

Imputation Approaches for Top-Coded Wages in the German IAB Employment Sample

Thomas Büttner¹, Susanne Rässler²

¹ Institute for Employment Research, e-mail: thomas.buettner@iab.de

² University of Bamberg, e-mail: susanne.raessler@uni-bamberg.de

Abstract

In many large data sets of economic interest, some variables, as wages, are top-coded or right-censored. In order to analyze wages with the German IAB employment sample we first have to solve the problem of censored wages at the upper limit of the social security system. We treat this problem as a missing data problem and use multiple imputation approaches to impute the censored wages by draws of a random variable from a truncated distribution based on Markov chain Monte Carlo techniques. In general, the variation of income is smaller in lower wage categories than in higher categories and the assumption of homoscedasticity in an imputation model is highly questionable. Therefore, we suggest a new multiple imputation method which does not presume homoscedasticity of the residuals. A first simulation study shows that this approach is superior to approaches assuming homoscedasticity. In this paper, we perform a simulation study, using uncensored wage information from a survey (German structure of earnings survey) to compare the imputation approaches and to confirm the validity of the new approach.

JEL codes: C24, C15.

KEY WORDS: top coding, missing data, censored wage data, Markov chain Monte Carlo

1 Introduction

For a large number of research questions, like analyzing the gender wage gap or measuring overeducation with earnings frontiers, it is interesting to use wage data. To address this kind of questions two types of data are usually used: surveys and process generated data, i.e. administrative data. Administrative data have several advantages, like a large number of observations, no nonresponse burden and no problems with interviewer effects or survey bias. Unfortunately, in many large administrative data sets of economic interest some variables, such as wages, are top-coded or right-censored. This problem is very common with administrative data from social security systems like the IAB employment sample (IABS). This data set represents approximately 80 percent of the employees in Germany. The IABS includes, among others, information on age, sex, education, wage and the occupational group (see Bender et al. 2000) and is based on the register data of the German social insurance system. The contribution rate of this insurance is charged as a percentage of the gross wage. Is the gross wage higher than the current contribution limit, however only the amount of the ceiling is liable for the contribution. In 2009 the contribution limit in the unemployment insurance system is fixed in Western Germany

at a monthly income of 5,400 euros. As therefore wages are only recorded up to the contribution limit, the wage information in this sample is censored at this limit. (Figure 1 shows the distribution of wages in the IAB employment sample in 2000).

In order to analyze wages with the IAB employment sample, we first have to solve the problem of the censored wages (see Rässler 2006). We treat this problem as a missing data problem and use imputation approaches to impute the censored wages. Gartner (2005) proposes a non-Bayesian single imputation approach to solve the problem of the censored wages. Another approach - a multiple imputation method based on draws of a random variable from a truncated distribution and Markov chain Monte Carlo techniques - is suggested by Gartner and Rässler (2005).

These two approaches assume homoscedasticity of the residuals. But on the contrary of this assumption, the variance of income is smaller in lower wage categories than in higher categories and assuming homoscedasticity in an imputation model is highly questionable. Therefore, we suggest a new multiple imputation method allowing for heteroscedasticity. A simulation study (Büttner and Rässler 2008) shows that in case of heteroscedasticity this approach is superior to the two approaches assuming homoscedasticity. Moreover it does not matter if the algorithm considering heteroscedasticity is chosen in a homoscedastic case, since it just represents a generalization of the homoscedastic approach and therefore works well in case of homoscedasticity. In this paper we additionally perform a simulation study to compare the two multiple imputation approaches using the German structure of earnings survey (GSES) 2001, which contains uncensored wage information, in order to confirm the validity of the two multiple imputation approaches, especially the new approach considering heteroscedasticity.

The paper is organized as follows: The next section describes the two multiple imputation approaches. In Section 3 we provide a description of the simulation study, followed by the results. Finally, Section 4 concludes.

