Where have all the data gone? Stochastic production frontiers with multiply imputed German establishment data

UWE JENSEN* & SUSANNE RÄSSLER†

Abstract

In this paper, stochastic production frontier models are estimated with IAB establishment data from waves 2002 and 2003 to find important determinants of productivity and inefficiency. The data suffer from nonresponse in the most important variables (output, capital and labor) leading to the loss of 25 % of the observations and possibly imprecise estimates and invalid test statistics. Therefore, the missing values are multiply imputed. Analyzes of the estimation results show that, particularly in the inefficiency submodel, working with multiply imputed data reveals some interesting and plausible results which are not available when ignoring missing observations.

Keywords: Efficiency and productivity measurement, establishment data, Markov chain Monte Carlo, multiple imputation

JEL codes: C15, C24, C81, D24

*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.: +49-911-179-3084, email: Susanne.Raessler@iab.de

Contents

1 Introduction **1**

2 Data and nonresponse **1**
2.1 Data and response behavior . . . 1
2.2 Nonresponse and imputation . . . 3

3 Analyst's model **4**
3.1 Stochastic production frontiers . . . 4
3.2 Analyst's model selection . . . 6

4 Imputer's model: data augmentation **7**
4.1 Introduction to multiple imputation . . . 7
4.2 Data augmentation using the normal/Wishart model . . . 9
4.3 Data preparation . . . 10

5 Results **12**
5.1 The controversial results . . . 13
5.2 The unanimous results . . . 14

6 Conclusions **15**

References **17**

Appendix: Data preparation, variable construction **19**
Variables in the questionnaire (to be transformed) . . . 19
Variables in the regressions . . . 20
Data transformation for MI procedure . . . 21

Tables **22**

1 Introduction

In this paper, stochastic production frontier models are estimated with German establishment data to find important determinants of productivity and inefficiency. We are confronted with missing values in our data set, a typical situation in empirical research. A closer look to the data reveals 4 % to 15 % of missing values particularly in the most important variables: output, capital and labor. Ignoring this would reduce the complete data records available for any multivariate analysis considerably. Whereas information from 18447 observations from the panel waves of 2002 and 2003 is collected in principle, only 13969 observations of them can be used when inference is based only on the complete cases. Ignoring the missing values would certainly lead to lower precision of the estimates. And the question arises whether the remaining data are still representative for the population of interest. If not, the resulting test statistics are no longer valid and the resulting estimates may be biased.

Biases can be expected to occur particularly in the establishment's inefficiency estimates of the stochastic production frontier. Because frontier estimates depend on the extreme efficient establishments in the sample and because the inefficiency estimates are derived from the estimation residuals, the latter are extremely sensitive to any kind of misspecification in the model – see e.g. Jensen (2005). That is why it is the aim of this paper to demonstrate in an empirical application the dangers of ignoring missing data or the gains of properly imputing them when estimating a stochastic production frontier with establishment data.

The article is structured as follows. In the next section, the data and the response behavior in the panel are described. In section 3, the stochastic production frontier model and the selection steps to the analysts's model are presented. In the following section, a short introduction to multiple imputation is provided. We describe the imputation process as well as the preparations and transformations of the variables to be used in the imputer's model. In the fifth section, the estimation results using the imputed data are given and compared with the results based only on the complete data. Finally, section six summarizes the paper.

2 Data and nonresponse

2.1 Data and response behavior

Our data are taken from two waves (2002 and 2003) of the Establishment Panel of the Institute for Employment Research of the Federal Labor Service (Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit, IAB). The basis for the panel is the employment statistics register of the Federal Employment Service, conducted within the framework of the 1973 revisions to the social insurance system. Each year, all employers are required, under sanction, to report levels of and changes in the number of their employees who are subject to the compulsory social security scheme. The register covers all dependent employment in the private and public sector and accounts for almost 85% of total employment in Germany. The survey unit of the register is the establishment or local production unit, rather than the legal and commercial entity of the company.

The IAB Establishment Panel draws a stratified random sample of units from the register, the selection probabilities depend on the employment frequency of the respective stratum. The strata comprise some 20 industries and 10 establishment size intervals covering all sectors and employment levels. The overall and size-specific response rates including firms that are interviewed for the first time exceed 60 percent, and, for repeatedly interviewed establishments, more than 80 percent.

The panel is designed to meet the needs of the Federal Labor Service. Basically, it focusses on employment-related matters. Much of the information in the panel concerns worker characteristics and qualifications as well as levels of and changes in establishment employment. There is also information on the training of employees and their working time. Additionally, information on certain establishment policies, business developments, and investment is collected on an annual basis. Other information is collected biennially or triennially. Each year the panel also addresses a specific topic.

We exclude all establishments from the sample that do not use turnover as an output measure. This affects non-profit organisations, public offices, banks and insurances. Thus, an unbalanced sample of 13969 observations remains without any item-nonresponse on the variables used in this study. Multiple imputation provides 18447 data records for 2002 and 2003 from 9462 establishments.

Unfortunately, we do not have exact information about the reasons for unit-nonresponse and drop-out in the data. It is commonly assumed that next to the general attitude to take part in a survey there are two main reasons for nonresponse. First, there are questions that are too difficult to understand or the information wanted is not easily available and, second, there are questions that concern sensitive information. In both cases, the interviewee is not willing to participate in the panel. A study for earlier waves of the panel comes to the result that only a few items influence the willingness of firms to participate significantly (see Hartmann and Kohaut, 2000).

Mainly, item-nonresponse in the data is found only in few variables, particularly those used to construct output, labor and capital. Output is measured as value added, capital by the replacement investment and labor by earnings (see subsection 3.2 and the data appendix for the correct definitions). Table 1 gives the variables in the questionnaire with the highest item-nonresponse rates. All the other variables used in our study are distinctly below the rates shown there.

Table 1: Variables with the highest nonresponse (in %)

Variable	2002	2003
Turnover	13.69	15.05
Input of materials, goods and services	11.99	12.67
Total gross monthly wages in June	11.07	12.78
Investment to enlarge capital	8.38	6.92
Investment	4.19	4.51

2.2 Nonresponse and imputation

First formalized by Rubin (1976), in modern statistical literature (see Little and Rubin 1987, 2002, p. 12) the missing data mechanisms are commonly distinguished according to the probability of response yielding the following three cases:

- The missing data are said to be missing completely at random (MCAR), if the nonresponse process is independent of both unobserved and observed data.
- If, conditional on the observed data, the nonresponse process is independent only of the unobserved data, then the data are missing at random (MAR). This is the case, e.g., if the probability of answering the turnover question varies according to the size of the company, and the size is observed.
- Finally, data are termed not missing at random (NMAR), if the nonresponse process depends on the values of the variables that are actually not observed. This might be the case for turnover reporting, where companies with higher turnover tend to be less likely to report their turnover.

In the context of likelihood-based inference and when the parameters describing the measurement process are functionally independent of the parameter describing the nonresponse process, MCAR and MAR are said to be ignorable; otherwise we call it nonignorable missingness which is the hardest case to deal with analytically because the missingness mechanism has to be modeled itself.

As mentioned above, the highest amount of missing values occurs in the most important variables: output, capital and labor. A further analysis of the amount of data missing per variable shows that item-nonresponse is higher the larger the companies are. So, the establishment size in terms of the number of employees seems to be a good predictor of missingness. Therefore, we assume that the missing values of the variables used in the productivity model are missing at random (MAR). As it is often the case, the missing values are spread around in the data set. If we estimate our model by any econometric software, we loose 25 % of the observations which still contain hard-earned information.

