due to the small sample size of the data set (German Socioeconomic Panel, GSOEP) used in that study. Since essentially the same individual and job characteristics determine (potential) earnings and overeducation, it seemed to be impossible to find significant metric variables determining overeducation in that study. And the parameter estimates of some makeshift dummies used instead show unexpected signs, as for age and tenure.

The results in table 3 for the overeducation effects submodel (10) reveal the rich information available in the large IAB employment sample (made available by multiple imputation) and show the explanatory power of the income ratio approach when applied to such a data set. Whereas the well-established objective or subjective overeducation measures focus on some (ordinal) matching aspects, the income ratio measure allows the simultaneous modeling of potential income and overeducation and reveals the (metric) income and efficiency aspects of overeducation. Our result is that overeducation decreases – see (14) – with more experience, higher tenure, higher education and lower status.

Since the functional forms of the relations between experience or tenure and overeducation are polynomials of order 3, figures 1 and 2 in the appendix demonstrate the forms and the sizes of these relations. Note that the model – see (11) – only guarantees $u_i \geq 0$. But since there are n different individuals with n different vectors z_i there is no sensibly

defined zero for the vertical axis in these plots. Therefore, the functions are simply shifted such that minimum overeducation is zero. The range of experience and tenure in the plots is the same as in the data.

The cubic relation between experience and overeducation is in line with theories that are well-known in the literature on overeducation (see Büchel, 2001), viz job matching theory and career mobility theory. Overeducation is very high at the beginning of the working career and then decreases very fast. This can be explained by the career mobility theory by Sicherman and Galor (1990): "Individuals may choose an entry level in which the direct returns to schooling are lower than those in other feasible entry levels if the effect of schooling on the probability of promotion is higher in this entry level" (p. 177). So, young employees waive some present income in favor of a promotion option. And the employers get the opportunity to test their young employees with reduced wages.

Job matching theory, for example in the form of Jovanovic (1979), interprets jobs as 'experience goods'. The information I on labor market conditions, the wage distribution and job search methods should increase with experience. And, as outlined in subsection 2.2, higher information I should lead to better matches between workers and jobs, and lower overeducation. Figure 1 shows that this development comes to an end with an experience of roughly 20 years. After that time, overeducation slightly increases again, maybe due to the higher age of the workers leading to a growing difference between experience and productivity.

Job matching theory also helps to understand the cubic relation between tenure and overeducation: Job tenure is often simply taken as indicator for matching quality because only sufficiently good matches between workers and jobs should endure (if reasonable alternatives are available). Nevertheless, employees need a certain search time to find a better match and many of them search without success, maybe due to low skills. Therefore, individuals with higher tenure show lower overeducation. But this relation stops at a tenure of roughly 10 years. This could mean that a considerable part of bad matches (but better a bad match than no match) has to remain because no better alternatives are available or because it might be a bad signal to employers – see below – when jobs are changed too often. This phenomenon is also compatible with the segmentation theory by Doeringer and Piore (1971) assuming that labor markets are divided into sub-markets or segments (a primary segment with well-paid and promising jobs and a secondary peripheral segment with low-paid and unskilled jobs) with little permeability between the segments. Interestingly, at a tenure of roughly 20 years, the tenure-overeducation-profile starts to fall again. This might happen because, then, many bad matches are finished by (early) retirement.

With equations (10) and (14), the interpretation of the education dummy parameter

estimates is straightforward. For example, studying successfully at a university means a reduction of overeducation by

$$(1 - e^{-0.4830 \cdot 1}) \cdot 100 = 38.3\% \tag{37}$$

ceteris paribus and on average. It turnes out that higher education goes in line with lower overeducation. This can be understood with the help of other theories that are also well-known in the literature on overeducation, viz the theories of signaling, screening and job competition.

Following the theories of signaling (Spence, 1973) and screening (Stiglitz, 1975), employers have problems to discern the latent future productivity of their applicants. Therefore, they screen them by taking the quality of their diploma as a signal for this latent future productivity and some other latent desired properties (higher self-discipline and motivation, better health, lower future training costs). Therefore, in the competition (Thurow, 1975) for good adequate jobs, applicants with higher education are preferred by employers. Applicants with lower education are ousted to jobs with lower requirements (see for example Muysken and ter Weel, 2000). Skill-biased technological change and organizational change add to this development because they lead to an upgrading of formerly 'simple' jobs with the same effect on less qualified job applicants.

The standard deviation ratio λ (see (8)) and the average overeducation of 0.2222 (see (14)) show that a considerable amount of total variation of income is due to overeducation.

