Multiple Imputation of Censored Variables

Dissertation

zur Erlangung des akademischen Grades eines
Doktors der Sozial- und Wirtschaftswissenschaften
(Dr. rer. pol.)

an der Fakultät Sozial- und Wirtschaftswissenschaften
der Otto-Friedrich-Universität Bamberg

vorgelegt von
Thomas Büttner

Bamberg, im Juli 2010

Datum der Disputation: 17. Juni 2010

Promotionskommission:
Professor Dr. Susanne Rässler (Erstgutachter)
Professor Trivellore E. Raghunathan, Ph.D. (Zweitgutachter)
Professor Dr. Johannes Schwarze
Professor Dr. Henriette Engelhardt-Wölfler

Danksagung

Diese Dissertation wäre nicht möglich gewesen ohne die Unterstützung einer Vielzahl von Kollegen und Freunden, die entscheidend zum Gelingen der Arbeit beigetragen haben. Zuallererst gilt mein Dank meiner Doktormutter Prof. Dr. Susanne Rässler für ihre langjährige Unterstützung. Vor allem möchte ich Ihr dafür danken, dass sie mich in die Welt der multiplen Imputation eingeführt hat und dafür, dass sie mir die Möglichkeit eröffnet hat, dieses Dissertationsprojekt überhaupt anzugehen. Ein entscheidender Grund, dass diese Dissertation nun abgeschlossen vorliegt, ist sicherlich, dass sie auch in schwierigen Phasen, oftmals deutlich stärker als ich selbst, immer an den Erfolg dieses Projekts geglaubt hat. Mein ausdrücklicher Dank gilt auch meinem Zweitbetreuer Prof. Trivellore Raghunathan, Ph.D., vor allem für einige wertvolle Tipps, die mir aus mehreren hoffnungslos erscheinenden Situationen geholfen haben.
Diese Dissertation ist während meiner Tätigkeit am Institut für Arbeitsmarkt- und Berufsforschung (IAB) entstanden und wurde im Rahmen des IAB/WiSo-Graduiertenprogramms (GradAB) mit einem Stipendium gefördert. Mein Dank gebührt daher auch einer Reihe von Kolleginnen und Kollegen am IAB, insbesondere jenen im Kompetenzzentrum empirische Methoden (KEM). Besonders hervorzuheben sind hier Dr. Johannes Ludsteck, Prof. Dr. Hans Kiesl und Dr. Jörg Drechsler. Johannes Ludsteck und Hans Kiesl dafür, dass sie mir von Anfang an mit Rat und Tat zur Seite standen. Nur durch ihre Hilfe war es möglich, dass die eine oder andere Wissenslücke meinerseits dem Erfolg des Dissertationsvorhabens nichts anhaben konnte. Jörg Drechsler dafür, dass er den oftmals steinigen Weg der Promotion mit mir gemeinsam gegangen ist. Allen drei danke ich für hilfreiche Diskussionen über die verschiedensten Themen, seien sie wissenschaftlich oder auch eher privat, und natürlich auch für die unvergesslichen gemeinsamen Reisen zu Konferenzen in der ganzen Welt. Bedanken möchte ich mich auch bei den Kolleginnen und Kollegen im Graduiertenprogramm, besonders bei Katrin Hohmeyer und Eva Kopf, die immer ein offenes Ohr auch für die irrelevantesten Probleme hatten.

Nicht zuletzt möchte ich mich bei meinen Eltern Christa und Ulrich Büttner und meiner Schwester Susanne Büttner bedanken, die mich bei diesem Vorhaben aber auch in allen anderen Lebenslagen stets unterstützt haben. Kaum möglich wäre das Vorhaben auch ohne die Unterstützung meiner Freundin Ariadna Ripoll Servent gewesen. Ich kann ihr gar nicht genug danken für ihre beeindruckenden Motivationskünste, ihr Verständnis und ihre Hilfe in den verschiedensten Situationen, nicht nur bei der Korrektur des englischen Textes. Abschließend möchte ich mich bei allen bedanken, die ich hier nicht explizit aufführen konnte, die mich aber dennoch in vielfältiger Weise unterstützt haben, sei es nur, dass sie meinen in den letzten Wochen vor Abschluss der Arbeit stetig wechselnden Gemütszustand klaglos hingenommen haben.

Nürnberg, Juli 2010
Thomas Büttner

Contents

List of Figures

List of Tables

Chapter 1

Introduction and Motivation

Censoring of the dependent variable is a very common problem with micro-data. In case of a censored variable, all values in a certain range are reported as a single value, which means the variable is partly continuous but has multiple observations at one point. This often occurs when the variable is zero for a significant part of the population but many different positive outcomes can be observed for the rest of the population. Common examples for this situation are vacation expenditures, automobile expenditures, hours of work, or charitable contributions. Wooldridge (2002, p. 517) calls this kind of variables 'corner solution outcome'. In such cases standard estimation techniques, like, e.g., ordinary least squares, are inconsistent because these methods fail to account for the difference between limit observations and continuous observations.

Wooldridge (2002) defines a second category of censoring: data censoring. In case of data censoring we have a variable with quantitative meaning, y^*. Due to a data problem y^* is censored from above and/or below and therefore cannot be observed for some part of the population. If y^* was observed for the entire population, standard estimation techniques could be applied, but due the censoring specific censored data models have to be adapted. Censoring from below, also called left-censoring, frequently appears with environmental data due to detection limits of laboratory assay procedures (see, e.g., Helsel (1990) or Newton and Rudel (2007)). Censoring from above or right-censoring is a common problem of survey data. An important example are the top-coded income variables in the U.S. Current Population Survey (CPS) conducted by the Census Bureau. Here, censoring is used as a measure to ensure confidentiality of the respondents. Therefore, if earnings are to be analyzed from these

data, standard models cannot be applied.

Generally, the problem of data censoring concerning wage and income variables occurs frequently in all fields of economics and sociology, where these variables are in the center of interest of many studies. For a large number of research questions, like analyzing the gender wage gap, assessing the determinants of wage returns to education, evaluating the effects of changes in the institutional and legal framework or several other applications, 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 over survey data, 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 or sociological interest some variables are not entirely available. This applies prevalently to wage and earnings information, which are often top-coded or right-censored due to manifold reason. The data may not be available due the data collection process, artificially censored to ensure confidentiality, or just not reliable because high wage earners tend above average not to answer income questions.

An important example for this problem is the German IAB Employment Sample (IABS), which represents administrative data coming from the social security systems. Here, right-censoring of wages occurs due to the contribution limit in the German social security system. 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. Therefore, if the gross wage is higher than the current contribution limit only the amount of the ceiling is liable for the contribution. In 2010, the contribution limit in the unemployment and pension insurance system is fixed at a monthly income of 5,500 euros in West Germany and at 4,650 euros in East Germany. Therefore, since wages are only recorded up to the contribution limit, the wage information in the sample is censored at this limit.

Due to its importance for all kind of researchers in Germany, the thesis focuses on the right-censored wage variable in the IAB Employment Sample. Nevertheless, all suggested approaches are generally valid for all kind of data sets

faced with censoring from above or below.

In the literature a wide range of models to handle censored data is proposed. The most famous is without any doubt the censored regression model first proposed by Tobin (1958). Other models include Powell's (1984) censored least absolute deviation method (CLAD) or the iterative linear programming algorithm by Buchinsky (1994). While most of these models are intended to be used for direct estimation, we use an alternative approach. We treat the problem of censored wages as a missing data problem and impute the censored wages using multiple imputation. The theory and principle of multiple imputation originates from Rubin (1978) and involves replacing each missing value by a number of imputed values yielding to m imputed data sets. This number may be rather small; usually $m = 5$ times can be regarded as an adequate number. Here, the goal is not to provide an estimation method that is applicable to get the estimates of interest for a particular research question, but to provide a complete data set that can be used by researcher to examine a variety of research questions. Once the data are imputed, these analyses can be performed applying standard methods and models. Therefore, multiple imputation has the advantage that analysts do not have to familiarize themselves with multiple imputation or other models for censored data. As the data can be analyzed like any complete data set, multiply imputed data create new potential for a wide range of research questions. Even research questions, for which no applicable models for the analysis of incomplete data exist, can be easily examined using multiply imputed data and standard estimation techniques.

Gartner (2005) proposes a non-Bayesian single imputation approach to solve the problem of censored wages in the IAB Employment Sample. As it will be discussed later, single imputation has some serious drawbacks. The main criticism is that single imputed data yield biased variance estimates making multiple imputation generally preferable (see, e.g., Little and Rubin (1987, 2002)). The main argument to impute missing values multiply is to be able to calculate correct variance estimates. Here, the uncertainty due to the imputation can be reflected in the final variance estimates by adding a correction term based on the variance between the results of the m different imputations.

A multiple imputation method for right-censored wages based on draws of a random variable from a truncated distribution and Markov chain Monte Carlo techniques is suggested by Gartner and Rässler (2005). Both approaches that are suggested in the literature to solve the censoring in the IABS assume ho-

moscedasticity of the residuals. But contrary to this assumption, the variance of income is usually smaller in lower wage categories than in higher categories, thus assuming homoscedasticity in an imputation model is highly questionable. This becomes evident if one thinks of the wage dispersion within education groups. While in lower groups, there is generally little wage inequality, wages of highly skilled employees, for example holding an university degree, may differ significantly. Therefore, in this thesis new imputation methods allowing for heteroscedasticity are suggested. In a first step a single imputation procedure is developed. Furthermore a new multiple imputation approach will be presented. First simulation studies show 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.

Whereas one goal of this thesis is to present new imputation approaches that are applicable for right-censored wages, a main objective will be also to confirm the validity of multiple imputation approaches in general and to show the superiority of the new approach considering heteroscedasticity in a wide range of situations. In a series of simulation studies different approaches are evaluated to confirm the quality of the multiply imputed data. Besides simulated data, uncensored wage information of the German Structure of Earnings Survey (GSES) 2001 is employed to assess the quality of imputation. Later, the external complete wage information is also used for the imputation model. The first reason to do so is to try to develop an even more robust imputation technique, the second is to have a benchmark for the proposed approaches, that work without external information.

The thesis is organized as follows. Chapter 2 gives an overview on German databases that are applicable to analyze research questions concerning wages. First, we distinguish between survey and register data. Second, the data stemming from the German Federal Employment Agency, including the IAB Employment Sample is presented and its potential for analyses discussed. In Chapter 3, the German social insurance system is briefly described in order to explain why censoring occurs in the IAB Employment Sample. This explanation is followed by some examples of other wage data affected by censoring in order to illustrate that the necessity of appropriate solutions to handle censored data is not restricted to the German data. On the contrary, the imputation

approaches addressed here are applicable to various surveys and other kind of data sets whose potential is hindered by censoring. Chapter 4 discusses censored models applicable to the analysis of various research questions. To assess the potential of multiply imputed wages in the IAB Employment Sample, Chapter 5 gives an overview of studies based on the wage data of the IABS. These studies are presented to illustrate the variety of analyses that are performed using the IABS and the multitude of techniques that are applied to handle the censoring. This overview shows that multiply imputed wages generate new potential in various fields. Beyond, by means of this overview one can easily see that multiple imputed data simplify the analysis of wages in the IAB Employment Sample. Before specific imputation approaches for right-censored wages are presented, Chapter 6 offers an introduction to multiple imputation in general. The chapter starts with the explanation of different missing-data mechanism, continues by exposing rather simple imputation approaches and finally addresses the theory of multiple imputation. Chapter 7 introduces imputation approaches for right-censored wages. This chapter starts with explaining approaches assuming homoscedasticity of the residuals and later presents new approaches considering heteroscedasticity. Chapter 8 to 10 evaluate these approaches and confirm the superiority of the new multiple imputation approach considering heteroscedasticity. Chapter 8 describes a series of simulation studies to compare the different approaches. The first two simulation studies are based on simulated wage data generated using the IABS, the following simulation studies are based on the German Structure of Earnings Survey, which contains uncensored information on wages. In Chapter 9, alternative approaches to the approach considering heteroscedasticity are suggested and evaluated. Finally Chapter 10 presents some real world examples. The first part of the chapter compares results of three research questions used as examples. Results based on original complete data, censored data, and multiply imputed complete data are compared to demonstrate once more the validity of imputed data. The second part reviews recent studies based on one of the imputation approaches, that were discussed in this thesis. The conclusion summarizes the main findings and gives an outlook towards future steps. These involve providing access to the proposed imputation algorithms and multiply imputed versions of the IAB Employment Sample to researchers both at the IAB and other research institutions.

Chapter 2

Wage Data

By definition, wage is the financial compensation a worker receives in exchange for his labor, hence it is a central element of the labor market and examining wages is a central issue in labor economics and labor market research. For that reason several data sources exist, that cover the broad range of different aspects related to the analysis of wages. This chapter gives an overview on this kind of data sources in Germany starting with survey data and followed by register or administrative data sets. Finally, the Chapter *'Wage Data'* introduces the register data of the German Federal Employment Agency, that are stored, edited and released to researchers at the Institute for Employment Research.

2.1 Wage Information in Surveys and Register Data in Germany

To address questions concerning wages, two types of data are usually used: surveys and process generated data, i.e., administrative data. In Germany, several data sources for both types of data exist. In order to be able to classify advantages and disadvantages of administrative data in general and the data of German Federal Employment Agency in particular, this section briefly describes the most important ones. Some of the data sets cover several sources of income and are not restricted to wages or labor earnings. Many report income at the individual and household level. As we are interested in data to analyze wages, we report here mainly surveys and administrative data that admit to analyze individual income from the labor market.

2.1.1 Surveys

As Lewis-Beck et al. explain "The social survey is a widely used method of collecting and analyzing social data for academic, government, and commercial research" (Lewis-Beck et al., 2004, p. 1102). Surveys are widely accepted as a means of collecting information about populations, but also face criticism due to some shortcomings. For, instance methods of collecting survey data may be subject to error due to sampling problems and flawed data collection instruments and methods. Especially the reliability of high wages is questionable in surveys. In a study examining consistency of income in 2002 across eight major U.S. surveys, Czajka and Denmead (2008) found out that a large percentage of yearly incomes is divisible by 5,000, suggesting that many respondents are rounding when reporting income. Nevertheless, we briefly describe the most important German surveys containing wage and income information.

German Socio-Economic Panel Study (GSOEP)

The German Socio-Economic Panel Study (GSOEP) is intended to offer microdata for research in the social and economic sciences. It is not restricted to the field of employment and wages, but includes as well information on other fields such as living conditions, values, or willingness to take risks. The GSOEP is not only used for basic academic research but also for policy-related social reports. It is conducted annually as a longitudinal study of private households since 1984 in West Germany and since 1990 in East Germany. In 1984, 5,921 households with 12,290 individual respondents participated in the 'SOEP West', in 2007 3,337 households with 5,963 respondents were still participating. In the 'SOEP East' sample, 2,179 house-holds with 4,453 members responded in the first year 1990; in 2007, 1,654 households and 3,067 individuals still participated. The GSOEP contains, apart from other sources of income like social security transfers, information on the gross and net monthly labor market income of all household members. Since 2002, a subsample of high income households which is selected independently from all other subsamples is added in order to oversample these households. Originally, the selection scheme required that the responding household had a monthly income of at least 7,500 DM (3,835 euros) to be relevant this subsample. From 2003, only households with a net monthly income of at least 4,500 euros were included. Further advantages of this survey are its panel design and the information on

the household context. Besides, it is referred to as the largest survey of foreigners and immigrants in Germany. As it is conducted as a survey the problems concerning the reliability of the wage information applies to this data set as well. More information on the survey and current results can be found in Headey and Holst (2008) or Haisken-DeNew and Frick (2005). A scientific use file is released by the research data center of the GSOEP at the German Institute for Economic Research (DIW) in Berlin.

Income and Expenditure Survey (IES)

The Income and Expenditure Survey (IES) is a data source applicable to the analysis of the different components of household income, income tax, welfare contributions and benefits received, savings, and the structure and development of household consumption. It has been conducted since 1962/63 in West Germany and since 1994 in East Germany. Since the wave of 1973, it is carried out every five years. About 0.2 percent of all households in Germany participate in each wave. The IES is a proportional sample as households are chosen according to a quota plan. The aim of this survey is to cover in-depth data on income and expenditure of private households. It is mainly used for income analysis, but provides information on a wider range of research fields such as the composition of households, participation in professional life, consumer goods consumption, wealth, level of assets and debt of private households and, as previously mentioned, type and level of income, including labor market earnings. A problem of this survey is that households with a monthly net income above 18,000 euros are not included because these data are considered as not statistically reliable. Another drawback is that foreign citizens in Germany are not sampled representatively. Hence, this data do not allow to study income of foreigners or to compare income of foreign and German citizens. The data can only be accessed by appointment with the Federal Statistical Office by members of independent German research institutions.

Microcensus

The Microcensus is an official survey conducted by the Federal Statistical Office and is intended to give a snapshot of the entire population by questioning one part of it. Its purpose is to provide statistical information on the economic and social situation of the population as well as on employment, the

labor market, and education in order to update the results of the population census. The Microcensus is a representative one percent random sample of all households in Germany, which are about 390,000 households with 830,000 persons in total, including about 150,000 persons in about 72,000 households in East Germany. It is carried out once a year since 1957 (Schwarz, 2001). Every household stays in the sample for four years and every year 25 percent of the included households are exchanged. All members of the household are interviewed, information for other household members is permitted only under specific premises. The details provided - especially those on employment - refer to a specific report week, normally the last week of April. Main topics of the Microcensus are sociodemographic characteristics (age, sex, nationality, etc.), economic and social situation of individual, household and family contexts, labor market status, questions on general and vocational level of qualification. It also contains information on income, but restricted to the total individual and household net income, including all sources of income. Another disadvantage is that income is asked in classes of 200 euros. A drawback of the Microcensus is that the access is restricted since it is not a voluntary survey. Therefore, the original data of Microcensus is de facto anonymized. In the form of a scientific use file, which contains an anonymized 70 percent sample of the 1 percent sample and just represents a cross-section, it can be obtained by German research institutions.

The German Structure of Earnings Survey (GSES)

The German Structure of Earnings Survey was conducted in 1990, 1992, 1995, and 2001 in establishments of the manufacturing industry and the service sector. For 2006 it reports wages from all sectors. The data for 2001 can be obtained as a scientific use file from the research data center of the German Statistical Office. All other years can only be accessed on-site. The German Structure of Earnings Survey is designed as 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). Since the collection of the GSES is

performed at the individual level, the latter provides a comprehensive data set to analyze possible merits to the workplace and personal characteristics. The GSES includes all employees covered by social insurance. The survey is conducted in establishments with at least 10 employees. Thus, the sample covers approximately 90 percent of all workers.
The survey is therefore suitable to examine a broad range of questions concerning wages. For more details see Forschungsdatenzentrum der Statistischen Landesämter (2006). This survey will play an important role later, when we perform simulation studies to compare different imputation approaches for censored wages. As the structure of this survey is very similar to the variables in the IAB Employment Sample and as it contains uncensored wage information for all employees it is especially appropriate to evaluate the performance of imputation approaches.

Further Surveys

Apart from these surveys several other surveys include questions on earnings on income. One example is the German General Social Survey (ALLBUS/GGSS), which is similar to the American General Social Survey (GSS). Its intention is to collect and disseminate high quality information on attitudes, behavior, and social structure in Germany. Since 2004, the European Union Statistics on Income and Living Conditions (EU-SILC) is conducted in 13 member states of the European Union and includes questions on income as well. In the German wage literature these surveys do not play an important role compared to the surveys discussed previously.

2.1.2 Register Data

Register data, also called administrative or process-generated data, have several advantages, like a large number of observations, no nonresponse burden and no problems with interviewer effects or survey bias. Especially when data are collected for official reasons, for example for taxation or for calculating unemployment benefits, there is a high interest and relevance for all involved persons to report accurate information and generate correct data. This applies especially to wages and other sources of income, for which reason register data are especially suitable to address questions concerning wages and earnings. Sometimes, e.g., in the German social insurance, some additional variables

are asked to the employers concerning job classification, education, nationality or other characteristics of their employees, which increase the value of an administrative data for research issues. One shortcoming of this additional information may be that it is not of primary interest to calculate contributions and benefits, but only asked for statistical reasons. If information is collected for statistical reasons only, it may not be as reliable as those variables collected for the official process. A further advantage of register data is the almost complete absence of panel mortality.

Wage and Income Tax Statistics

The German Wage and Income Tax Statistics report detailed information on all persons liable to income tax as well as on the amount, distribution, and taxation of their income with liability to taxation. Its primary aim is to assist political and fiscal decisions and to allocate tax revenues to the states ('Länder') and communities, but it is also distributed as a public use file and a scientific use file through the research data centers of the German Statistical Office. It is conducted every three years as a secondary statistic from the taxation records of the state revenue authorities. Public and scientific use files are currently available for the years 1992, 1995, 1998, and 2001. The Wage and Income Tax Statistics are a census with about 30 million records, comprising up to 400 variables on about 40 million persons and therefore are the largest secondary statistic on income in Germany (Merz et al., 2005). The data contain information, for example, on taxable wages and income, income tax, social transfer income, but also on socio-demographic characteristics like sex, age, religion, children, location, industry or profession of the tax payers. More details on this data source can be found in Statistische Ämter des Bundes und der Länder (2009)(only partly in English). Comprising a large number of items, the German Wage and Income Tax Statistics represent an applicable data set for a broad range of research questions. This involves not only fiscal questions, but also questions related to the income distribution. A main advantage is that it covers also recipients of high incomes in a very accurate way as it is based on the records of the revenue authorities. Another advantage is that not only the wages of employees can be examined, but also the income of self-employed. Serious drawbacks of this data source are that it is conducted only every three years and that different years can not easily be compared

due to frequent changes in the income tax law. Hence, it is mainly useful for regional comparisons.

Further Register Data

Further administrative data in Germany containing wage information are for example the Social Welfare Statistics and the Housing Allowance Statistics. Moreover, the branches of the German social security insurance system record administrative data to be able to satisfy their duties. Some of these data are edited and released for researchers. One of these administrative data are the data of the German Federal Employment Agency, which stem from the employment notifications of employers to the employment agency. Edited data sets based on these notifications are provided by the Research Data Center of the German Federal Employment Agency, which is located at the Institute for Employment Research (IAB), the research institute of the Federal Employment Agency. These data and their advantages and disadvantages are discussed in detail in the next section.

2.2 Register Data of the German Federal Employment Agency at the Institute for Employment Research

The Institute for Employment Research provides via its Research Data Center data on individuals, households, and establishments, as well as data that comprise both establishment and personal information. Some of the data come from surveys like for example the IAB Establishment Panel or the panel study 'Labour Market and Social Security' (PASS). Most of the data are process generated and originate from two different sources: One part of the data are collected in the notification process of the social security system, the other part comes from the internal procedures of the Federal Employment Agency for computer-aided benefit allowance, job placement, and administration of employment and training measures.
The IAB files the social security notifications and provides these data in the form of a history data set known as the Employment History (BeH). Another database, the Benefit Recipient History (LeH), originates from the internal

data processing modules of the Federal Employment Agency. These databases, BeH and LeH, are linked to form the Employee and Benefit Recipient History (BLH), from which several specific samples are generated:

- The Establishment History Panel (BHP) which is an aggregation of the BLH to the establishment level.
- The linked employer-employee data of the IAB (LIAB) that are formed by matching data from the BLH with the IAB Establishment Panel.
- The Integrated Employment Biographies sample of the IAB (IEBS), generated by matching spells of Employment History (BeH), the Benefit Recipient History (LeH), participants in measures and the applicants pool.
- The IAB Employment Samples (IABS) which are drawn from the Employee and Benefit Recipient History (BLH).

While most of the administrative or process-generated data of the Institute for Employment Research can be accessed only by internal researchers or onsite at the Research Data Center of the Federal Employment Agency at the Institute for Employment Research, the IABS is also provided in several versions as a scientific use file. It is therefore an important database for many studies of economic interest concerning the German labor market conducted by researchers of the Institute for Employment Research as well as by external researchers.[1] All data sets that are based (or partly based) on the Employment History (BeH) coming from the social security notifications contain information on wages. In principle the problem of censoring occurs in all these administrative data sets based on these notifications that contain wage information. Even if all proposed imputation procedures are applicable for all administrative data sets provided by the Institute of Employment Research, due to its importance for all kind of researchers, in the following, the focus will be on the IAB Employment Samples.

[1]More details on the data sets and on the ways to access them can be found at the website of Research Data Centre of the Federal Employment Agency at the Institute for Employment Research (http://fdz.iab.de/en.aspx).

2.3 The IAB Employment Sample (IABS)

As mentioned, the German IAB Employment Samples (IABS) are random samples drawn from the IAB Employee History with additional information on benefit recipients and hence are samples of all employees covered by social security. Consequently, self-employed, family workers, and civil servants are not included and therefore the data represent approximately 80 percent of all employees in Germany (see Bender et al. (2000)). Since 1999, also marginal employment ('Geringfügige Beschäftigung') with earnings of 400 euros or less per month, which is not fully liable to social insurance, is included. The IAB Employment Samples comprise a continuous flow of data on employment subject to social security as well as on receipt of unemployment benefits, unemployment assistance, and maintenance allowance, and contain additionally a number of establishment characteristics. Key variables are for example:

- gender
- age
- nationality
- marital status
- number of children
- school education and professional qualifications
- type of employment (especially differentiation between employment covered by social security and marginal employment)
- person group
- gross earnings subject to social security
- profession
- occupational status (including full or part-time employment)
- start and end date of employment
- industry

The different versions of the IAB Employment Sample

	Basic file 75-95	**Regional file 75-97**	**Regional file 75-04**	**Weakly anonymized version 1975-04**
Description	1% random sample Employees covered by social security, benefit recipients 559,540 individuals	1% random sample Employees covered by social security, benefit recipients 588,827 individuals	2 % random sample Employees covered by social security (including marginal employment since 1999), benefit recipients 1,360,949 individuals	2 % random sample Employees covered by social security (including marginal employment since 1999), benefit recipients 1,360,949 individuals
Period covered	West: 1.1.1975 to 31.12.1995 East: 1.1.1992 to 31.12.1995	West: 1.1.1975 to 31.12.1997 East: 1.1.1992 to 31.12.1997	West: 1.1.1975 to 31.12.2004 East: 1.1.1992 to 31.12.2004	West: 1.1.1975 to 31.12.2004 East: 1.1.1992 to 31.12.2004
Time reference	Employment biographies on a day-to-day basis	Employment biographies on a day-to-day basis	Employment biographies on a day-to-day basis	Employment biographies on a day-to-day basis
Regional structure	West/ East Germany	West/East Germany, federal states (Bundeslaender), 348 regions	West/East Germany, federal states (Bundeslaender), 348 regions	Employment agency (Arbeitsagentur), districts (Kreis)
Topics	Socio-demographic characteristics Employment-related characteristics Benefit-related characteristics Detailed occupation and industry classification	Socio-demographic characteristics Employment-related characteristics Benefit-related characteristics Aggregation of occupation and industry classifications	Socio-demographic characteristics Employment-related characteristics Benefit-related characteristics Aggregation of occupation and industry classifications	Socio-demographic characteristics Employment-related characteristics Benefit-related characteristics Detailed occupation and industry classification
Access	Scientific Use File	Scientific Use File	Scientific Use File	On-site use and remote data access

Table 2.1: Versions of the IAB Employment Sample

- establishment location
- establishment size

The IABS is provided as a scientific use file in three versions and one weakly anonymized version that can only be accessed on-site and subsequently by remote data access. Table 2.1 gives an overview over of these four different versions. The main difference between the versions is the anonymization process. In the scientific use files some variables are aggregated in order to prevent the identification of individuals. In the basic file 75-95 the regional variable is highly aggregated and allows only to separate between East and West Germany. In the regional aggregation anonymization concerns the industry and occupation variables. The weakly anonymized version is not aggregated. Because the samples are drawn from the longitudinal processed database of employment notifications, all version contain not only cross-sectional information, but represent panel data. A detailed description of the employment sample can be found in Drews (2007, 2008) or Schönberg (2009). In the following chapters, the weakly anonymized version will be considered as the IAB Employment Sample.

Originating from the employer notifications, the IABS has one big advantage such that it covers all employees subject to social security in Germany for a long time period. It contains very reliable information on a broad range of variables and therefore is optimally qualified for the analysis of various research questions. The main advantage for wage analysis is that information on the employment history and especially wages is measured more precisely than in surveys like the GSES or GSOEP.

One important disadvantage is caused by the contribution limit of the German social security system. The contribution rate of the insurance is charged as a percentage of the gross wage. If the gross wage is higher than the current contribution limit only the amount of the ceiling is liable for the contribution. In 2010, the contribution limit in the unemployment insurance system is fixed in West Germany at a monthly income of 5,500 euros. Therefore as wages are only recorded up to the contribution limit, the wage information in this sample is censored at this limit. To illustrate this problem, Figure 2.1 shows the distribution of wages in the IAB Employment Sample in 2000. To be able to analyze wages based on this data set and to be able to access the whole potential of the data, one has to find appropriate techniques that yield

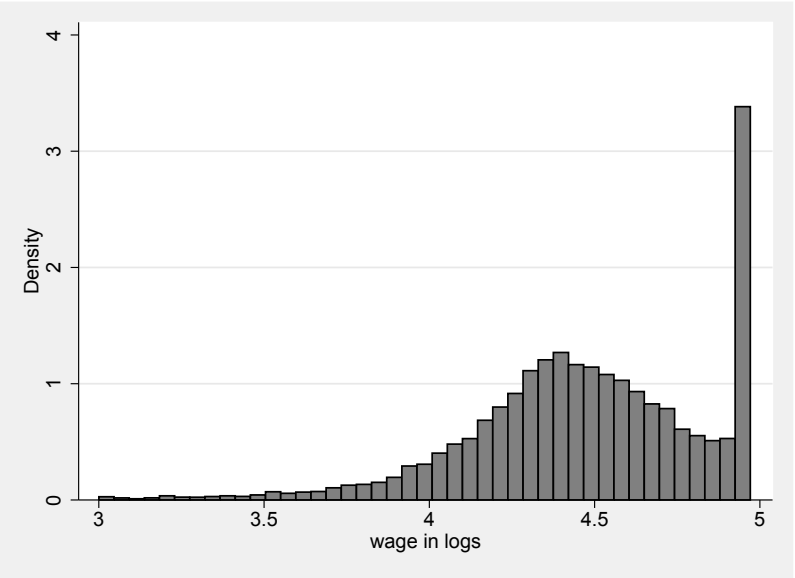

Figure 2.1: Distribution of daily wages in logs in the IAB Employment Sample (IABS) in West Germany 2000.

unbiased results in the case of censoring. The censoring and its impacts are discussed in detail in the next chapter.

Chapter 3

Censoring in Wage Data

Many data sets collected by economists or social scientists are incomplete in some way for different reasons. Two specific cases of incomplete data are truncation and censoring. Truncation occurs if all observations for both the dependent and explanatory variables lying outside some range are completely missing. An important example in the literature is the New Jersey negative income tax experiment. Only families with incomes lower than 1.5 times the 1967 poverty were included in this study, families with higher incomes were not selected (see, e.g., Robins (1985) or Maddala (2001)). Therefore, the data contain no information at all for these families. If we have information on the explanatory variable for all individuals, but the dependent variable is missing for some individuals, censoring occurs. In the case of censoring the distribution of the data on the dependent variable is cut off outside of some range. Therefore we observe multiple observations at the endpoint of that range. The advantage in this case is that we know the number of (missing) observations and the value for all explanatory variables, even if the dependent variable is censored. Li and Racine describe the situation of censoring as follows: "Strictly speaking, a sample has been censored if no observations have been systematically excluded, but some of the information has been suppressed. Envision a censor who reads your mail and blacks out part of it - you still get your mail, although some parts of it are illegible" (Li and Racine, 2007, p. 331). Many examples in the empirical literature deal with dependent variables that are zero for a significant fraction of the observations. In this case conventional regression methods are not able to account for the difference between limit (zero) and nonlimit (continuous) observations. The first important studies dealing with this kind

of problem are Tobin (1958) examining household purchases of durable goods or Fair (1977, 1978) analyzing the number of extramarital affairs.
In the situation of the register data of the German Federal Employment Agency coming from the German social security insurance system we find a censoring of wages, one of the key variables for many research questions of economic interest. The wages are not left-censored at zero as discussed in the examples above, but we observe data censoring at the right. The following chapter describes this situation in detail.