Moreover, basing inference only on the complete cases in our application implicitly assumes that the data are missing completely at random (MCAR) which obviously is not the case. To ensure the MAR-assumption and allow to estimate a sophisticated econometric model with missing data, we decided to use a multiple imputation procedure. Using a single imputation technique such as mean imputation, hot deck, or regression imputation, in general results in confidence intervals and p-values that ignore the uncertainty due to the missing data, because the imputed data were treated as if they were fixed known values. Thus, basing standard complete data inference on singly imputed data will typically lead to standard error estimates that are too small, p-values that are too significant, and confidence intervals that undercover – see, e.g., Rässler et al. (2003). To correct for these effects using singly imputed data, special variance estimation techniques have to be applied. For a very recent discussion of the merits and demerits of single and multiple imputation see Groves et al. (2002).

Notice that the ignorability assumption can never be contradicted by the observed data. However, Schafer (2001) provides evidence that even the erroneous assumption of MAR might have only minor impact on estimates and standard errors using a proper multiple imputation strategy. Only when NMAR is a serious concern, it is obviously necessary to jointly model the data and the missingness, although such models are based on other untestable assumptions. Therefore, a multiple imputation procedure seems to be the best alternative at hand in our situation to account for missingness, to exploit all valuable information, and to get statistically valid subsequent analyses based on standard complete data inference.

3 Analyst's model

3.1 Stochastic production frontiers

This subsection summarizes the theory on stochastic production frontiers necessary in the following.

In microeconomic theory, economic 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.

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

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

or in logs

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

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,

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \tag{3}$$

is estimated maximum possible output (in logs) for given inputs. The log output difference

$$u_i = \hat{y}_i - y_i \geq 0 \tag{4}$$

or better the output ratio

$$0 \leq TE_i = \exp(-u_i) = \frac{Y_i}{\hat{Y}_i} \leq 1 \tag{5}$$

is interpreted as technical inefficiency of firm no. i. Finally, the composed error term e_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 log-likelihood function is $l(\beta, \sigma, \lambda, \mu) =$

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

with

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

and the 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 e and not u, the inefficiencies must be estimated indirectly with the help of the minimum mean-squared error predictor

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

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

Independence of x_{ij} and u_i may be a hard assumption. That is why Reifschneider and Stevenson (1991) 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

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

The ML estimators $\hat{\beta}_j$, $\hat{\delta}_j$, $\hat{\sigma}$ and $\hat{\lambda}$ are derived simultaneously using iterative ML techniques. The inefficiency terms u_i have to be estimated indirectly again.

See the given references for the likelihood function of the full model etc. and see the surveys in Coelli et al. (1998), Greene (1997) or Jensen (2001a) for more details on frontiers.

3.2 Analyst's model selection

This subsection documents the model selection steps in the derivation of the specification of the estimated model.

The first decision for the analyst is on the functional form for the relation between output, capital and labor. In order to avoid the well-known hard restrictions of simpler functions like Cobb-Douglas, we have chosen the rather general translog production function.

The second decision was on the measurement of output, capital and labor. Output is measured by the value added (see the appendix on variable construction for exact definitions). We excluded all establishments from the sample that do not use turnover as output measure. This affects non-profit organisations, public offices, banks and insurances. In the imputed data-sets, 3 distinct outliers in the output variable had to be eliminated because – particularly with a frontier function – they would significantly bias the estimates.

A reasonable measure for labor input should take account of skill and productivity differences between employees, among others. For labor, the data set provides two possible approximations: full-time equivalents (total number of employees minus 0.5 times total number of part-time employees) or earnings. The first choice would implicitly assume e.g. that all employees are equally skilled and productive whereas the second choice implicitly assumes that earnings are a good proxy for skills and productivity, among others. We decided for the latter because that assumption seems to be more reasonable.

The capital variable is notorious for the difficulties any approximation to the latent value of the capital stock causes in the estimation. With time series data, the capital variable approximated by the perpetual inventory method often shows low variation and non-stationarity. In this paper, with cross-section data covering two years, we decided to proxy capital by the replacement investment in the current year. Of course, this choice implicitly assumes that capital is replaced uniformly and sufficiently, among others. An alternative would be to approximate capital by the average replacement investment of several years. But since firms are born and die, this approximation would lead to even more missing values or firms.

In subsection 2.1, we have shown that replacement investment is one the variables suffering from many missing values. This problem will be soothed by multiple imputation. But another problem is that a large part (7888 of 18447) of the values on investment in the sample are zero. There is some evidence that many of these firms are simply not able or not willing to provide exact non-zero investment numbers. That is why one important contribution of our paper is the suggestion to multiply impute these zeroes as well. Section 5 will show the consequences of this additional imputation of the capital variable.

After these fundamental decisions, the covariates of labor and capital in the production function and the inefficiency determinants in submodel (10) had to be found. It is well-known that forward and backward variable selection procedures can lead to very different results when the regressors are correlated. That is why a very detailed data analysis including a factor analysis to examine the correlation structure of the regressors was conducted. Then, in a large-scale model selection procedure combining several forward and backward runs (using both the imputed data and only the observed data), the final sets of variables for the production function and the submodel were fixed. Every variable had several opportunities to enter the production function and the submodel. A variable is included in all regressions if it was significant in at least one of the 11 regressions (5 + 5 auxiliary regressions with imputed data and one with only the observed data). The appendix on variable construction shows the exact definitions of all variables and the tables show the use of the variables.

4 Imputer's model: data augmentation

4.1 Introduction to multiple imputation

Multiple imputation (MI), introduced by Rubin (1978) and discussed in detail in Rubin (1987), is a Monte Carlo technique replacing missing values by $m > 1$ simulated versions, generated according to a probability distribution or, more generally, any density function indicating how likely imputed values are given the observed data. MI therefore is an approach that retains the advantages of imputation while allowing the data analyst to make valid assessments of uncertainty. The concept of multiple imputation reflects uncertainty in the imputation of the missing values through wider confidence intervals and larger p-values than under single imputation. Typically m is small, with $m = 3$ or $m = 5$. Each of the imputed and thus completed data sets is first analyzed by standard methods. Then, the results are combined or pooled to produce estimates and confidence intervals that reflect the missing data uncertainty.

The theoretical motivation for multiple imputation is Bayesian. Let Y_{obs} denote the observed components of any uni- or multivariate variable Y, and Y_{mis} its missing components. Basically, MI requires independent random draws from the posterior predictive distribution

$$f(y_{\text{mis}}|y_{\text{obs}}) = \int f(y_{\text{mis}}, \psi|y_{\text{obs}})\, d\psi = \int f(y_{\text{mis}}|y_{\text{obs}}, \psi)\, f(\psi|y_{\text{obs}})\, d\psi \tag{12}$$

of the missing data Y_{mis} given the observed data Y_{obs} with parameter vector ψ. Since $f(y_{\text{mis}}|y_{\text{obs}})$ itself often is difficult to derive, we may alternatively perform

- random draws of the parameters according to their observed-data posterior distribution $f(\psi|y_{\text{obs}})$ as well as
- random draws of the missing data according to their conditional predictive distribution $f(y_{\text{mis}}|y_{\text{obs}}, \psi)$ given the drawn parameter values.

For many models the conditional predictive distribution $f(y_{\text{mis}}|y_{\text{obs}}, \psi)$ is rather straightforward due to the data model used. On the contrary, the corresponding observed-data posterior

$$f(\psi|y_{\text{obs}}) = L(\psi; y_{\text{obs}})\frac{f(\psi)}{f(y_{\text{obs}})} \tag{13}$$

(with the likelihood function $L(\psi; y_{\text{obs}}) = f(y_{\text{obs}}|\psi)$) usually is difficult to derive, especially when the data have a multivariate structure and different, non-monotone missing data patterns. The observed-data posteriors often are not standard distributions from which random numbers could easily be generated. Therefore, simpler methods have been developed to enable multiple imputation on the grounds of Markov chain Monte Carlo (MCMC) techniques. They are extensively discussed by Schafer (1997). In MCMC, the desired distributions $f(\psi|y_{\text{obs}})$ and $f(y_{\text{mis}}|y_{\text{obs}})$ are achieved as stationary distributions of Markov chains which are based on the complete-data distributions which are more easily computed. Creating m independent draws from such chains can be used as imputations of Y_{mis} from their posterior predictive distribution $f(y_{\text{mis}}|y_{\text{obs}})$.

