SSOAR

Open Access Repository
www.ssoar.info

Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models

Kölling, Arnd; Rässler, Susanne

Veröffentlichungsversion / Published Version
Monographie / monograph

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:
SSG Sozialwissenschaften, USB Köln

Empfohlene Zitierung / Suggested Citation:
Kölling, A., & Rässler, S. (2004). *Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models.* (IAB Discussion Paper: Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung, 5/2004). Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesagentur für Arbeit (IAB). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-330754

Nutzungsbedingungen:
Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.
Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

Terms of use:
This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.
By using this particular document, you accept the above-stated conditions of use.

gesis Leibniz-Institut für Sozialwissenschaften

Mitglied der Leibniz-Gemeinschaft

IAB Discussion Paper

Beiträge zum wissenschaftlichen Dialog aus dem Institut für Arbeitsmarkt- und Berufsforschung

No. 5/2004

Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models

Arnd Kölling & Susanne Rässler

Bundesagentur für Arbeit

Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models

Arnd Kölling & Susanne Rässler

Auch mit seiner neuen Reihe „IAB-Discussion Paper“ will das Forschungsinstitut der Bundesagentur für Arbeit den Dialog mit der externen Wissenschaft intensivieren. Durch die rasche Verbreitung von Forschungsergebnissen über das Internet soll noch vor Drucklegung Kritik angeregt und Qualität gesichert werden.

Also with its new series "IAB Discussion Paper" the research institute of the German Federal Employment Agency wants to intensify dialogue with external science. By the rapid spreading of research results via Internet still before printing criticism shall be stimulated a quality shall be ensured.

Editing and multiply imputing German establishment panel data to estimate stochastic production frontier models

Arnd Kölling & Susanne Rässler*

Abstract

This paper illustrates the effects of item-nonresponse in surveys on the results of multivariate statistical analysis when estimation of productivity is the task. To multiply impute the missing data a data augmentation algorithm based on a normal/Wishart model is applied. Data of the German IAB Establishment Panel from waves 2000 and 2001 are used to estimate the establishment's productivity. The processes of constructing, editing, and transforming the variables needed for the analyst's as well as the imputer's models are described. It is shown that standard multiple imputation techniques can be used to estimate sophisticated econometric models from large-scale panel data exposed to item-nonresponse. Basis of the empirical analysis is a stochastic production frontier model with labour and capital as input factors. The results show that a model of technical inefficiency is favoured compared to a case where we assume different production functions in East and West Germany. Also we see that the effect of regional setting on technical inefficiency increases when inference is based on multiply imputed data sets. This could have influence on the economic and regional policies in Germany in the future.

Keywords: Data augmentation, Markov chain Monte Carlo, establishment panel data, productivity.

JEL-classification: C15, C81, D24

+ **Acknowledgement:** An earlier draft of this article was presented at the DataClean2002 Conference, Jyväsylä, Finland, May 30-31, 2002. The authors wish to thank the editor and two unknown referees for very detailed comments which helped improving this paper considerably. Moreover, we are very grateful for advice by Donald B. Rubin

* Institute for Employment Research (IAB), Regensburger Str. 104, 90478 Nürnberg, Germany, phone (+49) 911 / 179 / 3084, email:susanne.raessler@iab.de

1 Introduction

In this paper stochastic production frontier models are estimated to figure out whether there are significant differences in the use of input factors between East and West German establishments. As it is a typical situation in empirical research, we are confronted with missing values in our data set. A closer look to the data reveals 5% to 30% of missing values in a few variables, reducing the complete data records available for any multivariate analysis considerably. Whereas information from 17294 observations from the panel waves of 2000 and 2001 is collected in principle, only 10223 observations of them can be used when inference is based on the complete cases. Then, at a minimum, precision of estimates is lost, at the worst, the resulting estimates will be biased. So the questions arise whether the remaining data are still representative for the population of interest and how (multiple) imputation can be implemented successfully and easily with large-scale establishment panel data while a sophisticated econometric model is to be estimated.

Rubin (1987) and Little and Rubin (1987) once classified the nonresponse phenomenon 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). A nonresponse process that is neither MCAR nor MAR is called nonrandom or missing not at random (MNAR); i.e., the probability of a variable being observed depends on the variable itself. 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.

Investigating the variables that are used in the estimation process, we find the highest amount of missing data especially with variable *input of material, services, and goods*, variable *turnover*, and variable *investment*. Moreover, analyses of the amount of data missing per variable show that item-nonresponse on input of material, services, and goods, turnover, and

investment as well as wage and salary information and working overtime is higher the larger the companies are. Especially 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 more than 40% 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, i.e., the dependent and independent regression variables, are missing completely at random (MCAR) which obviously is not the case. To ensure the MAR-assumption and allow estimating 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., Rubin and Schenker (1998) or Rässler et al. (2003). To correct for these effects using singly imputed data, special variance estimation techniques have to be applied. For the time being, these techniques are restricted to special univariate statistics; for a very recent discussion of the merits and demerits of single and multiple imputation see Groves et al. (2002).

Furthermore, 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 MNAR is a serious concern, it is obviously necessary to jointly model the data and the missing ness, 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.

The investigation of differences in productivity between East and West Germany is a challenging area of research because of several reasons. First, since the reunification of both parts of Germany has happened in 1990, several billions of Euros have been transferred to the eastern part to help the former socialist regime to become a modern capitalist economy. This includes a converging respective increasing productivity. Therefore, it is very important to investigate whether these transfers show the intended results or not. On the other side, the data set used in this project gives a unique opportunity to estimate the productivity on the establishment level for Germany. Most of other studies on this topic rely on aggregated sectoral data or do not contain sufficient information to estimate a production function. The IAB Establishment Panel overcomes these problems and also allows taking into account firm specific effects.