3.1 The German Social Insurance System

In Germany nearly 90 percent[1] of the population is covered by either compulsory or voluntary social insurance (Deutsche Sozialversicherung, 2009) which consists of five branches. The social insurance schemes are primarily financed through contributions paid by employees and employers. The branches of social insurance include:

- Statutory unemployment insurance: insures employees' livelihood in case of unemployment,
- Statutory pension insurance: insures aged members and cases of reduced earning capacity. Upon an employees' death, it insures his or her survivors as well,
- Statutory health insurance: supports maintenance and restoration of good health and eases the financial consequences of illness,
- Statutory accident insurance: helps an employee regain his earning ability after a (work-related) accident,
- Statutory long-term care insurance: provides financial support for those dependent on care and assistance from others.

The social insurance funds are generally financed equally by contributions from insured fund members and their employers. Contributions are calculated as

[1] Even if only 80 percent of employees are covered by social insurance, nearly 90 percent of the population are covered by social security, because children are insured without contribution if at least one parent is covered and families are over-represented as insurants in the social security system.

percentage of the gross wage, but only up to a contribution limit. For higher earnings the contribution rate remains the same. As the exact wage is not needed to calculate the contribution, wages are in those cases only recorded up to this limit and are consequently censored on the IAB Employment Sample. The level of the contribution limit differs from branch to branch. Decisive for the extent of censoring in the IAB Employment Sample are the limits in the unemployment and pension insurance branches, which are identical. The ceilings of the unemployment and pension branch are decisive, because these insurances have the highest ceilings. The relevant limits are shown in the following section for the years 1975 to 2010. Constantly updated figures can be found in Deutsche Rentenversicherung (2010).

3.2 Contribution Limits and Censoring

The contribution limits are constantly adjusted, typically every year. Table 3.1 and Table 3.2 show the upper contribution limits in the statutory pension insurance of workers and employees for West Germany from 1975 and for East Germany from 1990, the year of the reunification. Until 2001, the ceilings are shown in German mark (DM), since 2002 in euros (€)[2]. In 2010, the current contribution limit in West Germany is fixed at a yearly wage of 66,000 euros and a monthly wage of 5,500 euros. In East Germany it is fixed at a yearly wage of 55,800 euros and a monthly wage of 4,650 euros. Daily values were calculated by division of the yearly values by the number of calendar days (i.e., 365, 366 in leap years).

An exception is the statutory pension insurance for miners, where the contribution limits are higher. For 2010, it is fixed for West Germany at 81,600 euros per year and for East Germany at 68,400 euros. This additional contribution limit is relevant in only very few cases, which are difficult to identify. Because these cases cannot be distinguished from misreporting of wages that are higher than the actual contribution limit, these special cases are normally disregarded. Instead the limits of the pension insurance of workers and employees are used for all cases. Misreported wages and contributions liable to the miners insurance are accordingly cut off at this ceiling.

The wage is reported by the employer for the entire period of employment in

[2]The relation of the German mark to the euro is officially fixed at 1.95583.

Contribution Limits West Germany

	Upper earnings limits		
	Year	Month	Day
	DM	DM	DM
1.1. to 31.12.1975	33,600	2,800	92.05
1.1. to 31.12.1976	37,200	3,100	101.64
1.1. to 31.12.1977	40,800	3,400	111.78
1.1. to 31.12.1978	44,400	3,700	121.64
1.1. to 31.12.1979	48,000	4,000	131.51
1.1. to 31.12.1980	50,400	4,200	137.70
1.1. to 31.12.1981	52,800	4,400	144.66
1.1. to 31.12.1982	56,400	4,700	154.52
1.1. to 31.12.1983	60,000	5,000	164.38
1.1. to 31.12.1984	62,400	5,200	170.49
1.1. to 31.12.1985	64,800	5,400	177.53
1.1. to 31.12.1986	67,200	5,600	184.11
1.1. to 31.12.1987	68,400	5,700	187.40
1.1. to 31.12.1988	72,000	6,000	196.72
1.1. to 31.12.1989	73,200	6,100	200.55
1.1. to 31.12.1990	75,600	6,300	207.12
1.1. to 31.12.1991	78,000	6,500	213.70
1.1. to 31.12.1992	81,600	6,800	222.95
1.1. to 31.12.1993	86,400	7,200	236.71
1.1. to 31.12.1994	91,200	7,600	249.86
1.1. to 31.12.1995	93,600	7,800	256.44
1.1. to 31.12.1996	96,000	8,000	262.30
1.1. to 31.12.1997	98,400	8,200	269.59
1.1. to 31.12.1998	100,800	8,400	276.16
1.1. to 31.12.1999	102,000	8,500	279.45
1.1. to 31.12.2000	103,200	8,600	281.97
1.1. to 31.12.2001	104,400	8,700	286.03
	€	€	€
1.1. to 31.12.2002	54,000	4,500	147.95
1.1. to 31.3.2003	61,200	5,100	167.67
1.4. to 31.12.2003	61,200	5,100	167.67
1.1. to 31.12.2004	61,800	5,150	168.85
1.1. to 31.12.2005	62,400	5,200	170.96
1.1. to 31.12.2006	63,000	5,250	172.60
1.1. to 31.12.2007	63,000	5,250	172.60
1.1. to 31.12.2008	63,600	5,300	173.77
1.1. to 31.12.2009	64,800	5,400	177.53
since 1.1.2010	66,000	5,500	180.82

Table 3.1: Contribution limits West Germany

Contribution Limits East Germany

	Upper earnings limits		
	Year	Month	Day
	DM	DM	DM
1.7. to 31.12.1990	32,400	2,700	88.77
1.1. to 30.6.1991	36,000	3,000	98.63
1.7. to 31.12.1991	40,800	3,400	111.78
1.1. to 31.12.1992	57,600	4,800	157.38
1.1. to 31.12.1993	63,600	5,300	174.25
1.1. to 31.12.1994	70,800	5,900	193.97
1.1. to 31.12.1995	76,800	6,400	210.41
1.1. to 31.12.1996	81,600	6,800	222.95
1.1. to 31.12.1997	85,200	7,100	233.42
1.1. to 31.12.1998	84,000	7,000	230.14
1.1. to 31.03.1999	86,400	7,200	236.71
1.4. to 31.12.1999	86,400	7,200	236.71
1.1. to 31.12.2000	85,200	7,100	232.79
1.1. to 31.12.2001	87,600	7,300	240.00
	€	€	€
1.1. to 31.12.2002	45,000	3,750	123.29
1.1. to 31.12.2003	51,000	4,250	139.73
1.1. to 31.12.2004	52,200	4,350	142.62
1.1. to 31.12.2005	52,800	4,400	144.66
1.1. to 31.12.2006	52,800	4,400	144.66
1.1. to 31.12.2007	54,600	4,550	149.59
1.1. to 31.12.2008	54,000	4,500	147.54
1.1. to 31.12.2009	54,600	4,550	149.59
since 1.1.2010	55,800	4,650	152.88

Table 3.2: Contribution limits East Germany

one year. If the person is employed the whole year, the reporting refers to the entire year, if the employment is shorter, to the period the person was employed within the current year (of course several periods of employment within one year are possible). If the wage for the reported period exceeds the income threshold, it will be censored. In this case, the employer reports only the amount up to the ceiling in accordance with reporting rules. In some cases the reported earnings may lie above the income threshold as since 1984 employers have to include special payments for the year in the notifications and add them to the wage. As the wage refers to the entire period of employment, the daily wage as it can finally be found in the IABS represents an average daily wage over the reported period. This information is important because the wage may vary over the year for example if there is a raise of salary. In rare special cases the average daily wage may be biased due to a change from an uncensored wage to a censored wage during the reporting period. Misreporting of wages due to other cases than described above on the other hand is very unlikely, even if erroneous messages can never be prevented completely. But since the notifications are relevant to calculate security allowances, however, the error rate can be expected to be rather small. An additional problem with the wage information is that the change of the reporting system in 1984 (inclusion of bonus payments) leads to a structural break.

Because the data contain all employment spells of the persons included in the sample, for every individual several independent spells may be observed in one year. Therefore, researcher usually create cross-sections in every year for a reference date, e.g., June 30. Then, the average wage for the particular year, is the average wage of the employment spell that covers the reference date.

	<25	25-34	35-44	45-54	55+
Low/intermed. school	0	.003	.008	.012	.17
Vocational training	.001	.021	.068	.116	.150
Upper school	.010	.110	.232	.331	.371
Upper school and vocational training	.003	.110	.283	.393	.470
Technical college	.024	.190	.450	.558	.604
University degree	.056	.256	.549	.686	.769

Table 3.3: Fractions of censored wages in the IAB Employment Sample (Males in West Germany)

To illustrate the problem of censoring, Table 3.3 shows descriptive information about the fraction of censored incomes of six educational and five age groups among male West German residents holding a full-time job covered by social security on June 30th 2000. The figures show the necessity to impute the missing wage information (or adjust for missingness in a different way) in order to obtain unbiased results. While, in total, 11 percent of all employees have censored wage observations, in some subgroups the fraction of missing wages may be much higher. Especially for analyzing high-skilled employees (with technical college degree or university degree), the table clearly indicates the necessity to correct for the censoring, best to impute.

3.3 Censored Wage Data in Other Countries

The problem of censored wage or income variables is not only known with the German IAB data, but is a common problem in several data sets. These problems originate not necessarily from a contribution limit in the social security system. Most researchers are familiar with the top-coding of income variables in the U.S. March Current Population Survey (CPS) conducted by the Census Bureau. In the CPS censoring is used as a measure to ensure confidentiality. In Austria on the other hand, where a social security insurance system similar to the German exists, wages recorded in order to release administrative data sets of economic interest are censored due to a contribution limit as well.

3.3.1 U.S. Current Population Survey (CPS)

The U.S. Current Population Survey is a survey conducted by the United States Census Bureau. It is a representative sample of all households in the United States and is collected since 1942 by the U.S. Census Bureau. It is the primary data source used by public policy researchers and administrators to investigate yearly trends in average income and its distribution in the United States (Larrimore et al., 2008). It is also used by the Bureau of Labor Statistics to monthly report the employment situation and contains, among others, questions on the employment status and on weekly and hourly earnings. In every month of March it contains additional questions on income in the previous calendar year. Unlike the IAB Employment Sample, the Current Population Survey comprises not only one source of income but, starting in 1975, 11

sources and since 1987 24 sources of income are recorded (Burkhauser et al., 2008). In the case of the Current Population Survey, wages and other sources of income are not censored due to the process of collecting the data, as in the case of the IAB Employment Sample. Since the CPS is conducted as a survey, high values are not censored, because they are not asked or not reported, but are topcoded before publishing the data as a public use file in order to ensure the confidentiality of the respondents. To protect the confidentiality of its respondents the Census Bureau topcodes the highest values from each source of income that it collects (Burkhauser and Larrimore, 2008). In the public use file, the highest values are topcoded for each source of household income, not simply the high total household income values. One drawback of this proceeding is that it complicates the aggregation of multiple income sources to the total household income, because each of the sources may be topcoded. Another problem is that the topcode values are inconsistently defined over years. Therefore, the proportion of individuals with topcoded household income in each CPS ranges between 2.1 percent and 5.7 percent over the period from 1995 to 2005 (Jenkins et al., 2009), which leads to artificial increases and decreases in mean income. This drawback is to some extent reduced since the introduction of cell means which are provided since 1995 based on the internal data. Until 1994, the topcode value defined for the specific source of income was assigned to all observations above this value. Since 1995, all high values in the public use data are substituted by a cell mean value derived from the internal data (Burkhauser and Larrimore, 2008; Burkhauser et al., 2008). The introduction cannot solve the problem of topcoding completely as the internal data are themselves censored, even if to a lesser degree. Initially the internal data were censored due to data-storage limitations in the computing systems of the 1970s. Therefore, written records were truncated to 5 digits (Burkhauser et al., 2008). Even if these storage limitations are not a constraint anymore, the Census Bureau continues this censoring practice. In 1985, values higher than 250,000 U.S. dollars in each source of income were still censored, mainly due to concerns about data reliability of individuals who report an extremely high income. From then the limits were increased constantly to keep the percentage of censored individuals in the internal data below 1 percent. Burkhauser et al. (2008) also mention that despite the Census Bureau's attempt to alleviate the problem of topcoding, their cell means have generally been ignored by researchers, since time-inconsistencies arise from using unadjusted public use

data for 1995 and before and CPS data with imputed cell means from 1996. Some solutions that are used to analyze the CPS public use data - even if there inconsistencies between different years (apart from using cell means) - are for example measuring inequality with the ratio between the 90th and the 10th percentile of the wage distribution or artificially truncating the data by removing the highest and lowest two percent of observations. Another method is to artificially lower the topcodes in the data for each year to create a series with constant percentage of people with topcoded data in each year, which is referred to as the 'consistent topcoding method'. This method is intended to solve at least the problem of inconsistent censoring points over the years. All these solutions have their drawbacks, but are preferable to using unadjusted data. More sophisticated approaches to handle the presence of censoring, including multiple imputation, will be discussed later.

3.3.2 U.S. Social Security Administration Earnings Records (SSA)

The problem of censoring appears also in another U.S. database, where the reason of the censoring is similar to the German data. The Social Security Administration (SSA) collects data on social security earnings coming from the social security tax records. One shortcoming of the SSA earnings data is that many records are censored at the maximum taxable earnings level of the social security. At least 32 percent of the observations in the sample are censored, with a maximum of censored records reaching more than 50 percent in some years (Chay and Powell, 2001). In this database not only censoring from the right, but also from the left occurs, as the SSA earnings data also contain records that are censored at zero. This situation occurs with individuals earning a rather low income, which is not in the taxed sector, or individuals out of labor force or experiencing a year-long spell of unemployment. According to Chay and Honoré (1998) about 15 percent of the sample has no earnings in the covered sector. The SSA administrative records supply accurate information on income but compared to a survey lack demographic information (Fisher, 2007). Therefore, in a joint project of the Census Bureau and the Social Security Administration, respondents to the March Current Population Survey were matched to the SSA earnings history for some years using the social security number (Chay and Honoré, 1998).

3.3.3 Austrian Social Security Database (ASSD)

In Austria data for all workers, called the Austrian Social Security Database (ASSD), are collected based on the employer notifications by the social security authority. It is provided as a matched firm-worker data set, containing the labor market history of almost 11 million individuals and 2,2 million firms. The data contain information on all workers, except for self-employed, civil servants, and marginal workers. These data cover longitudinal (earnings and employment) information necessary to assess the pension benefits and are provided by the Austrian Social Security Agency ('Hauptverband der österreichischen Sozialversicherungträger'). The data set comprises the individual's detailed employment and earnings history, a worker's (anonymized) social security number, and a limited set of socio-demographic characteristics (such as age, sex, and broad occupation). The ASSD covers all employees in the private sector in Austria from January 1972. As in Germany, the data are mainly collected for reasons of social security insurance. Hence, the wage information is censored due to contribution ceiling in the social insurance system, which resembles the German system. In 2007, the ceiling was fixed at 53,760 euros per year. Following Hofer and Weber (2002), the data set contains at most 15 percent censored wage observations per year. For further details on the Austrian data see Humer et al. (2007) or Zweimüller et al. (2009).

Chapter 4

Modeling Censored Data

Applying standard estimation methods to censored data leads to seriously biased estimations results. Chay and Powell (2001), for example, show that the results of an ordinary least squares (OLS) analysis of censored data implies only little decrease of the earnings gap between black and white workers in the United States in the 1960s. On the other hand, estimates from models that account for censoring suggest that the difference between earnings of black and white decreased significantly after 1964. In case of censoring, the mean of a censored dependent variable in the observed data differs from the actual mean of that variable, which cannot be observed due to the censoring. Consequently, the variation of the dependent variable in the observed data will understate the true variation and the application of ordinary least squares methods (or other classical methods) will, in general, yield parameter estimates that are biased towards zero (see, e.g, Li and Racine (2007)). Figure 4.1 illustrates how OLS estimation based on a right-censored dependent variable tends to underestimate the slope of the regression line and accordingly the parameter estimates as well if the observations above $y = a$ are omitted. If all observations above the ceiling are set to the value of this ceiling, it is also obvious that the regression line will be shifted towards zero. Burkhauser et al. (2008) show in a study analyzing the income inequality in the United States over three decades (1975-2004), that using unadjusted topcoded wage information from the public use data of the CPS leads to lower estimated levels of inequality and potentially affects estimates of trends over time. In the study they compare estimation results from different methods for addressing topcoding (e.g., using the cell mean imputation series) and from different versions of the CPS, including the

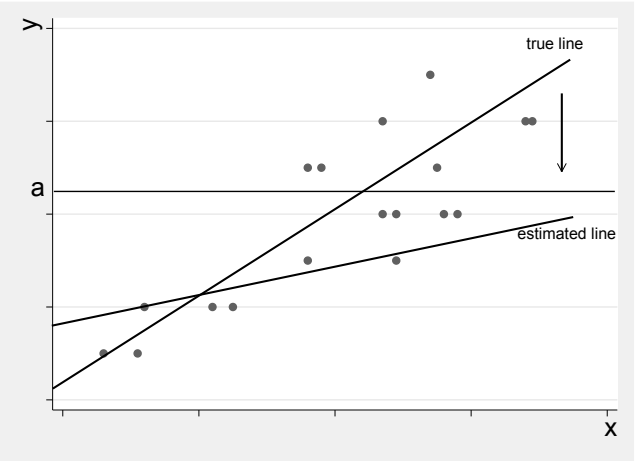

Figure 4.1: Bias of estimation based on censored data

internal version. Further examples that illustrate the problem of estimation results biased towards zero can be found in Greene (2008).
In the last decades a large number of innovative approaches has been proposed to handle the presence of censoring. Before multiple imputation is considered in detail as an approach to deal with censoring and the advantages are discussed, this section gives an overview on parametric, semiparametric, and nonparametric approaches, that have been suggested in the literature. The discussion of these methods is followed by an overview on quantile regression methods that are applicable to censored data.

4.1 Parametric Approaches

Parametric approaches provide an adjustment mechanism that overcomes the bias that would arise from the direct application of standard methods, like for example ordinary least squares, in the presence of censoring. The regression model for censored data is referred to as the censored regression model or the tobit model. The model was first proposed by Tobin (1958) and usually is described for the case of left-censoring at zero, but can be easily adapted for presence of right-censoring. Detailed instructions on implementing this model can be found for instance in Greene (2008) or Li and Racine (2007).
The starting point for the model is that there is a latent variable y^* which cannot be observed in some cases (in this example all cases with $y^* < 0$) even

if the covariates in x are observable. In a truncated distribution, only the part of the distribution above $y^* = 0$ would be relevant, as we would confine our attention only to the observed observations. When data are censored, the distribution is a mixture of discrete and continuous distributions. To analyze this distribution, we define a new random variable y transformed from the original one, y^*. We consider the 'latent variable model' given by

$$y_i^* = x_i^{'}\beta + \varepsilon_i, \quad i = 1, ..., n \tag{4.1}$$

where β is a vector of parameters, x is a vector of observed explanatory variables and ε_i is a mean zero disturbance term with $\varepsilon_i \sim N(0, \sigma^2)$. As we handle a censored variable, we do not observe y_i^*, rather we observe y_i given by

$$y_i = \begin{cases} y_i^* \text{ if } y_i^* > 0 \\ 0 \text{ if } y_i^* \leq 0 \end{cases} \tag{4.2}$$

where $i = 1, ..., n$. It is obvious that estimating the parameter β by regressing the observed y_i on x_i, the resulting ordinary least squares estimator is biased and inconsistent.
Censored regression models are generally applicable for three situations:

- left-censoring at a non-zero limit
- left-censoring at zero ('corner solution outcome')
- right-censoring at a non-zero limit

The three different situation are now described in detail starting with left-censoring at a non-zero limit. We can estimate β and σ^2 for all three situations by maximum likelihood. For this, we need the density of the uncensored observations, which is the same as that for y_i^*,

$$f(y_i) = f_N(y_i; \mu = x_i'\beta, \sigma^2) = \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\frac{1}{\sigma} \quad \text{if } y_i > a, \tag{4.3}$$

For the censored observations, we need the probability that y_i equals the censoring value a, given x_i,

$$f(y_i) = P(y_i^* \leq a) = \Phi\left(\frac{a - x_i'\beta}{\sigma}\right) \quad \text{if } y_i = a. \tag{4.4}$$

These two parts can be combined to obtain the density of y_i^*, given x_i and a. The censored regression model is incorporated in all important software packages and can be estimated via maximum likelihood. The maximum likelihood function is given by

$$L(\beta, \sigma^2) = \prod_{y_i > a} \frac{1}{\sqrt{2\pi\sigma^2}} \; e^{-\frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2} \prod_{y_i = a} \Phi\left(\frac{a - x_i'\beta}{\sigma}\right). \tag{4.5}$$

$\Phi(.)$ is the standard normal cumulative distribution function and $\phi(.)$ is the standard normal density function. For an observation randomly drawn from the population,

$$\begin{aligned} E(y_i|x_i) &= a \cdot P(y_i = a) + E(y_i > a) \cdot P(y_i > a) \\ &= a \cdot \Phi(a^*) + (x_i'\beta + \sigma\lambda(a^*))(1 - \Phi(a^*)) \\ \text{where} \quad \lambda(a^*) &= \frac{\phi(a^*)}{1 - \Phi(a^*)} \quad \text{and} \quad a^* = \frac{a - x_i'\beta}{\sigma}. \end{aligned} \tag{4.6}$$

The log-likelihood for observation i can be obtained by taking the natural log of the density of each i. Then, the log-likelihood for the censored regression model for left-censoring at a is

$$\begin{aligned} ln\, L(\beta, \sigma^2) &= \sum_{y_i > a} -\frac{1}{2}\left(ln(2\pi) + ln\, \sigma^2 + \frac{(y_i - x_i'\beta)^2}{\sigma^2}\right) \\ &+ \sum_{y_i = a} ln\left(\Phi\left(\frac{a - x_i'\beta}{\sigma}\right)\right). \end{aligned} \tag{4.7}$$

The first part corresponds to the classical regression for the nonlimit observations and the second part to the relevant probabilities for the limit observations. The likelihood is a nonstandard type because it represents a mixture of discrete and continuous distributions. Amemiya (1973) proves that this likelihood estimator suggested by Tobin for this model is consistent and maximizing log L produces an estimator with all desirable properties attained by maximum likelihood estimation.

This general censored regression for left-censoring can be applied the special case of censoring at zero as well. Although censored regression models often generally are referred to as tobit models, the following model originally describes the tobit model. We assume that a variable is zero for a significant

part of the population, but many different positive outcomes can be observed for the rest of the population. Then, the density of the uncensored observations is

$$f(y_i) = \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\frac{1}{\sigma} \quad \text{if } y_i > 0, \tag{4.8}$$

and the probability that y_i equals the censoring value 0 is

$$f(y_i) = P(y_i^* \leq 0) = \Phi\left(\frac{0 - x_i'\beta}{\sigma}\right) \quad \text{if } y_i = 0. \tag{4.9}$$

Combing the two parts again yields the likelihood function

$$\begin{aligned} L(\beta, \sigma^2) &= \prod_{y_i>0} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2} \prod_{y_i=0} \Phi\left(\frac{-x_i'\beta}{\sigma}\right) \\ &= \prod_{y_i>0} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2} \prod_{y_i=0} \left(1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right) \end{aligned} \tag{4.10}$$

and the expected value of y_i given x_i

$$\begin{aligned} E(y_i|x_i) &= E(y_i > 0) \cdot P(y_i > 0) \\ &= (x_i'\beta + \sigma\lambda(a^*))\left(1 - \Phi\left(\frac{-x_i'\beta}{\sigma}\right)\right) \\ &= (x_i'\beta + \sigma\lambda(a^*))\Phi\left(\frac{x_i'\beta}{\sigma}\right) \\ \text{where} \quad \lambda(a^*) &= \frac{\phi\left(\frac{-x_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{-x_i'\beta}{\sigma}\right)} = \frac{\phi\left(\frac{x_i'\beta}{\sigma}\right)}{\Phi\left(\frac{x_i'\beta}{\sigma}\right)}. \end{aligned} \tag{4.11}$$

The log-likelihood for the tobit model is given by

$$\begin{aligned} ln\ L(\beta, \sigma^2) &= \sum_{y_i>0} -\frac{1}{2}\left(ln(2\pi) + ln\ \sigma^2 + \frac{(y_i - x_i'\beta)^2}{\sigma^2}\right) \\ &+ \sum_{y_i=0} ln\left(1 - \Phi\left(\frac{x_i'\beta}{\sigma}\right)\right). \end{aligned} \tag{4.12}$$

In the situation of wages in the IAB Employment Sample, we find a right-censoring at a non-zero censoring point. The model for this situation is now

described in detail. Here, the density of the uncensored observations is given by

$$f(y_i) = f_N(y_i; \mu = x_i'\beta, \sigma^2) = \phi\left(\frac{y_i - x_i'\beta}{\sigma}\right)\frac{1}{\sigma} \quad \text{if } y_i < a \tag{4.13}$$

and the probability that y_i equals the censoring value a is

$$\begin{aligned} f(y_i) &= P(y_i^* \geq a) = P(y_i = a) = 1 - P(y_i^* < a) \\ = 1 - P(y_i < a) &= 1 - \Phi\left(\frac{a - x_i'\beta}{\sigma}\right) \quad \text{if } y_i = a. \end{aligned} \tag{4.14}$$

The likelihood function is given by

$$L(\beta, \sigma^2) = \prod_{y_i<a} \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2}\left(\frac{y_i - x_i'\beta}{\sigma}\right)^2} \prod_{y_i=a}\left(1 - \Phi\left(\frac{a - x_i'\beta}{\sigma}\right)\right), \tag{4.15}$$

and the expected value is

$$\begin{aligned} E(y_i|x_i) &= E(y_i|y_i < a) \cdot P(y < a) + a \cdot P(y_i = a) \\ &= (x_i'\beta + \sigma\lambda(a^*))\Phi(a^*) + a(1 - \Phi(a^*)) \\ \text{where} \quad \lambda(a^*) &= \frac{\phi(a^*)}{1 - \Phi(a^*)} \quad \text{and} \quad a^* = \frac{a - x_i'\beta}{\sigma}. \end{aligned} \tag{4.16}$$

The log-likelihood for right-censoring at a non-zero threshold is

$$\begin{aligned} ln\ L(\beta, \sigma^2) &= \sum_{y_i<a} -\frac{1}{2}\left(ln(2\pi) + ln\ \sigma^2 + \frac{(y_i - x_i'\beta)^2}{\sigma^2}\right) \\ &+ \sum_{y_i=a} ln\left(1 - \Phi\left(\frac{a - x_i'\beta}{\sigma}\right)\right). \end{aligned} \tag{4.17}$$

Heckman (1979) suggests an alternative two-step estimation, here described for censoring at zero. It involves first estimating the unobserved term λ_i via maximum likelihood using a probit model with outcome 0 if the observation is censored and 1 otherwise. In the second step $y_i = x_i'\beta + \sigma\widehat{\lambda}_i + \varepsilon_i$ is estimated using only the observations for which $y_i > 0$. Details on all estimation approaches for censored regression models or tobit models can be found in Amemiya (1984, 1985).

The censored regression or tobit estimator is a parametric estimator because it specifies a functional form for both the regression equation and for the distribution of the error process. A drawback of this model is that maximum likelihood estimators are potentially inconsistent when ε_i is heteroscedastic. In many empirical problems, the distribution of the errors is not known or is subject to heteroscedasticity of unknown form (Chay and Powell, 2001). Thus semiparametric estimation methods that provide consistent estimates for censored data even when the error distribution is nonnormal or heteroscedastic have been developed. These approaches are discussed in the following section.

4.2 Semiparametric Approaches

Generally, semiparametric estimators for censored data can be computed by alternating between a 'recensoring' step, in which the data are 'trimmed' to compensate for the censoring problem, and a 'regression' step using the trimmed data to obtain coefficient estimates (Chay and Powell, 2001). A complete discussion of various alternative estimators can be found in Powell (1994). Semiparametric estimators assume a functional form for the regression but no functional form for the error process and therefore have the advantage that no assumption on the error term is needed. As such, they are robust to nonnormality and heteroscedasticity.

Powell (1984) proposed the censored least absolute deviations (CLAD) estimation method. For the linear model, the method of least absolute deviations obtains regression coefficient estimates by minimizing the sum of absolute residuals. It is based on a generalization of the sample median to the regression context as least squares is a generalization of the sample mean to the linear model. If the latent variable y^* was observed, the median of this variable would be the function $x'\beta$ under the condition that the errors have a zero median. In this case, the least absolute deviations method could be applied to estimate the unknown coefficients by minimizing the sum of absolute residuals. In the case of censoring, the median is unaffected as long as the regression function $x'\beta$ is in the uncensored region. On the other hand, the estimation may get a bit complicated if the regression function $x'\beta$ is below the lower threshold (or above the upper threshold in case of right-censoring) and consequently more than 50 percent of the distribution accumulate at the censoring point. A solution for this case, but not limited to this case, can be found in Buchinsky

(1994).

Buchinsky proposes the iterative linear programming algorithm (ILPA) to obtain Powell's estimator. ILPA is based on iterating between the deletion of observations of the regression function $x'\beta$ that are outside the uncensored region and estimating the coefficients using least absolute deviations applied to the remaining observations. In the first step, a median regression is computed using the whole data set. Based on the estimated parameters, all observations with a censored predicted value are deleted. From this truncated data set, the median regression is estimated again. In the third step, go back to the whole data set and delete again all observations with a censored predicted value using the updated parameters and repeat the sample truncating step. The iteration process can be stopped when two sets of consecutive iterations are the same. Buchinsky (1994) shows that a local minimum is guaranteed if the number of iterations is finite. A disadvantage of this approach is that additional observations have to be removed in order to analyze the desired research question. A discussion of and extensions to Powell's and Buchinsky's methods can be found in Berg (1998) or Paarsch (1984).

Based on a symmetric trimming idea, the symmetrically censored least squares (SCLS) estimation method is another approach to handle censored data, proposed by Powell (1986). The idea of this approach is to restore symmetry by 'symmetrically censoring' the dependent variable y from below the point $2x'\beta - a$, where a is the censoring point and the censoring appears in the upper part. We assume that the latent variable y^* is symmetrically distributed around the regression function $x'\beta$. That means the data are trimmed, so that the regression function is equidistant from both censoring points. Chay and Powell (2001) explain that since the 'recensored' dependent variable is now symmetrically distributed around the regression function, the regression coefficients can be estimated by ordinary least squares. Afterwards, iterating between censoring the dependent variable symmetrically using the current estimates and least squares estimation of the regression coefficients using the trimmed data yields the SCLS estimator.

The motivation of identically censored least absolute deviations (ICLAD) and identically censored least squares (ICLS) estimation methods is similar to the 'symmetric trimming' idea. The latter are proposed by Honoré and Powell (1994). In contrast to the SCLS estimator, these estimators involve recensoring the dependent variable for pairs of observations. Then, the regression

coefficients can be estimated by finding the value of β that minimizes the sum of absolute (ICLAD) or squared (ICLS) differences of the identically censored residuals across all distinct pairs of observations. As with SCLS the ICLS and ICLAD estimators can be calculated by iterating the identically censoring step and the least squares or least absolute deviations regression.
A further estimation method is proposed by Newey (1991) based on GMM that also allows for heavy censoring of the data. Chen and Kahn (2000) consider semiparametric estimation procedures with nonparametric heteroscedasticity. Kaplan and Meier (1958) provide estimators to determine the distribution of survival times after receiving treatment when data are censored for example due to loss of contact to individuals.