Based on these m imputed data sets we calculate m complete data statistics $\hat{\theta}^{(r)}$ and their variance estimates $\hat{\mathrm{V}}(\hat{\theta}^{(r)})$, $r = 1, \ldots, m$. The complete-case estimates are combined according to Rubin's rule such that the MI point estimate $\hat{\theta}_{\text{MI}}$ for parameter θ is the average

$$\hat{\theta}_{\text{MI}} = \frac{1}{m}\sum_{r=1}^{m} \hat{\theta}^{(r)} \tag{14}$$

Its estimated total variance T is calculated according to the analysis of variance principle:

$$\begin{aligned}
\text{'between-imputation variance':}\quad B &= \frac{1}{m-1}\sum_{r=1}^{m}\left(\hat{\theta}^{(r)} - \hat{\theta}_{\text{MI}}\right)^2 \\
\text{'within-imputation variance':}\quad W &= \frac{1}{m}\sum_{r=1}^{m}\hat{\mathrm{V}}\left(\hat{\theta}^{(r)}\right) \\
\text{'total variance':}\quad T &= W + \left(1 + \frac{1}{m}\right)B
\end{aligned} \tag{15}$$

For large sample sizes, tests and two-sided interval estimates can be based on the Student's t-distribution

$$\frac{\hat{\theta}_{\mathrm{MI}} - \theta}{\sqrt{T}} \dot{\sim} t(v) \quad \text{with} \quad v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2 \tag{16}$$

degrees of freedom. For a comprehensive overview of MI see Schafer (1999a)

Multiple imputation is in general applicable when the complete-data estimates are asymptotically normal or t distributed; e.g., see Rubin and Schenker (1986), Rubin (1987), Barnard and Rubin (1999), or Little and Rubin (2002). Notice that the usual maximum-likelihood estimates and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions. In this paper we use ML estimation for the analyst's model.

4.2 Data augmentation using the normal/Wishart model

For the creation of the multiple imputations we use the stand alone software NORM provided for free by Schafer (1999b).

We assume a k-dimensional normal distribution for all the k variables in the imputer's model. Moreover we assume to have n independent observations from this data model; i.e., for every observable variable Y_i of each unit i holds that $Y_i \sim N(\mu, \Sigma)$, $i = 1, \dots, n$.

As prior distribution $f(\mu, \Sigma)$ for the location and scale parameters, the common uninformative prior distribution

$$f(\mu, \Sigma) \approx f(\mu)\, f(\Sigma) \approx c\, |\Sigma|^{-(k+1)/2} \propto |\Sigma|^{-(k+1)/2} \tag{17}$$

is chosen; i.e., μ and Σ are assumed to be approximately independent – for details see Schafer (1997). As long as no identification problems occur, the assumption of a noninformative prior distribution seems to be the most 'objective' choice.

Under this prior distribution (17), the complete-data posterior distribution $f(\mu, \Sigma | y)$ of the parameters given the complete data is a normal distribution for μ given Σ and the data and an inverted-Wishart distribution for Σ given the data

$$\begin{aligned} \Sigma | y &\sim W^{-1}(n-1, (nS(\bar{y}))^{-1}) \\ \mu | \Sigma, y &\sim N(\bar{y}, \Sigma/n) \end{aligned} \tag{18}$$

with the sample covariance matrix

$$S(\bar{y}) = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})(y_i - \bar{y})', \quad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{19}$$

and $y_i = (y_{i1}, \dots, y_{ik})'$. According to the data model, the conditional predictive distribution of the missing data given the observed data and the parameters is a conditional normal distribution

$$Y_{\mathrm{mis}} | y_{\mathrm{obs}}, \mu, \Sigma \sim N(\mu_{\mathrm{mis|obs}}, \Sigma_{\mathrm{mis|obs}}). \tag{20}$$

The data augmentation algorithm proceeds iteratively in two steps, the so-called imputation step and the posterior step.

I-step: For each unit i with missing values, random draws are performed for the missing data from their conditional predictive distribution $f(y_{\text{mis}}|y_{\text{obs}}, \theta)$, see (20), given the observed data and an actual draw of the parameters $\mu^{(t)}$ and $\Sigma^{(t)}$; i.e., random values are generated according to

$$Y_{\text{mis}}^{(t)}|y_{\text{obs}}, \mu^{(t)}, \Sigma^{(t)} \sim N\left(\mu_{\text{mis}|\text{obs}}^{(t)}, \Sigma_{\text{mis}|\text{obs}}^{(t)}\right) \tag{21}$$

P-step: Using the completed data $y^{(t)} = \left(y_{\text{obs}}, y_{\text{mis}}^{(t)}\right)$, actual values for the mean vector $\bar{y}^{(t)}$ and the covariance matrix

$$S(\bar{y}^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(t)} - \bar{y}^{(t)}\right)\left(y_i^{(t)} - \bar{y}^{(t)}\right)' \tag{22}$$

are calculated. Then, new actual values for the parameters $\mu^{(t)}$ and $\Sigma^{(t)}$ are drawn according to their complete-data posterior distribution (18)

$$\begin{aligned} \Sigma^{(t+1)}|y^{(t)} &\sim W^{-1}\left(n-1, \left(nS\left(\bar{y}^{(t)}\right)\right)^{-1}\right) \\ \mu^{(t+1)}|\Sigma^{(t+1)}, y^{(t)} &\sim N\left(\bar{y}^{(t)}, \Sigma^{(t+1)}/n\right) \end{aligned} \tag{23}$$

Such random draws of $\mu^{(t)}$ and $\Sigma^{(t)}$ are considered to be the Bayesianly stochastic counterpart of maximizing the complete-data likelihood being performed in the M-step of the EM algorithm. Analogous to the EM, which uses the complete-data likelihood, data augmentation makes use of the complete-data posterior, which often is more attractive than the observed-data posterior.

Using some starting values $\mu^{(0)}$ and $\Sigma^{(0)}$, the two steps with (21), (22), and (23) are repeated many times until independence from the starting values is achieved and convergence of the Markov chain can be assumed. For $t \to \infty$, the Markov chain $\left\{\left(\mu^{(t)}, \Sigma^{(t)}, Y_{\text{mis}}^{(t)}\right) | \, t = 0, 1, \ldots\right\}$ converges in distribution to $f(y_{\text{mis}}, \theta|y_{\text{obs}})$. Thus, $Y_{\text{mis}}^{(t)}$ converges to a draw from the desired posterior predictive distribution $f(y_{\text{mis}}|y_{\text{obs}})$ given in (12). After assessing convergence, e.g. every $t + 100$, $t + 200$, $\ldots$ value can be used to produce m independent multiple imputations. Data augmentation techniques have been used in practice and provide rather flexible tools for creating multiple imputations from parametric models. A very detailed description of this data augmentation algorithm is given by Schafer (1997).