The article is structured as follows. In the next section, the data and the response behaviour in the panel are described. In the third section, a short introduction to the multiple imputation paradigm is provided. There we discuss and describe the imputation process as well as the preparations and transformations of the variables to be used in the imputer's model. In section four, the stochastic production frontier models to be estimated are presented as well as the preparation and editing of the variables to fit for the analyst's model. In the fifth section, the estimation results using imputed data are given and compared with the results based only on the complete data. Finally, section six summarises the work.

2 Data and response behaviour

Our data are taken from two waves (2000 & 2001) of the Establishment Panel of the Institute for Employment Research of the German Federal Employment Agency (*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 employ-

ment in the private and public sector, and accounts for almost 80 percent of total employment in Western Germany. The survey unit of the register is the establishment or local production unit, rather than the legal and commercial entity of the company.

For its part, the *IAB Establishment Panel* draws a stratified random sample of units from the register, the selection probabilities depending 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 times exceed 60 percent, and, for repeatedly-interviewed establishments, more than 80 percent.

The first wave of the establishment panel in 1993 contains data on 4,265 establishments. Since 1993 the panel has been augmented regularly to reflect establishment mortality, other exits, and newly-founded units. In 1996 a panel was started for Eastern Germany with an initial sample of 4,313 establishments. Currently, the overall number of establishments in the sample approximates 15,000 with the addition of Eastern Germany and other regional samples.

The panel is designed to meet the needs of the Federal Labour Service, so that its focus is on employment-related matters – although its scope is wider than the parent register. 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 and further training of employees, working time, and overtime. Additionally, information on certain establishment policies, business developments, and investment is similarly collected on an annual basis. Other information is collected biennially or triennially. Examples include works council status (first asked in 1996 and then every other year), organisational changes, and use of public employment subsidies. Finally, each year the panel also addresses a specific topic; in 2000, for example, that topic was shortages of qualified manpower.

We exclude all establishments from the sample that do not use turnover as an output measure. This affects in principal non-profit organisations, public offices, banks and insurances. For inference based on the complete

cases we work with an unbalanced panel for both years. In this unbalanced sample we have 10223 observations from firms with complete interviews and without any item-nonresponse. If we would use a balanced sample for the complete case analysis we would loose even more data and were finally left with only 6988 data records from 3494 establishments which have observations on all variables in 2000 and 2001. For the imputation process we could stay with the balanced sample of originally 17294 data records for 2000 and 2001 from 8647 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 & Kohaut 2000). The most important reason for nonresponse seems to be the change of the interviewer or of the firm representative. This shows that a successful panel survey should have constant structures to reduce nonresponse rates.

Mainly, item-nonresponse in the data is found by only a few variables, especially the two that are used to construct the endogenous variable. Output is defined as the log of turnover minus input of materials, goods, and services (value added). Input has an item-nonresponse rate of 31.79% in 2000 and 12.32% in 2001. This remarkable reduction in the two waves is due to a change in the questionnaire. In 2000 the interviewed firm representatives could answer with the special category "I don't know". In the following year this category was dropped from the questionnaire. It is known that it takes some time and effort to give an exact answer to the question about input materials. Thus, dropping the exit category "I don't know", an "easy" way to answer the question does not exist anymore and the representatives are expected to give, at least, a guess of the correct value. Although large changes in the response behaviour often mean that the content of the question may has changed seriously, we do not expect this to be the case here. Therefore, we assume that the answers are com-

parable to that in the previous wave. The item-nonresponse rate for turnover lies between 10% and 11% in both waves. Next to these two variables, three other questions dealing with the firm's investment behaviour have an item-nonresponse greater than 2%. The values are figured in Table 1. All the other variables used in our study are below that limit.

Table 1: Variables with the highest item-nonresponse (%)

	2000	2001
Input of material, goods and services	31.79	12.32
Turnover	10.62	10.80
Investment to enlarge capital	5.81	4.59
Sum of investment	2.58	2.05
Investment in ICT	2.29	1.70

Source: IAB Establishment Panel 2000 & 2001

3 Imputer's model: data augmentation

3.1 Introduction to the multiple imputation principle

Multiple imputation (MI), introduced by Rubin in 1978 and in detail proposed by Rubin (1987), is a Monte Carlo technique replacing the missing values by $m>1$ simulated versions, generated according to a probability distribution or, more generally, any density function indicating how likely are imputed values given the observed data. Typically m is small, with m=3 or 5. Each of the imputed and thus completed data sets is first analysed by standard methods; the results are then combined or pooled to produce estimates and confidence intervals that embed the missing data uncertainty.

The theoretical motivation for multiple imputations is Bayesian. Basically, MI requires independent random draws from the posterior predictive distribution

$$f(y_{mis}|y_{obs}) = \int f(y_{mis},\zeta|y_{obs})d\zeta = \int f(y_{mis}|y_{obs},\zeta)f(\zeta|y_{obs})d\zeta \quad (7)$$

of the missing data given the observed data. Since $f(y_{mis}|y_{obs})$ itself often is difficult to derive, we may alternatively perform:

1. random draws of the parameters according to their observed-data posterior distribution $f(\zeta|y_{obs})$ as well as

2. random draws of the missing data according to their conditional predictive distribution $f(y_{mis}|y_{obs},\zeta)$ given the drawn parameter values.

For many models the conditional predictive distribution $f(y_{mis}|y_{obs},\zeta)$ is rather straightforward due to the data model used. On the contrary, the corresponding observed-data posterior $f(\zeta|y_{obs}) = L(\zeta;y_{obs})f(\zeta)/f(y_{obs})$ usually is difficult to derive, especially when the data have a multivariate structure and different not monotone missing data patterns. The observed-data posteriors are often not standard distributions from which random numbers can 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 $f(\zeta|y_{obs})$ and $f(y_{mis}|y_{obs},\zeta)$ are achieved as stationary distributions of Markov chains which are based on the easier to compute complete-data distributions. Creating m independent draws from such chains can be used as imputations of Y_{mis} from their posterior predictive distribution $f(y_{mis}|y_{obs})$.