4.3 Nonparametric Approaches

Besides these approaches, several nonparametric approaches for censored regression have been proposed, which are described again for the case of left-censoring. Lewbel and Linton (2002) suggest a censored regression model of the form $y_i = max\{a, g(x_i) + \varepsilon_i\}$, where $g(\cdot)$ is the conditional expectation for the uncensored population and a the censoring point, which is presumed to be a known constant. If $E(\varepsilon_i) = 0$, the function $g(x_i)$ equals the regression function of the uncensored population. Lewbel and Linton (2002) propose a two-step procedure to estimate the function $g(\cdot)$. Further details on this approach can be found in Lewbel and Linton (2002) or Li and Racine (2007).
Another nonparametric approach is proposed by Chen et al. (2005), which is an extension of the nonparametric location-scale model which is usually of the form

$$y_i = g(x_i) + \sigma(x_i)\varepsilon_i, \tag{4.18}$$

to handle censored data. It is motivated by problems in which main interest lies in the estimation of a location function in regions where it is less than the censoring point (Li and Racine, 2007). Chen et al. consider the model

$$y_i^* = g(x_i) + \sigma(x_i)\varepsilon_i, \tag{4.19}$$

$$y_i = max\{y_i^*, 0\}, \tag{4.20}$$

where y_i^* is an unobserved latent variable, y_i is the observed dependent variable equal to y_i^* if it exceeds the censoring point and equals zero otherwise. x_i is an observed q-dimensional random vector and ε_i is a mean zero, random disturbance term that is distributed independently of x_i. Under some conditions $g(x_i)$ can be identified and estimated after imposing a local restriction, namely that the median of ε_i is zero and that $g(x)$ can be identified on the entire support of x, not just the region exceeding the censoring point. An overview on nonparametric estimation of censored and truncated regression models can be found in Chen (2010). A drawback of these nonparametric solutions compared to other approaches is that they cannot be easily implemented by researchers using standard software packages.

4.4 Censored Quantile Regression

Quantile regression was introduced in the 1970s by Koenker and Basset (1978) and was further developed by a series of researchers. In the field of quantile regression several estimation methods that can be employed in case of censoring have been proposed. Censored quantile regression is based on Powell (1984, 1986). Starting with the model

$$Q_\tau(y_i|x_i) = max\{a_i, x_i'\beta(\tau)\} \tag{4.21}$$

where a_i is again the censoring point. The censoring point can be different for individuals and not necessarily has to be 0. τ in parentheses denotes the dependence on the corresponding quantile with $0 < \tau < 1$. The Powell estimator minimizes

$$\sum \rho_\tau(y_i - max\{a_i, x_i'\beta(\tau)\}, \tag{4.22}$$

where $\rho_\tau(\varepsilon) = (\tau - 1(\varepsilon \leq 0))$. As all other approaches the Powell estimator can be redefined as well for the case of right-censoring. The median or censored absolute deviation estimator (CLAD), which is discussed in Section 4.2, is a special case of this estimator with $\tau = 1/2$. Under weak regularity conditions, Powell's estimator has desirable large sample properties, but undesirable properties in small samples. In addition, numerical optimization based on the Powell estimator is arduous, even with modern computers (see for example Haupt and Ludsteck (2007)). To avoid these computational problems the semiparametric two-step estimators have been developed. A further

suggestion of Chernozhukov and Hong (2002), which is based on Buchinsky and Hahn (1998) and Khan and Powell (2001) uses a three-step estimation procedure. This approach reaches the asymptotic efficiency of Powell's estimator, but avoids its difficulties. In the first step a logit or probit regression explaining not-censoring is estimated with the form

$$\delta_i = p(x_i'\gamma) + \varepsilon_i \tag{4.23}$$

where δ_i is the indicator of not-censoring. Now, a sample $J_0 = \{i : p(x_i'\widehat{\gamma}) > 1-\tau+c\}$ is selected, where c is strictly between 0 and τ and not too small. The practical choice of c is discussed in Chernozhukov and Hong (2002). The idea behind this approach is similar to the ILPA method proposed by Buchinsky (1994). The sample is here restricted by removing observations with a high probability to be censored. The goal of the first step is to select some, not necessarily the largest, subset of observations to obtain a consistent but inefficient initial estimator $\widehat{\beta}_0(\tau)$. Then, in the second step the initial estimator $\widehat{\beta}_0(\tau)$ can be obtained by the standard quantile regression

$$min \sum_{i \epsilon J_0} \rho_\tau(y_i - x_i'\beta). \tag{4.24}$$

In the next step, select $J_1 = \{i : x_i'\widehat{\beta}_0(\tau) > a_i + \delta_n\}$, where δ_n is a small positive number and a_i the censoring point. That means again a percentage of the observations is discarded. In the third step finally the quantile regression is performed with J_1 in place of J_0 to obtain the final estimation results.

4.5 Advantages and Disadvantages of Models for Censored Data

The models discussed in this chapter provide solutions for the problem of censored wages for many research questions. But we have to consider that these models cannot be applied for every possible research question and require very specific knowledge to be able to perform them. A serious drawback is that most of them are not implemented in standard software packages. Moreover, these models are only applicable for direct analysis of a specific research question and do not generally provide potential for a wide range of research questions.

Besides, many of the models have the disadvantage that additional observations have to be discarded in order to be able to perform unbiased estimation or that they require a lot of computational power.
In some studies researchers therefore simply avoid the problem of censoring by using a restricted sample for the analysis. In these studies, for example, only young people just entering the labor market are examined and it is argued that they normally start with wages significantly lower than the censoring point (see, e.g., Stevens (2007)). In other studies, the analysis sample is restricted to employees without university or technical college degree (see, e.g., Möller (2005a,b)). The same argumentation could be used for example if only women were in the focus of a research question. Certainly, the fraction of censored wages may be lower in these groups, but this approach cannot solve the problem of censoring completely and it is surely not a solution that can be used as a general guideline for the analysis of censored data like, e.g., wages in the IAB Employment Sample. Another simple approach that cannot be recommended as a general valid solution and is used to handle censored wages mainly in U.S. studies based on the CPS survey, is to replace censored wages by the ceiling times a factor, e.g 1.33 (Devereux, 2002; Juhn et al., 1993), 1.4 (Lemieux, 2006) or 1.5 (Autor et al., 2008; Katz and Murphy, 1992).
Later we present multiple imputation approaches for censored wages. We will show that multiple imputation can ease the treatment of censored variables because it represents a flexible technique that allows the application of standard estimation techniques also for data sets with censoring or other kinds of missing values. Once the data are imputed, the analysts are free to perform any desired analysis using standard complete data models. Then, researchers not necessarily require specific knowledge about missing data techniques to be able to analyze originally censored data. Additionally, we present a series of simulation studies that confirm the validity of the suggested multiple imputation approaches.

Chapter 5

Selected Studies Based on Censored IAB Data

Apart from (multiple) imputation and the approaches described in the preceding chapter, variations of these approaches and further methods are used in the literature to handle censored data. Before we discuss in detail multiple imputation and the special case of imputation for censored wages, this chapter gives an overview on studies that are based on censored wage data. The goal is to discuss different methods that are applied to analyze censored wages in order to classify the advantages and disadvantages of our approaches and to illustrate the analytical potential of the IAB data. This chapter seeks to demonstrate the variety of wage analyses and related topics that could benefit from the availability of properly imputed wages. Studies are described from recent years that are based on administrative data of the German Federal Employment Agency and the Institute for Employment Research and a broad range of research questions dealing with wages are examined. For these studies, the method applied, the database, and the main findings are briefly summarized. For studies based on the weakly anonymized version of the IAB Employment Sample, we refer to the latter just as IAB Employment Sample.

5.1 Gender Wage Gap

The analysis of wage differences between men and women has a long tradition in Germany and most other developed countries. The question of the gender wage gap is examined comprehensively in the social sciences as well as in eco-

nomics. Hence, a lot of studies concerning Germany using the IAB data can be found in the literature, which are (apart from the problem of censoring) an excellent database to examine this kind of questions. These studies apply various solutions to the problem of censoring. The first and probably least ambitious solution is to simply restrict the analysis to groups of persons where censoring does not play a role or at least not an important role. Most of these studies restrict the sample to persons with low and medium education and simply leave out highly skilled persons. Another option would be as well to drop some industries or occupations with a high wage level from the sample to reduce the percentage of censoring. However, by doing this one loses information and receives estimation results which are only representative for the selected subgroups.

An example for this approach is a study of Black and Spitz-Oener (2007), who examine the changes in the gender wage gap based on the IAB Employment Sample applying an approach that uses direct measures of job tasks and gives a characterization of how work for men and women has changed in recent decades. They find out that the differences between men and women are less pronounced in recent years and argue that a relative task change explains a substantial fraction of the reduction of the gender wage gap. According to this study women have witnessed relative increases in non-routine analytic tasks and non-routine interactive tasks, which are associated with higher skill levels. Due to the censoring they restrict the wage analysis to employees with low and medium levels of education only and argue that "(t)he impact of this restriction is less severe than it might first appear. The reason is that relative changes in task inputs across the genders were most pronounced for low and medium educated employees; hence, they appear to be the most interesting groups to look at" (Black and Spitz-Oener, 2007, p. 11).

Jurajda and Harmgart (2004) compare the importance of occupational gender segregation for the gender wage gap in East and West Germany using the IAB Employment Sample. As the wages are censored from above they focus their descriptive analysis on median wage gaps instead of mean wage gaps. Furthermore the study examines the impact of possible sources of the observed wage gap using logarithmic wage regressions. Specifically, they account for occupational segregation, worker and firm characteristics and finally estimate a logarithmic ordinary least squares regression. The effect of gender segregation on wages is captured by conditioning on the 'femaleness' of the occupation,

which is measured by the percentage of females in a given group of employees, e.g., in one occupation. Their main findings are that segregation is not related to the West German wage gap, but in East Germany wages of both men and women are higher in predominantly female occupations. In an additional step they check the sensitivity of the OLS estimates to the top-coding of wages in the IAB data. To do so, they compare the OLS results to those based on the censored least absolute deviation (CLAD) estimator proposed by Powell (1984) and claim that the new estimates show "little material difference" (Jurajda and Harmgart, 2004, p. 17). They conclude that ignoring right-censoring has a negligible quantitative effect on the estimated parameters, even if for example the coefficient 'Fraction of females in occupation' changes from 0.037 to 0.043 for men in West Germany and from 0.124 to 0.097 for men in East Germany.

Another possibility that is used in a broad range of studies concerning gender is to apply a tobit model. Heinze and Wolf (2006) apply a tobit model to the linked employer-employee data of the IAB and show that the mean gender wage gap within firms is smaller than the average overall gender wage gap and that firms with formalized co-determination (workers' council) and those covered by collective wage agreements are more likely to have a smaller gender wage gap. A further finding is that the wage differential between men and women decreases with firm size and increases with the wage level. In another study, Heinze and Wolf (2007), applying the same data set and using a tobit model as well, calculate firm-specific gender wage gaps accounting for differences in individual characteristics and show that innovative human resources practices tend to limit the wage differential between men and women. Furthermore, in a similar study Heinze (2009) examines the impact of the proportion of women working within an establishment upon individual wages.

A number of studies use a simple (single) imputation procedure according to Gartner (2005) to be able to analyze differences in wages. The basic principle of this procedure is to first estimate a tobit model, where the dependent variable is the log wage and the independent variables are those included in the desired analyses. In the second step, for every censored observation a random value is drawn from a normal distribution left-truncated at the social security contribution ceiling (with predicted log wage as mean, and standard deviation as estimated from the tobit model). Achatz et al. (2004) use the linked employer-employee database (LIAB) for East and West Germany in 2000 and this procedure to estimate a decomposition of the wage differential proposed

by Blinder (1973) and Oaxaca (1973). They find that only one tenth of the gender gap in wages is explained by human capital differences between men and women. Furthermore, with increasing proportions of women within job cells they observe decreasing wage levels for men and women but with higher rates of decline for women than for men. Besides, the presence of workers' councils has a positive impact on wage levels. Using the same data set and the same imputation approach, Gartner and Stephan (2004) find as well that the gender wage gap in Germany is smaller in firms covered by collective contracts or having a workers' council. The authors argue that these findings can be explained in part by the fact that these institutions are associated with lower unobserved productivity differences and less wage discrimination, and in part because they compress the distribution of wage residuals.

Kluve and Schaffner (2007) apply a Blinder-Oaxaca method for tobit models proposed by Bauer and Sinner (2005) to the IABS to decompose the gender wage gap. The main result is that part of the observed gender wage gap can be explained by segregation into more and less secure jobs. Since women select themselves into more secure jobs than men and since workers with high injury risks are compensated for the risk, including the injury risk, the explained part of the gender wage gap increases by about three percentage points and amounts to up to 12 percent of the whole explained part.

An overview of older studies concerning the gender wage gap based on data of the Institute for Employment Research can be found in Hübler (2003).

5.2 Wage Inequality

Besides wage differentials between men and women, wage inequality in general is another subject that is in the center of a number of studies based on the wage information contained in the data of the Institute for Employment Research. Many studies analyze the development of wage differences between and within certain groups over several years.

Möller (2005a,b) investigates the wage dispersion between employees working full-time in the lower and upper part of the wage distribution using the regional file of the IAB Employment Sample. As a measure of wage dispersion he compares the ratio of the 90th percentile to the median and the ratio of the median to the 10th percentile in different years. This analysis is done separately for men and women and three educational groups:

- Low-skilled: with no vocational degree
- Medium-skilled: with vocational degree
- High-skilled: with university or technical college degree

Due to the censoring, results can only be shown for the groups of low and medium skilled persons. In these groups, the 90th wage percentile is uncensored and results can easily be calculated and reported. The 90th wage percentile of high skilled employees is censored as approximately 45 percent of wages of men in this group are censored. In this case, it is impossible to obtain results without any correction for the censoring. In Chapter 10, it will be shown that multiple imputation could be implemented in this case to receive valid results for all groups. The main finding of the study is that from 1984 to 2001 a rising wage inequality can be observed in Germany in the examined educational groups. This development is somewhat higher for low skilled employees than for medium skilled and somewhat more pronounced for women than for men.

Dustmann et al. (2009) analyze the wage structure during the 1980s and 1990s and find that wage inequality increased in the 1980s, but only at the top of the distribution. In the early 1990s, wage inequality started to rise also at the bottom of the distribution. They show that changes in the education and age structure can explain a substantial part of the increase in inequality, in particular at the top of the distribution. They additionally argue that, for example, about one third of the increase in lower tail inequality in the 1990s can be related to de-unionization and that fluctuations in relative supply play an important role in explaining trends in the skill premium. The analysis is based on the IAB Employment Sample. Due to the missing wage information the 85th percentile is used as descriptive measure instead of the 90th percentile and semi-parametric censored quantile regressions are applied.

Also based on the IAB Employment Sample Kohn (2006) studies the wage structure in the German labor market for the years 1992 to 2001. The findings are similar to the studies described above: While wage dispersion generally rose, the increase was more pronounced in East Germany and occurred predominantly in the lower part of the wage distribution for women and in the upper part for men. To reveal diverse age and skill patterns, censored quantile wage regressions are used. Adapting a decomposition proposed by Machado

and Mata (2005) to the case of censoring, Kohn finds that differences in the composition of the work force had only a small impact on the observed wage differentials between East and West Germany.
In a study comparing the structure of wages in different countries Lazear and Shaw (2007) use the IAB linked employer-employee data to analyze the case of Germany. Here, the missing wage information is imputed according to Gartner (2005).

5.3 Central Wage Bargaining and Union Wages

Another central issue of studies based on wage data is to analyze the impact of trade unions and workers' councils on wages. Fitzenberger and Kohn (2006) examine the relationship between the level of union organization as a measure of union power and the wage structure within and between segments of the German labor market for 1985 to 1997 based on the IAB Employment Sample. To the IAB data individual probabilities of membership in an union are merged which were estimated in Beck and Fitzenberger (2004). The authors group the data according to socio-demographic characteristics of the employees and characteristics of their jobs and form cells with the dimensions time, age, and industries. The specific wage level of each cell is estimated using a tobit regression. The main findings are that a higher level of qualification wage differentials can be found in segments with strong unions. In accordance with a minimum wage character of union negotiated wages, the compression of the wage distribution is more pronounced in the lower part of the wage distribution. Union effects also vary with age of workers and over time.
Fitzenberger et al. (2001) apply a cohort analysis using censored quantile regression to the IAB Employment Sample to test for uniform wage trends in West Germany. Their results can be summarized as follows: Wages of workers with medium skill level deteriorated slightly compared to high and low skill levels during the 1970s and 1980s. However, compared to other countries, the German wages were fairly stable.
A study of Braun and Scheffel (2007) is focuses on the effect of outsourcing on the wage premium of collective bargaining agreements. It is based on the linked employer-employee data (LIAB) and the missing wage information is

imputed based on a tobit model (Gartner, 2005). They find that low skilled workers experience a decline in the union wage premium when working in industries with high outsourcing intensities, which applies to both firm- and sector-level agreements. On the other hand, outsourcing has no negative effect on the wages of employees not covered by collective bargaining agreements. In contrast to low skilled workers, wages of medium skilled workers are not affected by outsourcing, and highly skilled workers employed in industries with a high level of outsourcing even gain rising wages.

5.4 Wage Rigidity

Bauer et al. (2007) examine real and nominal wage rigidities in West Germany using the regional file of the IAB Employment Sample. Due to the censoring they drop all individuals with a wage observation at the threshold or slightly below and argue: "While this approach is common practice, it is important to note that it changes the skill composition of the sample. High skilled workers are removed more than proportionally. This might cause another selection bias in our rigidity measures, if wage rigidity is correlated with the skill (or wage) level" (Bauer et al., 2007, p. F513). Based on this restricted sample Bauer et al. find that a substantial fraction of workers faces wage increases that are caused by nominal and particularly real wage rigidity. Furthermore, the extent of real rigidity rises with inflation and falls with regional unemployment; for nominal rigidity the opposite holds. The conclusion of their findings is that the incidence of wage rigidity, which accelerates unemployment growth, is most likely minimized in a moderate inflation environment.

In another study, Knoppik and Beissinger (2003) examine downward nominal wage rigidity, right-censored observations are dropped as well. The authors admit that "(s)ince this leads to a substantial change in the skill structure of the sample, where high-skilled employees are no longer properly represented, the analysis is confined to unskilled and skilled male employees" (Knoppik and Beissinger, 2003, p. 638). They conclude that there is a high degree of downward nominal wage rigidity in the IAB Employment Sample.

5.5 Labor Supply

An issue examined by Hirsch et al. (2006, 2008) in studies based on IAB data is the question of labor supply to firms. In Hirsch et al. (2006), they use the linked employer-employee data and wages imputed based on a tobit model. Applying a structural estimation procedure based on a dynamic model of new monopsony, the authors estimate the long-run wage elasticity of firms' female and male labor supplies. The estimated elasticities were found to be small (0.9-2.4), whereas women's elasticity is only about half the size of men's. Hirsch et al. argue that an implication of these findings is that the gender pay gap could be the result of wage discrimination by profit-maximizing monopsonistic employers.

In a second version of the paper (Hirsch et al., 2008), the estimation is not based on imputed data. The authors here admit that using the censored wage data without any correction would bias the estimates and add as explanation: "However, any imputation of the censored values cannot completely remedy this problem since it will introduce, by construction, some measurement error. This will cause inconsistent estimates of wages if they are used as an explanatory variable" (Hirsch et al., 2008, p. 16). As a consequence, the analysis is carried out only for individuals whose wages were below the threshold during the examined period, which reduces the samples for men by 21.8 percent, while for women only by 8.0 percent. The conclusion remains more or less the same: labor supply elasticities are still small, but vary now from 1.9 to 3.7 and women's labor supply to the firm is again less elastic than men's. In the paper, there is no evidence that the mentioned measurement error would affect the estimation results actually more severe than the bias due to the restriction of the sample.

5.6 Regional Studies

As the IAB Employment Sample distinguishes 348 regional labor markets over a long period, it is used for a series of regional studies as well. Three examples dealing with regional wages will be discussed here. The first study, conducted by Lehmer and Möller (2008), analyzes effects of inter-regional mobility on earnings for different groups. The database for this paper is the regional file of the IAB Employment Sample. Because of the censoring, a tobit estimation

method is used. The authors find negative wage differentials of movers in the year before migration and strong evidence for significant wage gains through mobility. Additionally, a decomposition of Blinder-Oaxaca based on tobit estimates is employed to reveal different group-specific rewards effects suggesting a positive post-mobility wage differential of movers over the incumbent workforce for some groups irrespective of the region of destination.
Lehmer and Ludsteck (2008) analyze extensively the effects of inter-regional mobility on the earnings of skilled workers. The basic idea of this study is to interact returns to inter-regional migration with employer changes to separate the two effects. Lehmer and Ludsteck find that inter-regional mobility results in positive additional returns compared to job mobility within a region in general. The study is based on the IAB employment history data (BeH). Due to the high proportion of censored wages in the group of highly-skilled workers, the earnings analysis is restricted to the medium qualification group. Summing up the main results, they find that both job mobility and regional mobility lead to a wage increase in the year after changing firm relative to the group of immobile workers. In addition, they find out, that contemporaneous return for people moving to a different region is statistically significantly larger in the aggregate level than for job movers that stay in the same region.
Lehmer and Möller (2009) review interrelations between the urban wage premium and firm-size wage differentials. A tobit estimation method is again used to account for top-coding in the data, here the regional file 1975 to 2004 of the IAB Employment Sample. The authors find clear evidence for the existence of an urban wage premium in Germany. The raw wage premium amounts to 15.5 percent, controlling for personal characteristics, it can be reduced to approximately 13.5 percent. Firm-size categories in the econometric specification additionally lower the magnitude of the urban wage premium by roughly one fourth. However, firm-size differences between rural and urban areas explain a significant part of the interregional wage differential.

5.7 Other Wage Analyses

To illustrate the wide range of question that can be examined by using the rich IAB data and the multitude of solutions that can be applied to the problem of censored wages, this last section summarizes further studies from various fields of wage analysis.

Based on the IAB Employment Sample, Stevens (2007) investigates the role of economic conditions at entry into the labor market. As only men who do not take higher education and enter the labor market in West Germany before the age of 19, i.e., mainly at age between 16 and 18, are examined, censoring really seems not to be a problem in this case. The study reports that overall around 1 percent of earnings observations are censored from above in the used sample (for males approximately 1.6 percent). As the examined group is only rarely affected by censoring, standard regression methods are applied, where the local unemployment rate at entry is a regressor. According to this paper small but significant adverse effects of economic conditions at entry on earnings are found. Moreover, this negative effect gains in strength throughout working life.

Another topic of analyses based on IAB data are wage effects of immigration. An example for this topic is a paper by Bonin (2005) based on the regional file of the IAB Employment Sample (IABS-R) for the period 1975 to 1997. Using single imputed wages, the study shows that penetration of migrants into skill cells has no significant negative effect on the earnings and employment opportunities of native men. Following Bonin, the results indicate that a 10 percent rise of the share of immigrants in the workforce would reduce wages by less than one percent and not increase unemployment. For less qualified and older workers, however, the effects appear to be stronger.

Ludsteck (2008) reviews the aggregate wage cyclicality and the wage curve for establishment stayers and movers using the IAB Employment Sample for the years from 1985 to 2004. The study finds that movers' wage responses to aggregate unemployment rate changes exceed those of stayers by about 30-40 percent and that the increments of movers over stayer responses to regional unemployment shocks are considerably greater and amount to about 150 percent. Ludsteck explains this finding by the importance of centralized wage bargaining in Germany. In order to check if the censoring causes significant bias in the analysis, the author implements the consistent Honoré (1992) fixed effects GMM estimator as well, which can be thought of as a generalization for the idea behind Powell's trimmed least squares estimators for tobit models (without fixed effects). This estimator is like Powell's estimators semi-parametric and it is not necessary to assume a parametric form for the disturbances nor is it necessary to assume homoscedasticity. As differences between the Honoré estimates and conventional OLS turned out to be negligible, the computation-

ally less demanding OLS is applied.

Based on the panel information of the IAB Employment Sample Baltagi et al. (2009) consider the West German wage curve. The authors choose to use the censored wage information in the data set, i.e., the value of the threshold which is reported for high income groups. To justify this approach it is stated that "(t)ests were carried out using refined methods of dealing with this kind of problem, i.e., multiple imputation of wages above the threshold. Using panel data on a shorter time period these tests showed only very small changes in the results on the wage curve" (Baltagi et al., 2009, p. 48). The main findings of this study are that the wage equation is highly autoregressive but far from unit root and moreover that the unemployment elasticity is significant but relatively small.

Schönberg (2004) compares the sources of wage growth of young workers in the United States and Germany, two countries with very different labor market institutions. The analysis for Germany is based on the IAB Employment Sample. Because of the censoring the empirical analysis is mainly restricted to unskilled workers and workers with an apprenticeship and because of the research question on those individuals who are observed from their entry in the labor market onwards. This means that those individuals have to be at most 15 years old in 1975. The main findings are that in both countries and for all educational groups general human capital accumulation is the most important source of wage growth. 60 percent of total wage growth can be attributed to human capital accumulation after ten years spent in the labor market. The second main reason for wage growth for all education groups in both countries is job search. Interestingly, wage growth due to job switching is roughly similar for German apprentices and for US high school dropouts and graduates. The analysis is done using a method of decomposing total wage growth into wage growth due to general human capital accumulation, firm-specific human capital accumulation, and job search proposed in the paper.

Schank et al. (2004) use the LIAB to demonstrate that exporting firms do not pay higher wages compared to other firms. Existing wage premia disappear when individual characteristics of the employees and of the work place are controlled for. Due to the censoring, a tobit model is applied to estimate the effects at the individual level. At the plant level, OLS is used, as the authors argue that the distribution of the average wages analyzed at that level is not censored. This becomes a problem if individual wages are aggregated to the

plant level. The authors justify the use of aggregated wages by arguing that "[o]nly one plant in the regression sample employs solely workers with censored wages (and hence, only for this plant the average wage is censored). In other plants, some of the workers earn wages that are censored, so that the average reported wage is smaller than the average of the actual wages. However, we have ignored any (small) bias arising from this underreporting since the bias should be correlated with individual qualification for which we control in our estimations and since there is no clear cut truncation point which could be taken into account in the plant-level estimations" (Schank et al., 2004, p. 8).

Based on the linked employer-employee data Bauer and Bender (2001) examine the effects of flexible workplaces on wages as well as on the wage structure within firms. The empirical results suggest that workers benefit from flexible workplace systems through higher wages and that there is an increase of within-firm wage inequality through a relative increase in the wages at the upper parts of an establishment's wage distribution. The analysis is based on the censored wages which "should bias the estimated coefficients on our variables indicating the use of flexible workplace systems towards zero, particulary so for high-skilled workers" (Bauer and Bender, 2001, p. 15). To prevent this bias, tobit models are applied for the estimation.

Binder and Schwengler (2006) propose a procedure to adjust not the wage at the individual level, but the mean of the gross wage for each region. The aim of this study is to facilitate the comparison of mean earnings between regions. As all incomes above the limit are censored in the IAB data, the yearly gross wage per employee in a region in the data is lower than the 'actual' mean. To represent the actual earning potential in the various regions as accurately as possible, they suggest to correct the error induced by the censoring of some individual wages. For each region a hypothetical income distribution curve is searched, since the actual distributions above the threshold are not known. This hypothetical distribution function is then on the cut-off point 'extended' under the assumption of a log normal distribution. Applying this method, persons who are in the censored income class, are distributed according to the log-normal distribution above the threshold in order to reach a realistic upward correction of the regional average wage.

Table 5.7 briefly summarizes the studies described in this chapter. The table additionally shows whether these studies make use of the longitudinal structure of the data or use only cross-sectional information for one year. The overview

underlines that the IAB data contain research potential for various fields of economics and social sciences. From 2004 to 2008, 105 publications where produced using the IABS by authors not affiliated to the IAB. Among them 12 were published in Social Sciences Citation Index (SSCI) listed journals and 11 in other refereed journals (Heining, 2010). For researchers affiliated to the IAB no information on the number of publications is available, but the data are also extensively used by these researchers. Accordingly, there is a broad range of researchers that could benefit from multiply imputed wage data. The following chapter introduces imputation techniques in general and multiple imputation in particular.

Overview: Recent studies based on IAB data

Year of publication	Authors	Publication	Research question	Database	Years	Solution for the censoring
2001	Fitzenberger et al.	Empirical Economics	Testing for uniform wage trends in West Germany	IABS	1976 to 1984	Censored quantile regression
2001	Bauer and Bender	IZA Discussion Paper	Effects of flexible workplaces on wages and the wage structure within firms	LIAB	1993, 1995, 1997	Analysis based on censored wages
2003	Knoppik and Beissinger	The Scandinavian Journal of Economics	Examining downward nominal wage rigidity	IABS	1976 to 1995	Dropping all individuals with a wage observation at the threshold or slightly below
2004	Jurajda and Harmgart	IZA Discussion Paper	Importance of occupational gender segregation for the gender wage gap in East and West Germany	IABS	workers employed in 1992 and 1995	Median instead of mean for descriptive analysis, OLS (CLAD as sensitivity analysis)
2004	Achatz et al.	IAB Discussion Paper	Blinder-Oaxaca wage differential decomposition	LIAB	2000	Single Imputation
2004	Gartner and Stephan	IAB Discussion Paper	Effects of collective contracts and works councils on the gender wage gap	LIAB	2001	Single Imputation
2004	Schönberg	Society for Economic Dynamics	Sources of wage growth of younger workers	IABS	first entries 1975 to 1995	Sample restricted to employees with low and medium levels of education
2004	Schank et al.	IZA Discussion Paper	Wages of exporting firms	LIAB	1995 to 1997	Censored regression (tobit model) at the individual level, OLS at the plant level
2005	Möller	Book chapter	Wage dispersion between employees in the lower and upper part of the wage distribution	IABS-R	1984 to 2001	Restriction to low and medium skilled employees
2005	Bonin	IZA Discussion Paper	Wage effects of immigration	IABS-R	1975 to 1997	Single Imputation
2006	Heinze and Wolf	ZEW Discussion Paper	Gender wage gap and firm size, codetermination, and collective wage agreements	LIAB	1997 to 2001	Censored regression (tobit model)
2006	Kohn	ZEW Discussion Paper	Wage structure in the German labor market	IABS	1992 to 2001	Machado and Mata (2005) decomposition for censoring
2006	Fitzenberger and Kohn	ZEW Discussion Paper	Examining the relationship between the level of union organization and the wage structure	IABS	1985 to 1997	Censored regression (tobit model)

2006	Hirsch et al.	IZA Discussion Paper	Labor supply elasticities	LIAB	2000	Single Imputation
2006	Binder and Schwengler	IAB Discussion Paper	Average earnings in regions	IABS		Searching a hypothetical wage distribution for every region
2007	Black and Spitz-Oener	IZA Discussion Paper	Changes in the gender wage gap between 1979 and 1999	IABS/ BIBB	1979 and 1999	Sample restricted to employees with low and medium levels of education
2007	Heinze and Wolf	ZEW Discussion Paper	Firm-specific gender wage gaps	LIAB	1993 to 2003	Censored regression (tobit model)
2007	Kluve and Schaffner	Ruhr Economic Paper	Blinder-Oaxaca wage differential decomposition	IABS	1995 to 2001	Blinder-Oaxaca decomposition for censored regression (tobit model)
2007	Shaw and Lazear	NBER Working Paper	Comparing the structure of earnings in different countries	LIAB	1993, 1995, 2000	Single Imputation
2007	Braun and Scheffel	SFB 649 Discussion Paper	Effects of outsourcing on the wage premium of collective bargaining agreements	LIAB	1995 to 2000	Single Imputation
2007	Bauer et al.	The Economic Journal	Examining real and nominal wage rigidities in West Germany	IABS-R	1975 to 2001	Dropping all individuals with a wage observation at the threshold or slightly below
2007	Stevens	IZA Conference Paper	Effects of the economic conditions at entry into the labor market	IABS	1980 to 2001	Sample restrictions
2008	Hirsch et al.	Laser Discussion Paper	Labor supply elasticities	LIAB	2000	Dropping all individuals with a wage observation at the threshold or slightly below
2008	Lehmer and Möller	Regional Studies	Effects of inter-regional mobility on earnings of different groups	IABS-R	1996/1997	Censored regression (tobit model)
2008	Lehmer and Ludsteck	IAB Discussion Paper	Returns to job mobility and inter-regional migration	BeH	1995, 1996, 2000	Restriction to medium skilled employees
2008	Ludsteck	IAB Discussion Paper	Wage cyclicality and the wage curve for establishment stayers and movers	IABS	1985 to 2004	OLS, Honoré (1992) fixed effects GMM estimator
2009	Heinze	ZEW Discussion Paper	Assessing the impact of the proportion of women working within an establishment upon individual wages	LIAB	2002	Censored regression (tobit model)
2009	Dustmann et al.	The Quarterly Journal of Economics	Changes in the wage structure during the 1980s and 1990s	IABS	1975 to 2001	85th percentile as a descriptive measure instead of the 90th percentile, censored quantile regression
2009	Lehmer and Möller	Annals of Regional Science	Interrelations between the urban wage premium and firm-size wage differentials	IABS	1990 to 1997	Censored regression (tobit model)
2009	Baltagi et al.	Labour Economics	The wage curve for West Germany	IABS	1980 to 2004	Analysis based on censored wages

Table 5.1: Recent studies based on IAB data

Chapter 6

Multiple Imputation

In general, multiple imputation is a statistical technique for analyzing incomplete data sets, e.g., data sets for which some values are missing. The missingness can appear due to several reasons: Among many other reasons, subjects may fail to provide data, individuals may drop out from an observational study or some information may just not be reported because of legal reasons, like it occurs in the case of censored wages. Application of the technique requires three steps: imputation, performing the analysis m times, and combining the results. The chapter gives an overview on different imputation approaches and finally gives a brief introduction to multiple imputation. A detailed description about analysis of missing data can be found in Little and Rubin (1987, 2002) and an overview is given, e.g., in Rässler et al. (2008). Besides, Reiter and Raghunathan (2007) describe some of the main adaptations of multiple imputation.