4.3 Data preparation

In the normal/Wishart model, we assume a multivariate normal distribution for the data. Clearly, our survey data are not normally distributed: some are bounded between zero and one, others are skewed and some have large proportions of zeros; the latter are called semi-continuous variables. A way to handle non-normality of the data is by applying suitable transformations to the variables which is done in our application. Moreover, if non-normal variables (such as discrete or binary ones) are completely observed, then it is quite plausible to still use the multivariate normal model because incomplete variables are modeled as conditional normal given a linear function of the complete variables – see, e.g., Schafer (1997). The variables and their transformations used in our models are listed in the appendix.

When a variable is treated as being semi-continuous, then it has a proportion of responses at the fixed value of, e.g., zero and a continuous distribution among the remaining observations. Subject to an approach published by Schafer and Olsen (1999), one may encode each semi-continuous variable Y to a binary indicator W (with $W = 1$ if $Y \neq 0$ and $W = 0$ if $Y = 0$) and a continuous variable V which is treated as missing whenever $Y = 0$. See table 2 for an illustration.

Table 2: Example: preparation of semi-continuous variables

Y		W	V
2		1	2
0	$\rightarrow$	0	NA
NA		NA	NA

Notice that a relationship between W and V would have little meaning and could not be estimated by the observed data. However, we aim at generating plausible imputations for the original semi-continuous variable Y and, thus, are only interested in the marginal distribution for W and the conditional distribution for V given $W = 1$. Data augmentation algorithms have been shown to behave well in this context with respect to the parameters of interest – see Schafer and Olsen (1999).

When the values of the variables Y (or the remaining V) are bounded between zero and one representing probabilities, a conventional logit-transformation (see Greene, 2003) works quite well:

$$g(Y) = \frac{Y}{1-Y} \quad \text{for} \quad Y \in (0,1) \tag{24}$$

For positively skewed Y, an ordinary log transformation $g(Y) = \ln(Y)$ often is a good choice. Another useful transformation is the Box-Cox transformation

$$g(Y) = \frac{Y^{\theta} - 1}{\theta} \quad \text{for} \quad \theta \neq 0 \tag{25}$$

However, theoretically, we should transform the data to achieve multivariate normality. Practically, such transformations are not yet available: the usual transformations are performed on a univariate scale. Investigations show that such deviations from normality (for the variables to be imputed) should not harm the imputation process too much – see

Schafer (1997) or Gelman et al. (1998). A growing body of evidence supports the claim to use a normal model to create multiple imputations even when the observed data are somewhat non-normal. The focus of the transformations is rather to achieve a range for continuous variables to be imputed that theoretically have support on the whole real line than to achieve normality itself. Even for populations that are skewed or heavy-tailed, the actual coverage of multiple imputation interval estimates is reported to be very close to the nominal coverage. The multiple imputation framework has been shown to be quite robust against moderate departures from the data model – see Schafer (1997). Caution is required if the amount of missing information is very high; i.e., beyond 50% – which is not the case in this paper. Thus, we may proceed further with these transformed data.

With NORM 2.03, the imputations are created very easily. After a burn-in period of 2000 iterations, every further 200 iterations the imputed data sets are stored. Finally, $m = 5$ multiply imputed data sets are used for our analysis. Investigations of time-series and autocorrelation plots did not suggest any convergence problems. Notice that in the imputer's and the analyst's model the same set of input data, i.e., variables and observations, is used to avoid problems of misspecification – see Meng (1995) or Schafer (2001).

5 Results

The stochastic production frontier (2) with inefficiency submodel (10) has been estimated with the IAB German establishment data described in subsection 2.1. The production function has translog form in capital and labor and includes further variables given in the appendix where the variables of the inefficiency submodel are given as well. As described in subsection 3.2, 11 regressions have been run for 3 approaches:

- Approach MISS: One regression with only the observed data. See tables 3 and 3a for the results.
- Approach MIC0: $m = 5$ auxiliary regressions with the full data set where all missing values have been filled by multiple imputation (see section 4) but where the zeroes in the capital variable are maintained. Tables 5 and 5a provide the results of the auxiliary regressions, tables 3 and 3a provide the pooled results.
- Approach MIMI: $m = 5$ auxiliary regressions with the full data set where all missing values and the zeroes in the capital variable have been filled by multiple imputation. Tables 4 and 4a provide the results of the auxiliary regressions, tables 3 and 3a provide the pooled results.

Estimation has been performed with LIMDEP 8.0.

5.1 The controversial results

In the following, 'significance' means 'significance on the 5 % level' if not otherwise stated. We begin comparing the results on the production frontier in table 3. Here, all 3 approaches perform rather similar – with one important exception. In the MIC0 approach, labor is insignificant, even with changing signs in the auxiliary regressions (see table 5). This certainly is a severe drawback of this approach.

Apart from that, it strikes that higher export activity leads to higher productivity only when missing observations remain missing whereas, after multiple imputation, the export parameter becomes insignificantly or weakly significantly negative. See the next subsection for the relation between export activity and efficiency.

Another interesting difference is the effect of collective agreements on productivity. With multiply imputed data, there is evidence for reduced productivity whereas, with missing observations, the parameter is insignificantly positive. The net effect of collective bargaining on productivity is an open question in labor economics (see e.g. Filer et al., 1996, p. 513). Studies with German data mostly seem to have not found effects of collective bargaining on productivity (see e.g. Schnabel, 1991). But this might be caused by too many missing observations...

More striking differences between the approaches are found in the results on the inefficiency submodel in table 3a. With multiply imputed data,

- labor has a weakly significantly positive effect on u, i.e. a weakly significantly negative effect on efficiency – see (5) – whereas, with missing observations, higher wage costs significantly increase efficiency. It is interesting to see that, with multiply imputed data, the univariate relation between efficiency and labor is positive. This means that the covariates are more influential on this relation in these approaches. The negative effect of wages on efficiency could be explained by standard arguments from labor economics, namely shirking theory (Lazear, 1981): Larger firms with many employees have problems with monitoring the work effort of their employees. The solution are higher relative wages and the threat of being discharged, a powerful disciplinary threat. But, of course, this might be inefficient.

- higher exports significantly raise efficiency whereas the influence is weakly significantly negative with missing observations. The parameters of the production frontier (2) and inefficiency submodel (10) are jointly estimated (see subsection 3.1). Thus, substitution between effects on productivity and efficiency may occur. Whereas the MISS approach finds a positive effect of exports on productivity (see the previous subsection), the MIC0/MIMI approaches see a positive effect on efficiency.

- collective agreements (weakly) significantly coincide with higher efficiency whereas the influence is insignificantly negative with missing observations.

- firms receiving relatively more wage subsidies are significantly less efficient. Employees receiving wage subsidies might not work efficiently. This effect is only weakly significant with missing observations.

- firms supporting relatively more on-the-job-training cases are less efficient. This can make sense because the returns to the firm costs of on-the-job-training might not be sufficient. This effect is insignificant with missing observations, where firms supporting the use of PCs for on-the-job-training cases are significantly less efficient.

- the variance ratio λ in (8) is distinctly higher than with missing observations meaning that noise, i.e. the denominator in (8), constitutes a relatively larger part of total variance in the latter case.

- mean technical efficiency – see (5) – is distinctly higher (55 %) than with missing observations (48 %).

- most parameter estimates are drastically higher than with missing observations.

Since we are working with real data and not with simulated data, we don't know anything about the true parameter values. Hence, we are not able to say which results come closer to the truth. Nevertheless, particularly in the inefficiency submodel, working with multiply imputed data seems to reveal some interesting and plausible results which are not available with missing observations. And, summarizing the performance of the two multiple imputation approaches, the MIC0 approach suffers from the serious drawback of counterintuitively producing an insignificant labor parameter in the production function. So, we have a small but significant preference for the results obtained with multiple imputation where the capital zeroes are imputed as well.

5.2 The unanimous results

In this subsection, a larger part of the unanimous and significant results are interpreted. We start with the results on the production function.