The MI principle assumes that the estimate $\hat{\theta}$ of any quantity θ and its variance estimate $\hat{V}(\hat{\theta})$ can be regarded as an approximate complete-data posterior mean and variance for θ with $\hat{\theta} \approx E(\theta \mid y_{obs}, y_{mis})$ and $\hat{V}(\hat{\theta}) \approx V(\theta \mid y_{obs}, y_{mis})$ based on a suitable complete-data model and prior; see also Schafer (1997). Moreover, we must assume that with complete data, tests and interval estimates based on the normal approximation $(\hat{\theta}-\theta)\Big/\sqrt{\hat{V}(\hat{\theta})} \sim N(0,1)$ should work well; the relaxation of this assumption allowing a *t*-distribution is given by Barnard and Rubin (1999). Notice that the usual maximum-likelihood estimates and their asymptotic variances derived from the inverted Fisher information matrix typically satisfy these assumptions. Sometimes it is necessary to transform the estimate $\hat{\theta}$ to a scale for which the normal approximation can be applied.

Supposing now that the data are missing at random, we create *m>1* dependent simulated imputations. Based on these *m* imputed data sets we calculate *m* complete data statistics $\hat{\theta}^{(i)}$ and their variance estimates $\hat{V}(\hat{\theta}^{(i)})$, *i* = 1, …, *m*. According to the MI principle, the MI point estimate $\hat{\theta}_{MI}$ for a parameter θ is the average $\hat{\theta}_{MI} = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}^{(i)}$. Its estimated total variance *T* is calculated according to the analysis of variance principle.

- "Between-imputation" variance $B = \frac{1}{m}\sum_{i=1}^{m}(\hat{\theta}^{(i)} - \hat{\theta}_{MI})^2$,
- "within-imputation" variance $W = \frac{1}{m-1}\sum_{i=1}^{m}\hat{V}(\hat{\theta}^{(i)})$,
- "total-variance" $T = W + (1+\frac{1}{m})B$.

Tests and two-sided interval estimates may be based approximately on the Student's *t*-distribution $(\hat{\theta}_{MI} - \theta)/\sqrt{T} \sim t_v$ with degrees of freedom $v = (m-1)\left(1 + \frac{W}{(1+m^{-1})B}\right)^2$. Hence we realise that the multiple imputation interval estimate is expected to produce a larger but valid interval than an estimate based only on single imputation because the interval is widened to account for the missing data uncertainty and simulation error. For a good introduction to the MI paradigm see Schafer (1999a) or Brand (1999). Notice that confidence intervals under MI can be shorter than confidence intervals based only on the complete or available cases. This is especially true if the imputed sample is substantially larger than the complete case sample. The following Tables 2 and 3 indicate that the analysis of the imputed sample is, at least, more precise than the analysis on the complete cases. Therefore, the possibility to use all valuable information is also an important argument for applying MI here.

3.2 Data augmenation 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), see Website www.stat.psu.edu/~jls.

We assume an *r*-dimensional normal distribution for all the *r* 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, ..., *n*.

As a prior distribution $f(\mu,\Sigma)$ for the parameters of location and scale the common uninformative prior distribution is chosen with

$$f(\mu,\Sigma) \approx f(\mu) f(\Sigma) \approx c|\Sigma|^{-(r+1)/2} \propto |\Sigma|^{-(r+1)/2}; \tag{8}$$

i.e., μ and Σ are assumed to be approximately independent, for details see Box and Tiao (1992) or Schafer (1997). As long as no problems of identification occur, the assumption of a noninformative prior distribution seems to be the most "objective" choice.

Under this prior distribution (8), 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; i.e.,

$$\begin{aligned} \Sigma|y &\sim W^{-1}(n-1,(nS(\bar{y}))^{-1}), \\ \mu|\Sigma,y &\sim N(\bar{y},\Sigma/n), \end{aligned} \tag{9}$$

with the sample covariance matrix $S(\bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})(y_i-\bar{y})'$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $y_i = (y_{i1}, y_{i2}, ..., y_{ir})'$. 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, i.e.,

$$Y_{mis}|y_{obs},\mu,\Sigma \sim N(\mu_{mis|obs},\Sigma_{mis|obs}). \tag{10}$$

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 to their conditional predictive distribution $f(y_{mis}|y_{obs},\zeta)$, see (10), given the observed data and an actual draw of the parameters $\zeta^{(t)}=(\mu^{(t)},\Sigma^{(t)})$; i.e., random values are generated according to

$$Y_{mis}{}^{(t)}|y_{obs},\mu^{(t)},\Sigma^{(t)} \sim N(\mu_{mis|obs}{}^{(t)},\Sigma_{mis|obs}{}^{(t)}) \tag{11}$$

P-step: Using the completed data $y^{(t)}=(y_{obs},y_{mis}{}^{(t)})$ actual values for the mean vector $\bar{y}^{(t)}$ and the covariance matrix $S(\bar{y}^{(t)})=\frac{1}{n}\sum_{i=1}^{n}(y_i{}^{(t)}-\bar{y}^{(t)})(y_i{}^{(t)}-\bar{y}^{(t)})'$ are calculated. Then new actual values for the parameters $\mu^{(t)}$ and $\Sigma^{(t)}$ are drawn according to their complete-data posterior distribution (9); i.e.,

$$\begin{aligned}\Sigma^{(t+1)}\left|y^{(t)}\right. &\sim W^{-1}(n-1,(nS(\bar{y}^{(t)}))^{-1}),\\ \mu^{(t+1)}\left|\Sigma^{(t+1)},y^{(t)}\right. &\sim N(\bar{y}^{(t)},\Sigma^{(t+1)}/n).\end{aligned} \tag{12}$$