6.1 Missing-Data Mechanisms

Before we discuss different imputation strategies, the first section distinguishes various missing-data mechanisms. These mechanisms describe to what extent missingness depends on the observed and/or unobserved data values and were formalized first by Rubin (1976). Following this work, Little and Rubin (1987, 2002) distinguish three cases: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). To be able to distinguish these cases formally, let Y represent the $N \times P$ matrix of complete data and R represent the $N \times P$ matrix of indicator values for observed and

missing values in Y. The missing-data mechanism gives the probability of the matrix of indicator variables R, given Y and possible parameters governing this process, $\xi : f(R|Y, \xi)$.

Missing data for which missingness does not depend on any of the data values, neither missing nor observed, is referred to as MCAR. In this case, the probability that data on a particular variable can be observed does not depend on the value of that or any other variable: $f(R|Y, \xi) = f(R|\xi)$. In many cases the MCAR assumption is unrealistically restrictive and can be contradicted by the observed data (see, e.g., Rässler et al. (2008)).

A different situation appears if the missingness can be explained by observed values in the data, like for example gender, age, or social status. If the probability of units responding to items depend on observed values, but not on any missing values then according to Little and Rubin (1987, 2002) the missing data are MAR, but not necessarily MCAR because of the following dependence: $f(R|Y, \xi) = f(R|Y_{obs}, \xi)$, where Y_{obs} are observed values in Y, $Y = (Y_{obs}, Y_{mis})$, and Y_{mis} being the missing values in Y.

The data are finally NMAR, if, even given the observed values, missingness still depends on data values that are missing. In addition to the concept of MCAR, MAR and NMAR, Rubin (1976) introduced the concept of ignorability. He shows that if the data are MAR and the parameters of the data distribution, ψ, and the missing-data mechanism, ξ, are distinct, then valid inferences about the distribution of the data can be obtained using a likelihood function that does not contain a factor for the missing-data mechanism and is simply proportional to $f(Y_{obs}|\psi) = \int f(Y|\psi)dY_{mis}$. He finds that in this case the missing-data mechanism may be 'ignored' for likelihood or Bayesian inference.

Often, it is reasonable to assume that the parameters of the data distribution and the missing-data mechanism are distinct and the question of whether the missing-data mechanism is ignorable often reduces to a question of whether the missing data are MAR. Even when the ignorability assumption is not known to be correct, it is common to make this assumption in analyses of incomplete data as it can be advantageous to do so for a variety of reasons (Rässler et al., 2008): The most convincing reason is that it can simplify analyses greatly. Another reason is that the MAR assumption is often reasonable, especially when there are fully observed covariates available in the analysis to 'explain' the reasons for the missingness and that MAR cannot be contradicted by the

observed data without the incorporation of external assumptions. Besides, even if the missingness is not MAR but NMAR, an analysis based on the assumption of MAR can be helpful in reducing bias by imputing missing data using relationships that are observed. The last reason that is mentioned, e.g., by Rässler et al. (2008) is that it is usually not easy to specify a correct nonignorable model, even if the missing data are NMAR. The main problem here is that any evidence concerning the relationship of missingness to the missing values is absent because the missing values are (by definition) not observed.

6.2 Handling Missing Data

Little and Rubin (1987, 2002) categorize methods for analyzing incomplete data into four main groups:

- Simple methods like complete-case analysis and available-case analysis
- Weighting procedures
- Imputation-based procedures
- Multiple Imputation

In the following sections the basics of these approaches will be presented and advantages and disadvantages will be discussed. An extensive introduction and discussion of these approaches can be found in Rässler et al. (2008).

6.2.1 Simple Approaches

The simplest way to deal with missing data is the complete-case analysis. Here, all cases with at least one missing value are deleted and only complete cases are used for the analysis. This method therefore is sometimes called 'listwise deletion'. This procedure is generally biased if the missing data are not MCAR. The degree of bias depends on different factors like the amount of missing data, the degree to which the assumption of MCAR is violated and the particular analysis that is implemented. Another disadvantage of this approach is that even if it is unbiased, it can be highly inefficient, especially with multivariate

data sets, where a large fraction of units may be subject to deletion even if there are only missing values in some variables.
An alternative to the complete-case method is the available-case method, where all units that have complete data on the variables that are needed for the analysis are considered. In Rässler et al. (2008) this approach is called 'complete-case analysis restricted to the variables of interest'. The advantage of this method is that an equal or higher number of data values are retained compared to the complete-case analysis, but this becomes problematic when more than one quantity is estimated and different estimates are combined or are supposed to be comparable, as the sample base changes from estimation to estimation. As complete-case and available-case analysis are often the default treatments of missing data in software packages, like, e.g., STATA, they are easy to implement, but may have the discussed serious drawbacks. As they are the default treatment in some software packages, sometimes it may occur that analysts are not even aware of the bias that may arise, e.g., if units with single missing values are automatically deleted.

6.2.2 Weighting Adjustments

Weighting adjustment can be interpreted as a modification of complete-case analysis to remove bias when the missing data are not MCAR. It can be applied for example in case of unit nonresponse in surveys. Here, complete cases are weighted based on background information that is available for all units in the survey. One simple possibility to perform weighting adjustment is as follows: When a nonrespondent matches a respondent with respect to background variables that are observed for both, the weight of the nonrespondent can be simply added to the matching respondents weight and the nonrespondent can be discarded (Rässler et al., 2008). As the matching is performed using observed variables, this kind of weight adjustment implicitly assumes MAR. The disadvantage of this approach is that it nearly always arises new problems, mainly because discarding incomplete cases discards additional observed data that are not used in creating the weighting adjustment.

6.2.3 Single Imputation

By applying single imputation, one value is imputed for each missing value. Little and Rubin (2002, p. 72) summarizes guidelines for (single) imputations.

They should be:

(1) conditional on observed variables

(2) multivariate to reflect associations among missing variables and

(3) randomly drawn from predictive distributions rather than means to ensure that correct variability is reflected.

The main advantage of single imputation is that the imputed data set is straightforward to analyze using standard complete-data methods. Rässler et al. (2008) describe a number of single imputation approaches.
The simplest single imputation method is to replace each missing value with the mean of the observed values of the variable. This method meets none of the guidelines formed by Little and Rubin (1987, 2002). Another method, regression imputation refers to replacing the missing values with values predicted from a regression of the variable containing missing values on other variables. This method satisfies the first two guidelines. A special case of regression imputation is the cell mean imputation. Here, missing values are replaced with the mean of that variable calculated within cells defined by categorical variables. Another method that can meet all three guidelines for single imputation, when done properly, is stochastic regression imputation. Here, random noise is added to the predicted value. An example for this approach is the method proposed by Gartner (2005), which is addressed in the preceding chapter. The last single imputation method is referred to as 'hot-deck imputation'. Each missing value is replaced here with a random draw from a pool of donors. The donor pool consists of observed values of that variable stemming from units similar to the unit with the missing value. They can be selected by choosing units with complete data and similar observed values to the unit with missing values, for example by exact matching on their observed values or using a distance measure (metric) on observed variables to define 'similar' (Rässler et al., 2008). A special case of hot-deck imputation is the so-called 'predictive mean matching'. Here, the distance is defined as the difference between units on the predicted value of the variable to be imputed (Rubin, 1986). Supposing that it is properly done, hot-deck imputation can also satisfy the three guidelines for single imputation. If the single imputations have been done following the guidelines of Little and Rubin (1987, 2002), then, according to Rässler et al. (2008), analyzing the imputed data set with standard complete-data

techniques is straightforward and can lead to approximately unbiased point estimates under ignorability. An important disadvantage, however, is that the analyses will nearly always result in estimated standard errors that are too small, confidence intervals that are too narrow and p-values for hypothesis tests that are too significant. The reason is that imputed data are treated by standard complete-data analyses as if they were known with no uncertainty (Little and Rubin, 1987, 2002). As a consequence single imputation is almost always statistically invalid if it is followed by a complete-data analysis that does not distinguish between real and imputed values.
Nevertheless, a series of special methods for variance estimations following single imputations have been developed. The problem with these techniques is that they are only appropriate for specific imputation procedures and estimation problems, but are not generally applicable for all estimation problems. Here, an imputed data set cannot be used for all kind of research questions without having detailed information on the imputation method used and knowledge on imputation techniques. An alternative approach that is broadly applicable but computationally intensive is to use replication techniques like jackknife or bootstrap for variance estimation with separate imputation procedures for each replication.
Multiple imputation (MI) on the other hand, is a generally valid alternative, which is compared to specific estimation procedures generally applicable and compared to replication techniques less computationally intensive. Hence, it is particulary useful in the context of creating data sets shared by many users, as it could be the case with the IAB Employment Sample. The theory and principle of multiple imputation originates from Rubin (1978) and involves repeating the drawing of single imputations several times, but its exact validity requires that the imputations are 'proper' (Rubin, 1987).

6.3 Principles of Multiple Imputation

Multiple imputation is an approach to complete missing data and to reflect the added uncertainty due to the fact that the imputed values are not the actual values. A big advantage is that it permits to analyze the imputed data sets using standard complete-data methods. Rässler et al. (2008) argue that in general, only MI and direct analysis can lead to valid inferences and add that valid inferences have to satisfy the following three criteria:

(a) approximately unbiased estimates of population estimates (e.g., means, correlation coefficients),

(b) interval estimates with at least their nominal coverage (e.g., 95% intervals for a population mean should cover the true population mean at least 95% of time) and

(c) tests of significance should reject at their nominal level or less frequently when the null hypothesis is true (e.g., a 5% test of a zero population correlation should reject at most 5% of the time when the population is zero).

Resampling methods, like jackknife and bootstrap are able to satisfy criteria (b) and (c) asymptotically, but cannot help to satisfy (a) in the presence of missing data. Hot-deck imputation for example can satisfy criterion (a), but fails to satisfy (b) and (c). As we want to develop a solution for the missing wage information in the IAB Employment Sample, where the once imputed censored wages can be used by several researchers for a broad range of research questions applying standard methods and that fulfills all criteria discussed above, MI is the most useful approach in this case. That is why, from now, we focus on the advantages of MI.
MI was introduced by Rubin (1978) and discussed in detail in Rubin (1987, 2004b,a). It is a simulation technique that replaces the missing values Y_{mis} with $m > 1$ plausible values and therefore reveals and quantifies uncertainty in the imputed values. For notational simplicity, we assume here ignorability of the missing-data mechanism, even though this assumption is not necessary for MI to be appropriate. Generally, a set of m imputations (i.e., each single imputation for Y_{mis}) creates m complete data sets: $Y^{(1)}, ..., Y^{(m)}$, where $Y^{(m)} = (Y_{obs}, Y_{mis}^{(m)})$. Typically m is fairly small, $m = 5$ is a standard number of imputations to use. Each of the m imputations is done by properly drawn (single) imputations. Such a proper imputation can be obtained by a random draw from the 'posterior predictive distribution' of the missing data given the observed data $f(Y_{mis}|Y_{obs})$. Often it is not possible to specify this distribution directly. But it can be formally written as $f(Y_{mis}|Y_{obs}) = \int f(Y_{mis}, \psi|Y_{obs})d\psi = \int f(Y_{mis}|Y_{obs}, \psi)f(\psi|y_{obs})d\psi$. This expression effectively gives the distribution of the missing values, Y_{mis}, given the observed values, Y_{obs}, under a model for Y governed by the parameter ψ, $f(Y|\psi)f(\psi)$, where $f(\psi)$ is the prior distribution

of ψ. The distribution $f(Y_{mis}|Y_{obs})$ is called 'posterior' because it is conditional on the observed Y_{obs} and 'predictive' as it predicts the missing Y_{mis}.
For simple patterns of missing data, like with only one variable subject to missingness, a two-step procedure is then relatively straightforward to implement:

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

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

The imputation can be made proper in Rubin's sense if it reflects all uncertainty, including in parameter estimation, by taking draws of ψ from its posterior distribution, $f(\psi|Y_{obs})$, before using ψ to impute the missing data, Y_{mis}, from $f(Y_{mis}|Y_{obs}, \psi)$. Imputation methods are labeled as 'improper' by Rubin (1987), if they do not account for all sources of variability. An example for an improper method would be fixing ψ at a point estimate $\widehat{\psi}$ and then drawing m imputations for Y_{mis} independently from its posterior distribution, $f(Y_{mis}|Y_{obs}, \widehat{\psi})$.
Finally, each of the m imputed data sets is analyzed as if there were no missing data and the results of the m analyses have to be combined using combining rules that will be discussed later.
If there are missing values in more than one variable, it is only straightforward to draw random samples from $f(Y_{mis}|Y_{obs})$ if the missing data follow a monotone pattern. This situation appears for example in clinical trials, when data are missing due to a patient dropout. Where once a patient drops out, the patient never returns (Rässler et al., 2008). Here, the imputation can be started by fitting an appropriate model to predict the variable with the fewest missing values from all variables with no missing values. Then the missing values for the variable with the second fewest missing values can be imputed using the variables with no missing values and the first imputed variable. Now we continue to impute the next most complete variable until all missing values have been imputed. According to Rässler et al. (2008), imputation is proper under this model and the collection of univariate prediction models defines the implied full imputation model, $f(Y_{mis}|Y_{obs})$.
In a case where the missing data are not monotone, iterative computational methods are generally necessary. Here, creating imputations generally involves

iteration because it is often difficult to draw from the distribution $f(Y_{mis}|Y_{obs})$ directly. In this case, the data-augmentation algorithm (Tanner and Wong, 1987) is often straightforward to implement. This algorithm briefly involves iterating between randomly sampling missing data given a current draw of the model parameters and randomly sampling model parameters given a current draw of the missing data and a Markov Chain whose stationary distribution $f(Y_{mis}|Y_{obs})$ is formed by the draws of Y_{mis}.

Markov chain Monte Carlo (MCMC) in general is a method based on drawing values of ψ from approximate distributions and then correcting those draws to better approximate the target posterior distribution, $f(\psi|y)$. Based on the distribution of the sampled draws depending on the last value drawn, the samples are drawn sequentially and the draws form a Markov chain. Hence, a Markov chain is a sequence of random variables $\psi^{(1)}, \psi^{(2)}, ...$, for which for any t, the distribution $\psi^{(t)}$ given all previous $\psi's$ depends only on the most recent value, $\psi^{(t-1)}$. According to Gelman and Hill (2007) the key to the success of this method is that the approximate distributions are improved at each step in the simulation, converging to the target distribution

Further algorithms that apply Markov chain Monte Carlo methods to impute missing values are the Gibbs sampler (Geman and Geman, 1984) and the Metropolis-Hastings algorithm (Metropolis and Ulam, 1949; Hastings, 1970). The Gibbs sampler is a special case of Markov chain simulation algorithms that can be used to iteratively estimate parameters in any statistical model. Markov chain simulation and the Gibbs sampler in particular can be thought of as iterative imputation of unknown parameters (Gelman and Hill, 2007). The Gibbs sampler updates the parameters one at a time or in batches using their conditional distributions.

Alternatively to performing an imputation under one specified model, imputation can be done under potentially incomplete models like a potentially incomplete Gibbs sampler. These iterative simulation methods run a regression on each variable that contains missing data on all other variables using previously imputed values for these other variables. The regression can be for example least squares, logistic etc. The regression and imputation step is then cycled through all variables with missing values. These imputation methods, which are not necessarily derived from a joint distribution for all of the data, provide very flexible tools for imputations. Computational guidance on creating multiple imputations under a variety of models can be found in Schafer (1997).

6.3.1 Combining Rules for Multiply Imputed Data

As mentioned before, multiple imputation consists of imputation, analysis, and combination of the results. Imputation approaches have been discussed above and, as we have discussed, analysis can be performed afterwards by applying standard complete data methods. The m results of the analysis step then have to be combined following combining rules first described by Rubin (1987). To illustrate these rules, let θ represent the estimand (scalar) of interest and $\widehat{\theta}$ the standard complete data estimator of θ and let $\widehat{V}(\widehat{\theta})$ represent the standard complete-data estimated variance of $\widehat{\theta}$. Multiple Imputation has been used to create m completed data sets. Accordingly, we receive m complete data statistics applying the complete data analysis to each data set, say $\widehat{\theta}_l$ and $\widehat{V}_l$, where $l = 1, ..., m$. The m sets of statistics are combined to obtain the final point estimate $\widehat{\theta}_{MI} = m^{-1} \sum_{l=1}^{m} \widehat{\theta}_l$ and the corresponding variance $T = W + (1 + m^{-1})B$, where $W = m^{-1} \sum_{l=1}^{m} \widehat{V}_l$ is the 'within-imputation' variance, $B = (m-1)^{-1} \sum_{l=1}^{m} (\widehat{\theta} - \widehat{\theta}_{MI})^2$ is the 'between-imputation' variance and the factor $(1+m^{-1})$ reflects the fact that only a finite number of completed-data estimates $\widehat{\theta}, l = 1, ..., m$ are averaged together to obtain the final point estimate. Additionally, $\widehat{\gamma} = (1 + m^{-1})B/T$ is introduced, which estimates the fraction of information about θ that is missing due to the missing data.

Based on $\widehat{\theta}_{MI}$, T and a student's t reference distribution, inferences from the multiply imputed data can be calculated. Interval estimates for θ for example have the form $\widehat{\theta}_{MI} \pm t(1-\alpha/2)\sqrt{T}$, where $t(1-\alpha/2)$ is the $(1-\alpha/2)$ quantile for the t distribution. Following Rubin and Schenker (1986) the degrees of the t distribution can be approximates by the value $\nu_{RS} = (m-1)\widehat{\gamma}^{-2}$, under the assumption that with the complete data, a normal reference distribution would have been appropriate. To allow for a t reference distribution with complete data, Barnard and Rubin (1999) proposed the value $\nu_{BR} = (\nu_{RS}^{-1} + \widehat{\nu}_{obs}^{-1})^{-1}$ for degrees of freedom in the MI analysis, where $\widehat{\nu}_{obs} = (1 - \widehat{\gamma})(\nu_{com})(\nu_{com} + 1)(\nu_{com} + 3)$, and ν_{com} is the complete-data degrees of freedom. We can see that the MI interval estimate is expected to produce a larger interval than an estimate based only on a single imputation. The MI interval estimates are widened to account for the missing data uncertainty. For additional combining rules, e.g., for significance levels, see Rubin and Schenker (1991) or Little and Rubin (1987, 2002).

6.3.2 Advantages of Multiple Imputation

The main advantage of imputation, either single or multiple, that gives this kind of procedures great inherent flexibility and makes imputation especially attractive, when an imputed data set is supposed to be used by many different users, is that the implicit or explicit model used for the imputation need not necessarily be the same as the explicit or implicit model applied by the data users in their analyses using the completed data (Rässler et al., 2008). Thus, this feature makes multiple imputation a very appropriate approach to handle censored wages in the IAB Employment Sample. Once the missing wage information has been imputed, analysts are free to explore a variety of models for analyzing the completed data. The same applies to the question of releasing public and scientific use files in general. The analysts do not have to worry about a possible bias due to the censoring (or other missing data problems) or applying special censored data methods anymore, but can apply all kind of standard methods using standard software packages. Many software packages, like STATA (which is the main software used by researchers to analyze the IAB Employment Sample), already include tools to apply the MI combining rules.

One important restriction to the general applicability, which should not be concealed, is that the formal derivation of procedures for analyzing multiply imputed data is based on the assumption that the imputer's and analyst's models are compatible. According to Meng (1994), for the resulting analyses to be fully valid, the imputer's and analyst's model have to be 'congenial'. Uncongeniality refers to the situation when the model used by the analyst of the data differs from the model used for the imputation. This can lead to biased results, if the analyst's model is more complex than the imputation model and the imputation model omitted important relationships present in the original data. When the imputer and the analyst are the same person or at least communicate with each other, congeniality can easily be enforced. It gets more complicate in the context of shared data sets. Thus, to promote near-congeniality of the imputers's and user's implicit model, so that analyses based on multiply imputed data will be at least approximately valid, the imputer should include as rich a set of variables in the imputation model as possible in order to accommodate the variety of analyses that might be carried out by users (Rässler et al., 2008). If the analyst's model is a sub-model of the

imputer's model, i.e., the imputation model contains a larger set of covariates than the analyst's model and the covariates are good predictors for the missing values, then MI inference is superior to the best inference possible using only the variables in the analyst's model. Rubin (1996) calls this property super-efficiency. If the imputation model does not contain all important correlates of variables with missing data, which are used in the analyst's model, the results will be biased. In the case of the IAB Employment Sample, where the total number of variables in the data set is manageable, all variables should be included in the imputation model, especially when the aim is to multiply impute the censored wages in order to produce a scientific use file. If the intention is to produce a complete data set for a specific research question, a restricted imputation model following the analyst's model can be used. For research questions where additional information has to be merged to the IAB Employment Sample, like for example regional unemployment information or data stemming from the IAB establishment panel, larger imputation models have to be applied. The user's possibility to merge the IAB Employment Sample with other sources makes it difficult to find an imputation model that is generally valid for all purposes. Even in this more complicated case, MI can be easily applied, but an individual imputation model will have to be found in those special cases.

6.3.3 Multiple Imputation for Censored Variables

Comparing the advantages and drawbacks of multiple imputation with other approaches to handle censored data, MI provides an excellent and flexible solution for censored wages in the IAB Employment Sample, although, the situation with censored wages is slightly different to other missing data problems concerning wages. In most surveys with nonresponse concerning income or wages the problem of missingness appears in high income groups as individuals with high incomes or wages tend to higher nonresponse rates. Here, in most cases the imputation can still be performed using standard imputation software, as the information is not missing completely from a certain point. Regression-based imputation as well as other imputation methods like hot-deck imputation could be applied for those kind of problems. In case of censoring, we find a situation where standard programmes cannot be easily applied, because virtually no information on high wages is available. For this

situation, special methods have to be adapted. In the following chapter we present regression-based approaches to apply the technique of multiple imputation to the censored wages and finally perform a series of simulation studies to confirm the necessity as well as the validity of these approaches.

Chapter 7

Imputation for Right-Censored Wages

Applications of multiple imputation for right-censored or truncated wages are very rare in the literature so far. Apart from an approach proposed by Gartner and Rässler (2005), that will be discussed later in detail, only few approaches are noteworthy in this context. Jenkins et al. (2009) suggest a multiple imputation approach for censored observations in the U.S. CPS to measure income inequality using draws from generalized beta of the second kind distributions to provide data sets that can be analyzed using complete data methods. The approach is applied to the internal and public data series, but in both cases the fraction of censored income is significantly lower than in the IAB Employment Sample. The procedure consists of five steps. First, an imputation model with a parametric functional form that is presumed to describe the income distribution including right-censored observations is fitted. Second, a value is drawn from the implied distribution using a randomization procedure for each censored observation. Third, inequality indices and associated variances are estimated based on complete data methods using the distribution comprising imputed incomes for censored observations and observed incomes. In step four the preceding steps are repeated 100 times and finally, the results from each of the 100 data sets are combined. In this paper, Jenkins et al. (2009) show that using CPS public use data with cell mean imputation may lead to incorrect inferences about inequality differences, but also admit that researchers using the public use data could build more sophisticated imputation models to improve the quality of estimates derived. This is necessary to allow for

example for subgroup differences by allowing for covariates in the estimation of the parametric model.

Another study that is close to the censoring in the IAB Employment Sample is by An and Little (2007). Here, multiple imputation is not applied due to missing wages, but as a method of statistical disclosure control. They use hot-deck multiple imputation and parametric multiple imputation based on lognormal and power-transformed normal distributions in order to create synthetic data for individuals with high incomes in the 1995 Chinese household income project. Using the non-parametric hot-deck imputation procedure, high income values are replaced with values randomly drawn with replacement from the set of the deleted values. The parametric method is based on Bayesian statistics and assumes a model for the data, draws model parameters from their posterior distribution, and then imputes the deleted values with random draws from the posterior predictive distribution. The context is different to the IAB Employment Sample, as here multiple imputation is not a measure to impute missing wages but a method to avoid artificial censoring due to data protection and statistical disclosure control requirements. Therefore, the values that have to be replaced are generally known and therefore can be used for the imputation model. To be able to release data to the public, high income data classified as sensitive, i.e., all observations from a certain cut-off point are deleted, and MI is applied to fill in these values again. Then multiple imputed data sets can be released to the public.

Heitjan and Rubin (1991) develop a generalization of the condition missing at random (MAR) for coarsened data, which includes as special cases censored, rounded, heaped, and partially categorized data. Rubin and Heitjan introduce coarsened at random (CAR) to generalize the ideas of MAR and ignorable missing-data to coarsened data. According to Heitjan (1994) the censoring mechanism is CAR but not MAR, if the censoring does not depend on the values of the outcome, although it can depend on the values of the covariates. Generally, Little and Rubin (1987, 2002) define censored normal data as an important special case of grouped normal data and describe Bayesian inference using the Gibbs's sampler for data where some observation are grouped into categories. These approaches can be easily transformed for the case of right-censoring. Little and Rubin discuss for example in detail an approach by Heeringa et al. (2002), who develop multiple imputations of coarsened and missing data for 12 assets and liability variables in the U.S. Health and Retire-

ment Survey, where data are a mixture of actual valued responses, bracketed (or interval-censored) replies, and completely missing data. Here, an attractive feature of the Gibbs sampler is used: draws of the missing values can be generated one variable at a time, conditionally on current draws of the parameters and the observed or drawn values of all the other variables. Since the conditional distribution of any one variable given the others is normal, interval-censored information about that variable is easily incorporated in the draws. For the imputation based on Gibbs sampling, Gibbs' sequences are created with 20 different random starts to yield 20 multiply imputed data sets. Comparing this approach to other missing data techniques, Heeringa et al. (2002) show that complete-case as well as mean imputation analysis markedly underestimates the distribution of household net worth. Hot-deck imputation produces lower estimated values for the mean and upper quantiles of the distribution than the Bayes method.

Apart from applications for censored wage and income variables, several methods for the imputation of censored or coarsened variables are proposed in the literature. An example for left-censored data are concentrations of pollutants in the arctic, which are coarsened in the sense of being either fully missing or below detection limits. Hopke et al. (2001) propose multiple imputation for multivariate data with missing and below-threshold measurements to facilitate scientific analysis in this case and create complete data by filling in missing values so that standard complete-data methods can be applied. Multiple imputation is also used to analyze data in coarse categories, as it occurs with age heaping. Heitjan and Rubin (1990) multiply impute heaped ages in a Tanzanian demographic data set with plausible true ages using different models, i.e., a simple naive model and a complex model that relates true age to the observed values of heaped age, sex, and anthropometric variables. Pan (2000) proposes an iterative semiparametric method based on multiple imputation for cox regression with interval-censored data, which can be easily applied by taking advantage of routines for right-censored data that are implemented in standard software packages. In addition to posterior computations for censored regression data (Wei and Tanner, 1990), Wei and Tanner (1991) present semiparametric multiple imputation approaches for the analysis of censored regression data by implementing two approximations to the data augmentation algorithm (Tanner and Wong, 1987) to the context of censored regression data.

In the following sections of this chapter, we describe imputation approaches for the right-censored wages in the IAB Employment Sample. We apply two approaches assuming homoscedasticity of the residuals, which we will be discussed in detail in the next section. We will show that the assumption of homoscedasticity is highly questionable with wage data as the variance of income is smaller in lower wage categories than in higher categories. That is why we furthermore suggest a new single imputation approach and a new multiple imputation approach allowing to control for heteroscedasticity.

Before we describe the approaches and develop the new imputation approaches, we first need some notation that is valid for all methods that will be discussed later. All of them assume that the wage y for every person i is given by

$$y_i^* = x_i'\beta + \varepsilon_i \qquad \text{where} \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), i = 1, ..., n, \tag{7.1}$$

where x is a vector of covariates such as education, gender or age. Notice that we will use a log transformation of the wages as well as further transformations, like, e.g., a cube root transformation, that can (and will be) also used to perform the imputation. 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^*:

$$y_i = \begin{cases} y_{obs,i} \text{ if } y_i^* \leq a \\ a \text{ if } y_i^* > a \end{cases} \tag{7.2}$$

To be able to analyze wages with our data set, we first have to impute the wages above a. We define $y_z = (y_{obs}, z)$, where z is a truncated variable in the range (a, ∞).

According to Gartner and Rässler (2005) we regard the missingness mechanism as not missing at random (NMAR, according to Little and Rubin (1987, 2002)) as well as missing by design: The former because the missingness depends on the value itself; if the limit is exceeded, the true value will not be reported but the value of the limit a. The latter occurs because the data are missing due to the fact that they were not asked.

7.1 Homoscedastic Imputation Approaches for Right-Censored Wages

7.1.1 Homoscedastic Single Imputation

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). This kind of imputation method comes along with the advantages of regression-based imputation, and is easy to implement. However, it is a sort of an ad-hoc method, i.e., not realizing draws from the proper posterior distribution as described in Section 6.3. Some studies based on this approach are discussed in Chapter 5. Applying this method, a tobit (or censored regression) model is used to estimate the parameter β and σ^2 of the imputation model. According to the estimated parameters the censored wage z can be imputed by draws of a random value. As we know that the true value is above the contribution limit, we have to draw a random variable from a truncated normal distribution

$$z_i \sim N_{trunc_a}(x_i'\widehat{\beta}, \widehat{\sigma}^2) \text{ if } y_i = a \text{ for } i = 1, ..., n. \tag{7.3}$$

This means we add an error term ε to the expected wage (see Gartner (2005) for a description of drawings from a truncated distribution in STATA):

$$z_i = x_i'\widehat{\beta} + \varepsilon_i \text{ if } y_i = a \text{ for } i = 1, ..., n \tag{7.4}$$

As already mentioned before, 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 the imputation should rather be done in a multiple and Bayesian way according to Rubin (1978). Therefore, it is preferable to use multiple imputation approaches to impute the missing wage information.