- Except the capital parameter in the MIC0 approach (see the previous subsection), the capital and labor parameters show the expected signs and ranges.

- OUTPROGP/OUTPROGN: If turnover is expected to increase (decrease), it seems to be rather low (high). Thus, an expected increase (decrease) goes in line with lower (higher) productivity.

- DEVELOP: If the technical condition of a firm is up to date, the productivity is higher.

- NEWWORK: Firms with relatively many new hires (having little firm-specific human capital) are less productive.

- SKSEARCH: Firms searching relatively many skilled employees as of now are producing on the efficient frontier and would like to expand.

- FLUCT: Stronger production fluctuations lead to lower productivity.

- EAST: Enterprises which are by majority in East German property are less productive, a well-known result.

- TRAIND/TRAINPC: Firms supporting on-the-job-training (with or without PCs) are more productive.

- PROP1: Firms offering many jobs for whom experience is important do not seem to operate on the technological frontier and hence are less productive.

Finally, two stable significant results on the inefficiency submodel are:

- SKILL: Firms with relatively many skilled employees are producing more efficiently.

- PROP4: Firms offering many jobs for whom creativity is important might be exposed to relatively many production risks leading to lower efficiency.

6 Conclusions

In this paper, we have demonstrated in an empirical application the gains of properly imputing missing data when estimating a stochastic production frontier with establishment data. Frontier estimates and particularly establishment's inefficiency estimates are known to react extremely sensitive to any kind of misspecification.

In conventional empirical research concerning econometric issues, often missing data are simply ignored and analysis is based on the complete cases only. Omitting valuable information that is already in the data is statistically inefficient and often leads to substantially biased inferences when the data are not missing completely at random (MCAR), which is the case in most typical settings. In general, multiple as well as single imputation techniques can be used under a less restrictive MAR-assumption. However, with single imputation, standard complete-case analysis can often not be applied directly, because it leads to standard errors that are too small, p-values that are too significant, and confidence intervals that undercover. Especially when inference is drawn from a multivariate and complex model, we regard multiple imputation as the most flexible tool to get valid inference if the data are exposed to nonresponse.

A further contribution of this paper is the double imputation of the capital variable proxied by the replacement investment in the current year. Replacement investment suffers from many missing values and from the fact that a large part of its values in the sample are zero. Since there is some evidence that many of these firms are simply not able or not willing to provide exact non-zero investment values we have suggested to multiply impute these zeroes as well.

Having worked with real data, we are not able to say which results come closer to the truth. But, particularly in the inefficiency submodel, working with multiply imputed data seems to reveal some interesting and plausible results which are not available with missing observations. And, comparing the performance of the two multiple imputation approaches, the approach which maintained the zeroes in the capital variable suffers from counterintuitively producing an insignificant labor parameter in the production function. Thus, we have a small but distinct preference for the results obtained with multiple imputation where the capital zeroes are imputed as well.

Missing values are a typical problem in empirical research. We hope that our study helps raising the probability that proper multiple imputation tools will be more widespread in standard econometric software as soon as possible.

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.

Barnard, J. and Rubin, D.B. (1999): Small-Sample Degrees of Freedom with Multiple Imputation, Biometrika 86, 948 - 955.

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

Filer, R.K., D.S. Hamermesh and A.E. Rees (1996): The economics of work and pay, 6th ed., Harper Collins (New York).

Gelman, A., G. King, and C. Liu (1998): Not asked and not answered – multiple imputation for multiple surveys (with discussion), Journal of the American Statistical Association 93, 846 - 869.

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

Greene, W.H. (2003): Econometric analysis, 5th ed., Prentice Hall (Upper Saddle River).

Groves, R.M., D.A. Dillman, J.L. Eltinge and R.J.A. Little (2002): Survey nonresponse, Wiley (New York).

Hartmann, J. and S. Kohaut (2000): Analysen zu Ausfällen (Unit-Nonresponse) im IAB-Betriebspanel, Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 39, 609 - 618.

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

Jensen, U. (2005): Misspecification preferred: The sensitivity of inefficiency rankings, Journal of Productivity Analysis 23/2, 219 - 240, to appear.

Lazear, E. (1981): Agency, earnings profiles, productivity and hours restrictions, American Economic Review 71, 606 - 620.

Little, R.J.A. and D.B. Rubin (1987, 2002): Statistical analysis with missing data, Wiley (New York).

Meng, X.L. (1995): Multiple-imputation inferences with uncongenial source of input (with discussion), Statistical Science 10, 538 - 573.

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

Rässler, S., D.B. Rubin and N. Schenker (2003): Imputation, in: Bryman, A., M. Lewis-Beck and T.F. Liao (eds.), Encyclopedia of social science research methods, Sage.

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

Rubin, D.B. (1976): Inference and missing data, Biometrika 63, 581 - 592.
Rubin, D.B. (1978): Multiple Imputation in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse, Proceedings of the Survey Research Methods Sections of the American Statistical Association, 20 - 40.
Rubin, D.B. (1987): Multiple imputation for nonresponse in surveys, Wiley (New York).
Rubin, D. B. and Schenker, N. (1986): Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. Journal of the American Statistical Association 81, 366 - 374.
Schafer, J.L. (1997): Analysis of incomplete multivariate data, Chapman & Hall (London).
Schafer, J.L. (1999a): Multiple imputation – a primer, Statistical Methods in Medical Research 8, 3-15.
Schafer, J.L. (1999b): Multiple imputation under a normal model, version 2, Software for Windows 95/98/NT, http://www.stat.psu.edu/ jls/misoftwa.html.
Schafer, J.L. (2001): Multiple imputation in multivariate problems when the imputation and the analysis models differ, in: Bethlehem, J. and S. van Buuren (eds.), Missing values proceedings of a symposium on incomplete data, Utrecht, 1 - 21.
Schafer, J.L. and M.K. Olsen (1999): Modeling and imputation of semi-continuous survey variables. Technical report 00-39, Pennsylvania State University.
Schnabel, C. (1991): Trade unions and productivity: the German evidence, British Journal of Industrial Relations 19, 15 - 24.
Stevenson, R.E. (1980): Likelihood functions for generalized stochastic frontier estimation, Journal of Econometrics 13, 57 - 66.

Appendix: Data preparation, variable construction

Variables in the questionnaire (to be transformed)

SALE	turnover in EUR
INPUT	input of materials, goods and services in % of turnover
INVEST	investment in EUR
ADDINV	investment to enlarge capital in % of investment
EMP	total number of employees
NOVERTIM	total number of employees with paid overtime in previous year
EXPORT	export in EUR
NSKILL	total number of highly skilled employees
NONEWHIR	dummy: NONEWHIR = 1 if no new hires in first half-year
WOULD	dummy: WOULD = 1 if employer wanted to hire new employees
NNEWHIR	total number of new hires in first half-year
QUIT	total number of quits in first half-year
NTERMIN	total number of terminations by employees in first half-year
NSKSEARC	total number of skilled employees searched as of now
NSUBSIDL	total number of employees supported by wage subsidies in previous year
NSHORT	total number of short-time workers in first half-year
NTRAINP	total number of employees in on-the-job-training in first half-year
NTRAINC	total number of on-the-job-training cases in first half-year