2 Imputation approaches for censored wages

This section of the paper describes the two different multiple imputation approaches to impute the missing wage information in the IAB employment sample, which were already mentioned. We assume that the wage in logs y for every person i is given by

$$y_i^* = x_i' \beta + \varepsilon_i \quad \text{where} \quad \varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2), i = 1, \dots, n. \quad (1)$$

X are covariates such as education, gender or age. As the wages in the IAB employment sample are censored at the contribution limit a we observe the wage $y_{obs,i} = y_i^*$ only if the wage is lower than the threshold a . If the wage is censored, i.e. has a value greater or equal to a , then we observe the limit a instead of the true wage y_i^* :

$$y_i = \begin{cases} y_{obs,i} & \text{if } y_i^* \leq a \\ a & \text{if } y_i^* > a \end{cases} \quad (2)$$

To be able to analyze wages with our data set, we first have to impute the wages above

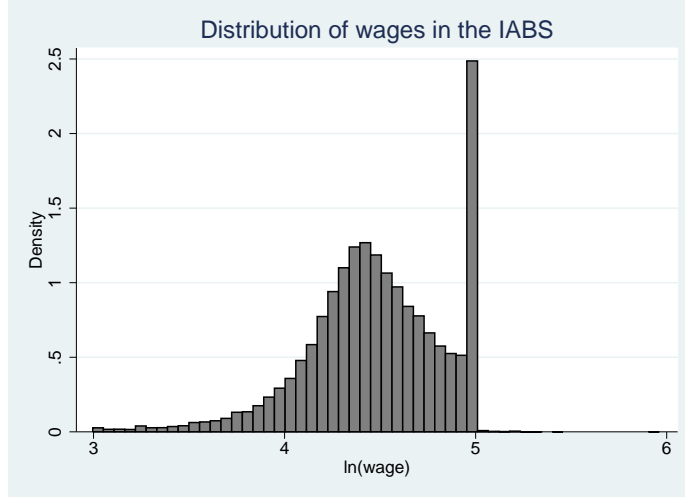


Figure 1: Distribution of daily wages in logs in the IAB employment sample (IABS) in Western Germany (2000).

a. We define $y_z = (y_{obs}, z)$, where z is a truncated variable in the range (a, ∞) .

2.1 Multiple imputation assuming homoscedasticity

One possibility to impute the missing wage information is using a single imputation approach. A homoscedastic single imputation based on a tobit model is proposed by Gartner (2005). Using a single imputation approach, we have to consider that this method may lead to biased variance estimations. Thus, Little and Rubin (1987, 2002) suggest that imputation should rather be done in a multiple and Bayesian way according to Rubin (1978). Therefore, we better use multiple imputation approaches to impute the missing wage information. Multiple imputation is discussed in detail in Rubin (1987, 2004a, 2004b) or Rässler et al. (2007). For computational guidance on creating multiple imputations see Schafer (1997).

To start with, let $Y = (Y_{obs}, Y_{mis})$ denote the random variables concerning the data with observed and missing parts. In our specific situation this means that for all units with wages below the limit a each data record is complete, i.e., $Y = (Y_{obs}) = (X, wage)$. For every unit with a value of the limit a for its wage information we treat the data record as partly missing, i.e., $Y = (Y_{obs}, Y_{mis}) = (X, ?)$. X is observed for all units. Thus, we have to multiply impute the missing data $Y_{mis} = wage$. The theory and principle of multiple imputation originates from Rubin (1978) and is based on independent random draws from the posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$ of the missing data given the observed data. As it may be difficult to draw from $f_{Y_{mis}|Y_{obs}}$ directly, a two-step procedure for each of the m draws is useful:

- (a) First, we perform random draws of Ξ according to the observed-data posterior distribution $f_{\Xi|Y_{obs}}$, where Ξ is the parameter vector of the imputation model.