7.1.2 Multiple Imputation Assuming Homoscedasticity (MI-Hom)

Gartner and Rässler (2005) propose a multiple imputation approach based on Markov chain Monte Carlo techniques. To perform multiple imputation

for the censored values we need independent random draws from the posterior predictive distribution $f(Y_{mis}|Y_{obs})$ of the missing data given the observed data. Since it is often difficult to draw from $f(Y_{mis}|Y_{obs})$ directly, we could rather apply the two-step procedure of drawing ψ from $f(\psi|Y_{obs})$ and in the second step drawing Y_{mis} from $f(Y_{mis}|Y_{obs}, \psi)$ to achieve imputations of Y_{mis} from their posterior predictive distribution as discussed in the preceding chapter.

In many situations the conditional predictive distribution $f(Y_{mis}|Y_{obs}, \psi)$ is rather straightforward; where in contrast, the corresponding observed-data posteriors $f(\psi|Y_{obs})$ are usually difficult to derive for the units with missing data. This is the case especially when the data have a multivariate structure or a variable is censored like in the IAB Employment Sample. Then, the observed-data posteriors are often no standard distributions from which random draws can easily be generated. That is the reason to apply MCMC techniques based on the Gibbs sampler to achieve the desired distributions $f(Y_{mis}|Y_{obs})$ and $f(\psi|Y_{obs})$ as stationary distributions of Markov chains, which are based on the complete-data distributions and therefore are easier to compute. By adapting starting values for ψ, we are able to start with draws for Y_{mis} from the conditional predictive distribution $f(Y_{mis}|Y_{obs}, \psi)$ and to start the Markov chain.

Chib (1992) proposes a Monte Carlo approach for tobit models that combines the data augmentation strategy of Tanner and Wong (1987), which iterates between randomly drawing missing data given a current draw of the model parameters and randomly drawing model parameters given a current draw of the missing data, and with a Gibbs sampler, which iteratively imputes unknown parameters to yield an elegant solution to censored data problems, which can be applied to the problem of censored wages: 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}. Let the index z denote estimates based on the imputed data after z is drawn and added, i.e., $Y_z = (X, Z)$.

The conditional predictive distribution for observations with missing wage in-

formation z is given by

$$f(z|y,\beta,\sigma^2) = \frac{f_N(z|x'\beta,\sigma^2)}{1-\Phi(\sigma^{-1}a-\sigma^{-1}x'\beta)} \tag{7.5}$$

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

$$f(\beta|y,z,\sigma^2) = f_N(\beta|\widehat{\beta}_z,\sigma^2(X'X)^{-1}) \tag{7.6}$$

$$f(\sigma^2|y,z,\beta) = f_G(\sigma^2|n/2,\sum_{i=1}^{n}(y_z-x'\beta)^2/2) \tag{7.7}$$

where $\widehat{\beta}_z = (X'X)^{-1}X'y_z$ is the usual OLS estimate based on the completed data set and f_G a gamma distribution. To receive valid imputations and random draws of the parameters from their observed data distribution, Gartner and Rässler (2005) propose the following MCMC technique.

Imputation model

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, comparable to the first step of single imputation approach assuming homoscedasticity. Second, in the imputation step, values for the missing wages are randomly drawn from a truncated distribution in analogy to the single imputation procedure

$$z_i^{(t)} \sim N_{trunc_a}(x_i'\beta^{(t)},\sigma^{2(t)}) \text{ if } y_i = a \text{ for } i = 1,...,n. \tag{7.8}$$

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

$$\widehat{\beta}_z^{(t)} = (X'X)^{-1}X'y_z^{(t)}. \tag{7.9}$$

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:

$$g \sim \chi^2(n-k) \tag{7.10}$$

$$\sigma^{-2(t+1)} = \frac{g}{RSS} \tag{7.11}$$

where RSS is the residual sum of squares $RSS = \sum_{i=1}^{n}(y_{zi}^{(t)} - x_i'\widehat{\beta}_z^{(t)})^2$ and k is the number of columns of X.
Now new random draws for the parameter β can be performed

$$\beta^{(t+1)}|\sigma^{2(t+1)} \sim N(\hat{\beta}_z^{(t)}, \sigma^{2(t+1)}(X'X)^{-1}). \tag{7.12}$$

We perform repeatedly the imputation and the posterior-steps (7.8) to (7.12) and create a Gibbs sampler. We start the Gibbs sampler with different values $\beta^{(0)}$ and $\sigma^{2(0)}$ and let m independent chains run. In that case, we take the endpoints as imputations. Another possibility is to monitor convergence and dependence structure of the chain and after a burn-in period, we take every 1,000th imputation to obtain m complete data sets. For more details see Gartner and Rässler (2005) or Jensen et al. (2010).

7.2 Heteroscedastic Imputation Approaches for Right-Censored Wages

Regression disturbances whose variances are not constant across observations are heteroscedastic (Greene, 2008). Heteroscedasticity arises in numerous applications. For example, even after accounting for firm size, greater variation in the profits of large firms than in those of small firms can be expected. According to Greene (2008), the variation of profits might also depend on product diversification, research and development expenditure, and industry characteristics. Therefore, variance might also vary across firms of similar size. Another example where heteroscedasticity might arise is analyzing family spending patterns. Here, greater variation in expenditure on high income families can be found compared to low income ones due to greater dispersion allowed by higher incomes (Prais and Houthakker, 1955). The same applies to wages, where the variation in high income groups might also be higher than in low income groups. Figure 7.1 plots the residuals against the fitted values of the observed wages in the IAB Employment Sample for the year 2000 to illustrate this problem. A linear regression of daily wages for males in West Germany on a constant, age, squared age, nationality, six dummies for education levels, and four categories of job level is applied to produce this plot. Figure 7.2 shows the plot for daily wages in logs. These figures confirm that

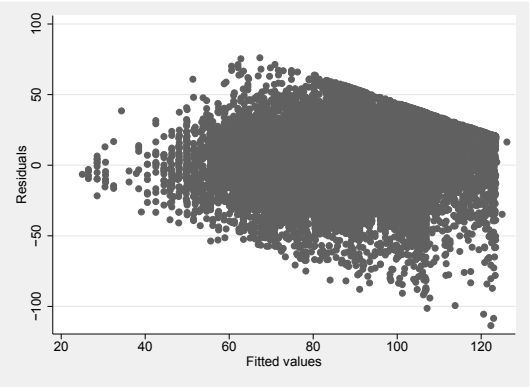

Figure 7.1: Residuals against fitted values of observed daily wages in the IAB Employment Sample.

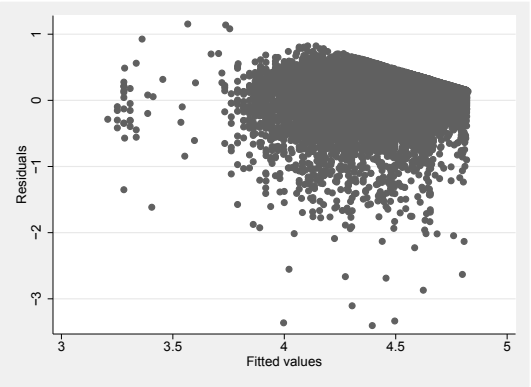

Figure 7.2: Residuals against fitted values of observed daily wages in logs in the IAB Employment Sample.

the assumption of homoscedasticity is highly questionable with this data set. Even if the ordinary least squares estimator β is unbiased, consistent, and asymptotically normal distributed in presence of heteroscedasticity (Greene, 2008), imputation results based on a regression model might be affected by heteroscedasticity, especially as we add a residual term based on the estimated variance by drawing a value from a truncated distribution. By applying a tobit model, this becomes even more problematic, because here we assume the same variation for the censored observations as for the observed observations, which is highly questionable.

7.2.1 Single Imputation Considering Heteroscedasticity

Since we obviously have to assume that the variation of income is smaller in lower wage categories than in higher categories, we extend our approach of Section 7.1.1 to consider heteroscedasticity. Therefore, we first use another single imputation procedure based on the first single imputation approach, a method that does not presume homoscedasticity of the residuals.

We assume that the error variance is related to a number of exogenous variables, gathered in a vector w (not including a constant). We use a generalized least squares model (GLS) for censored variables (e.g., *intreg* in STATA) to estimate the parameters of the imputation model, β, like in the first approach, and furthermore γ, here describing the functional form of the heteroscedasticity. Then, the imputation can be done by draws from a truncated normal

distribution, similar to the first approach,

$$z_i \sim N_{trunc_a}(x_i'\widehat{\beta}, \widehat{\sigma}_i^2) \quad \text{where} \quad \widehat{\sigma}_i^2 = e^{w_i'\widehat{\gamma}} \text{ if } y_i = a \text{ for } i = 1, ..., n, \quad (7.13)$$

where w is a vector of observed variables that is a function of x, e.g a subset of x variables. To consider the heteroscedastic structure of the residuals, we use here individual variances for every person to draw a random value. This solution takes into consideration the existence of heteroscedasticity, yet it does not solve the problem of biased variance estimations. Therefore, we have to derive the Bayesian solution considering heteroscedasticity.

7.2.2 A First Simulation Study

Since we assume the necessity of an approach that does not presume homoscedasticity and since Little and Rubin (1987, 2002) among others show that single imputation approaches may lead to biased variance estimations, we extend the MI-routine to a new multiple imputation approach. A first simulation study using the first three approaches shows the need for this approach as well. This simulation study points out that, in case of a homoscedastic structure of the residuals, the multiple imputation leads to better results than a single imputation approach. However, in case of heteroscedasticity, the single imputation considering heteroscedasticity is superior to the multiple imputation approach suggested by Gartner and Rässler (2005). This indicates the necessity to develop another approach that combines these two properties: an approach performing multiple imputation and considering heteroscedasticity.

7.2.3 Multiple Imputation for Right-Censored Wages Considering Heteroscedasticity (MI-Het)

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.

Imputation model

We now start the imputation by adapting starting values for $\beta^{(0)}$ and $\gamma^{(0)}$ from a GLS estimation for truncated variables like in the heteroscedastic single

imputation approach. Then we draw values z_i for the missing wages from a truncated distribution using individual variances $\sigma_i^2 = e^{w_i'\gamma}$ and use them as imputations, again like in the heteroscedastic single imputation model:

$$z_i^{(t)} \sim N_{trunc_a}(x_i'\beta^{(t)}, \sigma_i^{2(t)}) \quad \text{where} \quad \sigma_i^{2(t)} = e^{w_i'\gamma^{(t)}} \text{ if } y_i = a \text{ for } i = 1, ..., n. \tag{7.14}$$

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 $\widehat{\beta}_z^{(t)}$ and $\widehat{\gamma}^{(t)}$. Additionally, we estimate the variance-covariance-matrix of $\widehat{\gamma}^{(t)}$, $V(\widehat{\gamma}^{(t)})$, to be able to perform the following steps. We produce new random draws for the parameters according to their complete data posterior distribution. As we now consider the existence of heteroscedasticity, some modifications of the algorithm are necessary. In the next steps, we draw the variance $\sigma^{2(t+1)}$ according to

$$g \sim \chi^2(n-k) \tag{7.15}$$

$$\sigma^{-2(t+1)} = \frac{g}{RSS} \tag{7.16}$$

where

$$RSS = \sum_{i=1}^{n} exp(ln\,\widehat{\varepsilon}_i^2 - w_i'\widehat{\gamma}^{(t)}) = \sum_{i=1}^{n} \frac{(y_{z_i}^{(t)} - x_i'\widehat{\beta}^{(t)})^2}{e^{w_i'\widehat{\gamma}^{(t)}}}. \tag{7.17}$$

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

$$\gamma^{(t+1)} \sim N(\widehat{\gamma}^{(t)}, \widehat{V}(\widehat{\gamma}^{(t)})) \tag{7.18}$$

Consequently, the parameters β can be drawn like in the Gartner and Rässler (2005) approach, again with a slight modification compared to the homoscedastic multiple imputation:

$$\beta^{(t+1)}|\gamma^{(t+1)}, \sigma^{2(t+1)} \sim N(\widehat{\beta}_z^{(t)}, \sigma^{2(t+1)} \left(\sum_{i=1}^{n} \frac{x_i x_i'}{e^{w_i'\gamma^{(t+1)}}} \right)^{-1}). \tag{7.19}$$

Again, we repeatedly perform the steps (7.14) to (7.19) and create a Gibbs sampler to obtain m complete data sets as described in Section 7.1.2.

All approaches described in this chapter are generally also applicable for left-censoring. To perform an imputation for a left-censored variable, in the first step a tobit model for left-censoring at a non-zero limit has to be estimated. Then, the following steps can be performed as described above. An additional adjustment is only necessary concerning the draws for the missing values from a truncated normal distribution. Instead of using a normal distribution truncated at the left, a distribution truncated at the right has to be applied.
Having finally developed this multiple imputation approach considering heteroscedasticity, we are able to apply four different approaches to impute censored wage information in the IAB Employment Sample: A single imputation and a multiple imputation approach assuming homoscedasticity of the residuals and moreover a single imputation and a multiple imputation approach considering heteroscedasticity. As the results of the first simulation study have revealed to use an approach that multiply imputes the missing wages and does not assume homoscedasticity, we expect our new approach to have advantages in the imputation quality compared to the other approaches. To examine this hypothesis and the imputation quality of these approaches in general, in the following chapters several simulation studies are presented.

Chapter 8

Validation of the Approaches

To evaluate the different approaches and to show the relevance (and superiority) of the new approach, it seems to be an appropriate proceeding to perform a series of simulation studies. The aim of this chapter is to demonstrate that estimation of, e.g., an OLS regression based on multiply imputed wages leads results comparable to an estimation based on the complete data before deletion. The first simulation is based on the IAB Employment Sample itself. A serious drawback of this simulation study is that the IAB Employment Sample is censored and we cannot compare the imputation results with results based on the original data before censoring. Therefore, we use this data set to create complete (control) populations with different characteristics. In the second step, we use the uncensored wage information of the German Structure of Earnings Survey as complete population, which will artificially be censored and the deleted wages will be imputed applying the different approaches. Finally, the results of the imputation procedures can be compared with results based on the original (complete) data. Based on the GSES, we perform several simulation studies to compare the different imputation approaches (considering heteroscedasticity vs. assuming homoscedasticity) under different imputation models and different transformations of the wage data (i.e., log and cube root transformation). To improve the readability, Table 8.1 gives an overview of all simulation studies that will be presented in this chapter.

	Data set	**Main question**	**Results in Table**
1	IABS	SI and MI for a homoscedastic data set	8.2
2	IABS	SI and MI for a heteroscedastic data set	8.3
3	GSES	MI based on a lognormal transformation	8.5
4	GSES	MI based on a cube root transformation	8.6
5	GSES	MI and GLS estimation in the analysis step	8.7
6	GSES	MI using a limited set of variables	8.8
7	GSES	MI in education groups	8.9
8	GSES	Large imputer's model and small analyst's model - Example 1	8.10
9	GSES	Large imputer's model and small analyst's model - Example 2	8.11
10	GSES	Differing imputer's and analyst's model	8.12
11	GSES	Log transformation in the imputation step and cube root transformation in the analysis step	8.13
12	GSES	Cube root transformation in the imputation step and log transformation in the analysis step	8.14

Table 8.1: Simulation studies in Chapter 8

8.1 Simulation Study using the IAB Employment Sample

Before we use uncensored wage information from an income survey, the first simulation study is based on the IAB Employment Sample. In a first step, this simulation study is intended to confirm the necessity of the new multiple imputation approach considering heteroscedasticity.

8.1.1 Creating a Complete Population

To perform the simulation study based on the IAB Employment Sample, a complete, i.e., the true population is created in order to be able to compare the results of the different approaches with a complete database. As the wage information in our sample is right-censored, we first have to impute our sample to obtain this database. The fact that the data set has to be imputed before starting the simulation study allows us to create two true populations with different characteristics: We create one data set where homoscedasticity is existent and another with heteroscedasticity of the residuals. To obtain the first data set (data set A) we use the homoscedastic single imputation procedure as described in Section 7.1.1 to impute new wages for every person regardless if the wage was originally censored or not, according to

$$y_{new} \sim N(x_i'\beta, \sigma^2), \tag{8.1}$$

again with β and σ^2 from a tobit estimation based on the right-censored sample. To receive the second data set (data set B), the heteroscedastic single imputation method described in Section 7.2.1 is used in order to receive a control population with heteroscedasticity of the residuals[1], according to

$$y_{new} \sim N(x_i'\beta, \sigma_i^2). \tag{8.2}$$

These two data sets will later be used as complete populations where random samples are repeatedly drawn from. The random samples will be censored and the different approaches will be applied. Since we know the "truth" from our constructed population, we can compare the results based on the uncensored samples and the imputations with it.

[1]Performing a Breusch-Pagan-test for heteroscedasticity in the applied model using data set B, the hypothesis of homoscedasticity is rejected.

8.1.2 Simulation Study

Having created a complete data set without censored wages, we define a new limit and delete the wages above this limit. Afterwards, the missing wages are imputed using the different approaches.
To simplify the simulation design, we restrict the data for the simulation to male West-German residents. We use all workers holding a full-time job covered by social security effective on June 30th 2000. The data set contains 214,533 persons: 23,685 or 11 percent of them with censored wages.
The analysts model for the simulation study is in principle based on the well-known Mincer wage equation, which dates back to Mincer (1958) and models the statistical relationship between market wages, education, and experience. Our model contains additional variables in order to underline the applicability of multiply imputed wage for the analysis including a broad range of variables. For the simulation study we assume a model - simulating an analysis which is typically done with wage data - containing the wages in logs as dependent variable and as covariates:

X=(age, age^2, 7 education categories, 5 job level categories, nationality (German/Non-German)).

For the categorial variable education 'education missing' is used as reference category, for the variable job level the category 'trainee'. In many studies based on the IAB Employment Sample, units with missing education information are dropped or imputed using correction rules proposed by Fitzenberger et al. (2006). Other studies create an additional category for units with missing education information (e.g., Dustmann et al. (2009) for analyses with the LIAB). As in the first step, a new wage is created for every individual to receive a complete population, there is no essential need to drop these units and therefore the latter approach is applied here. As imputation model we apply the same model as the analysis model. Describing the heteroscedasticity, we assume a model containing the same set of variables.

Step 1: Drawing of a random sample
In the first step a random sample of n=21,453 persons is drawn without replacement from the population of N=214,533 persons (equivalent to 10 percent). This 10 percent random sample is kept to illustrate the results of the

different imputations later. For the simulation study we define a new threshold. To point out the differences between the four approaches we choose a limit lower than in the original IAB Employment Sample (censoring the highest 30 percent of incomes appears adequate) and delete the wages above this limit.

Step 2: Imputation of the missing wage information
The deleted wage information above the threshold of this (now again right-censored) sample is imputed by using the four different approaches described above:

- Homoscedastic single imputation
- Heteroscedastic single imputation
- Multiple imputation assuming homoscedasticity
- Multiple imputation considering heteroscedasticity

For the multiple imputation approaches we set m=5, i.e., applying one of the single imputation methods, one complete data set is obtained and applying one of the multiple imputation methods, m=5 complete data sets are obtained. These imputed data sets can now be used to evaluate the quality of the different approaches by comparing them with the original complete population.

Step 3: Analysis of the results
To analyze the results of the four approaches, we run OLS regressions using the analysis model on the imputed data sets and the 10 percent complete random sample on the one side, as well as on the complete 'true' population on the other side. Afterwards, we are able to evaluate which approach delivers the best imputation quality compared to the original complete data. Therefore, we compare $\widehat{\beta}$ - estimated based on the imputed data sets - with the parameter β of the regression on the complete population; we calculate the corresponding confidence intervals and count the so-called coverage, which contains the information, whether the 'true value' lies within the central 95 percent confidence interval of the estimated values.
Since the multiple imputation approaches lead to five complete data sets, the estimations have to be done five times as well. Afterwards, the results have to be combined using the combining rules first described by Rubin (1987) and

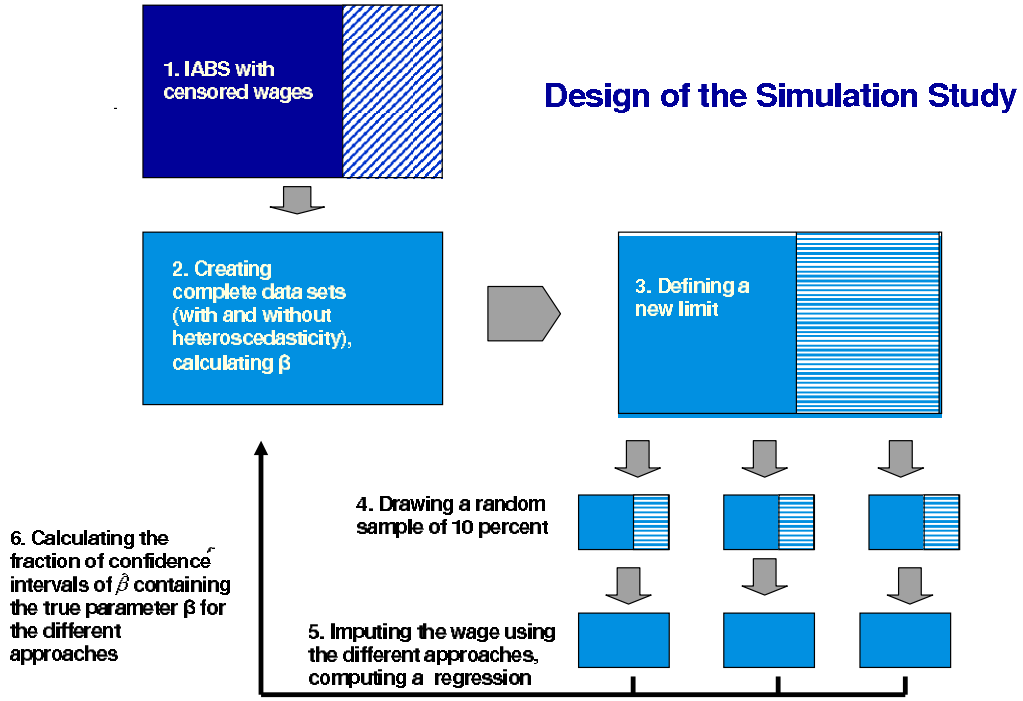

Figure 8.1: Design of the simulation study

that are shortly described in Section 6.3.1 To make it explicit, the multiple imputation point estimate for β is the average of the $m = 5$ point estimates

$$\widehat{\beta}_{MI} = \frac{1}{m}\sum_{t=1}^{m}\widehat{\beta}^{(t)}. \tag{8.3}$$

The variance estimate associated with $\widehat{\beta}_{MI}$ has two components. The within-imputation-variance is the average of the complete-data variance estimates,

$$W = \frac{1}{m}\sum_{t=1}^{m}\widehat{var}(\widehat{\beta}^{(t)}). \tag{8.4}$$

The between-imputation variance is the variance of the complete-data point estimates

$$B = \frac{1}{m-1}\sum_{t=1}^{m}(\widehat{\beta}^{(t)} - \widehat{\beta}_{MI})^2. \tag{8.5}$$

Subsequently the total variance is defined as

$$T = W + \frac{m+1}{m}B. \tag{8.6}$$

For large sample sizes, tests and two-sided $(1-\alpha)*100\%$ interval estimates for multiply imputed data sets can be calculated based on Student's t-distribution according to

$$(\widehat{\beta}_{MI} - \beta)/\sqrt{T} \sim t_v \qquad \text{and} \qquad \widehat{\beta}_{MI} \pm t_{v,1-\alpha/2}\sqrt{T} \tag{8.7}$$

with the degrees of freedom

$$v = (m-1)(1 + \frac{W}{(1+m^{-1})B})^2. \tag{8.8}$$

We save for every approach in every iteration the estimate $\widehat{\beta}$ (or $\widehat{\beta}_{MI}$ in case of the multiple imputation approaches) and the corresponding standard error of $\widehat{\beta}$, as well as the 95 percent confidence interval based on $\widehat{\beta}$. Besides, we keep the information if the confidence interval based on $\widehat{\beta}$ contains the parameter β of the original data set.

Step 4: 1000 iterations
The whole simulation procedure - consisting of drawing a random sample, imputing the data using the different approaches, running a regression on the different imputed data sets and calculating the confidence intervals - is repeated 1000 times. Finally, the fraction of confidence intervals based on $\widehat{\beta}$ or $\widehat{\beta}_{MI}$ containing the true parameter β can be calculated for the different approaches. The results of these iterations are described in the following section.

8.1.3 Results

This section contains tables showing the results of the simulation study comparing the four different approaches. The first column presents the true parameters β of the original complete population. The following columns show the estimates $\widehat{\beta}$ (here the average of the 1000 iterations) of the regression using the 10 percent complete random samples ('before censoring') and the regressions using the data sets imputed by the different approaches. The tables show as well the fraction of iterations where the 95 percent confidence interval based on $\widehat{\beta}$ contains β, i.e., the so-called coverage.

		before censoring		single homosc.		single heterosc.		multiple homosc.		multiple heterosc.	
	true β	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage
education1	0.1068	0.1069	0.959	0.1074	0.951	0.1073	0.950	0.1074	0.958	0.1073	0.958
education2	0.1791	0.1790	0.965	0.1792	0.953	0.1790	0.952	0.1792	0.965	0.1790	0.961
education3	0.1305	0.1310	0.954	0.1317	0.939	0.1330	0.935	0.1318	0.955	0.1330	0.957
education4	0.2621	0.2623	0.963	0.2624	0.928	0.2654	0.888	0.2624	0.957	0.2653	0.949
education5	0.4445	0.4446	0.948	0.4409	0.868	0.4466	0.759	0.4410	0.944	0.4469	0.922
education6	0.5098	0.5096	0.962	0.5064	0.852	0.5121	0.719	0.5065	0.953	0.5118	0.929
level1	0.5449	0.5441	0.949	0.5440	0.952	0.5447	0.950	0.5440	0.949	0.5446	0.950
level2	0.6517	0.6512	0.950	0.6515	0.954	0.6524	0.951	0.6515	0.952	0.6523	0.951
level3	0.8958	0.8950	0.948	0.8973	0.950	0.8958	0.936	0.8976	0.948	0.8959	0.954
level4	0.8962	0.8956	0.953	0.8961	0.950	0.8962	0.949	0.8962	0.951	0.8963	0.951
age	0.0498	0.0498	0.955	0.0500	0.943	0.0500	0.930	0.0500	0.964	0.0500	0.957
sqage	-0.0005	-0.0005	0.958	-0.0005	0.936	-0.0005	0.922	-0.0005	0.962	-0.0005	0.960
nation	-0.0329	-0.0327	0.962	-0.0334	0.948	-0.0334	0.942	-0.0335	0.953	-0.0334	0.955
cons	2.4424	2.4433	0.953	2.4406	0.945	2.4405	0.932	2.4411	0.951	2.4406	0.949

Table 8.2: Results of the homoscedastic data set

Homoscedastic data set

Table 8.2 shows the results of the simulation based on the homoscedastic data set A. As expected, the simulation study shows the necessity of a multiple imputation approach, since the coverage of the two multiple imputation approaches is higher compared to the single imputations throughout almost all variables. Using a homoscedastic data set, the results do not show serious differences between the homoscedastic and the heteroscedastic multiple imputation. We receive a coverage for both of these approaches of around 95 percent (between 0.922 and 0.965) - similar to the coverage received by the estimations using the complete random samples before censoring (between 0.948 and 0.965) - which refers to a good imputation quality. The coverage of the single imputations is for most of the variables lower than 0.95 - which indicates underestimated variances. Consequently, it can be concluded that, in any case, it is advisable to use a multiple imputation approach. Moreover, it does not matter if the algorithm considering heteroscedasticity is chosen in the homoscedastic case, since it just represents a generalization of the homoscedastic approach and therefore works also in case of homoscedasticity.

Heteroscedastic data set

The results based on the heteroscedastic data set B (Table 8.3) show a different situation. The results recommend as well the use of a multiple imputation approach, since the coverage of the single imputation approaches is again lower than 0.95 for all variables. Concerning the heteroscedastic structure of the residuals, it reveals the necessity of an approach considering heteroscedasticity. The homoscedastic approaches lead in several cases to a considerably lower coverage than the procedures that consider heteroscedasticity. The coverage of the heteroscedastic multiple imputation approach amounts again to around 95 percent and is similar to the coverage based on the complete samples before censoring (the coverage ranges between 0.917 and 0.97, except the dummy for the highest education level where the coverage is 0.896). Thus we see that, in this case, the coverage of the multiple imputation approach assuming homoscedasticity is lower (between 0.478 and 0.948, for some variables even lower than the coverage received by the heteroscedastic single imputation approach, where the coverage ranges between 0.718 and 0.948). Therefore, the results suggest the use of an approach considering heteroscedasticity to impute the

		before censoring		single homosc.		single heterosc.		multiple homosc.		multiple heterosc.	
	true β	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage	$\widehat{\beta}$	Coverage
education1	0.1141	0.1145	0.952	0.1271	0.794	0.1136	0.945	0.1272	0.804	0.1136	0.955
education2	0.1912	0.1915	0.955	0.2075	0.616	0.1903	0.948	0.2076	0.632	0.1903	0.955
education3	0.1442	0.1444	0.961	0.0947	0.745	0.1406	0.942	0.0952	0.769	0.1420	0.963
education4	0.2685	0.2686	0.961	0.2753	0.913	0.2688	0.922	0.2754	0.937	0.2689	0.960
education5	0.4433	0.4435	0.963	0.4790	0.366	0.4372	0.761	0.4796	0.478	0.4377	0.917
education6	0.5241	0.5248	0.954	0.5117	0.785	0.5164	0.718	0.5121	0.869	0.5161	0.896
level1	0.5422	0.5426	0.955	0.5415	0.946	0.5422	0.947	0.5416	0.946	0.5417	0.953
level2	0.6405	0.6411	0.950	0.6430	0.944	0.6412	0.944	0.6430	0.947	0.6407	0.950
level3	0.8856	0.8864	0.945	0.8780	0.941	0.8845	0.945	0.8782	0.948	0.8838	0.952
level4	0.8903	0.8908	0.952	0.8737	0.941	0.8919	0.943	0.8737	0.941	0.8913	0.951
age	0.0432	0.0431	0.955	0.0457	0.645	0.0431	0.948	0.0457	0.679	0.0431	0.970
sqage	-0.0004	-0.0004	0.960	-0.0005	0.590	-0.0004	0.941	-0.0005	0.623	-0.0004	0.968
nation	-0.0223	-0.0218	0.961	-0.0297	0.872	-0.0222	0.945	-0.0296	0.882	-0.0222	0.954
cons	2.5858	2.5865	0.947	2.5318	0.909	2.5868	0.945	2.5315	0.914	2.5875	0.952

Table 8.3: Results of the heteroscedastic data set

missing wage information in case of either an homoscedastic or heteroscedastic structure of the residuals.
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 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 results shown in Table 8.3 confirm the necessity of our new approach.