Variables in the regressions

Y	output: SALE * (1 - INPUT/100)
C	capital: INVEST * (1 - ADDINV/100), C = 1 if no investment
L	labor: total gross monthly wages in June
YEAR	dummy: YEAR = 1 if observation in 2003
OVERTIM	NOVERTIM/EMP
OUTPROGP	dummy: OUTPROGP = 1 if turnover is expected to increase
OUTPROGN	dummy: OUTPROGN = 1 if turnover is expected to decrease
EXP	EXPORT/SALE
DEVELOP	ordinal: Rating of technical condition of enterprise
	(0 = completely out-of-date, 4 = up to date)
COLLECT	dummy: COLLECT = 1 for collective agreements
SKILL	NSKILL/EMP
NOLABSUP	dummy: NOLABSUP = NONEWHIR * WOULD
NEWWORK	NNEWHIR/EMP
TERMIN	NTERMIN/QUIT
SKSEARCH	NSKSEARC/EMP
SUBSIDYL	NSUBSIDL/EMP
FLUCT	dummy: FLUCT = 1 for stronger production fluctuations in previous year
EAST	dummy: EAST = 1 if enterprise by majority in East German property
SHORTTIM	NSHORT/EMP
TRAIND	dummy: TRAIND = 1 if employer has supported on-the-job-training in first half-year
TRAINPER	NTRAINP/EMP
TRAINCAS	NTRAINC/EMP
TRAINPC	dummy: TRAINPC = 1 if employer supports use of PCs for on-the-job-training
TYPE1	dummy: TYPE1 = 1 for independent enterprise without any establishments elsewhere
TYPE2	dummy: TYPE2 = 1 for head office of an enterprise with establishments elsewhere
TYPE3	dummy: TYPE3 = 1 for branch establishment of a larger enterprise
TYPE4	dummy: TYPE4 = 1 for intermediate authority of a larger enterprise
PROP1	dummy: PROP1 = 1 if experience is important for most jobs in the firm
PROP2	dummy: PROP2 = 1 if physical endurance is important for most jobs in the firm
PROP4	dummy: PROP4 = 1 if creativity is important for most jobs in the firm
PROP5	dummy: PROP5 = 1 if discipline is important for most jobs in the firm
PROP6	dummy: PROP6 = 1 if flexibility is important for most jobs in the firm
PROP8	dummy: PROP8 = 1 if superior workmanship is important for most jobs in the firm
PROP9	dummy: PROP9 = 1 if theoretical knowledge is important for most jobs in the firm
PROP11	dummy: PROP11 = 1 if loyalty is important for most jobs in the firm
PROP12	dummy: PROP12 = 1 if willingness to learn is important for most jobs in the firm

Data transformation for MI procedure

Y	Box-Cox
C	log, dummy*
L	Box-Cox
OVERTIM	logit
EXP	log, dummy*
DEVELOP	no transformation
SKILL	logit
NEWWORK	Box-Cox
TERMIN	logit
SKSEARCH	Box-Cox
SUBSIDYL	Box-Cox
SHORTTIM	Box-Cox
TRAINPER	Box-Cox
TRAINCAS	Box-Cox

1. Variables marked with an asterisk are treated as semi-continuous, i.e., a major part of the observations are at the minimum or the maximum of values. Therefore, we defined dummy variables that indicate whether an observation is at the respective minimum or maximum. The transformation procedure is performed only for the continuous part of the variable (see subsection 4.3).

2. All variables not mentioned in this list are dummies which remain untransformed (see subsection 4.3).

Tables

Table 3: Estimates of stochastic production frontier

	Imputed missing values, imputed capital zeroes		Imputed missing values, with capital zeroes		Non-missing values	
Variable	Coefficient	t value	Coefficient	t value	Coeff.	t value
Const.	7.8103	30.52	9.3327	28.72	8.6089	73.44
$\ln(C)$	0.1541	3.19	0.0253	3.44	0.0227	3.78
$\ln(L)$	0.1144	2.70	0.0125	0.46	0.1721	7.38
$(\ln(C))^2$	0.0115	3.25	0.0070	13.37	0.0064	16.54
$(\ln(L))^2$	0.0486	17.70	0.0399	27.79	0.0317	25.64
$\ln(C) \cdot \ln(L)$	-0.0309	-4.08	-0.0088	-11.15	-0.0079	-12.63
YEAR	-0.0386	-1.34	-0.0026	-0.19	0.0136	1.21
OVERTIM	-0.0453	-1.44	-0.0382	-1.23	0.0292	1.34
OUTPROGP	-0.0508	-2.54	-0.0517	-2.69	-0.0565	-3.55
OUTPROGN	0.0678	4.52	0.0701	4.59	0.0852	6.73
EXP	-0.0750	-1.34	-0.0999	-1.79	0.0781	5.39
DEVELOP	0.0708	4.02	0.0640	4.65	0.0567	5.13
COLLECT	-0.0473	-2.02	-0.0606	-2.36	0.0126	0.64
NEWWORK	-0.4025	-6.60	-0.4282	-7.63	-0.5289	-11.64
SKSEARCH	0.2057	2.31	0.1956	2.21	0.1725	3.42
FLUCT	-0.0436	-2.11	-0.0471	-2.27	-0.0411	-2.56
TYPE2	0.1652	4.77	0.1646	5.02	0.0783	3.14
TYPE3	0.3810	12.01	0.3922	13.21	0.3203	14.44
TYPE4	0.4094	5.80	0.4197	5.91	0.3707	7.35
EAST	-0.1681	-7.00	-0.1695	-7.11	-0.1657	-8.35
TRAIND	0.0662	3.08	0.0551	2.30	0.0682	3.61
TRAINPER	-0.0081	-1.79	-0.0117	-2.60	-0.0059	-1.20
TRAINPC	0.0796	3.63	0.0770	3.54	0.0766	4.38
PROP1	-0.0587	-2.69	-0.0614	-2.89	-0.0557	-3.41
PROP2	-0.0396	-2.08	-0.0317	-1.69	-0.0703	-5.40
PROP5	0.0412	1.96	0.0441	2.09	0.0458	3.05
PROP6	0.0362	1.62	0.0373	1.74	0.0238	1.38
PROP8	-0.0651	-2.95	-0.0651	-2.91	-0.0548	-2.83
PROP9	-0.0700	-3.85	-0.0726	-3.98	-0.0529	-3.65
PROP11	0.0470	2.69	0.0381	2.10	0.0511	3.81
PROP12	0.0416	2.29	0.0385	2.12	0.0444	3.29
Industry dummies	yes		yes		yes	
	18447 observations		18447 observations		13969 observations	

Source: own calculations, based on IAB data

Table 3a: Estimates of inefficiency submodel

	Imputed missing values, imputed capital zeroes		Imputed missing values, with capital zeroes		Non-missing values	
Variable	Coefficient	t value	Coefficient	t value	Coeff.	t value
Const.	-32.816	-2.15	-29.564	-2.30	-0.1646	-0.41
$\ln(L)$	0.809	1.61	0.793	1.77	-0.0874	-2.90
EXP	-32.826	-2.74	-31.184	-2.99	0.0633	1.71
DEVELOP	1.039	1.30	0.708	1.03	0.1195	1.82
COLLECT	-3.100	-1.85	-3.407	-2.14	0.0148	0.12
SKILL	-5.615	-2.11	-4.750	-2.28	-0.3442	-2.62
NOLABSUP	4.066	1.53	3.745	1.59	0.0907	0.48
TERMIN	-4.482	-1.74	-4.599	-1.83	-0.2006	-1.64
SUBSIDYL	6.064	2.46	5.599	2.71	0.2723	1.84
FLUCT	-2.667	-1.61	-2.313	-1.64	-0.1489	-1.55
TYPE1	-7.468	-2.58	-6.533	-2.76	-0.5958	-5.15
EAST	-2.135	-1.28	-2.174	-1.40	-0.2946	-2.48
SHORTTIM	-4.841	-1.17	-5.279	-1.40	-0.0541	-0.22
TRAIND	-1.671	-1.11	-1.698	-1.26	0.1614	1.37
TRAINCAS	0.207	2.72	0.202	3.07	0.0121	1.46
TRAINPC	1.721	1.22	1.797	1.27	0.2222	2.08
PROP1	-1.766	-1.25	-1.945	-1.48	-0.1842	-1.84
PROP4	4.423	2.22	4.042	2.46	0.4545	5.38
PROP6	2.876	1.62	2.422	1.63	0.1528	1.44
PROP8	-4.468	-1.38	-4.003	-1.55	-0.2531	-2.21
Industry dummies	yes		yes		yes	
λ	6.428	2.88	6.024	3.21	2.6818	26.78