6 Conclusions

This paper has made two contributions to the literature. First, we have removed one serious drawback of the IAB employment sample impeding its applicability to the estimation of earnings frontiers: the censoring of the income data, by multiple imputation.

Afterwards, we have shown that the income ratio approach for measuring overeducation works excellently when applied to such a sufficiently large data set. The income ratio measure is a valuable addition to the overeducation literature because the well-established objective or subjective overeducation measures focus on some (ordinal) matching aspects and ignore the (metric) income and efficiency aspects of overeducation. Notice that all previous attempts to implement the income ratio approach suffered from econometric or data problems. This study has now provided detailed evidence on the influence of experience, tenure, and education on overeducation.

References

Aigner, D.J. and S.-F. Chu, 1968. On estimating the industry production function. American Economic Review 58, 826 - 839.

Aigner, D.J., C.A.K. Lovell and P. Schmidt, 1977. Formulation and estimation of stochastic frontier production function models. Journal of Econometrics 6, 21 - 37.

Bender, S., A. Haas and C. Klose, 2000. IAB employment subsample 1975-1995: opportunities for analysis provided by the anonymised subsample. IZA Discussion Paper no. 117, Bonn.

Blossfeld, H.-P., 1985. Bildungsexpansion und Berufschancen. Campus, Frankfurt am Main.

Borghans, L. and A. de Grip, 2000. The debate in economics about skill utilization. In: Borghans, L. and A. de Grip (eds.), The overeducated worker? The economics of skill utilization. Edward Elgar, Cheltenham/UK, 3 - 23.

Büchel, F., 2001. Overqualification - reasons, measurement issues, and typological affinity to unemployment. In: CEDEFOP: Descy, P. and M. Tessaring (eds.), Training in Europe. Second report on vocational training research in Europe 2000. Background report. CEDEFOP reference document. Office for Official Publications of the European Communities, Vol. 2, Luxembourg, 453 - 560.

Büchel, F., A. de Grip and A. Mertens (eds.), 2003. Overeducation in Europe: Current issues in theory and practice. Edward Elgar, Cheltenham/UK.

Chib, S., 1992. Bayes inference in the tobit censored regression model. Journal of Econometrics 51, 79 - 99.

Coelli, T., D.S.P. Rao and G.E. Battese, 1998. An introduction to efficiency and productivity analysis. Kluwer, Boston.

Daneshvary, N., H.W. Herzog Jr., R.A. Hofler and A.M. Schlottmann, 1992. Job search and immigrant assimilation: An earnings frontier approach. Review of Economics and Statistics 74, 482 - 492.

Doeringer, P.B. and M.J. Piore, 1971. Internal labor markets and manpower analysis. Lexington, MA.

Gartner, H. and S. Rässler, 2005. Analyzing the changing gender wage gap based on multiply imputed right censored wages. IAB Discussion paper 5/05, Nürnberg.

Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin, 2003. Bayesian data analysis, 2. edn. Chapman & Hall/CRC, Boca Raton.

Greene, W.H., 1997. Frontier production functions. In: Pesaran, M.H. and P. Schmidt, Handbook of applied econometrics. Blackwell, 81 - 166.

Hamann, S., G. Krug, M. Köhler and W. Ludwig-Mayerhofer, 2004. Die IAB-Regionalstichprobe 1975-2001: IABS-r01. ZA-Information 55, 34 - 59.

Jensen, U., 2001a. Robuste Frontierfunktionen, methodologische Anmerkungen und Ausbildungsadäquanzmessung. Peter Lang, Frankfurt.

Jensen, U., 2001b. The simplicity of an earnings frontier. In: Zellner, A., H.A. Keuzenkamp and M. McAleer (eds.), Simplicity, inference and modelling. Cambridge University Press, 277 - 291.

Jensen, U., 2003. Measuring overeducation with earnings frontiers and panel data. In: Büchel, F., A. de Grip and A. Mertens (eds.), Overeducation in Europe: Current issues in theory and practice. Edward Elgar, 155 - 169.

Jensen, U., 2005. Misspecification preferred: The sensitivity of inefficiency rankings. Journal of Productivity Analysis 23/2, 223 - 244.

Jovanovic, B., 1979. Job matching and the theory of turnover. Journal of Political Economy 87/5, 972 - 990.

Little, R.J.A. and D.B. Rubin, 1987. Statistical analysis with missing data, 1. edn. John Wiley, New York.

Little, R.J.A. and D.B. Rubin, 2002. Statistical analysis with missing data, 2. edn. John Wiley, New York.