8.2 Simulation using External Data

For the simulation study described above data sets with synthetic wage information were used. That means we generated for every individual a wage using a single imputation approach and deleted this information again if the wage is above a ceiling. A disadvantage of this proceeding is that the data-generating process is known when we start to impute the deleted wage information again. One could argue in this case, that we do not simulate the situation we normally have when we impute the censored wages in the IAB employment register. In order to impute the missing wages in this register, we need to find an appropriate imputation model that is a good predictor for the wage. In contrast to the first simulation study, normally we do not already know a model that we can use as imputation model. That means finding a suitable imputation model is a very sensitive part of the imputation procedure. Since in the case of the first simulation study we have information on the data-generating process, we do not have to care about finding a suitable model.
To confirm that the proposed multiple imputation approach works even if a suitable imputation model is a priori unknown, we perform further simulation studies using data from an income survey (German Structure of Earnings Survey, GSES) with uncensored wage information, which was already addressed in Chapter 2. This data set allows us to compare the different imputation approaches again using a complete population. We truncate the wage variable at a ceiling and recover the deleted information using different approaches.

Afterwards we compare again the imputed data sets to the original complete data set in analogy to the first simulation study. The advantage of this proceeding is that we simulate a situation where the data-generating process is unknown and that we nevertheless have a complete population to compare the imputation approaches.

For the analyses and simulation studies, the GSES 2001 in the weakly anonymized version of the scientific use file is used. Fitzenberger and Reize (2002) compare in detail the IABS and the GSES. They conclude that due to the differences in the sampling design, there are some minor differences in the structure of wages between the two data sets. But qualitatively the results concerning the wage structure are fairly identical. To simplify the simulation design and to keep the sample comparable to the IAB Employment Sample, for the following simulation studies the sample is restricted again to male West-German residents holding a full-time job covered by social security. We exclude executive managers according to §5(3) of the German Industrial Constitution Act ('Betriebsverfassungsgesetz'). The first reason to drop this group is that nearly half of these persons have a reported social security contribution of zero, for which reason it is questionable whether all these persons are generally subject to statutory social security insurance and therefore are not necessarily covered by the IAB Employment Sample. Second and even more important, the data quality in this group seems to be very questionable. According to the German Industrial Constitution Act, there is a reference wage of 6871.76 euros, which is meant to indicate the minimum monthly wage of an executive manager according to §5(3) of the German Industrial Constitution Act. But on the contrary, only 38 percent of all persons in this group have a wage above this reference wage. Due to these problems, in preceding versions of the GSES this group was excluded or artificially censored. The version of 2001 is the first version to include this group without any restrictions. Hence this group, which represents less than 3 percent of all persons in the data set, is dropped.

We additionally exclude trainees undergoing an apprenticeship or professional training because these persons receive wages clearly lower than the contribution limit. The final sample contains 368,337 persons. The GSES reports the monthly gross wage. In analogy to the IAB Employment Sample we use the daily gross wage, which is calculated as the monthly gross wage divided by 31 (the wage in the GSES is reported for the month of October).

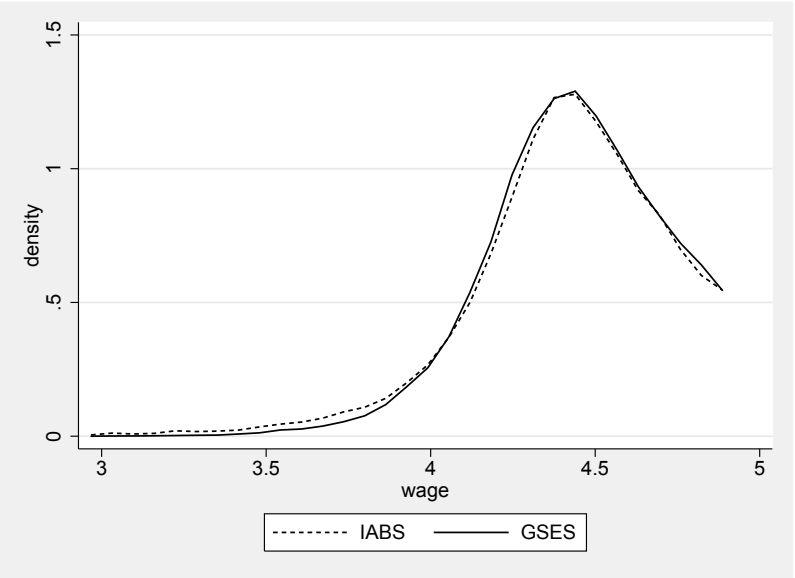

Figure 8.2: Kernel density estimates of wages up to the contribution limit in the IABS and GSES (2001)

To illustrate that the distribution of wages is fairly similar, Figure 8.2 shows the distribution of wages in the two samples in the region up to the ceiling, where the data sets are comparable before the imputation of the censored wages in the IAB Employment Sample. For the density plot the same sample restrictions as described above apply. Table 8.4 gives an impression on some descriptive statistics of the two data sets used for the simulation studies to show the comparability of the IAB Employment Sample and the German Structure of Earnings Survey. The table shows the shares of education groups and job level groups as well as the average age of employees in these two data sets. For the descriptions all observations in the IABS were used, equal if censored or not. These brief descriptions also underline the utility of the GSES for evaluating imputation approaches meant to solve the problem of censored wages in the IAB Employment Sample.

In the following sections, several simulation studies based on the GSES will be presented. The aim of these analyses is to confirm again the necessity and validity of the new multiple imputation approach considering heteroscedasticity under different situations and to be able to give a guideline, which wage transformations and imputation models are most appropriate to impute the missing wage information. Another intention to perform simulation studies

	IABS	GSES
Low/intermed. school	15.06	14.83
Vocational training	68.93	69.33
Upper school	0.81	0.78
Upper school and vocational training	3.72	3.77
Technical college	4.78	5.58
University degree	6.71	5.70
Blue collar level 1	23.35	24.58
Blue collar level 2	32.44	32.62
Blue collar level 3	2.86	2.95
White collar	41.32	39,85
Age (in years)	40.32	40.36

Table 8.4: Comparison of shares of education groups, shares of job levels groups, and average age (IABS and GSES 2001)

in several variations is to point out that the imputation procedures are robust to various situations. First, different transformation of wages, i.e., log and cube root transformation, are compared using an imputation model that contains a rich set of covariates. To evaluate the imputation results, ordinary least squares regression as well as generalized least squares regression will be applied. In the second step, the imputation model will be varied, e.g., simpler imputation models will be examined, to see if for some research questions a more limited model yields a sufficient imputation quality. Finally, the impact of differing imputation and analysis models will be examined, which reflects some simple cases of uncongeniality. Of course, this series of simulation studies cannot cover all possible imputation designs that might be intended to be performed by data distributing organization or researchers. But the variety of analyses that will be carried out underlines the applicability of multiple imputation in general and of the approach considering heteroscedasticity in particular to solve the problem of censored wages.

For the simulation studies, we truncate the wage variable at a ceiling (we delete the wages above the 85 percent quantile comparable to the top-coding in the IABS) and impute the deleted information using the two different multiple imputation approaches. As the first simulation study based on the IAB Employment Sample has already confirmed the hypothesis that multiple imputation is superior to single imputation, the following simulation studies focus on

the multiple imputation approaches. We delete the 15 percent highest wages instead of applying the real contribution limit of the current year, to have a certain percentage of censored wages that is independent from the chosen simulation sample. In the real world, the share of censored wages varies from research question to research question. If, for example, certain groups are excluded from the analysis, the share might be higher or lower. To eliminate any suspicion that samples with a low share of censored observations are chosen for evaluating the approaches, we choose the 85 percent quantile as ceiling. In our sample consisting of full-time employed males in West Germany the real share of censored observations would be even less than 15 percent.
The simulation studies consist again 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. The whole procedure is again repeated 1,000 times. Then we compare the imputed data sets with the complete population calculating the coverage as described before .

8.2.1 Simulation Study Based on a Log Transformation

For the first simulation study we assume an imputation model containing the wages in logs as dependent variable and a rich set of covariates:

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type).*

For the categorial variables always the first category is used as reference category. Note that, compared to the simulation studies based on the IABS, we have one education category less because persons with missing education information are dropped and one job level category less because trainees are dropped. 'Contract type' represents a dummy for fixed-term employment contracts. For the model describing the functional form of the heteroscedasticity, we assume for all simulation studies a subset of these variables:

W=*(age, age^2, 6 education categories, 4 job level categories, 4 region dummies, contract type).*

A log transformation implies that the wages are log-normal distributed, reflecting the right skewness of the distribution. This assumption is standard in

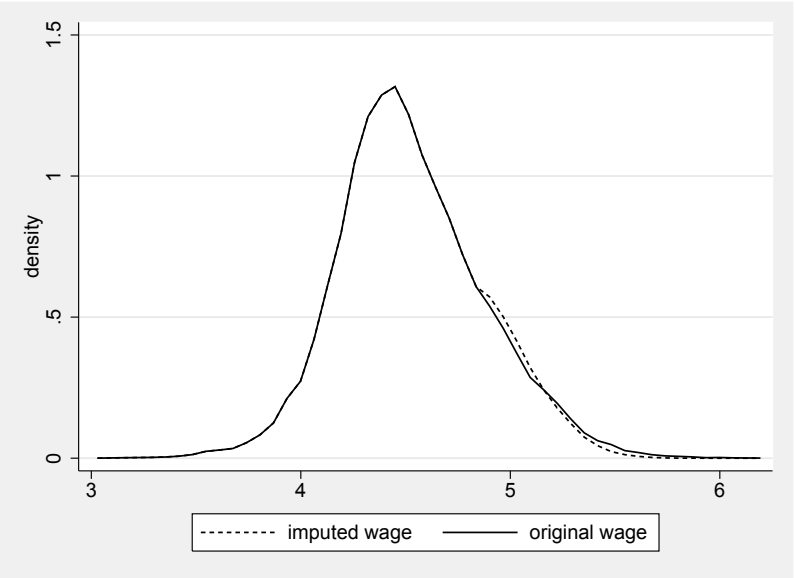

Figure 8.3: Kernel density estimates of original wage versus imputed wage

the German wage literature. To analyze the results we use the same model as the imputation model. In the first simulation study, we apply OLS regression in the analysis step. Therefore, the true parameters β are again obtained from an OLS regression based on the complete population using the analysis model. To give a first impression of the imputation quality, Figure 8.3 plots the original wages versus the imputed wages. The line referred to as imputed wage reflects the result of the first iteration of the simulation study. Up to the censoring point, the wages are identical as we only need to impute the censored values. The two lines that describe the censored part of the distribution indicate a good imputation quality, although they are not completely identical.

Table 8.5 shows the results of the corresponding 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 random samples before censoring - 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). To make it more explicit, the coverage is sometimes lower for industries with a rather little number of employees, i.e., industry 18, which refers to water supply, and industries with a high share of censored wages and a high dispersion

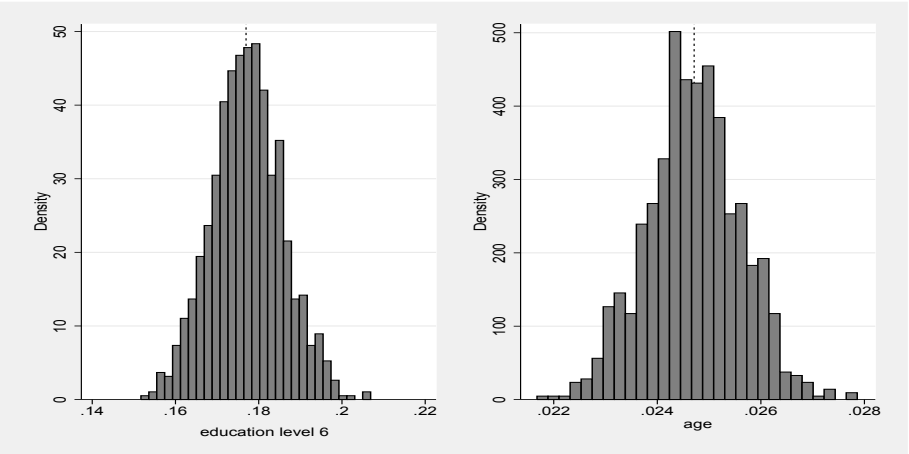

Figure 8.4: Distribution of $\widehat{\beta}_{MI}$ in the analysis step of the simulation study

of wages, i.e., industry 35, which refers to lawyers and market researchers. The highly aggregated region dummies in the GSES on the other hand can also be problematic for the imputation because there is a high wage dispersion within these regions. Nevertheless, taking into consideration the results of the first simulation study, it can be concluded that it is still advisable to use the multiple imputation approach considering heteroscedasticity to impute the missing wage information in the IABS.
Additionally, Figure 8.4 shows the distribution of the estimate $\widehat{\beta}_{MI}$, which is estimated for education level 6 (university degree) and age in the analysis step of each iteration of the simulation study using the wages imputed considering heteroscedasticity to illustrate the variation of this estimate over the 1000 iterations. The dashed line refers to the 'true' parameter based on the original complete data set, which is used as reference to calculate the coverage rate.

8.2.2 Simulation Study Based on a Cube Root Transformation

So far, we have assumed a log-normal distribution of the wages and have applied a log transformation of the wages because this transformation is standard in the German wage literature. To assume normality or log-normality of the distribution becomes especially problematic for the treatment of outliers. According to Gelman et al. (2003) the normal distribution is notoriously sensitive to outliers, what means that a single outlier can strongly affect the inference

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0345	0.0346	0.965	0.0352	0.962	0.0348	0.972
education3	0.0596	0.0592	0.963	0.0501	0.919	0.0530	0.943
education4	0.0713	0.0714	0.964	0.0639	0.860	0.0633	0.838
education5	0.1370	0.1374	0.962	0.1495	0.657	0.1462	0.799
education6	0.1770	0.1771	0.956	0.1776	0.951	0.1772	0.948
level2	0.0105	0.0106	0.962	0.0099	0.962	0.0095	0.967
level3	0.0371	0.0382	0.969	0.0399	0.961	0.0395	0.964
level4	0.0201	0.0212	0.977	0.0094	0.963	0.0055	0.940
group2	-0.0947	-0.0948	0.952	-0.0926	0.933	-0.0925	0.927
group3	-0.1899	-0.1897	0.947	-0.1866	0.902	-0.1866	0.891
group4	-0.3098	-0.3098	0.964	-0.3065	0.924	-0.3071	0.929
group5	0.3875	0.3863	0.969	0.3956	0.964	0.3866	0.967
group6	0.1412	0.1401	0.963	0.1488	0.964	0.1498	0.957
group7	0.0479	0.0469	0.971	0.0589	0.952	0.0613	0.943
group8	-0.1702	-0.1713	0.967	-0.1589	0.957	-0.1554	0.936
group9	-0.3394	-0.3400	0.969	-0.3280	0.954	-0.3253	0.945
age	0.0247	0.0247	0.961	0.0253	0.931	0.0247	0.976
sqage	-0.0003	-0.0003	0.958	-0.0003	0.866	-0.0003	0.969
region2	0.0369	0.0368	0.956	0.0463	0.300	0.0408	0.853
region3	0.0038	0.0037	0.947	0.0081	0.801	0.0062	0.930
region4	0.0517	0.0516	0.959	0.0529	0.944	0.0471	0.679
industry2	-0.0407	-0.0404	0.958	-0.0403	0.954	-0.0401	0.950
industry3	-0.1097	-0.1094	0.945	-0.1140	0.926	-0.1135	0.929
industry4	0.0053	0.0054	0.959	0.0044	0.964	0.0060	0.961
industry5	0.0765	0.0773	0.976	0.0729	0.950	0.0712	0.936
industry6	0.0788	0.0791	0.968	0.0827	0.950	0.0824	0.950
industry7	0.0636	0.0641	0.968	0.0701	0.860	0.0692	0.889
industry8	-0.0145	-0.0146	0.956	-0.0115	0.949	-0.0112	0.935
industry9	-0.0157	-0.0158	0.969	-0.0129	0.963	-0.0120	0.961
industry10	0.0252	0.0257	0.957	0.0301	0.903	0.0303	0.899
industry11	-0.0356	-0.0355	0.960	-0.0329	0.946	-0.0324	0.937
industry12	-0.0029	-0.0027	0.949	0.0015	0.890	0.0015	0.885
industry13	-0.0166	-0.0166	0.953	-0.0174	0.961	-0.0187	0.947
industry14	-0.0278	-0.0275	0.960	-0.0276	0.968	-0.0277	0.965
industry15	-0.0408	-0.0404	0.956	-0.0371	0.946	-0.0373	0.948
industry16	0.0341	0.0343	0.951	0.0369	0.950	0.0367	0.947
industry17	-0.0727	-0.0722	0.967	-0.0718	0.970	-0.0703	0.957
industry18	-0.0100	-0.0096	0.958	0.0047	0.426	0.0053	0.372
industry19	-0.0219	-0.0216	0.955	-0.0170	0.906	-0.0163	0.882
industry20	-0.1047	-0.1047	0.967	-0.1045	0.965	-0.1034	0.960
industry21	-0.0874	-0.0866	0.933	-0.0863	0.927	-0.0853	0.910
industry22	-0.1124	-0.1124	0.965	-0.1163	0.939	-0.1151	0.948
industry23	-0.0549	-0.0546	0.959	-0.0566	0.968	-0.0565	0.959
industry24	-0.1604	-0.1599	0.954	-0.1608	0.958	-0.1593	0.954
industry25	-0.2215	-0.2215	0.960	-0.2206	0.953	-0.2198	0.947
industry26	-0.0560	-0.0558	0.968	-0.0557	0.962	-0.0545	0.950
industry27	-0.0454	-0.0449	0.958	-0.0493	0.927	-0.0486	0.940
industry28	-0.0865	-0.0863	0.971	-0.0845	0.966	-0.0839	0.963
industry29	-0.0697	-0.0696	0.958	-0.0669	0.934	-0.0659	0.923
industry30	-0.0705	-0.0700	0.954	-0.0806	0.829	-0.0782	0.886
industry31	-0.0673	-0.0670	0.952	-0.0658	0.946	-0.0652	0.947
industry32	-0.0662	-0.0653	0.874	-0.0699	0.888	-0.0685	0.875
industry33	0.0112	0.0113	0.966	0.0114	0.955	0.0114	0.961
industry34	-0.0948	-0.0945	0.967	-0.0879	0.885	-0.0840	0.780
industry35	-0.0019	-0.0015	0.964	-0.0183	0.702	-0.0192	0.636
industry36	-0.2604	-0.2603	0.944	-0.2639	0.943	-0.2629	0.951
contract	-0.1114	-0.1117	0.941	-0.1139	0.948	-0.1111	0.958
cons	4.0440	4.0447	0.958	4.0331	0.936	4.0445	0.972

Table 8.5: Simulation results based on a lognormal transformation

for all the parameters in the model, even those with little connection to the outlying data point. This problem arises as well by using transformations of a normal distribution. When imputing right-censored wages, extremely high values are therefore an important issue. As it is questionable whether the log transformation is really an applicable assumption, we test another transformation, which is less sensitive to extreme values, the cube root of the wages. Schwartz (1985) for example shows that the cube root of income exhibits additional statistical properties that make it perhaps a more suitable transformation for multivariate analyses of income. Apart from applying this transformation the simulation design is the same as in the simulation study described before. That means we assume an imputation and analysis model containing the cube root of wages as dependent variable and as covariates

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type).*

Table 8.6 shows the results of the second simulation study. The coverage is for most variables again similar to the coverage received by the estimations using the complete random samples before censoring. But here we also have to state some coverage rates that are significantly lower. Especially the coverage of 0.151 for education level 6 received by the imputation approach assuming homoscedasticity of the residuals indicates a serious problem. It seems that this approach does not perform satisfyingly for this group with a high share of censored wages. Some rather low coverage rates can be again found for both approaches concerning some industry dummies. In conclusion, the coverage rates resulting from an imputation based on cube root transformed wages are somewhat lower than based on a log transformation. Consequently, the log transformation seems to be more appropriate for German wage data than the cube root transformation. Figure 8.5 confirms the finding that the log transformation is more appropriate for the German wage data. Normal Q-Q plots compare randomly generated, independent standard normal data on the vertical axis to the wage distribution of the different transformations and the original wages in the complete GSES on the horizontal axis. The linearity of the points suggests that the log transformed wages are approximately normally distributed, while the cube root transformed are a bit further away from being normally distributed.

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0489	0.0490	0.964	0.0493	0.966	0.0491	0.966
education3	0.0952	0.0947	0.959	0.0723	0.819	0.0800	0.890
education4	0.1106	0.1108	0.966	0.0902	0.544	0.0931	0.670
education5	0.2188	0.2195	0.959	0.2170	0.959	0.2259	0.920
education6	0.2893	0.2895	0.955	0.2560	0.151	0.2744	0.783
level2	0.0187	0.0189	0.967	0.0166	0.944	0.0165	0.942
level3	0.0633	0.0645	0.969	0.0641	0.971	0.0664	0.971
level4	0.0407	0.0419	0.972	0.0148	0.916	0.0116	0.905
group2	-0.1416	-0.1418	0.951	-0.1368	0.873	-0.1366	0.873
group3	-0.2767	-0.2764	0.951	-0.2693	0.823	-0.2690	0.817
group4	-0.4271	-0.4271	0.967	-0.4214	0.914	-0.4209	0.906
group5	0.6298	0.6285	0.968	0.5864	0.769	0.5982	0.869
group6	0.2103	0.2091	0.967	0.2225	0.960	0.2256	0.946
group7	0.0634	0.0623	0.970	0.0888	0.918	0.0912	0.905
group8	-0.2590	-0.2603	0.963	-0.2315	0.898	-0.2285	0.884
group9	-0.4817	-0.4822	0.970	-0.4547	0.925	-0.4521	0.911
age	0.0352	0.0352	0.956	0.0357	0.944	0.0352	0.960
sqage	-0.0004	-0.0004	0.954	-0.0004	0.846	-0.0004	0.947
region2	0.0485	0.0484	0.956	0.0640	0.134	0.0592	0.478
region3	0.0029	0.0029	0.945	0.0116	0.597	0.0074	0.863
region4	0.0765	0.0763	0.954	0.0737	0.916	0.0699	0.689
industry2	-0.0592	-0.0588	0.958	-0.0567	0.944	-0.0570	0.949
industry3	-0.1598	-0.1595	0.941	-0.1626	0.942	-0.1634	0.942
industry4	0.0055	0.0057	0.962	0.0052	0.964	0.0061	0.962
industry5	0.1316	0.1327	0.974	0.1112	0.832	0.1138	0.871
industry6	0.1214	0.1218	0.966	0.1246	0.958	0.1245	0.962
industry7	0.1009	0.1017	0.969	0.1039	0.952	0.1046	0.956
industry8	-0.0276	-0.0277	0.963	-0.0197	0.869	-0.0200	0.885
industry9	-0.0292	-0.0294	0.967	-0.0203	0.930	-0.0201	0.929
industry10	0.0330	0.0339	0.958	0.0430	0.862	0.0429	0.870
industry11	-0.0564	-0.0561	0.962	-0.0483	0.866	-0.0485	0.886
industry12	-0.0045	-0.0041	0.957	0.0040	0.835	0.0037	0.845
industry13	-0.0198	-0.0197	0.952	-0.0238	0.939	-0.0245	0.941
industry14	-0.0403	-0.0398	0.960	-0.0403	0.965	-0.0402	0.969
industry15	-0.0605	-0.0600	0.958	-0.0516	0.916	-0.0523	0.915
industry16	0.0479	0.0481	0.954	0.0521	0.947	0.0523	0.945
industry17	-0.1104	-0.1098	0.967	-0.1037	0.931	-0.1036	0.924
industry18	-0.0208	-0.0203	0.956	0.0099	0.111	0.0097	0.127
industry19	-0.0382	-0.0377	0.955	-0.0262	0.776	-0.0265	0.793
industry20	-0.1548	-0.1548	0.962	-0.1492	0.943	-0.1494	0.950
industry21	-0.1257	-0.1248	0.954	-0.1194	0.935	-0.1197	0.942
industry22	-0.1675	-0.1674	0.966	-0.1677	0.956	-0.1685	0.958
industry23	-0.0791	-0.0785	0.960	-0.0804	0.965	-0.0810	0.961
industry24	-0.2406	-0.2401	0.959	-0.2317	0.882	-0.2327	0.908
industry25	-0.3112	-0.3112	0.952	-0.3033	0.929	-0.3042	0.938
industry26	-0.0858	-0.0856	0.968	-0.0807	0.926	-0.0809	0.936
industry27	-0.0667	-0.0660	0.958	-0.0710	0.934	-0.0714	0.936
industry28	-0.1319	-0.1316	0.967	-0.1231	0.907	-0.1240	0.922
industry29	-0.1090	-0.1087	0.956	-0.0973	0.809	-0.0982	0.839
industry30	-0.1085	-0.1077	0.957	-0.1153	0.928	-0.1173	0.919
industry31	-0.1053	-0.1048	0.951	-0.0929	0.865	-0.0953	0.907
industry32	-0.0772	-0.0763	0.947	-0.0786	0.949	-0.0794	0.956
industry33	0.0307	0.0309	0.962	0.0258	0.935	0.0263	0.944
industry34	-0.1535	-0.1531	0.963	-0.1247	0.335	-0.1260	0.429
industry35	0.0101	0.0106	0.961	-0.0237	0.420	-0.0237	0.440
industry36	-0.3403	-0.3401	0.951	-0.3397	0.946	-0.3409	0.947
contract	-0.1462	-0.1464	0.961	-0.1476	0.956	-0.1473	0.966
cons	3.8243	3.8247	0.955	3.8172	0.953	3.8273	0.958

Table 8.6: Simulation results based on a cube root transformation

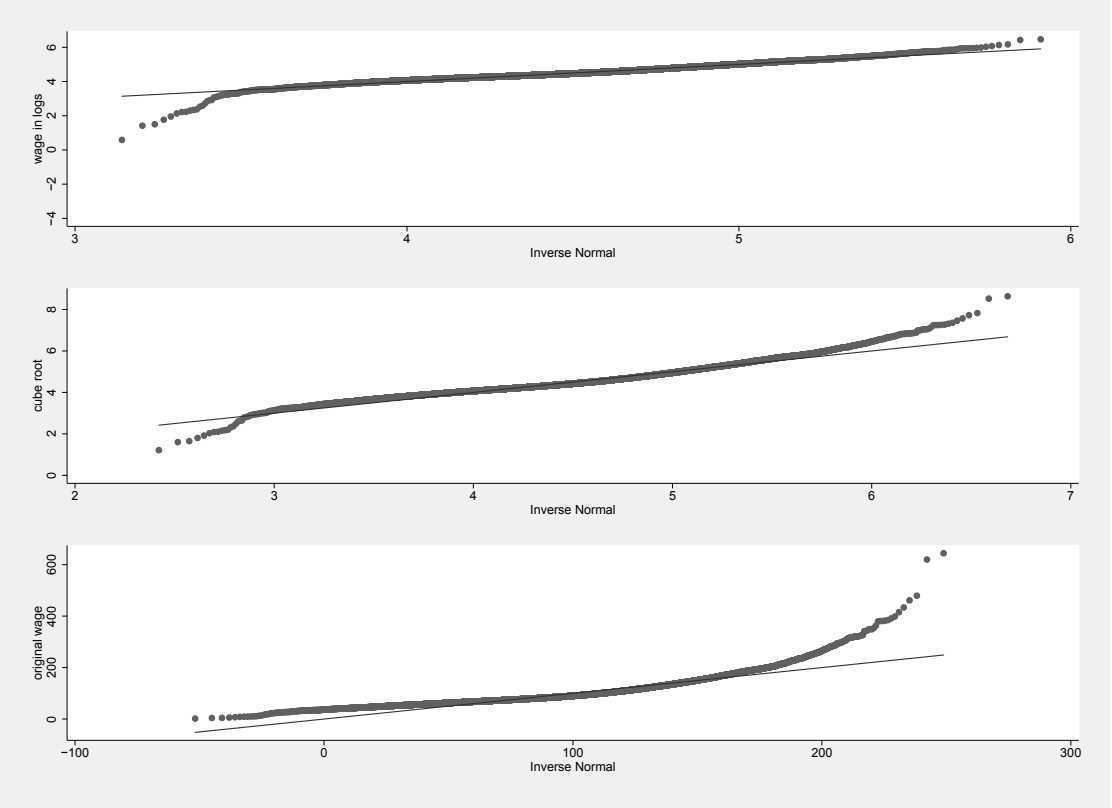

Figure 8.5: Normal Q-Q plot comparing randomly generated, independent standard normal data to the wage distribution

Additionally a simulation study using the original wages without any transformation was performed. Here, we receive an imputation quality that is considerably lower. These results indicate the relevance of an appropriate transformation. The corresponding simulation results can be found in Table A.1 in the appendix.