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

Because

$$f_{Y_{mis}|Y_{obs}}(y_{mis}|y_{obs}) = \int f_{Y_{mis}|Y_{obs},\Xi}(y_{mis}|y_{obs},\xi) f_{\Xi|Y_{obs}}(\xi|y_{obs}) d\xi \quad (3)$$

holds, with (a) and (b) we achieve imputations of Y_{mis} from their posterior predictive distribution $f_{Y_{mis}|Y_{obs}}$. Due to the data generating model being used, for many models the conditional predictive distribution $f_{Y_{mis}|Y_{obs},\Xi}$ is rather straightforward. That means it can be more or less easily formulated for each unit with missing data. In contrast, the corresponding observed-data posteriors $f_{\Xi|Y_{obs}}$ are usually difficult to derive for those units with missing data, especially when the data have a multivariate structure. In these cases, they often do not follow a standard distribution from which random numbers can easily be generated. However, simpler methods have been developed to enable multiple imputation based on Markov chain Monte Carlo (MCMC) techniques (Schafer 1997). In MCMC the desired distributions $f_{Y_{mis}|Y_{obs}}$ and $f_{\Xi|Y_{obs}}$ are achieved as stationary distributions of Markov chains based on the complete-data distributions, which are easier to compute.

Imputation model

Gartner and Rässler (2005) suggest an imputation approach based on Markov chain Monte Carlo techniques to multiply impute the right-censored wages in the IAB employment sample, which contains the following steps. To be able to start the imputation based on MCMC, we first need to adapt starting values for $\beta^{(0)}$ and the variance $\sigma^{2(0)}$ from a ML tobit estimation. Second, in the imputation step, values for the missing wages are randomly drawn from a truncated distribution:

$$z_i^{(t)} \sim N_{trunc_a}(x_i' \beta^{(t)}, \sigma^{2(t)}) \text{ if } y_i = a \text{ for } i = 1, \dots, n. \quad (4)$$

Then an OLS regression is computed based on the imputed data according to

$$\widehat{\beta}_z^{(t)} = (X'X)^{-1} X' y_z^{(t)}. \quad (5)$$

After this step, new random draws for the parameters can be produced according to their complete data posterior distribution. To draw the variance $\sigma^{2(t+1)}$ we need the inverse of a gamma distribution, which is produced as follows:

$$g \sim \chi^2(n - k) \quad (6)$$

$$\sigma^{-2(t+1)} = \frac{g}{RSS} \quad (7)$$

where RSS is the residual sum of squares $RSS = \sum_{i=1}^n (y_{z_i}^{(t)} - x_i' \widehat{\beta}_z^{(t)})^2$ and k is the number of columns of X.

Now new random draws for the parameter β can be performed

$$\beta^{(t+1)} | \sigma^{2(t+1)} \sim N(\hat{\beta}_z^{(t)}, \sigma^{2(t+1)}(X'X)^{-1}). \quad (8)$$

We repeat the imputation and the posterior-steps (4) to (8) 6,000 times and use $(z_i^{2000}, z_i^{3000}, \dots, z_i^{6000})$ to obtain 5 complete data sets. For more details see Gartner and Rässler (2005) or Jensen et al. (2006).

2.2 Multiple imputation considering heteroscedasticity

As we assume that the variation of income is smaller in lower wage categories than in higher categories, we suppose an imputation approach considering heteroscedasticity. We develop this new method based on the multiple imputation approach proposed by Gartner and Rässler (2005). The basic element of the new approach is that we need additional draws for the parameters γ describing the heteroscedasticity. We start now the imputation by adapting starting values for $\beta^{(0)}$ and $\gamma^{(0)}$ from a GLS estimation for truncated variables. Then we are able to draw values for the missing wages from a truncated distribution using individual variances $\sigma_i^2 = e^{w_i' \gamma}$:

$$z_i^{(t)} \sim N_{trunc_a}(x_i' \beta^{(t)}, \sigma_i^{2(t)}) \quad \text{where} \quad \sigma_i^{2(t)} = e^{w_i' \gamma^{(t)}} \quad \text{if } y_i = a \quad \text{for } i = 1, \dots, n. \quad (9)$$