Such random draws of $\mu^{(t)}$ and $\Sigma^{(t)}$ is considered to be the Bayesian stochastic counterpart of maximising 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 (11) and (12) 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 $\{(\mu^{(t)},\Sigma^{(t)},Y_{mis}{}^{(t)}):t=0,1,\ldots\}$ converges in distribution to $f(y_{mis},\zeta|y_{obs})$. Thus, $Y_{mis}{}^{(t)}$ converges to a draw from the desired posterior predictive distribution $f(y_{mis}|y_{obs})$ given in (7); e.g., after assessing convergence every *t* + 100, *t* + 200, … value can be used to produce m independent multiple imputations. Typically, *m* = 5 imputed data sets are created. Data augmentation techniques have been used in practice, and provide rather flexible tools for creating multiple imputations from parametric models. A very detailed introduction is given by Schafer (1997).

3.3 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 semicontinuous 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 semicontinuous, 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 semicontinuous 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; for an illustration, see Figure 1.

Figure 1: Preparation of semicontinuous variables

Unit no.	Y
1	12
2	NA
3	0
4	0
5	NA
...	...
n-1	3
n	0

⇨

Unit no.	W	V
1	1	12
2	NA	NA
3	0	NA
4	0	NA
5	NA	NA
...		...
n-1	1	3
n	0	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 semicontinuous 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).

Additionally, because the data augmentation is made with the original data from the survey and not the constructed variables in the estimation, we expect to have accounted for effects due to non-linearities, quadratic terms, or interactions in the analyst's model. The definition of the variables used in the analyst's model is given in the appendix.

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, 1997) works quite well with *g(Y) = Y/(1-Y)* for *Y* in (0,1). For skewed positive *Y* values an ordinary log transformation often is a good choice with *g(Y) = ln(Y)*. Another useful transformation is given by the Box-Cox transformation with $g(Y)=(Y^c-1)/c, c\neq 0$.

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. 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, 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).

4 Analyst's model: production frontier model (stochastic frontier model)

The analyst's model deals with one of the important issues in German economy. More than ten years after reunification it is very interesting to investigate whether or not the billions of Euros that have been transferred to the eastern part of Germany lead to converging economies. If still big differences in productivity occur, one should question the economic benefit of these transfers. The common way to estimate productivity is to regress a production function; several approaches are known from economic theory. The most famous one is a Cobb-Douglas-type production function, which has been very useful for analysing macroeconomic data. Nevertheless, there are some strong restrictions in that model, e. g., constant partial productivities independently of the use of input factors. Also, the sum of partial derivatives has to be one. To overcome these problems, generalised production functions are introduced to economic theory. In these kinds of models, most of the restrictions on the estimated parameters are abolished and many known production functions are special cases of these generalised ones. They are especially useful when microeconomic data with firm specific effects are to be analysed. Therefore, generalised production functions like the translog production function play an important role by explaining the amount of output of goods and services or the demand for different input factors (see Greene 2000). These theoretic formulations assume an ideal world where all factors are used efficiently. In reality the world is of course not perfect and there are deviations from the ideal input of capital and labor. These inefficiencies will lower the output and for any input x the observed amount of produced goods and services is less or equal to the theoretical value of the production function $f(x)$. Thus, the empirical formulation should differ from that in theory. This means, from theory we would expect a higher productivity. Instead, in reality for several reasons, there are deviations from an optimal use of production inputs. We do not know the reasons for the lower productivity per

se and possibly they are completely firm specific, but it is may be possible to find some patterns by using stochastic production frontier functions. When simple OLS is used to estimate the firm's productivity, the result for the constant and therefore for all dummy variables will be biased (see Greene 2000, 395). As we are highly interested in the differences between East and West Germany and we estimate these differences with dummy variables, we decide to use stochastic production frontier functions instead of an OLS approach. Since the stochastic production frontier function was independently developed by Aigner, Lovell, and Schmidt (1977) and Meeusen and van den Broeck (1977), a considerable number of theoretical and empirical studies are provided. According to Battese and Coelli (1995, 1996), a general model that fits these needs for panel data is given by:

$$(1) \qquad \ln Y_{it} = x'_{it}\beta + V_{it} - U_{it},\ i = 1,\ldots, n,\ t = 1, 2,$$

where Y_{it} denotes the output for establishment i at time t; x_{it} is a vector of input variables of production; β are unknown parameters to be estimated; V_{it} is a randomly distributed error term defined as $V_{it} \sim N(0, \sigma_v^2)$; U_{it} are non-negative random variables indicating technical inefficiencies of production. It is assumed that each U_{it} follows a normal distribution with $U_{it} \sim N_+(z'_{it}\delta, \sigma_U^2, 0)$, $i = 1,\ldots, n$, $t = 1, 2$, that is truncated at zero. The U_{it} and the V_{it} are independently distributed for all t's and i's. The z_{it} are a vector of exogenous variables associated with the technical inefficiency of production, δ is a vector of coefficients.

The term U_{it} violates the assumptions of a simple OLS-model. If the term is not recognised, at least the estimate of the constant term is biased (Greene 2000, 395). However, even if the estimates for the β's except the constant are consistent, it is not possible to detect sources of inefficiency with a simple least square estimation and also OLS does not account for panel data. One way to specify the effect of technical inefficiency of production is given by:

$$(2) \qquad U_{it} = z'_{it}\delta + W_{it},$$

where W_{it} is now defined by the truncation of a normal distribution with $W_{it} \sim N_+(0, \sigma_U^2, -z'_{it}\delta)$, $i = 1,\ldots, n$, $t = 1, 2$. Because U_{it} is positive and when the point of truncation is $-z'_{it}\delta$, W_{it} is always greater or equal to $-z'_{it}\delta$.