8.2.3 GLS Estimation in the Analysis Step

As we assume a heteroscedastic distribution of the residuals, one could argue that not only the imputation step has to be done considering heteroscedasticity, but the analysis step also has to be done based on generalized least squares regression. The impact of a questionable homoscedasticity assumption is somewhat less severe in this case, as censoring does not play a role in the analysis step anymore. Because we observe the entire wage distribution now, we do not have to make assumptions on the distribution of the residuals based only on the lower part of the distribution. Another reason to apply OLS in the analysis step is that in most studies concerning wages based on the IAB Employment Sample homoscedasticity of the residuals is assumed. Therefore, when we simulate an analysis that is typically done withe these data, it makes sense to apply OLS regression. Nevertheless, there are reasonable arguments to repeat the simulation study with this variation in the analysis step. Apart from the estimation based on GLS regression in the analysis step, the same simulation design as in the simulation study based on a log transformation is applied. As imputation and analysis model we use again the model containing the wages in logs as dependent variable and as covariates

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type).*

The true parameters β are now obtained from a GLS regression based on the complete population using the analysis model. Table 8.7 shows the results of this simulation study. The results indicate that there is no significant difference in the coverages whether an OLS regression or an GLS regression is applied in the analysis step. Qualitatively the results show the same imputation quality, whereas results of the approach considering heteroscedasticity are superior compared to the approach assuming homoscedasticity of the residuals. Some coverage rates turn out a little smaller in this simulation study compared to the

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0323	0.0328	0.945	0.0333	0.939	0.0328	0.952
education3	0.0570	0.0562	0.957	0.0490	0.929	0.0524	0.951
education4	0.0685	0.0685	0.963	0.0629	0.904	0.0625	0.885
education5	0.1347	0.1353	0.953	0.1484	0.572	0.1447	0.755
education6	0.1748	0.1749	0.947	0.1773	0.938	0.1760	0.938
level2	0.0143	0.0142	0.951	0.0135	0.944	0.0133	0.945
level3	0.0448	0.0456	0.938	0.0463	0.944	0.0463	0.945
level4	0.0264	0.0271	0.950	0.0156	0.925	0.0124	0.895
group2	-0.0940	-0.0941	0.961	-0.0921	0.953	-0.0916	0.941
group3	-0.1912	-0.1904	0.950	-0.1876	0.898	-0.1874	0.889
group4	-0.3082	-0.3073	0.931	-0.3038	0.874	-0.3039	0.876
group5	0.3809	0.3802	0.947	0.3919	0.911	0.3801	0.937
group6	0.1332	0.1326	0.942	0.1431	0.918	0.1431	0.912
group7	0.0458	0.0450	0.947	0.0564	0.916	0.0589	0.899
group8	-0.1769	-0.1779	0.943	-0.1646	0.915	-0.1615	0.884
group9	-0.3439	-0.3453	0.927	-0.3319	0.904	-0.3300	0.885
age	0.0257	0.0255	0.928	0.0263	0.885	0.0258	0.921
sqage	-0.0003	-0.0003	0.929	-0.0003	0.842	-0.0003	0.922
region2	0.0363	0.0365	0.948	0.0453	0.330	0.0391	0.885
region3	0.0046	0.0051	0.941	0.0091	0.789	0.0070	0.910
region4	0.0510	0.0508	0.947	0.0523	0.932	0.0465	0.668
industry2	-0.0425	-0.0425	0.947	-0.0406	0.944	-0.0395	0.930
industry3	-0.1107	-0.1106	0.940	-0.1142	0.947	-0.1122	0.952
industry4	0.0053	0.0052	0.961	0.0052	0.967	0.0074	0.960
industry5	0.0800	0.0808	0.936	0.0760	0.920	0.0731	0.884
industry6	0.0764	0.0771	0.952	0.0802	0.928	0.0792	0.938
industry7	0.0643	0.0645	0.980	0.0714	0.876	0.0713	0.870
industry8	-0.0210	-0.0207	0.975	-0.0172	0.941	-0.0170	0.939
industry9	-0.0147	-0.0152	0.976	-0.0125	0.979	-0.0113	0.974
industry10	0.0246	0.0250	0.961	0.0293	0.921	0.0296	0.919
industry11	-0.0345	-0.0345	0.969	-0.0320	0.957	-0.0307	0.933
industry12	-0.0017	-0.0017	0.961	0.0031	0.898	0.0039	0.855
industry13	-0.0156	-0.0162	0.958	-0.0154	0.966	-0.0151	0.962
industry14	-0.0289	-0.0293	0.967	-0.0272	0.966	-0.0259	0.958
industry15	-0.0397	-0.0398	0.971	-0.0355	0.953	-0.0344	0.937
industry16	0.0332	0.0335	0.957	0.0356	0.958	0.0352	0.960
industry17	-0.0732	-0.0727	0.963	-0.0711	0.960	-0.0692	0.938
industry18	-0.0041	-0.0030	0.967	0.0077	0.650	0.0081	0.617
industry19	-0.0230	-0.0226	0.971	-0.0173	0.922	-0.0164	0.897
industry20	-0.1077	-0.1079	0.961	-0.1053	0.961	-0.1030	0.940
industry21	-0.0874	-0.0854	0.919	-0.0846	0.919	-0.0829	0.909
industry22	-0.1118	-0.1117	0.957	-0.1147	0.945	-0.1126	0.955
industry23	-0.0547	-0.0547	0.966	-0.0558	0.967	-0.0547	0.958
industry24	-0.1623	-0.1619	0.953	-0.1635	0.956	-0.1626	0.957
industry25	-0.2215	-0.2210	0.918	-0.2191	0.911	-0.2167	0.901
industry26	-0.0574	-0.0569	0.961	-0.0573	0.963	-0.0566	0.960
industry27	-0.0495	-0.0486	0.958	-0.0524	0.949	-0.0521	0.949
industry28	-0.0838	-0.0837	0.976	-0.0834	0.972	-0.0824	0.970
industry29	-0.0634	-0.0637	0.971	-0.0616	0.967	-0.0605	0.962
industry30	-0.0800	-0.0790	0.943	-0.0852	0.940	-0.0827	0.957
industry31	-0.0679	-0.0672	0.959	-0.0651	0.947	-0.0640	0.941
industry32	-0.0563	-0.0542	0.798	-0.0603	0.771	-0.0566	0.806
industry33	0.0175	0.0171	0.937	0.0196	0.940	0.0205	0.926
industry34	-0.0864	-0.0864	0.956	-0.0805	0.918	-0.0767	0.810
industry35	-0.0005	-0.0005	0.944	-0.0170	0.679	-0.0173	0.631
industry36	-0.2482	-0.2480	0.847	-0.2504	0.841	-0.2455	0.814
contract	-0.1054	-0.1054	0.940	-0.1102	0.951	-0.1076	0.965
cons	4.0173	4.0206	0.954	4.0054	0.923	4.0141	0.948

Table 8.7: Simulation results based on GLS estimation in analysis step

one based on a log transformation, but this can be explained by the complete data coverage (before censoring), which is in general a little lower here.

8.2.4 Reduced Set of Variables in the Model

In Chapter 6, we have discussed that the imputer's model should include as rich a set of variables in the imputation model as possible in order to accommodate the variety of analyses that might be carried out by the analyst. That is why a rather rich set of variables was chosen for the first simulation studies using external data. Now, we will examine how a suitable imputation model may look like if it is known that the analyst only wants to analyze a limited set of variables. The usual advice for building up an imputation model is to use as many variables as are available (see, e.g., Rässler et al. (2008)). However, including variables with no influence on the missingness mechanism will add unnecessary noise and variation to the MI estimates. Therefore, we were interested in figuring out whether a smaller imputation model would work in this case, too.
In order to select an appropriate small model, we estimated a probit model with a dependent variable y, i.e., $y = 1$ if the observation is censored and $y = 0$ if not. The education categories, the contract type dummy, age, and age^2 turned out to be significant. Therefore, we suppose here the following, rather simple, model with wages in logs as dependent variable and

X_{small}=*(age, age^2, 6 education categories, contract type).*

In a first step we perform a simulation study using the restricted model as imputation and analysis model. Now, we apply again OLS regression in the analysis step. Of course, we receive the true parameters β now from an OLS regression based on the complete population using the small analysis model. We use again a lognormal transformation for the wages, because this transformations seems to be more appropriate. Besides these points we use the same simulation design as described before in this section.
Looking at the results, we find here coverage rates in a range comparable to the simulation studies presented before. We obtain again a higher coverage using the imputation approach considering heteroscedasticity compared to the approach assuming homoscedasticity (see Table 8.8). Especially the results concerning the dummies for highly-skilled employees, where the fraction of

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.1886	0.1888	0.970	0.1851	0.884	0.1858	0.920
education3	0.3275	0.3287	0.955	0.3098	0.890	0.3297	0.958
education4	0.4059	0.4058	0.962	0.3913	0.678	0.3989	0.898
education5	0.5780	0.5780	0.963	0.5934	0.557	0.5873	0.856
education6	0.6383	0.6385	0.967	0.6455	0.861	0.6466	0.913
age	0.0411	0.0411	0.965	0.0416	0.947	0.0408	0.976
sqage	-0.0004	-0.0004	0.970	-0.0004	0.903	-0.0004	0.984
contract	-0.2067	-0.2072	0.951	-0.2087	0.954	-0.2054	0.958
cons	3.5664	3.5661	0.968	3.5649	0.969	3.5718	0.976

Table 8.8: Results of a simulation study using a limited set of variables

censored wages is eminently high, are much better using the approach considering heteroscedasticity. In conclusion, a comprehensive imputation model containing all available variables seems not always to be necessary to impute the missing wage information, even when we want to analyze the effects of sensitive (in the sense of censoring) variables like for example education groups. To check whether the imputation results can still be improved, we modify the simulation design. We draw again randomly a 10 percent sample from the complete sample, define the threshold and delete the wages above this limit. Then we decompose the sample into three education groups:

- Low-skilled (Low/intermediate school or vocational training)
- Medium-skilled (Upper school with or without vocational training)
- High-skilled (Technical college or university degree)

The deleted wage information is now imputed again multiply using the same restricted imputation model, but separately in these subgroups. Afterwards the groups are combined again and the imputation quality is analyzed like in the simulation studies above. Table 8.9 shows the results of this simulation study, which can be summarized as follows: First, the coverage using the imputation approach considering heteroscedasticity is, as before, higher compared to the approach assuming homoscedasticity. But second, imputing the wage in groups does not improve the imputation results. Additional simulation studies

have shown, that if a cube root transformation is applied, this approach of decomposing the data set in subgroups is superior, even if the imputation quality is overall considerably lower using a cube root transformation. A drawback of the strategy to impute the wages in education groups is that it often cannot be performed if a set of dummies for industries or occupations on a disaggregated level is included in the imputation model. Here, often a situation appears where the share of persons of high (or low) education groups is zero for some industries or occupations.
Another possibility to impute the data for a research question, where a rather small analyst's model will be analyzed, is to use a set of variables as rich as possible. In a third simulation study we use the larger imputation model, which explains the wage by 66 percent[2]. This model contains as covariates again

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type).*

To analyze the imputed data we apply OLS regression using the smaller model containing

X_{small}=*(age, age^2, 6 education categories, contract type).*

That means the true parameters β are here obtained from an OLS regression using this smaller model. The results of this simulation study (Table 8.10) indicate that using a rich set of variables for the imputation model and then performing the desired analysis using a potentially smaller model is at first glance a less promising strategy in the simulated case. We again receive coverage rates of up to 97 percent, but for some education levels we receive for both approaches lower coverage rates (e.g., 0.501 in case of the heteroscedastic approach and 0.717 in case of the homoscedastic approach for education level 4). In general, it is noticeable that the coverage rates are somewhat lower compared to the larger analyst's model containing additional control variables and the superiority of the approach considering heteroscedasticity is less evident in this case. Summing up, we find that using a rich set of variables in the imputation model leads to a somewhat lower imputation quality if the analyst's

[2] R^2 of an OLS-regression using the original complete data set and the variables of the imputation model.

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.1886	0.1886	0.957	0.1848	0.868	0.1856	0.915
education3	0.3275	0.3276	0.967	0.3185	0.941	0.3426	0.954
education4	0.4059	0.4061	0.959	0.4031	0.952	0.4104	0.965
education5	0.5780	0.5783	0.954	0.5982	0.421	0.6166	0.448
education6	0.6383	0.6387	0.961	0.6516	0.744	0.6786	0.586
age	0.0411	0.0410	0.967	0.0419	0.921	0.0402	0.951
sqage	-0.0004	-0.0004	0.962	-0.0004	0.885	-0.0004	0.930
contract	-0.2067	-0.2069	0.947	-0.2089	0.962	-0.2068	0.956
cons	3.5664	3.5679	0.970	3.5588	0.961	3.5766	0.970

Table 8.9: Results of an imputation in education groups

interest is in examining a reduced model. On the other hand, an advantage of applying a rich set of variables is that the once imputed data can be used for several research questions. Based on the data completed using the imputation model that is applied in the last simulation study, a wide range of different models could be estimated.

Some further models were evaluated to check if the usually recommended procedure of using a rich set of variables in the imputation model for analyzing a smaller model is a generally appropriate approach also for the case of censoring. Table 8.11 shows the results of a simulation study where job levels categories were included in the analyst's model instead of the contract type dummy. The results indicate as well that applying a larger imputation model is a recommendable strategy, because the completed data can be used for various purposes, whereas only a little reduction of the imputation quality is potentially to be expected.

Accordingly, the best imputation strategy regarding censored data depends on the purpose of the imputed data. If the data are to be imputed for a single research question, a restricted imputation model might be sufficient. Note that this finding is not generally valid for the imputation of missing data. In the case of censoring the question whether a value is censored or not depends on the values itself, which is different from other cases of missing data. Nevertheless, to impute the censored values, we need to include covariates in the imputation model that have explanatory power for the missing wages. If the research

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.1886	0.1887	0.958	0.1873	0.949	0.1858	0.914
education3	0.3275	0.3273	0.952	0.3131	0.912	0.3118	0.906
education4	0.4059	0.4061	0.965	0.3929	0.717	0.3877	0.501
education5	0.5780	0.5782	0.953	0.5858	0.815	0.5749	0.946
education6	0.6383	0.6385	0.966	0.6329	0.866	0.6252	0.693
age	0.0411	0.0410	0.959	0.0418	0.926	0.0410	0.973
sqage	-0.0004	-0.0004	0.964	-0.0004	0.895	-0.0004	0.972
contract	-0.2067	-0.2067	0.941	-0.2081	0.941	-0.2042	0.940
cons	3.5664	3.5674	0.968	3.5598	0.967	3.5705	0.980

Table 8.10: Results based on a large imputation model and a small analyst's model - Example 1

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0700	0.0700	0.965	0.0699	0.966	0.0697	0.970
education3	0.1142	0.1141	0.961	0.1030	0.925	0.1050	0.935
education4	0.1578	0.1580	0.961	0.1484	0.853	0.1470	0.817
education5	0.3149	0.3151	0.952	0.3266	0.752	0.3198	0.940
education6	0.3674	0.3675	0.955	0.3659	0.948	0.3620	0.925
level2	0.0807	0.0808	0.961	0.0800	0.950	0.0799	0.954
level3	0.3548	0.3549	0.962	0.3649	0.805	0.3629	0.862
level4	0.3049	0.3050	0.960	0.3005	0.830	0.2957	0.514
age	0.0429	0.0429	0.958	0.0436	0.928	0.0428	0.965
sqage	-0.0004	-0.0004	0.958	-0.0004	0.896	-0.0004	0.968
cons	3.2956	3.2967	0.956	3.2880	0.949	3.3038	0.953

Table 8.11: Results based on a large imputation model and a small analyst's model - Example 2

question is not known to the imputer or the data are to be used for several analyses, all available variables should be used for the imputation model.

8.2.5 Differing Imputer's and Analyst's Models

In general, a situation where the analyst's and the imputer's model differ is called uncongeniality according to Meng (1994). In the preceding simulation study we examined a situation where the analyst is only interested in a subset of the variables of the imputation model like shown in Table 8.10. Another situation appears if the analyst wants to include variables in the analysis that were not used in the imputation model. Note that if the imputation model does not contain all important correlates of variables with missing data, i.e., variables that might explain the missing data mechanism or are correlated with variables with missing data, here the wage variable, the results will be biased. It is intuitively obvious that, if the imputation model does not contain variables of the analysis model, the correlation between these variables cannot be reflected in the imputed values. If the aim of a study is for example to examine the influence of the firm size on the individual wage level, in general the firm information should be used to impute the individual wages. Otherwise the impact of the firm size on the wage based on the imputed data might be biased towards zero.
We examine a special case of the situation described above. We use the large imputation model of the simulation study based on a log transformation containing the wages in logs as dependent variable and the covariates

$$X_{large}=(age,\ age^2,\ 6\ education\ categories,\ 4\ job\ level\ categories,\ 9\ performance\ groups,\ 4\ region\ dummies,\ 36\ industry\ dummies,\ contract\ type).$$

In the analysis step we drop the dummies for job levels, replace the industry dummies by occupation dummies and replace age and squared age by tenure and squared tenure. Tenure is defined as years employed in the current establishment. Accordingly, the true parameters β are here obtained from an OLS regression using this analysis model. The idea is to check whether a real differing imputation model still allows valid conclusions when the differing variables are highly correlated. As age is a good predictor for tenure and the occupation of the employee is also correlated with other variables like, e.g., the industry and education, the chosen imputation model might be also applicable for the

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
educ2	0.1035	0.1034	0.964	0.1033	0.964	0.1029	0.961
educ3	0.1534	0.1533	0.963	0.1436	0.937	0.1457	0.950
educ4	0.1938	0.1938	0.958	0.1858	0.876	0.1845	0.849
educ5	0.3075	0.3075	0.959	0.3148	0.907	0.3097	0.962
educ6	0.3745	0.3743	0.960	0.3724	0.941	0.3695	0.932
tenure	0.0188	0.0188	0.958	0.0186	0.953	0.0184	0.915
sqten	-0.0004	-0.0004	0.964	-0.0003	0.958	-0.0003	0.945
occupation2	0.0312	0.0306	0.951	0.0259	0.955	0.0265	0.954
occupation3	0.0678	0.0686	0.949	0.0643	0.957	0.0642	0.959
occupation4	0.1451	0.1450	0.950	0.1410	0.959	0.1408	0.960
occupation5	0.0664	0.0656	0.956	0.0612	0.960	0.0611	0.959
occupation6	0.0739	0.0727	0.956	0.0684	0.959	0.0683	0.963
occupation7	0.2221	0.2217	0.957	0.2151	0.963	0.2143	0.960
occupation8	0.1450	0.1449	0.949	0.1409	0.959	0.1409	0.958
occupation9	0.1179	0.1177	0.947	0.1129	0.957	0.1129	0.953
occupation10	0.0620	0.0614	0.953	0.0562	0.959	0.0564	0.959
occupation11	0.0912	0.0906	0.951	0.0849	0.954	0.0849	0.955
occupation12	0.1127	0.1122	0.950	0.1058	0.957	0.1057	0.955
occupation13	0.1069	0.1068	0.944	0.1025	0.955	0.1022	0.956
occupation14	0.0975	0.0965	0.958	0.0909	0.957	0.0907	0.956
occupation15	0.0763	0.0758	0.949	0.0710	0.955	0.0709	0.955
occupation16	-0.0453	-0.0460	0.963	-0.0506	0.965	-0.0507	0.963
occupation17	-0.0062	-0.0067	0.945	-0.0113	0.951	-0.0114	0.951
occupation18	0.1200	0.1197	0.948	0.1153	0.955	0.1152	0.957
occupation19	0.1005	0.0999	0.958	0.0958	0.960	0.0957	0.962
occupation20	0.1414	0.1409	0.951	0.1367	0.952	0.1364	0.952
occupation21	0.0506	0.0495	0.965	0.0449	0.970	0.0446	0.969
occupation22	0.0607	0.0597	0.959	0.0546	0.963	0.0546	0.963
occupation23	0.0069	0.0067	0.953	0.0024	0.962	0.0025	0.959
occupation24	0.0416	0.0411	0.951	0.0350	0.955	0.0352	0.956
occupation25	0.0653	0.0649	0.952	0.0614	0.962	0.0609	0.961
occupation26	-0.0620	-0.0627	0.950	-0.0677	0.954	-0.0683	0.950
occupation27	0.1682	0.1680	0.953	0.1650	0.955	0.1649	0.957
occupation28	0.4044	0.4040	0.951	0.4039	0.957	0.3960	0.956
occupation29	0.4128	0.4120	0.955	0.4184	0.961	0.4115	0.965
occupation30	0.3489	0.3486	0.949	0.3492	0.953	0.3445	0.955
occupation31	0.2321	0.2323	0.948	0.2331	0.958	0.2299	0.957
occupation32	0.2930	0.2926	0.948	0.2791	0.950	0.2740	0.940
occupation33	0.1145	0.1139	0.944	0.0981	0.945	0.0957	0.938
occupation34	0.4288	0.4290	0.956	0.4097	0.939	0.4021	0.906
occupation35	0.3177	0.3174	0.959	0.3038	0.949	0.2987	0.932
occupation36	0.1898	0.1897	0.958	0.1857	0.963	0.1841	0.959
occupation37	0.4077	0.4075	0.964	0.3786	0.922	0.3712	0.877
occupation38	0.2865	0.2851	0.959	0.2607	0.937	0.2563	0.931
occupation39	0.0720	0.0720	0.950	0.0676	0.958	0.0677	0.956
occupation40	0.1742	0.1742	0.961	0.1489	0.928	0.1475	0.921
occupation41	-0.0591	-0.0598	0.950	-0.0639	0.956	-0.0641	0.954
occupation42	0.0078	0.0075	0.948	0.0031	0.960	0.0029	0.955
occupation43	0.5205	0.5205	0.945	0.4639	0.687	0.4564	0.591
occupation44	0.4884	0.4885	0.964	0.4462	0.887	0.4388	0.827
occupation45	0.3470	0.3466	0.956	0.3397	0.959	0.3329	0.952
occupation46	0.4292	0.4287	0.948	0.4238	0.959	0.4165	0.948
occupation47	0.2463	0.2460	0.948	0.2411	0.959	0.2377	0.955
occupation48	0.1759	0.1746	0.957	0.1663	0.957	0.1632	0.952
occupation49	-0.0121	-0.0122	0.950	-0.0169	0.956	-0.0172	0.956
occupation50	0.5138	0.5164	0.943	0.4269	0.788	0.4196	0.745
occupation51	0.4542	0.4553	0.955	0.4445	0.965	0.4346	0.939
occupation52	0.2563	0.2552	0.953	0.2474	0.956	0.2438	0.957
occupation53	0.2829	0.2857	0.946	0.2617	0.960	0.2592	0.951
occupation54	0.1676	0.1673	0.955	0.1617	0.962	0.1612	0.959
occupation55	0.3110	0.3093	0.961	0.3230	0.965	0.3196	0.964
occupation56	0.4334	0.4356	0.950	0.4189	0.971	0.4083	0.947
occupation57	0.2546	0.2561	0.949	0.2741	0.957	0.2750	0.962
occupation58	-0.0601	-0.0609	0.962	-0.0670	0.957	-0.0678	0.961
occupation59	0.1214	0.1195	0.957	0.1237	0.955	0.1212	0.956
occupation60	-0.0303	-0.0312	0.953	-0.0359	0.961	-0.0362	0.958
occupation61	0.0649	0.0645	0.961	0.0592	0.962	0.0588	0.962
occupation62	0.0725	0.0716	0.955	0.0672	0.957	0.0672	0.957
occupation63	0.0109	0.0108	0.962	0.0064	0.964	0.0061	0.963
occupation64	0.0033	0.0031	0.952	-0.0006	0.957	-0.0009	0.957
occupation65	0.0760	0.0750	0.954	0.0715	0.958	0.0713	0.957
occupation66	0.0019	0.0016	0.956	-0.0027	0.962	-0.0027	0.961
contract	-0.1315	-0.1315	0.940	-0.1339	0.940	-0.1307	0.951
cons	4.2443	4.2448	0.947	4.2528	0.948	4.2506	0.957

Table 8.12: Results of a simulation study with differing imputation and analysis models

chosen analysis model. Table 8.12 shows the results of this simulation study, which affirm the expectation that there is no reduction in the data utility for the research (analysis) question in this case. The coverage rates range from 0.591 to 0.964, whereas most of them lie again around 0.95. It is noticeable that there is no particular difference in the coverage rate between variables that are only included in the analysis model and variables that are included in both models. In Appendix A.2, we use the measure of confidence interval overlap (see, e.g., Karr et al. (2006)) to examine situations where the analysis model contains variables that are not included in the imputation model. The examples in Appendix A.2 illustrate that a differing analysis model not necessarily has a negative influence on the quality of the estimation results, but in some cases it may lead to seriously biased results compared to results based on the original complete data set. The results based on multiply imputed data are also compared to results from a tobit estimation.

8.2.6 Different Transformations in the Imputer's and Analyst's Model

For this simulation studies, we applied mainly two different transformations of the wage: Log transformation and cube root transformation. The goal is to take that data set that is skewed to the right and transform it to a data set that is bell-shaped. Whereas the log transformation has generally more impact on skewness, the cube root transformation is less sensitive to outliers. The simulation studies showed that the log transformation is somewhat more appropriate in our case. To check if imputations based on these transformations are robust irrespective of the specific transformation, additional simulation studies are performed with differing transformations in the imputer's and analyst's model. To begin with, the missing wages are imputed based on a log transformation and the subsequent analysis step is performed with a cube root transformation. That means the wages were re-transformed after the imputation step. For this simulation study we use again the model containing the wages in logs and the covariates

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type)*

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0489	0.0490	0.964	0.0501	0.966	0.0493	0.968
education3	0.0952	0.0947	0.959	0.0776	0.899	0.0830	0.918
education4	0.1106	0.1108	0.966	0.0963	0.797	0.0951	0.752
education5	0.2188	0.2195	0.959	0.2425	0.566	0.2359	0.765
education6	0.2893	0.2895	0.955	0.2899	0.951	0.2888	0.940
level2	0.0187	0.0189	0.967	0.0175	0.963	0.0169	0.955
level3	0.0633	0.0645	0.969	0.0679	0.967	0.0670	0.971
level4	0.0407	0.0419	0.972	0.0198	0.940	0.0125	0.908
group2	-0.1416	-0.1418	0.951	-0.1377	0.911	-0.1376	0.905
group3	-0.2767	-0.2764	0.951	-0.2706	0.862	-0.2705	0.865
group4	-0.4271	-0.4271	0.967	-0.4208	0.909	-0.4219	0.918
group5	0.6298	0.6285	0.968	0.6443	0.949	0.6272	0.967
group6	0.2103	0.2091	0.967	0.2248	0.955	0.2268	0.945
group7	0.0634	0.0623	0.970	0.0846	0.941	0.0893	0.918
group8	-0.2590	-0.2603	0.963	-0.2369	0.940	-0.2304	0.900
group9	-0.4817	-0.4822	0.970	-0.4594	0.939	-0.4545	0.923
age	0.0352	0.0352	0.956	0.0364	0.872	0.0352	0.958
sqage	-0.0004	-0.0004	0.954	-0.0004	0.782	-0.0004	0.944
region2	0.0485	0.0484	0.956	0.0668	0.083	0.0565	0.705
region3	0.0029	0.0029	0.945	0.0115	0.671	0.0078	0.828
region4	0.0765	0.0763	0.954	0.0789	0.929	0.0679	0.538
industry2	-0.0592	-0.0588	0.958	-0.0586	0.955	-0.0583	0.952
industry3	-0.1598	-0.1595	0.941	-0.1681	0.920	-0.1672	0.913
industry4	0.0055	0.0057	0.962	0.0036	0.965	0.0066	0.961
industry5	0.1316	0.1327	0.974	0.1243	0.948	0.1211	0.928
industry6	0.1214	0.1218	0.966	0.1285	0.952	0.1281	0.947
industry7	0.1009	0.1017	0.969	0.1130	0.832	0.1112	0.876
industry8	-0.0276	-0.0277	0.963	-0.0218	0.930	-0.0213	0.925
industry9	-0.0292	-0.0294	0.967	-0.0239	0.963	-0.0223	0.956
industry10	0.0330	0.0339	0.958	0.0424	0.895	0.0427	0.884
industry11	-0.0564	-0.0561	0.962	-0.0511	0.934	-0.0501	0.914
industry12	-0.0045	-0.0041	0.957	0.0042	0.868	0.0041	0.850
industry13	-0.0198	-0.0197	0.952	-0.0207	0.962	-0.0232	0.955
industry14	-0.0403	-0.0398	0.960	-0.0400	0.971	-0.0401	0.970
industry15	-0.0605	-0.0600	0.958	-0.0533	0.944	-0.0538	0.950
industry16	0.0479	0.0481	0.954	0.0534	0.950	0.0530	0.945
industry17	-0.1104	-0.1098	0.967	-0.1089	0.974	-0.1060	0.950
industry18	-0.0208	-0.0203	0.956	0.0068	0.287	0.0079	0.208
industry19	-0.0382	-0.0377	0.955	-0.0289	0.878	-0.0277	0.839
industry20	-0.1548	-0.1548	0.962	-0.1544	0.969	-0.1522	0.964
industry21	-0.1257	-0.1248	0.954	-0.1240	0.956	-0.1222	0.953
industry22	-0.1675	-0.1674	0.966	-0.1747	0.917	-0.1724	0.941
industry23	-0.0791	-0.0785	0.960	-0.0825	0.968	-0.0824	0.960
industry24	-0.2406	-0.2401	0.959	-0.2417	0.958	-0.2387	0.957
industry25	-0.3112	-0.3112	0.952	-0.3095	0.956	-0.3078	0.947
industry26	-0.0858	-0.0856	0.968	-0.0856	0.966	-0.0832	0.958
industry27	-0.0667	-0.0660	0.958	-0.0746	0.910	-0.0732	0.924
industry28	-0.1319	-0.1316	0.967	-0.1279	0.962	-0.1268	0.956
industry29	-0.1090	-0.1087	0.956	-0.1033	0.932	-0.1015	0.906
industry30	-0.1085	-0.1077	0.957	-0.1286	0.781	-0.1238	0.844
industry31	-0.1053	-0.1048	0.951	-0.1028	0.952	-0.1014	0.951
industry32	-0.0772	-0.0763	0.947	-0.0855	0.941	-0.0828	0.951
industry33	0.0307	0.0309	0.962	0.0316	0.959	0.0315	0.969
industry34	-0.1535	-0.1531	0.963	-0.1400	0.868	-0.1326	0.706
industry35	0.0101	0.0106	0.961	-0.0221	0.597	-0.0237	0.516
industry36	-0.3403	-0.3401	0.951	-0.3468	0.924	-0.3449	0.939
contract	-0.1462	-0.1464	0.961	-0.1509	0.942	-0.1455	0.971
cons	3.8243	3.8247	0.955	3.8030	0.916	3.8245	0.962

Table 8.13: Results of a simulation study with log transformation in the imputation step and cube root transformation in the analysis step

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0345	0.0346	0.965	0.0348	0.961	0.0347	0.968
education3	0.0596	0.0592	0.963	0.0472	0.881	0.0514	0.922
education4	0.0713	0.0714	0.964	0.0606	0.693	0.0622	0.782
education5	0.1370	0.1374	0.962	0.1363	0.961	0.1411	0.927
education6	0.1770	0.1771	0.956	0.1601	0.348	0.1699	0.858
level2	0.0105	0.0106	0.962	0.0094	0.963	0.0094	0.963
level3	0.0371	0.0382	0.969	0.0378	0.968	0.0391	0.965
level4	0.0201	0.0212	0.977	0.0067	0.945	0.0050	0.937
group2	-0.0947	-0.0948	0.952	-0.0921	0.913	-0.0920	0.910
group3	-0.1899	-0.1897	0.947	-0.1858	0.881	-0.1857	0.875
group4	-0.3098	-0.3098	0.964	-0.3067	0.928	-0.3065	0.923
group5	0.3875	0.3863	0.969	0.3650	0.845	0.3713	0.901
group6	0.1412	0.1401	0.963	0.1475	0.965	0.1491	0.960
group7	0.0479	0.0469	0.971	0.0611	0.940	0.0624	0.940
group8	-0.1702	-0.1713	0.967	-0.1559	0.942	-0.1543	0.929
group9	-0.3394	-0.3400	0.969	-0.3254	0.946	-0.3240	0.942
age	0.0247	0.0247	0.961	0.0249	0.958	0.0247	0.963
sqage	-0.0003	-0.0003	0.958	-0.0003	0.917	-0.0003	0.961
region2	0.0369	0.0368	0.956	0.0449	0.401	0.0422	0.696
region3	0.0038	0.0037	0.947	0.0082	0.776	0.0059	0.913
region4	0.0517	0.0516	0.959	0.0502	0.920	0.0481	0.794
industry2	-0.0407	-0.0404	0.958	-0.0393	0.943	-0.0395	0.940
industry3	-0.1097	-0.1094	0.945	-0.1111	0.947	-0.1115	0.944
industry4	0.0053	0.0054	0.959	0.0052	0.964	0.0058	0.960
industry5	0.0765	0.0773	0.976	0.0660	0.867	0.0674	0.898
industry6	0.0788	0.0791	0.968	0.0806	0.959	0.0805	0.962
industry7	0.0636	0.0641	0.968	0.0654	0.949	0.0657	0.955
industry8	-0.0145	-0.0146	0.956	-0.0104	0.901	-0.0105	0.911
industry9	-0.0157	-0.0158	0.969	-0.0110	0.942	-0.0109	0.941
industry10	0.0252	0.0257	0.957	0.0304	0.883	0.0304	0.893
industry11	-0.0356	-0.0355	0.960	-0.0314	0.918	-0.0316	0.916
industry12	-0.0029	-0.0027	0.949	0.0014	0.882	0.0013	0.886
industry13	-0.0166	-0.0166	0.953	-0.0190	0.939	-0.0194	0.943
industry14	-0.0278	-0.0275	0.960	-0.0278	0.966	-0.0277	0.965
industry15	-0.0408	-0.0404	0.956	-0.0362	0.924	-0.0366	0.935
industry16	0.0341	0.0343	0.951	0.0363	0.950	0.0364	0.947
industry17	-0.0727	-0.0722	0.967	-0.0691	0.942	-0.0690	0.941
industry18	-0.0100	-0.0096	0.958	0.0063	0.258	0.0062	0.278
industry19	-0.0219	-0.0216	0.955	-0.0156	0.849	-0.0157	0.857
industry20	-0.1047	-0.1047	0.967	-0.1019	0.944	-0.1019	0.947
industry21	-0.0874	-0.0866	0.933	-0.0838	0.874	-0.0840	0.878
industry22	-0.1124	-0.1124	0.965	-0.1126	0.956	-0.1130	0.958
industry23	-0.0549	-0.0546	0.959	-0.0555	0.964	-0.0558	0.963
industry24	-0.1604	-0.1599	0.954	-0.1556	0.912	-0.1561	0.920
industry25	-0.2215	-0.2215	0.960	-0.2174	0.944	-0.2178	0.946
industry26	-0.0560	-0.0558	0.968	-0.0531	0.927	-0.0532	0.936
industry27	-0.0454	-0.0449	0.958	-0.0475	0.940	-0.0477	0.942
industry28	-0.0865	-0.0863	0.971	-0.0819	0.933	-0.0823	0.942
industry29	-0.0697	-0.0696	0.958	-0.0636	0.869	-0.0641	0.889
industry30	-0.0705	-0.0700	0.954	-0.0737	0.940	-0.0748	0.935
industry31	-0.0673	-0.0670	0.952	-0.0607	0.881	-0.0620	0.905
industry32	-0.0662	-0.0653	0.874	-0.0663	0.856	-0.0667	0.862
industry33	0.0112	0.0113	0.966	0.0083	0.934	0.0086	0.939
industry34	-0.0948	-0.0945	0.967	-0.0800	0.524	-0.0806	0.594
industry35	-0.0019	-0.0015	0.964	-0.0193	0.550	-0.0192	0.583
industry36	-0.2604	-0.2603	0.944	-0.2602	0.948	-0.2608	0.949
contract	-0.1114	-0.1117	0.941	-0.1122	0.943	-0.1120	0.945
cons	4.0440	4.0447	0.958	4.0405	0.952	4.0459	0.962

Table 8.14: Results of a simulation study with cube root transformation in the imputation step and log transformation in the analysis step

as imputation and analysis model. The results of the simulation study can be found in Table 8.13. Compared to the simulation study using a log transformation in the imputation and analysis step, we find coverage rates that are somewhat smaller, but only to a minor extent, when we apply the multiple imputation approach considering heteroscedasticity. Compared to using a cube root transformation in both steps, we find even a higher coverage rates for some variables, like, e.g., the key variable education level 6, where we observe the highest rate of censoring. Applying the multiple imputation approach assuming homoscedasticity the same conclusion applies.