Then a GLS regression is computed based on the imputed data set (comparable to the OLS regression in the homoscedastic multiple imputation approach) to obtain $\hat{\beta}_z^{(t)}$ and $\hat{\gamma}^{(t)}$. Additionally we estimate $V(\hat{\gamma}^{(t)})$ to be able to perform the following steps. Afterwards we produce new random draws for the parameters according to their complete data posterior distribution. As we consider now the existence of heteroscedasticity, some modifications of the algorithm are necessary. In the first step, we draw the variance $\sigma^{2(t+1)}$ according to

$$g \sim \chi^2(n - k) \quad (10)$$

$$\sigma^{-2(t+1)} = \frac{g}{RSS} \quad (11)$$

where

$$RSS = \sum_{i=1}^n \exp(\ln \hat{\varepsilon}_i^2 - w_i' \hat{\gamma}^{(t)}) = \sum_{i=1}^n \frac{(y_{z_i}^{(t)} - x_i' \hat{\beta}^{(t)})^2}{e^{w_i' \hat{\gamma}^{(t)}}}. \quad (12)$$

In an additional step, we have to perform random draws for γ

$$\gamma^{(t+1)} \sim N(\hat{\gamma}^{(t)}, \hat{V}(\hat{\gamma}^{(t)})) \quad (13)$$

Consequently the parameters β can be drawn like in the Gartner and Rässler approach, again with a slight modification compared to the homoscedastic multiple imputation:

$$\beta^{(t+1)} | \gamma^{(t+1)}, \sigma^{2(t+1)} \sim N(\hat{\beta}_z^{(t)}, \sigma^{2(t+1)} (\sum_{i=1}^n \frac{x_i x_i'}{e^{w_i' \gamma^{(t+1)}}})^{-1}). \quad (14)$$

Again, we repeat the steps (9) to (14) 6,000 times and use $(z_i^{2000}, z_i^{3000}, \dots, z_i^{6000})$ to obtain the 5 complete data sets. For more details on this approach see Büttner and Rässler (2008).

3 Simulation study

To evaluate the results delivered by these two approaches under different situations in order to show the relevance of the suggested multiple imputation approach, we use simulation studies. For a first simulation study (see Büttner and Rässler 2008) we created one data set where homoscedasticity is existent and another with heteroscedasticity of the residuals to compare the approaches under different situations in order to confirm the necessity and validity of the suggested new method. The results of the simulation study can be summarized as follows: The missing wage information should be imputed multiply, because single imputations may lead to biased variance estimations. Furthermore, the imputation should be done considering heteroscedasticity. As the assumption of homoscedasticity is highly questionable with wage data, the simulation study shows that it is preferable to use the new approach considering heteroscedasticity, as this approach is more general: In case of homoscedastic residuals the same quality of imputation results can be expected compared to the Gartner and Rässler (2005) approach. But if heteroscedasticity is existent the simulation study confirms the necessity of the new approach.

Here we perform a simulation study using data from an income survey (German structure of earnings survey, GSES) with uncensored wage information. This data set allows us to compare the different imputation approaches again using a complete data set. We truncate the wage variable at a ceiling (we delete the wages above the 85 quantile comparable to the top-coding in the IABS) and impute the deleted information using the two different approaches. Afterwards we compare the imputed data sets with the original complete data set.