The parameters β and δ are simultaneously estimated using the maximum likelihood method. The likelihood function and its partial derivatives with respect to the parameters are quite complex and presented in the appendix of Battese and Coelli (1993). Some useful variance parameter transformations are $\sigma^2 \equiv \sigma_V^2+\sigma_U^2$ and $\gamma \equiv \sigma_U^2/\sigma^2$. If γ is zero, the variance of the inefficiency is also zero, and the model reduces to a traditional mean response function where the z_{it} are directly included in the production function.

Our focus in this study is to estimate a production function for Germany, which allows for differences in the use of input factors between East and West Germany. We specifiy the theoretical model of (1) by using a translog production function, thus, the empirical model to be estimated is now defined as follows (indices are omitted to ease readability).

$$(3) \quad lnY = \beta_{10} + \beta_{11}lnN + \beta_{12}lnK + \beta_{13}lnN^2 + \beta_{14}lnK^2 + \beta_{15}lnNlnK + \sum_{k=1}^{8} \beta_{16k}BR_k$$

$$+ \sum_{l=1}^{9} \beta_{17l}DR_l + \beta_{18}YEAR + EW*\{\beta_{20} + \beta_{21}lnL + \beta_{22}lnK + \beta_{23}lnN^2 + \beta_{24}lnK^2$$

$$+ \beta_{25}lnNlnK + \sum_{k=1}^{8} \beta_{26k}BR_k + \sum_{l=1}^{9} \beta_{27l}DR_l + \beta_{28}YEAR\} + V - U.$$

The technical inefficiency effects are estimated by:

$$(4) \quad U_{it} = \delta_0 + \delta_1 EW + \delta_2 TECH + \delta_3 ORG + \delta_4 EXP + \delta_5 SHARE + \delta_6 COLL$$

$$+ \sum_{k=1}^{8} \delta_{7k}BR_k + \sum_{l=1}^{9} \beta_{8l}DR_l + \beta_9 YEAR + W.$$

The variables used are:

- Y output (value added),
- N labor (full-time equivalents),
- K capital (instrument: replacement investment),
- BR industries (8 dummies + reference group),
- DR degree of agglomeration (9 dummies + reference group),
- YEAR year of observation (two years used),
- EW east/west (dummy, 1 if establishment lies in west Germany),
- TECH investment in information and communication technologies (log),

- ORG organizational changes (dummy,1 if at least one out of four organizational changes[1] occurred in the last two years),
- EXP turnover obtained from export (log),
- SHARE profit or capital sharing (dummy, 1 if at least one of both exists),
- COLL collective agreement (dummy, 1 if collective agreement on regional or industrial level exists).

We estimated two versions of the production frontier model[2]. The first specification (a) in Table 2 assumes differences in the production function between East and West Germany according to (3) whereas the inefficiency model only consists of the constant δ_0:

(5) $$U_{it} = \delta_0 + W_{it}\,,\ i = 1,\ldots,\ n,\ t = 1,\ 2.$$

The second specification (b) in that Table assumes no differences between the productivity in East and West Germany; i.e., (indices omitted)

(6) $$\ln Y = \beta_{10} + \beta_{11}\ln N + \beta_{12}\ln K + \beta_{13}\ln N^2 + \beta_{14}\ln K^2 + \beta_{15}\ln N \ln K + \sum_{k=1}^{8} \beta_{16k} BR_k + \sum_{l=1}^{9} \beta_{17l} DR_l + \beta_{18} YEAR + V - U,$$

but allows for a elaborated model of technical inefficiency according to (4).

[1] Organizational changes: - reorganization of departments or sections, - delegation of decision making and responsibility to lower levels, - introduction of group work / units with own authority, - introduction of profit centers / units with cost and gain accounts.

[2] We used the statistical software FRONTIER V4.1 to estimate the production frontier model (Battese & Coelli 1996).

5 Results

5.1 Results based only on the complete cases

The results show that labour has a remarkably high marginal productivity near about 0.93, which is quite large compared to other studies. The productivity of both, capital and labor is increasing with the use of the same factor and decreasing with the other. None of the interaction variables and also the east/west-dummy are statistically significant. This is surprising as studies on differences between East and West Germany show productivity gaps of about 30% and 40%. Additionally, there is some evidence that technical inefficiency occurs (see Ragnitz 2001, Bellmann and Brussig 1999). The parameter estimate for technical inefficiency is highly significant and shows the expected sign. We assume that the differences in productivity among both parts of Germany are due to technical inefficiencies. The model of technical inefficiency also includes other variables like the use of the newest technology, the proportion of export, organisational changes, profit or capital sharing, and collective agreements, which all should have influence on the establishment's productivity. The estimates are shown in column (b) of Table 2. We deleted all of the interaction variables and the east/west-dummy from the production function because all parameters are statistically insignificant. Also a test on joint significance rejects the hypothesis that the interaction variables influence the results. We did not estimate a „mother model" including all variables, because this leads to serious problems with multicollinearity between the dummy variable that indicates differences among East and West Germany in the technical efficiency model and the interaction variables in the production function. From the theory of the stochastic frontier model we know that the results in column (a) are consistent. Thus, we deleted the interaction variables, as they show no significant influence on the dependent variable.