Meeusen, W. and J. van den Broeck, 1977. Efficiency estimation from Cobb-Douglas production functions with composed error. International Economic Review 18, 435 - 444.

Meng, X.-L., 1994. Multiple-imputation inferences with uncongenial sources of input (with discussion). Statistical Science 9, 538-573.

Mincer, J., 1974. Schooling, experience and earnings. New York.

Muysken, J. and B. ter Weel, 2000. Overeducation and crowding out of low-skilled workers. In: Borghans, L. and A. de Grip (eds.), The overeducated worker? The economics of skill utilization. Edward Elgar, Cheltenham/UK, 109 - 132.

Reifschneider, D. and R. Stevenson, 1991. Systematic departures from the frontier: A framework for the analysis of firm inefficiency. International Economic Review 32, 715 - 723.

Ritter, C. and L. Simar, 1997. Pitfalls of normal-gamma stochastic frontier models. Journal of Productivity Analysis 8, 167 - 182.

Rubin, D.B., 1978. Multiple imputation in sample surveys - a phenomenological Bayesian approach to non-response. Proceedings of the Survey Research Methods Sections of the American Statistical Association, 20 - 40.

Rubin, D.B., 1983. A case study of the robustness of Bayesian methods of inference: Estimating the total in a finite population using transformations to normality. In: Scientific inference, data analysis and robustness. Academic Press, Inc., New York, 213 - 244.

Rubin, D.B., 1987. Multiple imputation for non-response in surveys. Wiley, New York.

Rubin, D.B., 1996. Multiple imputation after 18+ years. Journal of the American Statistical Association 91, 473 - 489.

Schafer, J.L., 1997. Analysis of incomplete multivariate data. Chapman & Hall, London.

Sicherman, N. and O. Galor, 1990. A theory of career mobility. Journal of Political Economy 98/1, 169 - 192.

Spence, M., 1973. Job market signaling. Quarterly Journal of Economics 88, 355 - 374.

Stevenson, R.E., 1980. Likelihood functions for generalized stochastic frontier estimation. Journal of Econometrics 13, 57 - 66.

Stiglitz, J.E., 1975. The theory of screening, education and the distribution of income. American Economic Review 65/3, 283 - 300.

Thurow, L.C., 1975. Generating inequality - mechanisms of distribution in the US economy. New York.

Wang, H.-J. and P. Schmidt, 2002. One-step and two-step estimation of the effects of exogenous variables on technical efficiency levels. Journal of Productivity Analysis 18, 129 - 144.

Appendix: Tables

Table 1: Descriptive statistics of the variables in (12)

Variable	Mean	Std. dev.	Variable	Mean	Std. dev.
ln(wage)	4.497	0.387	**Type of region**		
Experience	20.121	8.979	Region type 1	0.316	0.465
Experience^2/100	4.855	3.722	Region type 2	0.170	0.375
Experience^3/1000	13.020	13.533	Region type 3	0.059	0.235
Tenure	7.726	7.383	Region type 4	0.017	0.128
Tenure^2/100	1.142	1.790	Region type 5	0.075	0.263
Tenure^3/1000	2.143	4.362	Region type 6	0.180	0.384
Qualification level			Region type 7	0.081	0.273
No qualification	0.115	0.319	Region type 8	0.103	0.304
Vocational training	0.703	0.457			
Abitur	0.009	0.096	**Industry**		
Abitur + voc. training	0.044	0.206	Agriculture	0.006	0.080
Tech. college	0.052	0.222	Energy	0.016	0.126
University	0.076	0.264	Mining	0.007	0.081
Job status			Quarrying	0.075	0.263
Unskilled worker	0.205	0.404	Machine building	0.253	0.435
Craftsman	0.324	0.468	Household goods	0.071	0.256
Master craftsman	0.028	0.164	Food	0.024	0.153
White collar worker	0.443	0.497	Construction	0.052	0.222
Size of firm			Building completition	0.041	0.198
1-5	0.065	0.247	Trade	0.115	0.319
6-20	0.142	0.349	Transport/communic.	0.061	0.239
21-50	0.132	0.338	Business service	0.135	0.342
51-100	0.114	0.318	Household service	0.020	0.141
101-500	0.266	0.442	Social service	0.077	0.266
501-1,000	0.094	0.292	Public service	0.048	0.214
1,001-2,000	0.069	0.254			
2,001-10,000	0.080	0.272			
>10,000	0.037	0.189			