3.1 The German Structure of Earnings Survey, GSES

The German structure of earnings survey is conducted every four years in establishments of the manufacturing industry and the service sector. To perform the simulation study we use the GSES for the year 2001, which is a linked employer-employee data set and contains information on about 22,000 establishments and more than 846,000 employees. The GSES includes information on the individuals (e.g. sex, age, education, children), on the job (e.g. occupation, job level, performance group, working times, tenure), on earnings (e.g. gross wage, net wage, income taxes, social security contributions) and additionally on the establishment (e.g. number of employees). The survey is therefore suitable to examine a broad range of questions concerning wages. For more details see Forschungsdatenzentrum

der Statistischen Landesämter (2006). To simplify the simulation design, we restrict the sample to male West-German residents holding a full-time job covered by social security. This sample contains 382,710 persons and is used as complete population for the following simulation study.

3.2 Simulation design

The simulation study consists of four steps. First we draw 10 percent random samples from the complete population repeatedly, delete the wages above the defined ceiling and impute the wages again using the two multiple imputation approaches. Then we compare the imputed data sets with the complete population. The whole simulation procedure - consisting of drawing a random sample, imputing the data using the different approaches, running a regression on the imputed data sets to be able to compare the results and calculating confidence intervals for the estimates - is repeated 1000 times. Finally the fraction of confidence intervals based on $\hat{\beta}_{MI}$ containing the parameter β (of the original complete data set) can be calculated for the different approaches (coverage).

For the simulation study we assume an imputation model containing the wages in logs as dependent variable and dummies for education levels, job levels and performance groups, age, age squared, a dummy for indefinite contracts of employment and dummies for regions and industries as independent variables. To analyze the results we use the same model as the imputation model. In every iteration we impute the samples $m=5$ times. That means performing each of the multiple imputation approaches, $m=5$ complete data sets are obtained and the regressions have to be done five times as well. Afterwards, the results have to be combined using the combining rules first described by Rubin (1987).

3.3 Results

Table 1 shows the results of this simulation study. We receive a coverage for both imputation approaches around 95 percent for most of the variables - similar to the coverage received by the estimations using the complete random samples - which refers to a good imputation quality. Only for some variables we find a considerably lower coverage. In these cases the coverage for both imputation approaches is lower (except for the dummy for Region 2, where the coverage of the homoscedastic approach is significantly lower). Interpreting these results, we find out, that both approaches deliver comparable, good imputation results, when we simulate to impute censored wages in the GSES. Taking into consideration the results of the first simulation study, nevertheless can be subsumed that it is still advisable to use the multiple imputation approach considering heteroscedasticity to impute the missing wage information in the IABS.

4 Conclusion

There is a wide range of ways to deal with censored wage data. We propose to use imputation approaches to estimate the missing wage information. Nevertheless, there are