Technical inefficiency estimates

Variable	Mean		Mean		Mean	
u_i	0.5924		0.5908		0.7433	
	18447 observations		18447 observations		13969 observations	

Source: own calculations, based on IAB data

Table 4: Estimates of stochastic production frontier
(5 auxiliary regressions: Imputed missing values, imputed capital zeroes)

Variable	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.
Const.	7.821	55.1	7.582	54.8	7.714	57.0	7.827	50.4	8.107	56.1
$\ln(C)$	0.136	5.8	0.176	7.7	0.212	9.5	0.118	5.2	0.130	5.7
$\ln(L)$	0.130	6.3	0.138	6.7	0.091	4.4	0.147	7.5	0.067	3.6
$(\ln(C))^2$	0.014	8.4	0.014	8.5	0.013	7.8	0.009	6.9	0.008	5.6
$(\ln(L))^2$	0.049	33.9	0.050	35.9	0.051	33.9	0.045	39.5	0.048	37.5
$\ln(C) \cdot \ln(L)$	-0.033	-13.0	-0.037	-14.3	-0.037	-13.3	-0.025	-13.1	-0.023	-10.4
YEAR	-0.045	-3.4	-0.032	-2.5	-0.001	-0.1	-0.061	-4.6	-0.053	-4.0
OVERTIM	-0.064	-3.1	-0.039	-1.9	-0.062	-3.1	-0.051	-2.5	-0.011	-0.5
OUTPROGP	-0.047	-2.6	-0.059	-3.3	-0.054	-3.0	-0.039	-2.2	-0.055	-3.0
OUTPROGN	0.070	4.7	0.063	4.3	0.068	4.6	0.068	4.6	0.070	4.8
EXP	-0.096	-1.9	-0.092	-1.8	-0.061	-1.2	-0.079	-1.5	-0.047	-0.9
DEVELOP	0.082	7.2	0.063	5.6	0.053	4.7	0.079	7.1	0.078	7.0
COLLECT	-0.057	-2.7	-0.041	-2.0	-0.049	-2.4	-0.056	-2.7	-0.033	-1.6
NEWWORK	-0.367	-7.0	-0.391	-7.2	-0.397	-7.4	-0.433	-8.2	-0.424	-7.7
SKSEARCH	0.214	2.8	0.158	1.9	0.177	2.2	0.251	3.4	0.228	2.9
FLUCT	-0.042	-2.3	-0.039	-2.2	-0.060	-3.4	-0.041	-2.4	-0.035	-2.0
TYPE2	0.185	6.4	0.178	6.2	0.160	5.5	0.163	5.7	0.140	4.8
TYPE3	0.392	17.1	0.379	16.4	0.408	18.2	0.371	16.1	0.356	15.2
TYPE4	0.403	5.9	0.411	6.0	0.405	6.1	0.441	6.7	0.388	5.9
EAST	-0.179	-8.2	-0.161	-7.3	-0.163	-7.5	-0.177	-8.2	-0.160	-7.4
TRAIND	0.060	2.9	0.077	3.8	0.063	3.1	0.063	3.1	0.069	3.5
TRAINPER	-0.009	-1.9	-0.007	-1.7	-0.007	-1.9	-0.009	-2.0	-0.009	-1.9
TRAINPC	0.086	4.5	0.071	3.8	0.067	3.6	0.091	4.9	0.084	4.5
PROP1	-0.058	-3.3	-0.057	-3.2	-0.054	-3.1	-0.078	-4.4	-0.046	-2.6
PROP2	-0.047	-3.1	-0.042	-2.8	-0.026	-1.8	-0.031	-2.1	-0.052	-3.5
PROP5	0.049	2.8	0.033	1.9	0.048	2.8	0.026	1.5	0.049	2.9
PROP6	0.036	1.9	0.024	1.3	0.037	2.0	0.054	2.9	0.029	1.6
PROP8	-0.066	-3.1	-0.063	-3.0	-0.075	-3.6	-0.061	-2.9	-0.060	-2.9
PROP9	-0.078	-4.6	-0.064	-3.9	-0.075	-4.6	-0.071	-4.4	-0.061	-3.7
PROP11	0.060	4.0	0.045	3.0	0.045	3.0	0.046	3.1	0.038	2.6
PROP12	0.050	3.2	0.051	3.3	0.029	1.9	0.039	2.6	0.040	2.6
Industry dummies	yes		yes		yes		yes		yes	

Source: own calculations, based on IAB data, 18447 observations

Table 4a: Estimates of inefficiency submodel
(5 auxiliary regressions: Imputed missing values, imputed capital zeroes)

Variable	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.
Const.	-36.30	-2.4	-29.00	-2.6	-39.77	-2.8	-21.76	-2.9	-37.25	-2.5
$\ln(L)$	0.71	2.2	0.70	2.4	1.26	2.9	0.37	1.8	1.01	2.7
EXP	-35.37	-2.5	-33.09	-3.1	-34.53	-3.2	-27.69	-3.5	-33.45	-2.6
DEVELOP	1.35	1.6	0.83	1.3	0.70	0.9	0.93	1.7	1.38	1.6
COLLECT	-3.68	-2.2	-2.32	-1.9	-4.00	-2.3	-2.39	-2.4	-3.11	-2.0
SKILL	-4.84	-2.0	-5.56	-2.5	-7.32	-2.6	-4.51	-2.8	-5.84	-2.2
NOLABSUP	5.84	2.1	3.43	1.6	3.88	1.5	4.27	2.4	2.91	1.2
TERMIN	-5.44	-2.2	-3.46	-2.1	-5.69	-2.4	-2.53	-2.1	-5.29	-2.2
SUBSIDYL	6.67	2.5	5.40	2.8	6.75	2.8	4.96	3.2	6.54	2.5
FLUCT	-2.80	-1.9	-1.87	-1.8	-4.02	-2.4	-1.85	-2.1	-2.80	-1.9
TYPE1	-6.93	-2.4	-7.13	-2.9	-8.59	-2.9	-6.43	-3.4	-8.26	-2.6
EAST	-3.16	-1.8	-2.24	-1.6	-1.83	-1.1	-1.99	-1.8	-1.46	-1.0
SHORTTIM	-5.62	-1.3	-3.62	-1.1	-3.87	-1.0	-6.87	-1.9	-4.23	-1.1
TRAIND	-1.85	-1.3	-0.64	-0.6	-1.78	-1.2	-1.41	-1.5	-2.68	-1.9
TRAINCAS	0.21	2.4	0.19	3.3	0.24	3.7	0.18	3.4	0.22	2.5
TRAINPC	2.84	2.0	1.32	1.3	0.99	0.8	1.39	1.6	2.06	1.7
PROP1	-2.14	-1.5	-1.78	-1.6	-1.84	-1.4	-2.41	-2.3	-0.66	-0.6
PROP4	4.29	2.3	4.86	2.7	5.38	2.7	3.12	2.9	4.48	2.3
PROP6	3.30	2.0	1.54	1.5	4.17	2.4	2.94	2.7	2.43	1.7
PROP8	-3.93	-1.3	-4.01	-1.5	-5.59	-1.5	-3.80	-1.8	-5.01	-1.3
Industry dummies	yes		yes		yes		yes		yes	
λ	6.65	2.7	6.07	3.1	6.90	3.1	5.73	3.7	6.80	2.7