The results of the second specification confirm the parameter estimates for the production function. The parameter for log capital becomes significant on a 10%-level, whereas the time dummy is now insignificant. Most of the variables in the technical inefficiency model show the expected influence. Inefficiency is decreasing when the investment in ICT is increasing, the firm achieves a higher amount of turnover from export, the establishment had experienced organisational changes or the employees participate at the firm's capital or profits. Additionally, collective agreements

on the regional or industry level have no influence on the efficiency of an establishment. The east/west-dummy is highly significant and negative indicating that inefficiency (U_{it}) is decreasing, when the firm is placed in the western part of Germany. The influence is much higher compared to the other exogenous variables in the model. At the average technical efficiency (see Battese and Coelli 1995, 327) an East German establishment has c. p. only 62% of the efficiency of a West German firm.

From the model we can conclude that differences in the productivity between East and West German firms are not due to different production functions but the result of a lower efficiency of East German firms even if the technical efficiency model controls for various other reasons of inefficiency.

Table 2: Estimates of a stochastic frontier production function for Germany (unbalanced panel data, 2000 - 2001, Battese & Coelli 1995)

Variables	(a)	(b)
Production function:		
Constant	4.893***	4.833***
	(119.341)	(123.125)
lnN	0.928***	0.946***
	(29.000)	(52.967)
lnK	-0.004	-0.004*
	(1.333)	(1.805)
lnN²	0.026**	0.016**
	(2.167)	(2.449)
lnK²	0.015***	0.018***
	(13.366)	(21.799)
lnN·lnK	-0.014***	-0.020***
	(4.667)	(10.468)
YEAR (2001 = 1)	0.018*	0.017
	(1.800)	(1.397)
9 industry dummies (BR)	yes	yes
8 agglomeration dummies (DR)	yes	yes
EW	0.027	-
	(0.444)	
EW·lnN	0.001	-
	(0.019)	
EW·lnK	0.006	-
	(1.406)	
EW·lnN²	-0.001	-
	(0.031)	
EW·lnK²	-0.002	-
	(0.960)	
EW·lnN·lnK	-0.002	-
	(0.450)	
EW·YEAR	-0.008	-
	(0.732)	
EW·BR_k	yes	-
EW·DR_l	yes	-
Technical inefficiency model:		
Constant	-1.606***	-1.350***
	(8.152)	(4.383)
TECH	-	-0.059***
		(6.556)
EW	-	-0.609***
		(7.709)
EXP	-	-0.162***
		(5.786)
ORG	-	-0.088***
		(3.520)
SHARE	-	-0.546***
		(7.000)
COLL	-	0.014
		(0.609)
YEAR	-	-0.054
		(1.636)
BR_k	-	yes
DR_l	-	yes
σ^2	0.708***	0.601***
	(12.207)	(8.838)
γ	0.911***	0.874***
	(101.222)	(54.625)
Mean inefficiency [=exp(-U_{it})]	0.763	0.796
Log. Likelihood	-3404.759	-3966.219
Obs.	10223	10223

Note: |t|-values in parentheses. ***, ** and * denote significance at the .01, .05, and .10 levels, respectively.

5.2 Results based on multiply imputed data

Table 3 contains the estimates for the regressions with the imputed data. From first sight, the parameters for the production function in column (a) have not changed very much. Only the time dummy is now insignificant and its parameter altered sign. The east/west-dummy and the variables that interact with this dummy still show no significant result, where as technical inefficiency seems to have influence on the results. Like in our first regressions with the unbalanced panel data, we conclude that there are no differences in the production function between East and West Germany. Therefore, we prefer the specification in column (b). Again, there are few changes in the results for the production function. Only the sign for the parameter of log capital alters from negative to positive in the regressions with the imputed data. Nevertheless, the absolute value of this parameter estimate stays small. Turning to the technical inefficiency model, more and explicit differences between the results of Tables 2 and 3 occur. In most of the cases, except the influence of organisational changes, the effect of the variables increases and leads to a higher technical efficiency.

Table 3: Estimates of a stochastic frontier production function for Germany (MI, balanced panel data, 2000 - 2001, Battese & Coelli 1995)

Variables	(a)	(b)
Production function:		
Constant	4.943***	4.897***
	(115.201)	(154.724)
lnN	0.932***	0.942***
	(28.998)	(62.737)
lnK	0.001	0.004*
	(0.304)	(1.933)
lnN²	0.027**	0.024***
	(2.242)	(4.480)
lnK²	0.014***	0.017***
	(12.487)	(21.829)
lnN·lnK	-0.015***	-0.022***
	(5.333)	(13.272)
YEAR (2001 = 1)	-0.004	0.012
	(0.279)	(1.513)
9 industry dummies (BR)	yes	yes
8 agglomeration dummies (DR)	yes	yes
EW	0.053	-
	(0.895)	
EW·lnN	-0.013	-
	(0.336)	
EW·lnK	0.005	-
	(0.947)	
EW·lnN²	0.003	-
	(0.215)	
EW·lnK²	-0.001	-
	(0.646)	
EW·lnN·lnK	-0.002	-
	(0.595)	
EW·YEAR	-0.009	-
	(0.781)	
EW·BR_k	yes	-
EW·DR_l	yes	-
Technical inefficiency model:		
Constant	-1.713***	-4.065***
	(14.632)	(9.971)
TECH	-	-0.063**
		(2.032)
EW	-	-1.012***
		(9.915)
EXP	-	-0.231***
		(13.236)
ORG	-	-0.075
		(0.598)
SHARE	-	-0.904***
		(7.611)
COLL	-	-0.145
		(1.591)
YEAR	-	-0.217**
		(2.088)
BR_k	-	yes
DR_l	-	yes
σ^2	0.812***	1.759***
	(15.710)	(16.230)
γ	0.904***	0.955***
	(95.102)	(288.108)
Mean inefficiency (e^{-Uit})	0.746	0.779
Obs.	17294	17294

Note: |t|-values in parentheses. ***, ** and * denote significance at the .01, .05, and .10 levels, respectively.