	complete data			MI homosc.		MI heterosc.	
	β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
educ2	0.0345	0.0346	0.962	0.0352	0.962	0.0348	0.972
educ3	0.0596	0.0592	0.947	0.0501	0.917	0.0530	0.942
educ4	0.0713	0.0714	0.948	0.0639	0.846	0.0633	0.834
educ5	0.1370	0.1374	0.949	0.1495	0.627	0.1462	0.790
educ6	0.1770	0.1771	0.940	0.1776	0.945	0.1772	0.936
level2	0.0105	0.0106	0.961	0.0099	0.962	0.0095	0.967
level3	0.0371	0.0382	0.967	0.0399	0.960	0.0395	0.964
level4	0.0201	0.0212	0.974	0.0094	0.963	0.0055	0.940
group2	-0.0947	-0.0948	0.944	-0.0926	0.933	-0.0925	0.927
group3	-0.1899	-0.1897	0.946	-0.1866	0.901	-0.1866	0.891
group4	-0.3098	-0.3098	0.961	-0.3065	0.924	-0.3071	0.929
group5	0.3875	0.3863	0.967	0.3956	0.964	0.3866	0.967
group6	0.1412	0.1401	0.963	0.1488	0.967	0.1498	0.957
group7	0.0479	0.0469	0.967	0.0589	0.950	0.0613	0.943
group8	-0.1702	-0.1713	0.966	-0.1589	0.957	-0.1554	0.936
group9	-0.3394	-0.3400	0.966	-0.3280	0.955	-0.3253	0.945
age	0.0247	0.0247	0.955	0.0253	0.934	0.0247	0.976
sqage	-0.0003	-0.0003	0.949	-0.0003	0.867	-0.0003	0.969
region2	0.0369	0.0368	0.949	0.0463	0.286	0.0408	0.851
region3	0.0038	0.0037	0.941	0.0081	0.798	0.0062	0.930
region4	0.0517	0.0516	0.939	0.0529	0.937	0.0471	0.679
industry2	-0.0407	-0.0404	0.952	-0.0403	0.952	-0.0401	0.950
industry3	-0.1097	-0.1094	0.925	-0.1140	0.920	-0.1135	0.929
industry4	0.0053	0.0054	0.959	0.0044	0.963	0.0060	0.961
industry5	0.0765	0.0773	0.956	0.0729	0.946	0.0712	0.936
industry6	0.0788	0.0791	0.962	0.0827	0.949	0.0824	0.950
industry7	0.0636	0.0641	0.956	0.0701	0.849	0.0692	0.887
industry8	-0.0145	-0.0146	0.955	-0.0115	0.946	-0.0112	0.935
industry9	-0.0157	-0.0158	0.969	-0.0129	0.958	-0.0120	0.959
industry10	0.0252	0.0257	0.952	0.0301	0.903	0.0303	0.899
industry11	-0.0356	-0.0355	0.959	-0.0329	0.941	-0.0324	0.937
industry12	-0.0029	-0.0027	0.940	0.0015	0.886	0.0015	0.885
industry13	-0.0166	-0.0166	0.931	-0.0174	0.950	-0.0187	0.946
industry14	-0.0278	-0.0275	0.948	-0.0276	0.965	-0.0277	0.965
industry15	-0.0408	-0.0404	0.949	-0.0371	0.943	-0.0373	0.948
industry16	0.0341	0.0343	0.946	0.0369	0.950	0.0367	0.947
industry17	-0.0727	-0.0722	0.964	-0.0718	0.971	-0.0703	0.957
industry18	-0.0100	-0.0096	0.958	0.0047	0.409	0.0053	0.371
industry19	-0.0219	-0.0216	0.954	-0.0170	0.902	-0.0163	0.881
industry20	-0.1047	-0.1047	0.959	-0.1045	0.964	-0.1034	0.960
industry21	-0.0874	-0.0866	0.923	-0.0863	0.924	-0.0853	0.910
industry22	-0.1124	-0.1124	0.950	-0.1163	0.938	-0.1151	0.948
industry23	-0.0549	-0.0546	0.953	-0.0566	0.965	-0.0565	0.959
industry24	-0.1604	-0.1599	0.935	-0.1608	0.952	-0.1593	0.954
industry25	-0.2215	-0.2215	0.953	-0.2206	0.948	-0.2198	0.947
industry26	-0.0560	-0.0558	0.961	-0.0557	0.958	-0.0545	0.950
industry27	-0.0454	-0.0449	0.945	-0.0493	0.925	-0.0486	0.939
industry28	-0.0865	-0.0863	0.963	-0.0845	0.965	-0.0839	0.963
industry29	-0.0697	-0.0696	0.950	-0.0669	0.931	-0.0659	0.923
industry30	-0.0705	-0.0700	0.897	-0.0806	0.815	-0.0782	0.885
industry31	-0.0673	-0.0670	0.926	-0.0658	0.937	-0.0652	0.946
industry32	-0.0662	-0.0653	0.846	-0.0699	0.887	-0.0685	0.875
industry33	0.0112	0.0113	0.944	0.0114	0.951	0.0114	0.955
industry34	-0.0948	-0.0945	0.963	-0.0879	0.878	-0.0840	0.777
industry35	-0.0019	-0.0015	0.932	-0.0183	0.682	-0.0192	0.625
industry36	-0.2604	-0.2603	0.936	-0.2639	0.943	-0.2629	0.951
contract	-0.1114	-0.1117	0.936	-0.1139	0.949	-0.1111	0.958
cons	4.0440	4.0447	0.955	4.0331	0.936	4.0445	0.972

Table 1: Simulation results

also different possibilities to impute the wages in the IAB employment sample, for example single and multiple imputation approaches. Another important question is whether the wages should be imputed considering heteroscedasticity or not.