Technical inefficiency estimates

Variable	Mean		Mean		Mean		Mean		Mean	
u_i	0.59		0.59		0.57		0.63		0.58	

Source: own calculations, based on IAB data, 18447 observations

Table 5: Estimates of stochastic production frontier
(5 auxiliary regressions: Imputed missing values, with capital zeroes)

Variable	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.
Const.	9.319	94.8	9.271	97.2	9.459	98.1	9.233	95.5	9.382	101.1
$\ln(C)$	0.023	3.5	0.031	4.9	0.022	3.4	0.025	4.0	0.025	4.0
$\ln(L)$	0.019	1.0	0.019	1.1	-0.001	-0.1	0.036	2.0	-0.010	-0.6
$(\ln(C))^2$	0.007	14.3	0.007	14.2	0.007	15.1	0.007	14.7	0.007	14.5
$(\ln(L))^2$	0.039	37.4	0.040	40.5	0.040	38.6	0.039	38.6	0.041	40.8
$\ln(C) \cdot \ln(L)$	-0.008	-11.6	-0.009	-13.3	-0.009	-12.3	-0.009	-12.6	-0.009	-12.6
YEAR	-0.001	-0.0	0.000	0.0	-0.004	-0.3	-0.007	-0.6	-0.001	-0.1
OVERTIM	-0.055	-2.6	-0.037	-1.8	-0.055	-2.7	-0.039	-1.9	-0.004	-0.2
OUTPROGP	-0.047	-2.5	-0.058	-3.2	-0.052	-2.8	-0.046	-2.5	-0.057	-3.1
OUTPROGN	0.073	4.9	0.066	4.4	0.072	4.8	0.067	4.6	0.072	4.9
EXP	-0.108	-2.1	-0.118	-2.3	-0.081	-1.6	-0.114	-2.2	-0.079	-1.5
DEVELOP	0.074	6.4	0.062	5.4	0.054	4.8	0.065	5.8	0.065	5.8
COLLECT	-0.076	-3.6	-0.052	-2.5	-0.062	-3.0	-0.072	-3.5	-0.042	-2.0
NEWWORK	-0.409	-7.7	-0.421	-7.7	-0.430	-7.9	-0.440	-8.3	-0.442	-8.0
SKSEARCH	0.217	2.8	0.124	1.6	0.206	2.7	0.219	2.9	0.213	2.8
FLUCT	-0.043	-2.4	-0.043	-2.4	-0.064	-3.6	-0.045	-2.6	-0.040	-2.3
TYPE2	0.177	6.2	0.176	6.1	0.164	5.6	0.164	5.7	0.142	4.9
TYPE3	0.401	17.2	0.387	16.5	0.414	18.2	0.390	16.8	0.369	15.7
TYPE4	0.413	5.9	0.407	5.8	0.421	6.2	0.450	6.8	0.407	6.1
EAST	-0.181	-8.1	-0.165	-7.4	-0.172	-7.9	-0.171	-7.8	-0.158	-7.3
TRAIND	0.048	2.3	0.068	3.3	0.066	3.2	0.042	2.1	0.052	2.6
TRAINPER	-0.012	-2.7	-0.011	-2.7	-0.011	-2.7	-0.013	-2.8	-0.012	-2.5
TRAINPC	0.081	4.2	0.066	3.5	0.067	3.6	0.088	4.7	0.082	4.4
PROP1	-0.064	-3.6	-0.061	-3.4	-0.055	-3.1	-0.078	-4.4	-0.049	-2.8
PROP2	-0.038	-2.5	-0.035	-2.3	-0.019	-1.3	-0.023	-1.6	-0.044	-2.9
PROP5	0.047	2.7	0.037	2.2	0.055	3.1	0.029	1.7	0.052	3.0
PROP6	0.043	2.3	0.028	1.5	0.044	2.3	0.046	2.5	0.027	1.4
PROP8	-0.064	-3.0	-0.063	-3.0	-0.076	-3.6	-0.061	-2.8	-0.061	-2.9
PROP9	-0.080	-4.7	-0.069	-4.2	-0.078	-4.6	-0.074	-4.5	-0.063	-3.8
PROP11	0.054	3.5	0.032	2.1	0.034	2.2	0.037	2.4	0.034	2.3
PROP12	0.046	3.0	0.049	3.1	0.028	1.8	0.037	2.4	0.033	2.2
Industry dummies	yes		yes		yes		yes		yes	

Source: own calculations, based on IAB data, 18447 observations

Table 5a: Estimates of inefficiency submodel

(5 auxiliary regressions: Imputed missing values, with capital zeroes)

Variable	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.	Coeff.	t val.
Const.	-33.82	-2.6	-26.89	-2.9	-32.62	-3.4	-19.10	-3.2	-35.38	-2.7
$\ln(L)$	0.67	2.3	0.74	2.7	1.14	3.5	0.37	2.0	1.04	2.9
EXP	-33.64	-2.7	-32.49	-3.6	-30.52	-4.2	-25.52	-3.9	-33.76	-2.9
DEVELOP	0.96	1.3	0.60	1.1	0.32	0.5	0.69	1.5	0.97	1.3
COLLECT	-4.16	-2.5	-2.76	-2.3	-4.08	-2.9	-2.58	-2.8	-3.45	-2.3
SKILL	-4.42	-2.1	-4.89	-2.7	-5.75	-3.0	-3.65	-2.9	-5.04	-2.3
NOLABSUP	5.47	2.2	3.25	1.7	3.12	1.5	3.72	2.5	3.16	1.4
TERMIN	-5.92	-2.4	-3.77	-2.4	-5.36	-2.8	-2.40	-2.3	-5.55	-2.5
SUBSIDYL	6.38	2.7	5.07	3.1	5.94	3.4	4.49	3.7	6.12	2.8
FLUCT	-2.26	-1.8	-1.68	-1.7	-3.50	-2.8	-1.62	-2.1	-2.51	-1.9
TYPE1	-6.14	-2.5	-6.44	-3.2	-7.25	-3.5	-5.36	-3.7	-7.48	-2.8
EAST	-3.15	-1.8	-2.38	-1.8	-1.75	-1.3	-2.06	-2.0	-1.53	-1.1
SHORTTIM	-5.95	-1.4	-4.12	-1.2	-4.89	-1.3	-6.18	-2.0	-5.25	-1.4
TRAIND	-1.89	-1.4	-0.83	-0.8	-1.89	-1.5	-1.34	-1.6	-2.55	-2.0
TRAINCAS	0.22	2.5	0.18	4.1	0.22	4.9	0.17	4.0	0.22	2.9
TRAINPC	2.97	2.1	1.40	1.5	0.97	1.0	1.30	1.7	2.33	1.9
PROP1	-2.57	-1.8	-1.97	-1.8	-1.87	-1.7	-2.34	-2.5	-0.98	-0.9
PROP4	4.16	2.4	4.48	3.0	4.34	3.1	2.90	3.2	4.33	2.5
PROP6	3.09	2.1	1.37	1.4	3.34	2.6	2.40	2.7	1.91	1.5
PROP8	-3.74	-1.4	-3.77	-1.7	-4.48	-1.8	-3.33	-2.0	-4.70	-1.5
Industry dummies	yes		yes		yes		yes		yes	
λ	6.40	2.9	5.75	3.5	6.21	4.0	5.27	4.3	6.49	3.0

Technical inefficiency estimates

Variable	Mean		Mean		Mean		Mean		Mean	
u_i	0.59		0.59		0.58		0.62		0.57	

Source: own calculations, based on IAB data, 18447 observations