Notes: 165,425 observations; own calculations, based on IABS 2001

Table 2: Stochastic earnings frontier (12), dependent variable: ln(daily wage)

Variable	Coefficient	t value
Constant	4.0745	492.97
Experience/10	0.2863	24.32
Experience^2/100	-0.0810	-12.78
Experience^3/1000	0.0078	7.62
Tenure/10	0.0506	13.53
Tenure^2/100	-0.0149	-9.90
Qualification level		
Vocational training	0.0512	17.36
Abitur	0.1336	16.32
Abitur + voc. training	0.1245	21.83
Tech. college	0.2225	37.19
University	0.3366	49.85
Job status		
Craftsman	0.0584	21.76
Master craftsman	0.3328	64.61
White collar worker	0.2893	61.58
Type of region		
Region type 1	0.0730	27.73
Region type 2	0.0868	36.86
Region type 3	0.0391	13.06
Region type 4	0.0364	7.40
Region type 5	0.0522	13.67
Region type 6	0.0412	15.25
Region type 7	0.0138	5.13
Size of firm		
1-5	-0.1338	-34.41
6-20	-0.1055	-39.86
21-50	-0.0632	-31.95
51-100	-0.0362	-12.85
501-1,000	0.0271	8.55
1,001-2,000	0.0575	15.83
2,001-10,000	0.0787	23.64
>10,000	0.1636	40.78
Industry dummies	yes	
Occupational dummies	yes	
165,425 observations		

Notes: reference group is no qualification, unskilled worker, rural region, firm size 101-500; coefficients and t-values calculated according to equation (18) and (21); own calculations, based on IABS 2001

Table 3: Overeducation effects submodel (10)

Variable	Coefficient	t value
Constant	0.0847	2.24
Experience/10	-0.6937	-11.75
Experience^2/100	0.2722	9.56
Experience^3/1000	-0.0332	-6.68
Tenure/10	-1.0279	-18.48
Tenure^2/100	0.7108	18.50
Tenure^3/1000	-0.1699	-13.56
Qualification level		
Vocational training	-0.2669	-21.50
Abitur + voc. training	-0.3902	-16.97
Tech. college	-0.4550	-17.68
University	-0.4830	-20.45
Job status		
Craftsman	-0.1508	-9.09
Master craftsman	0.2331	7.70
White collar worker	0.1773	8.61
Region dummies	yes	
Firm size dummies	yes	
Occupational dummies	yes	
Industry dummies	yes	
λ	2.2452	74.53
$\bar{u}$	0.2512	
Average overeducation	0.2222	
165,425 observations		

Notes: reference group is no qualification, unskilled worker; the coefficients of regional and firm size dummies are not presented, because many of them are insignificant; coefficients and t-values calculated according to equation (18) and (21); own calculations, based on IABS 2001

Table 4: Definition of regional types

Type of region	**Characterization**
Region type 1	Core cities in regions with major agglomerations
Region type 2	Very densely populated districts in regions with major agglomerations
Region type 3	Densely populated districts in regions with major agglomerations
Region type 4	Rural structured districts in regions with major agglomerations
Region type 5	Core cities in regions with features of conurbations
Region type 6	Densely populated districts in regions with features of conurbations
Region type 7	Rural structured districts in regions with features of conurbations
Region type 8	Rural regions