To figure out whether this increase in parameter estimates may be statistically significant, we apply two nonparametric methods, a sign-test and a signed-rank- (Wilcoxon-) test (Snedecor & Cochran, 1980). The parameter estimates in Table 3 are the means of the respective parameters of the five regressions with the imputed data. Therefore, we treat the five differences between the parameter values in Table 3 and the respective values in Table 2 as independent sample moments. Given, that the differences are distributed continuously and symmetrically around the median and that the probability of equal parameters is zero, a sample of five observations is enough to decide whether the parameters of the augmented regressions differ from that in Table 2 at least on a 10%-level.

Based on the results of these tests, we would reject the hypothesis that the effect of investment in ICT (TECH) changes when multiple imputations are used. The same is indicated for organisational changes (ORG), although the parameter becomes insignificant. All other variables in the technical inefficiency model experience significant alterations, when imputed data sets are used. Technical efficiency is increased by 6.3%, when investment in ICT doubles. As mentioned before this result does not differ statistically from the estimations in Table 2, where we find a 5.9% growth. The differences among the two parts of Germany (east and west) are now much higher compared to the regressions with the unbalanced panel data.

According to the results in Table 3, the average efficiency of an East German firm is only about 50% of that of a West German establishment. Using only the data without imputations leads to an average increase in efficiency of more than 10%-points. Export and profit or capital sharing also shows a higher impact on the technical efficiency of an establishment. Doubling the export activities lead to a growth of about 23% in technical efficiency. This result is 7%-points higher than before. The technical efficiency of a firm without profit or capital sharing decreases from more than 65% to less than 53% compared to a firm with profit or capital sharing. Organisational changes and collective agreements show no significant influences on firm's technical efficiency.

Using a multiple imputation technique leads to changes in the results for the technical inefficiency model. Whereas the directions of the estimated effects stay the same for almost all the cases, the size of the influence be-

comes remarkably larger for export activities, profit or capital sharing and the differences between East and West Germany. It is possible that these results also affect the direction and size of economic policies, especially for East Germany.

6 Conclusions

In this paper we analysed the effect of multiple imputations on the estimation of stochastic production frontier models. 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. This paper focuses on the imputation and editing process to show that multiple imputations can be created quite easily with standard multiple imputation techniques and multivariate real life panel data when a sophisticated econometric model is used for inference.

We apply a stochastic production frontier model as an example to show whether multiple imputations affect the size and the statistical significance of the parameters. Thus, we use German panel data from 2000 and 2001. One feature of the German economy still is the remarkable difference between the former two German states. From earlier studies, it is well known that the productivity in West Germany is much higher compared to East Germany. Therefore, we estimated two models. The first one assumes that the use of factors of production and thus the estimated parameters differ from each other. The second model supposes that technical inefficiencies are the reason for the empirical findings.

Estimations with the non-missing data only favours the hypothesis that the differences between East and West Germany are due to technical inef-

ficiencies. In average, an East German establishment has a technical efficiency of about 60% compared to one in West Germany. This result is in line with the outcome of other studies on this topic. On the other hand, there are no differences in the partial elasticities of the production inputs labour and capital. This means if the input of labour and capital is doubled the output of goods and services will increase the same percentage in both parts of Germany. But as Eastern Germany has a lower mean productivity, the absolute change in output will be lower than in the western part. Unfortunately, the question why the productivity in East Germany is so that lower cannot be answered with our analysis. Maybe, due to the transformation of economy, mainly firms in industries with a relatively low productivity have survived respectively have been established. Also, the firm structure in East Germany mostly consists of very small firms that cannot increase their productivity because of scale effects. Nevertheless, the investigation of these assumptions has to be subject of other studies on this topic. Turning to the multiply imputed panel data the technical inefficiency assumption is again a better description of the data. The parameters of the production function differ only slightly from those when we use the complete data only. Solely the result for log capital switched sign. The parameters of some variables in the technical inefficiency model become larger in absolute terms. The differences between East and West Germany increase about 10%-points, so that an East German establishment is only half as efficient as a West German establishment. The impact of other variables like investment in ICT, profit or gain sharing and export share also grows, whereas the mean inefficiency stays almost the same in both samples. The results of the estimations also indicate that the effects of some variables are not measured well when only the complete data records are used. This shows that MI is a possible way to increase the precision of empirical investigations and may affect the results of empirical analysis but also economic and regional policies that rely on such studies.

Appendix
Data preparation and construction of variables

Variables taken from the questionnaires (the questionnaires are available on request by the authors)

SALES: turnover in DM.

INPUT: input of materials, goods and services in % of turnover.

L: total number of employees in the establishment.

PART: number of part-time employees (PART = 0 if PDUM = 0).

PDUM: dummy whether the establishment has part-time employees (yes = 1).

INVEST: investment in DM.

ADDINV: investment to enlarge capital in % of investment.

NOINV: dummy whether the establishment invests or not (no investment = 1).

ICTINV: investment in information and communication technologies in % of investment.

EW: dummy whether the establishment is located in West or East Germany (West = 1).

EUEXP: export to countries in the European currency union in % of turnover.

NEUEXP: export to countries not in the European currency union in % of turnover.

REORG: dummy whether the establishment had reorganised sections or departments (yes = 1).

DES: dummy whether the establishment had delegated decision making and responsibilities to lower levels (yes = 1).

TEAM: dummy whether the establishment had introduced group work or units with own authority (yes = 1).

PC: dummy whether the establishment had introduced profit centers or units with own accounts (yes = 1).

PROF: dummy whether the establishment shares profits with employees (yes = 1).

KAP: dummy whether the establishment shares capital with employees (yes = 1).