We suggest to multiply impute the missing wage information above the limit of the social security in the IAB employment sample. We assume that the variance of income is smaller in lower wage categories than in higher categories and suggest a multiple imputation approach considering heteroscedasticity to impute the missing wage information. The basic element of this approach is to impute the missing wages by draws of a random variable from a truncated distribution, based on Markov chain Monte Carlo techniques. The main innovation of the suggested approach is to perform additional draws for the parameter γ describing the heteroscedasticity in order to be able to allow individual variances for every individual. As shown in Büttner and Rässler (2008), it is preferable to use the new approach considering heteroscedasticity, as this approach is more general: In case of homoscedastic residuals the same quality of imputation results can be expected compared to the Gartner and Rässler (2005) approach. But if heteroscedasticity is existent a simulation study shows the necessity of our new approach.

To confirm the validity of this new method we perform another simulation study and compare the different approaches using a survey with uncensored wage information. The results can be concluded as follows: Simulating to impute censored wages with the German structure of earnings survey we find the validity of both approaches confirmed. Both approaches deliver good imputation results, with some slight advantages for the approach considering heteroscedasticity.

5 References

- Bender, S., Haas, A. and Klose, C. (2000). *IAB Employment Subsample 1975-1995. Opportunities for Analysis Provided by Anonymised Subsample*. IZA Discussion Paper no. 117, Bonn.
- Büttner, T. and Rässler, S. (2008). *Multiple imputation of right-censored wages in the German IAB Employment Sample considering heteroscedasticity*. IAB Discussion Paper 44/2008, Nürnberg.
- Forschungsdatenzentrum der Statistischen Landesämter (2006). *Gehalts- und Lohnstrukturerhebung im Produzierenden Gewerbe und im Dienstleistungsbereich 2001*. Metadaten für das Scientific Use File, Wiesbaden.
- Gartner, H. (2005). *The imputation of wages above the contribution limit with the German IAB employment sample*. FDZ Methodenreport 2/2005, Nürnberg.
- Gartner, H. and Rässler, S. (2005). *Analyzing the changing gender wage gap based on multiply imputed right censored wages*. IAB Discussion Paper 05/2005, Nürnberg.
- Jensen, U., Gartner, H. and Rässler, S. (2006). *Measuring overeducation with earnings frontiers and multiply imputed censored income data*. IAB Discussion Paper 11/2006, Nürnberg.
- Little, R.J.A and Rubin D.R. (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Little, R.J.A and Rubin D.R. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, Hoboken.
- Rässler, S. (2006). *Der Einsatz von Missing Data Techniken in der Arbeitsmarktforschung des IAB*. Allgemeines Statistisches Archiv, 90, 527-552.

- Rässler, S., Rubin D.B., Schenker, N. (2007). *Incomplete data: Diagnosis, imputation and estimation*. In: de Leeuw, E., Hox, J., Dillman, D. (Eds.), *The international Handbook of Survey Research Methodology*. Sage, Thousands Oaks.
- Rubin, D.B. (1978). *Multiple imputation in sample surveys - a phenomenological Bayesian approach to nonresponse*. Proceedings of the Survey 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. (2004a). *Multiple Imputation for Nonresponse in Surveys*, 2nd ed. Wiley, New York.
- Rubin, D.B. (2004b). *The design of a general and flexible system for handling nonresponse in sample surveys*. The American Statistician, 58, 298-302.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman & Hall, New York.