Appendix: Figures

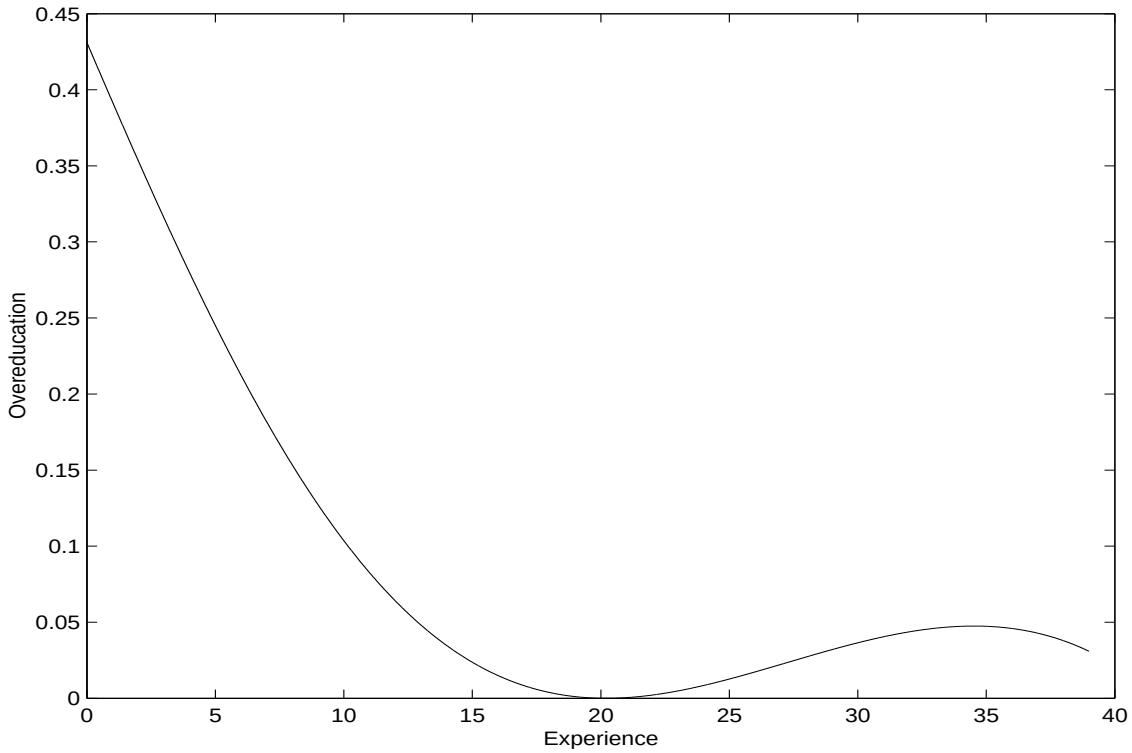

Figure 1: Relation between potential experience and overeducation

Notes: Experience in years; own calculations, based on IABS 2001

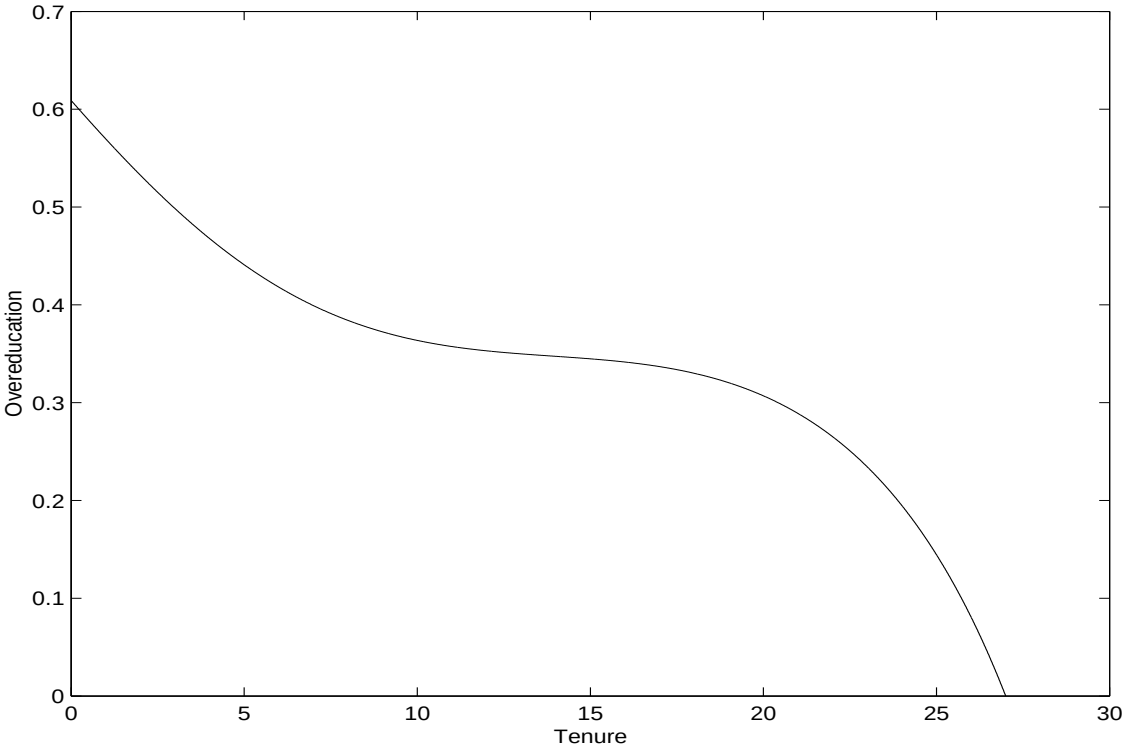

Figure 2: Relation between tenure and overeducation

Notes: Tenure in years; own calculations, based on IABS 2001