AGR: level of collective agreement (regional or industrial level = 1, firm or establishment level = 2, no collective agreement = 3).

BR_k: branches, k = 9.

DR_l: degree of agglomeration, l = 10.

Variables constructed for the regressions

Y (output):	SALES - SALES*(INPUT/100).
N (full-time equivalents):	L – 0.5*PART.
K (capital: instrumented by replacement investment):	INVEST - INVEST*(ADDINV/100), K = 0.001 if NOINV = 1 or if ADDINV = 100).
TECH (investment in ICT):	log(INVEST*(ICTINV/100)), TECH = log(0.001) if ICTINV = 0 or if NOINV = 1.
EW (east/west):	dummy, original variable (see above).
EXP (export):	log(SALES*(EUEXP + NEUEXP)/100), EXP = log(0.001) if EUEXP = 0 and NEUEXP=0.
ORG (organizational changes):	dummy, ORG = 1 if REORG = 1 or DES = 1 or TEAM = 1 or PC = 1.
SHARE (profit or capital sharing):	dummy, SHARE = 1 if PROF = 1 or KAP = 1.
COLL (collective agreements on regional or industrial level):	dummy, COLL = 1 if AGR = 1.
YEAR (year of observation):	dummy, YEAR = 1 if observation in 2001.
BR_k:	dummies from original categorical variable (see above).
DR_l:	dummies from original categorical variable (see above).

Data transformation for MI-procedure

SALES:	logarithmic
INPUT:	no transformation
L:	Box-Cox
PART:	logit, dummy*
PDUM:	dummy, no transformation
INVEST:	logarithmic
ADDINV:	logit
NOINV:	dummy, no transformation
ICTINV:	logit
EW:	dummy, no transformation
EUEXP:	logarithmic, dummy*
NEUEXP:	Box-Cox, dummy*
REORG:	dummy, no transformation
DES:	dummy, no transformation
TEAM:	dummy, no transformation
PC:	dummy, no transformation
PROF:	dummy, no transformation
KAP:	dummy, no transformation
AGR:	3 dummy variables
BR_k:	9 dummy variables
DR_l:	10 dummy variables

* We treated these variables as semicontinuous, 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.

References

Aigner, D., Lovell, C. and Schmidt, P. (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.

Battese, G. and Coelli, T. (1993) A stochastic frontier production function incorporating a model for technical inefficiency effects. *Working papers in econometrics and applied statistics*, **69**, Department of Econometrics, University of New England, Armindale, Australia.

Battese, G. and Coelli, T. (1995) A model for technical inefficiency effects in a stochastic frontier production function for panel data. *Empirical Economics*, **20**, 325 - 332

Battese, G. and Coelli, T. (1996) A guide to FRONTIER version 4.1: A computer program for stochastic frontier production and cost function estimation. *CEPA working paper*, **96/07**, Centre for Efficiency and Productivity Analysis, University of New England, Armindale, Australia.

Bellmann, L. and Brussig, M. (1999) Productivity differences between western and eastern German establishments. *IAB Topics*, **37**, Nürnberg, Germany.

Brand, J.P.L. (1999) *Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets.* Thesis Erasmus University Rotterdam. Enschede: Print Partners Ispkamp.

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

Greene, W. (2000) *Econometric analysis*. 4th edn. New York: Upper Saddle River.

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

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

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

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

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

Ragnitz, J. (2001) Produktivitätsrückstand der ostdeutschen Wirtschaft: Eine zusammenfassende Bewertung. *Wirtschaft im Wandel*, **7**, 181-189.

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

Rubin, D.B. (1987) *Multiple imputation for nonresponse in surveys*. New York: Wiley.

Rubin, D.B. and Schenker, N. (1998) Imputation. In *Encyclopedia of statistical sciences, update volume 2* (eds. S. Kotz, C.B. Read and D.L. Banks), 336 - 342. New York: Wiley.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data*. London: Chapman & Hall.

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 Missing values proceedings of a symposium on incomplete data (eds. J. Bethlehem, and S. van Buuren, 1 - 21, Utrecht.

Schafer, J.L. and Olsen, M.K. (1999) Modeling and imputation of semicontinuous survey variables. Technical report, 00 - 39, The Pennsylvania State University.

Snedecor, G. and Cochran, W. (1980) Statistical methods, 7th edn.. Iowa: Aimes.

In dieser Reihe sind zuletzt erschienen:

Recently published:

1	Bauer, Th. K., Bender, St., Bonin, H.	Dismissal Protection and Worker Flows in Small Establishments	7/2004
2	Achatz, J., Gartner, H., Glück, T.	Bonus oder Bias? Mechanismen geschlechtsspezifischer Entlohnung	7/2004
3	Andrews, M., Schank, Th., Upward, R.	Practical estimation methods for linked employer-employee data	8/2004
4	Brixy, U., Kohaut, S., Schnabel; C.	Do newly founded firms pay lower wages? First evidence from Germany	9/2004

Impressum

IABDiscussionPaper
No. 5 / 2004

Herausgeber
Institut für Arbeitsmarkt- und Berufsforschung
der Bundesagentur für Arbeit
Weddigenstr. 20-22
D-90478 Nürnberg

Redaktion
Regina Stoll, Jutta Palm-Nowak

Technische Herstellung
Jutta Sebald

Rechte
Nachdruck – auch auszugsweise – nur mit
Genehmigung des IAB gestattet

Bezugsmöglichkeit
Volltext-Download dieses DiscussionPaper
unter:
http://doku.iab.de/discussionpapers/2004/dp0504.pdf

IAB im Internet
http://www.iab.de

Rückfragen zum Inhalt an
Susanne Rässler, Tel. 0911/179-3084,
oder e-Mail: susanne.raessler@iab.de