Furthermore, the missing wages were imputed based on a cube root transformation and analyzed using a log transformation accordingly. These results can be found in Table 8.14. We find coverage rates that are generally lower to a certain extent compared to using a log transformation in both steps and higher to some extent compared to using a cube root transformation in both steps. As before, we receive rather low coverages for industry 18 and 35. We can conclude that the imputations approaches seem to be robust to different transformations used for the imputation model. If the analyst is interested in a model based on a log transformation, he can expect the more or less same imputation quality regardless if the imputer uses a log or a cube root transformation. If the analyst is interested in a model based on a cube root, an imputation based on a log transformation is even somewhat more appropriate compared to a cube root transformation.

Accordingly, in this chapter we have seen that multiply imputing censored wages is a flexible solution that yields valid estimation results for various research questions, when a suitable transformation of wages and an appropriate imputation model is chosen. A rather low imputation quality we find only concerning a few industry dummies. The same applies, but to a much smaller extent, to region dummies. Concerning the transformation, the simulation results recommend a log transformation. The imputation model should contain as many variables as possible if the imputed are to be used by different researchers or for different purposes. If wages are to be imputed for a specific research questions, the results show that the use of an imputation model close to the analysis model might also lead to valid results.

Chapter 9

Alternative Approaches

In the preceding chapters, several imputation approaches were proposed. We distinguished between single and multiple imputation approaches as well as between approaches assuming homoscedasticity of the residuals and considering heteroscedasticity. All these methods have in common to be based on multivariate regression with starting values from a tobit regression in the first step. In simulation studies the superiority of the multiple imputation procedures was confirmed. In this chapter, some alternative ideas will be presented, which can be distinguished by the quantity of external information required. First of all, an univariate (or unconditional) imputation idea will be addressed. Afterwards methods in the sense of file concatenation using uncensored wage information from external data (German structure of earnings survey, GSES) are discussed. In a last step, we assess the minimum amount of external information that is required in order to receive satisfying imputation results. These approaches are developed to present further alternatives to the approaches suggested in the preceding chapters, but also to assess their validity from another point of view. As in the previous chapter, we perform a series of simulation studies to compare the different imputation approaches again under different situations. The alternative approaches can be seen as a kind of benchmark that allows us assess the comparability or even superiority of the approaches suggested in the preceding chapters working without external information to approaches requiring additional information. For the different multiple imputation approaches the following abbreviations are used:

- MI-Hom: Multiple imputation assuming homoscedasticity based on a tobit model

- MI-Het: Multiple imputation considering heteroscedasticity based on GLS estimation for truncated variables
- MI-Ext: Multiple imputation based on combining the censored data with external data
- MI-Uni: Univariate or unconditional multiple imputation
- MI-Het(extern): Multiple imputation considering heteroscedasticity based on starting values from external data

9.1 Univariate Imputation

So far, multivariate regression-based imputation procedures were discussed and evaluated. As a first alternative approach, we suggest an univariate imputation approach. This approach is based on an example described in Greene (2008) and still works without additional information. In this example the number of tickets demanded for events at an arena is in the center of the interest. Whenever an event is sold out, only the actual number of sold tickets is known and we have only the information that the total number of demanded tickets was higher. The number of tickets demanded is censored when the number of tickets sold is used a proxy. In that example, Greene supposes that a particular arena has 20,000 seats and, in recent season, was sold out 25 percent of the time. The average attendance, including sellouts, was 18,000. The mean and the standard deviation of the demand for seats can be received as follows. According to the moments of the censored normal variable the average attendance of 18,000 is an estimate of

$$E[sales] = 20,000(1 - \Phi) + [\mu + \sigma\lambda]\Phi$$

and Greene provides the following solution to the question of the actual demand for tickets: "Since this is censoring from above, rather than below, $\lambda = -\phi(\alpha)/\Phi(\alpha)$. The argument of Φ, ϕ and α is $\alpha = (20,000 - \mu)/\sigma$. If 25 percent of the events are sellouts, then $\Phi = 0.75$. Inverting the standard normal 0.75 gives $\alpha = 0.675$. In addition, if $\alpha = 0.675$, then $-\phi(0.675)/0.75 = \lambda = 0.424$. This result provides two equations in μ and σ, (a) $18.000 = 0.25(20.000) + 0.75(\mu - 0,424\sigma)$ and (b) $0.675\sigma = 20.000 - \mu$. The solutions are $\sigma = 2426$ and $\mu = 18,362$" (Greene, 2008, p. 763f.).

		before censoring		MI-Uni		Mi-Het	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0345	0.0346	0.963	0.0351	0.967	0.0348	0.968
education3	0.0596	0.0590	0.955	0.0523	0.962	0.0525	0.933
education4	0.0713	0.0716	0.959	0.0644	0.921	0.0633	0.853
education5	0.1370	0.1373	0.960	0.1325	0.961	0.1459	0.818
education6	0.1770	0.1774	0.963	0.1532	0.147	0.1775	0.961
level2	0.0105	0.0105	0.966	0.0101	0.966	0.0095	0.966
level3	0.0371	0.0381	0.963	0.0457	0.961	0.0396	0.961
level4	0.0201	0.0207	0.960	0.0148	0.965	0.0050	0.926
group2	-0.0947	-0.0947	0.958	-0.0964	0.952	-0.0924	0.925
group3	-0.1899	-0.1899	0.964	-0.1923	0.949	-0.1867	0.913
group4	-0.3098	-0.3100	0.950	-0.3132	0.950	-0.3072	0.928
group5	0.3875	0.3866	0.959	0.3787	0.961	0.3868	0.959
group6	0.1412	0.1404	0.959	0.1557	0.941	0.1501	0.958
group7	0.0479	0.0473	0.955	0.0601	0.950	0.0616	0.930
group8	-0.1702	-0.1709	0.953	-0.1674	0.969	-0.1550	0.923
group9	-0.3394	-0.3398	0.961	-0.3341	0.962	-0.3251	0.938
age	0.0247	0.0247	0.960	0.0253	0.937	0.0247	0.976
sqage	-0.0003	-0.0003	0.963	-0.0003	0.901	-0.0003	0.964
region2	0.0369	0.0369	0.958	0.0435	0.673	0.0410	0.827
region3	0.0038	0.0036	0.967	0.0060	0.939	0.0061	0.945
region4	0.0517	0.0516	0.963	0.0493	0.920	0.0471	0.687
industry2	-0.0407	-0.0407	0.963	-0.0399	0.971	-0.0405	0.963
industry3	-0.1097	-0.1095	0.961	-0.1120	0.972	-0.1137	0.948
industry4	0.0053	0.0050	0.960	0.0047	0.979	0.0055	0.961
industry5	0.0765	0.0767	0.959	0.0724	0.981	0.0704	0.945
industry6	0.0788	0.0791	0.966	0.0850	0.941	0.0825	0.935
industry7	0.0636	0.0638	0.959	0.0640	0.978	0.0691	0.889
industry8	-0.0145	-0.0144	0.971	-0.0115	0.962	-0.0110	0.939
industry9	-0.0157	-0.0157	0.951	-0.0132	0.969	-0.0119	0.938
industry10	0.0252	0.0253	0.961	0.0292	0.939	0.0301	0.906
industry11	-0.0356	-0.0357	0.966	-0.0324	0.961	-0.0326	0.940
industry12	-0.0029	-0.0026	0.964	0.0022	0.915	0.0016	0.895
industry13	-0.0166	-0.0166	0.971	-0.0188	0.979	-0.0184	0.948
industry14	-0.0278	-0.0275	0.973	-0.0271	0.991	-0.0276	0.976
industry15	-0.0408	-0.0411	0.962	-0.0370	0.966	-0.0378	0.943
industry16	0.0341	0.0344	0.964	0.0357	0.970	0.0369	0.941
industry17	-0.0727	-0.0727	0.963	-0.0713	0.970	-0.0706	0.949
industry18	-0.0100	-0.0100	0.947	0.0018	0.665	0.0047	0.408
industry19	-0.0219	-0.0220	0.965	-0.0176	0.943	-0.0168	0.886
industry20	-0.1047	-0.1051	0.968	-0.1032	0.969	-0.1036	0.961
industry21	-0.0874	-0.0872	0.927	-0.0864	0.932	-0.0858	0.907
industry22	-0.1124	-0.1122	0.964	-0.1152	0.965	-0.1148	0.965
industry23	-0.0549	-0.0549	0.966	-0.0544	0.981	-0.0568	0.964
industry24	-0.1604	-0.1598	0.966	-0.1590	0.972	-0.1591	0.955
industry25	-0.2215	-0.2215	0.969	-0.2199	0.970	-0.2199	0.967
industry26	-0.0560	-0.0560	0.961	-0.0564	0.975	-0.0545	0.953
industry27	-0.0454	-0.0453	0.970	-0.0484	0.966	-0.0490	0.948
industry28	-0.0865	-0.0862	0.958	-0.0847	0.966	-0.0835	0.942
industry29	-0.0697	-0.0699	0.955	-0.0685	0.958	-0.0661	0.921
industry30	-0.0705	-0.0699	0.962	-0.0724	0.983	-0.0778	0.897
industry31	-0.0673	-0.0671	0.969	-0.0602	0.958	-0.0650	0.953
industry32	-0.0662	-0.0660	0.853	-0.0670	0.879	-0.0693	0.861
industry33	0.0112	0.0115	0.965	0.0117	0.984	0.0115	0.949
industry34	-0.0948	-0.0948	0.966	-0.0786	0.629	-0.0843	0.765
industry35	-0.0019	-0.0023	0.964	-0.0179	0.787	-0.0203	0.598
industry36	-0.2604	-0.2602	0.960	-0.2603	0.962	-0.2629	0.961
contract	-0.1114	-0.1115	0.955	-0.1134	0.959	-0.1109	0.967
cons	4.0440	4.0444	0.957	4.0407	0.953	4.0442	0.970

Table 9.1: Univariate imputation versus MI-Het

This solution can easily be applied to censored wages, as it simply represents an univariate or unconditional application of the tobit model. We have to assume that $y^* \sim N(\mu, \sigma^2)$ and just have to replace the percentage of sellouts by the percentage of censored wages (which is set here again to 15 percent), leading to $\Phi = 0.85$. Additionally we have to replace the size of the arena by the contribution limit and the average attendance by the censored mean. Then the uncensored mean μ and standard deviation σ can be calculated following the example described above. Afterwards, values for the censored observations can be drawn from $N(\mu, \sigma^2)$. As the drawn values have to be above the ceiling, lower imputed values are rejected and the imputation is repeated until all drawn values are above the ceiling. Alternatively drawings directly from a truncated distribution as performed before could be applied. In the multivariate case based on a tobit regression μ is replaced by $x_i'\beta$, which allows to keep the covariate structure in the data set.
In a simulation study, this approach is compared to the multivariate approach considering heteroscedasticity (MI-Het), which involves drawings from a conditional distribution. As imputation model for MI-Het and as analysis model for both approaches we assume again the model containing the wages in logs as dependent variable and as covariates

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type).*

In the analysis step, we apply OLS regression. Again, the true parameters β are obtained from an OLS regression using the original complete population and the analysis model. The results of this simulation study show that an unconditional imputation is not applicable to impute censored wages, especially for analyzing high wage groups with a high percentage of censored observations, like persons with university degree (education level 6, see Table 9.1). While results for most other variables are surprisingly good, we receive especially for education level 6 underestimated parameter estimates and a low coverage rate. If we consider that this imputation approach can also be seen as a case of uncongeniality, where the imputer's model contains none of the variables of the analyst's model, we find that apart from the problem concerning high-skilled employees, the imputation quality is much better than one could expect. Generally, here the same situation appears as when for example the effect of the establishment size is to be analyzed, but the establishment

size was not included in the imputation model. While in that case only the estimation of the effects of the establishment size might be biased, here the whole multivariate structure is not reflected in the imputed data.

9.2 Combining with External Data

A further alternative method which is feasible in the case of censored wages in the IAB Employment Sample is to concatenate external data with complete wage information. Rubin (1986) coined the term file concatenation for the situation of statistical matching which is similar to the idea presented here. If it is possible to find a database with a similar structure and similar variables, one could concatenate the complete data set with the data set with missings in order to obtain a missingness pattern, where some of the higher wages are missing, but the wage distribution is not completely censored from a certain ceiling. In such a case, standard imputation techniques and standard imputation software could be applied. Since the IABS and GSES have a similar structure and a common set of variables, this approach can be applied here. If the GSES is concatenated to the IABS, we obtain a situation where we have a common set of covariates Z and wage variable Y containing some missings in the upper part (see Figure 9.1).

In this case software packages, like, e.g., the standalone software IVEware, MICE in R or ICE in STATA, can be used to impute the missing part of wages. Figure 9.2 shows first results of a single imputation performed using IVEware[1] based on the IABS and GSES for 2001. The solid line refers to the original wages in the GSES, the dashed line to wages in the IABS (original wages up to the ceiling, imputed wages onwards). The vertical line indicates the ceiling in the IAB Employment Sample.

Although this approach is implemented in standard packages, to perform a simulation study, it appears more feasible to programme the procedure individually in STATA. Unfortunately, there is no adequate complete information to assess the imputation quality. The only way to get an idea about the imputation quality is again to perform a simulation study using the German Structure of Earnings Survey. To do so, the simulation procedure is adapted as follows: The complete data set is divided into two parts: one part is serving

[1]More information on IVEware can be found in Raghunathan et al. (2002)

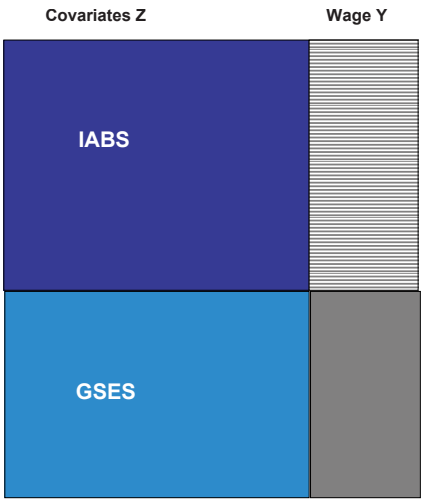

Figure 9.1: File concatenation of the IAB Employment Sample with external data

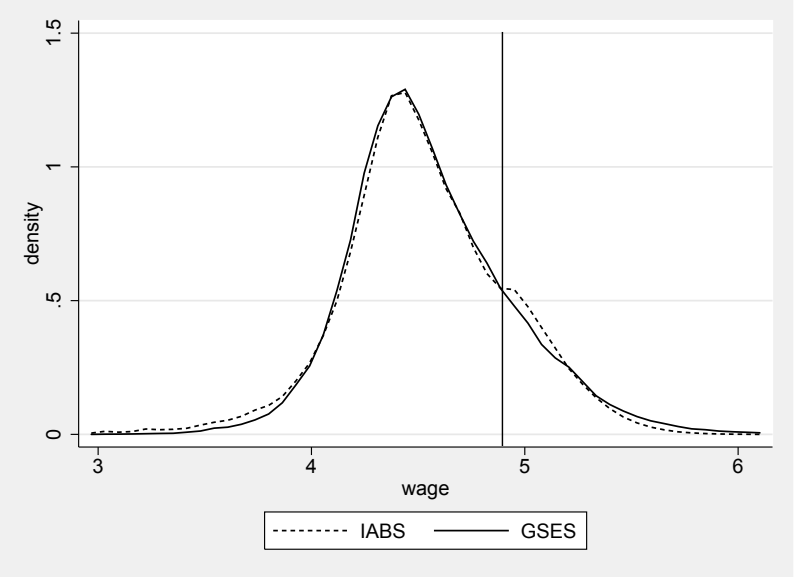

Figure 9.2: Kernel density estimates of imputed wages in the IABS and original wages in the GSES (2001)

as the complete data set (external data), the other part is artificially censored and serves as the data set with censoring. In every iteration we draw 10 percent samples from both data sets and concatenate these two samples. Then we impute the missing wages in the censored part applying the two-step procedure for patterns with only one variable subject to missingness as discussed in Section 6.3. As we now observe the entire wage distribution (with some missing values in the upper part), we can directly fit an OLS regression and perform random draws of the parameter ψ according to the observed-data posterior distribution $f(\psi|Y_{obs})$. Then, we perform random draws of Y_{mis} according to their conditional predictive distribution $f(Y_{mis}|Y_{obs}, \psi)$. This situation is similar to a classical missing data problem, where some information is missing, but only for one variable. In this case we do not need starting values to receive a first complete data posterior distribution and no iterations based on MCMC are necessary.

In particluar, we run an OLS regression using all units without missing wages from both parts to receive $\widehat{\beta}_{obs}$ and $\widehat{\sigma^2}_{obs}$. Then we perform random draws of β and σ^2 according to the observed-data posterior distribution. To draw the variance σ^2 we need again the inverse of a gamma distribution, which is produced as follows:

$$g \sim \chi^2(n-k) \tag{9.1}$$

$$\sigma^{-2} = \frac{g}{RSS} \tag{9.2}$$

where RSS is the residual sum of squares $RSS = \sum_{i=1}^{n}(y_{obs} - x_i'\widehat{\beta}_{obs})^2$ and k is the number of columns of X.

Now new random draws for the parameter β can be performed

$$\beta|\sigma^2 \sim N(\hat{\beta}_{obs}, \sigma^2(X'X)^{-1}). \tag{9.3}$$

Then we perform random draws of the missing wages according to their conditional predictive distribution

$$z_i|\beta, \sigma^2 \sim N_{trunc_a}(x_i'\beta, \sigma^2) \text{ if } y_i = a \text{ for } i = 1, ..., n \tag{9.4}$$

where z is again a truncated variable in the range (a, ∞). In every iteration of the simulation study we repeat the draws for the parameter and the draws of

the missing values $m = 5$ times to receive 5 complete data sets. Afterwards, we divide the two parts again and continue with the analysis step as in the simulation studies before. The whole procedure of drawing 10 percent samples, concatenating the data sets, performing the imputation and analysis steps is repeated again 1,000 times and the corresponding coverage rates are calculated. We perform parallely the imputation approach considering heteroscedasticity (MI-Het) to be able to compare the approaches. As imputation and analysis model for this simulation study we use again the model with

$$X_{large}=(age,\ age^2,\ 6\ education\ categories,\ 4\ job\ level\ categories,\ 9\ performance\ groups,\ 4\ region\ dummies,\ 36\ industry\ dummies,\ contract\ type).$$

In the analysis step, we apply again OLS regression. The corresponding results can be found in Table 9.2. These results indicate that the imputation quality otained by performing this alternative approach and MI-Het are very similar. Note, that we have divided the GSES into two parts. That means, the sample of the GSES serving as true complete population now contains only half of the observations used for the simulation studies in Chapter 8 (N=184,168). Therefore, the results for MI-Het in this chapter are not directly comparable with the corresponding results in Chapter 8 and in Section 9.1. For example, the coverage for industry 18 turns out considerably higher than in the preceding simulation studies, which might be due to the smaller simulation sample.
Instead of using an OLS regression in the first step, a GLS estimation could be applied in order to allow again for heteroscedasticity. Then, in the second step draws for γ would have to be performed additionally and in the imputation step, we could use individual variances again to draw values for the missing wages. The results based on the approach considering heteroscedasticity can be found in Table A.2 of the appendix. Using external information and considering heteroscedasticity leads to results that are very close to the results from the tobit-based MI approach allowing for heteroscedasticity (MI-Het). Accordingly, we conclude that the new approach MI-Het leads to results comparable to an approach that uses additional uncensored information from an external data set. Because the approach using external data can be seen as a kind of benchmark, the last results confirm again the validity of our new approach considering heteroscedasticity (MI-Het).
When discussing an approach based on combining data from the IAB Employment Sample and from the German Structure of Earnings Survey, we have to

		before censoring		MI-Ext		MI-Het	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0363	0.0363	0.951	0.0368	0.947	0.0367	0.950
education3	0.0682	0.0679	0.963	0.0555	0.948	0.0609	0.962
education4	0.0732	0.0736	0.965	0.0661	0.931	0.0673	0.929
education5	0.1364	0.1367	0.958	0.1421	0.945	0.1475	0.857
education6	0.1799	0.1801	0.966	0.1706	0.902	0.1778	0.952
level2	0.0097	0.0097	0.954	0.0086	0.950	0.0083	0.954
level3	0.0417	0.0429	0.969	0.0426	0.969	0.0419	0.970
level4	0.0237	0.0249	0.963	0.0109	0.956	0.0082	0.949
group2	-0.0943	-0.0943	0.967	-0.0922	0.959	-0.0919	0.953
group3	-0.1864	-0.1864	0.961	-0.1835	0.940	-0.1833	0.935
group4	-0.3095	-0.3095	0.961	-0.3069	0.941	-0.3068	0.938
group5	0.3854	0.3839	0.962	0.3803	0.960	0.3853	0.955
group6	0.1372	0.1359	0.957	0.1469	0.962	0.1472	0.960
group7	0.0442	0.0428	0.960	0.0571	0.951	0.0580	0.950
group8	-0.1719	-0.1731	0.964	-0.1588	0.946	-0.1563	0.938
group9	-0.3426	-0.3447	0.962	-0.3302	0.952	-0.3286	0.944
age	0.0249	0.0250	0.950	0.0255	0.945	0.0248	0.967
sqage	-0.0003	-0.0003	0.950	-0.0003	0.920	-0.0003	0.971
region2	0.0360	0.0363	0.965	0.0445	0.652	0.0400	0.921
region3	0.0039	0.0040	0.966	0.0084	0.881	0.0063	0.966
region4	0.0517	0.0519	0.963	0.0519	0.966	0.0474	0.869
industry2	-0.0459	-0.0458	0.948	-0.0439	0.955	-0.0445	0.955
industry3	-0.1091	-0.1091	0.955	-0.1115	0.953	-0.1124	0.942
industry4	0.0088	0.0087	0.963	0.0099	0.973	0.0110	0.963
industry5	0.0774	0.0771	0.957	0.0707	0.959	0.0734	0.944
industry6	0.0817	0.0815	0.962	0.0834	0.961	0.0842	0.952
industry7	0.0628	0.0624	0.961	0.0658	0.970	0.0673	0.949
industry8	-0.0144	-0.0144	0.958	-0.0103	0.950	-0.0105	0.943
industry9	-0.0170	-0.0173	0.965	-0.0129	0.960	-0.0130	0.955
industry10	0.0214	0.0212	0.973	0.0261	0.940	0.0260	0.935
industry11	-0.0354	-0.0359	0.957	-0.0322	0.951	-0.0323	0.945
industry12	-0.0030	-0.0032	0.960	0.0018	0.927	0.0021	0.922
industry13	-0.0138	-0.0143	0.963	-0.0173	0.973	-0.0174	0.959
industry14	-0.0302	-0.0305	0.952	-0.0295	0.955	-0.0296	0.943
industry15	-0.0416	-0.0421	0.965	-0.0376	0.961	-0.0384	0.956
industry16	0.0388	0.0385	0.959	0.0415	0.962	0.0419	0.956
industry17	-0.0808	-0.0808	0.963	-0.0781	0.965	-0.0779	0.960
industry18	-0.0088	-0.0088	0.966	0.0064	0.622	0.0069	0.603
industry19	-0.0201	-0.0203	0.957	-0.0140	0.904	-0.0138	0.901
industry20	-0.1055	-0.1053	0.959	-0.1036	0.957	-0.1035	0.951
industry21	-0.0906	-0.0916	0.910	-0.0897	0.876	-0.0898	0.880
industry22	-0.1134	-0.1137	0.959	-0.1163	0.949	-0.1168	0.937
industry23	-0.0527	-0.0532	0.961	-0.0536	0.979	-0.0542	0.966
industry24	-0.1616	-0.1617	0.963	-0.1603	0.971	-0.1609	0.970
industry25	-0.2182	-0.2180	0.952	-0.2139	0.944	-0.2142	0.947
industry26	-0.0579	-0.0580	0.956	-0.0562	0.958	-0.0561	0.952
industry27	-0.0477	-0.0476	0.957	-0.0505	0.957	-0.0507	0.952
industry28	-0.0904	-0.0909	0.956	-0.0881	0.954	-0.0885	0.950
industry29	-0.0702	-0.0704	0.964	-0.0659	0.945	-0.0663	0.946
industry30	-0.0690	-0.0690	0.952	-0.0775	0.915	-0.0787	0.887
industry31	-0.0657	-0.0661	0.959	-0.0633	0.970	-0.0650	0.960
industry32	-0.0596	-0.0605	0.921	-0.0636	0.929	-0.0631	0.927
industry33	0.0085	0.0082	0.967	0.0056	0.962	0.0072	0.951
industry34	-0.0960	-0.0958	0.962	-0.0848	0.869	-0.0846	0.855
industry35	-0.0061	-0.0064	0.959	-0.0201	0.876	-0.0227	0.809
industry36	-0.2607	-0.2607	0.948	-0.2610	0.948	-0.2616	0.952
contract	-0.1116	-0.1112	0.936	-0.1129	0.944	-0.1107	0.945
cons	4.0396	4.0375	0.945	4.0287	0.929	4.0395	0.967

Table 9.2: Imputation using external data versus MI-Het

note that there are some differences between these two samples. Whereas in the IAB Employment Sample all employees liable to social insurance are covered, in the GSES only employees in the manufacturing industry and service sector are covered. Accordingly the agriculture and fishing sector is excluded. Hence, we would have to exclude this sector from the IAB Employment Sample as well or we need to apply an imputation model that does not contain industry dummies. As the second solution may lead to an unappropriate imputation model, it is more feasible to exclude this sector, which does not play an important role in most studies anyway.
Besides, we have to find an imputation model that is a good predictor for wages and consists of variables that are available in both data sets. In the simulation study, we divided the GSES into two parts: One part simulating the complete data set, the other part simulating the censored data set. Because we aim to receive the best possible imputation results for this data set, we fit an imputation model, that seems to be appropriate in this case. But we have to note that if we actually want to apply this approach to the IABS, we would have to apply a different model. Table A.3 in the appendix shows the results when an imputation model restricted to variables that can be found in both data sets and the multiple imputation approach that allows for heteroscedasticity are applied.
Furthermore, this approach using external information involves some other drawbacks. The main disadvantage is that the IABS and the GSES have actually to be concatenated. Currently, a scientific use file of the GSES is only available for the year 2001, all other years can only be used on-site at the research data center of the German Statistical Office. The IABS on the other hand can be used only at the IAB. Accordingly, an imputation using this approach is possible for 2001 only at the moment. After the release of the scientific use file of the GSES 2006 it will applicable for two years.
That is why we present another approach based on using external data for the imputation in the IAB Employment Sample that is applicable for all years, the German Structure of Earnings Survey was conducted in.

9.3 Starting Values from External Data

The idea behind this second version of imputation approaches based on using external information is to improve the starting values. It is mainly based on

		before censoring		MI-Het (extern.)		MI-het (tobit)	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0363	0.0363	0.957	0.0368	0.967	0.0368	0.962
education3	0.0682	0.0686	0.961	0.0617	0.955	0.0616	0.955
education4	0.0732	0.0734	0.967	0.0672	0.926	0.0672	0.920
education5	0.1364	0.1367	0.965	0.1477	0.863	0.1477	0.862
education6	0.1799	0.1797	0.959	0.1776	0.957	0.1776	0.953
level2	0.0097	0.0099	0.958	0.0086	0.954	0.0086	0.955
level3	0.0417	0.0422	0.954	0.0414	0.954	0.0413	0.957
level4	0.0237	0.0236	0.951	0.0072	0.926	0.0072	0.921
group2	-0.0943	-0.0942	0.961	-0.0918	0.944	-0.0918	0.950
group3	-0.1864	-0.1863	0.958	-0.1832	0.930	-0.1832	0.930
group4	-0.3095	-0.3097	0.961	-0.3070	0.941	-0.3070	0.940
group5	0.3854	0.3856	0.943	0.3867	0.947	0.3867	0.939
group6	0.1372	0.1375	0.946	0.1486	0.934	0.1486	0.936
group7	0.0442	0.0444	0.946	0.0594	0.927	0.0594	0.928
group8	-0.1719	-0.1716	0.943	-0.1551	0.918	-0.1551	0.923
group9	-0.3426	-0.3411	0.950	-0.3255	0.933	-0.3255	0.936
age	0.0249	0.0249	0.958	0.0248	0.970	0.0248	0.974
sqage	-0.0003	-0.0003	0.957	-0.0003	0.971	-0.0003	0.973
region2	0.0360	0.0362	0.961	0.0396	0.919	0.0397	0.915
region3	0.0039	0.0039	0.967	0.0060	0.965	0.0060	0.966
region4	0.0517	0.0517	0.953	0.0471	0.837	0.0471	0.844
industry2	-0.0459	-0.0454	0.962	-0.0441	0.960	-0.0441	0.963
industry3	-0.1091	-0.1086	0.955	-0.1119	0.951	-0.1119	0.948
industry4	0.0088	0.0089	0.959	0.0113	0.951	0.0113	0.953
industry5	0.0774	0.0770	0.969	0.0731	0.950	0.0732	0.950
industry6	0.0817	0.0814	0.946	0.0841	0.935	0.0842	0.933
industry7	0.0628	0.0629	0.952	0.0678	0.929	0.0678	0.929
industry8	-0.0144	-0.0141	0.964	-0.0103	0.941	-0.0103	0.940
industry9	-0.0170	-0.0171	0.956	-0.0127	0.955	-0.0126	0.952
industry10	0.0214	0.0215	0.957	0.0264	0.914	0.0265	0.916
industry11	-0.0354	-0.0352	0.962	-0.0316	0.949	-0.0316	0.952
industry12	-0.0030	-0.0031	0.952	0.0019	0.915	0.0019	0.921
industry13	-0.0138	-0.0137	0.967	-0.0169	0.963	-0.0169	0.966
industry14	-0.0302	-0.0301	0.960	-0.0294	0.960	-0.0293	0.961
industry15	-0.0416	-0.0414	0.962	-0.0380	0.956	-0.0379	0.948
industry16	0.0388	0.0389	0.954	0.0422	0.950	0.0423	0.947
industry17	-0.0808	-0.0804	0.956	-0.0776	0.948	-0.0776	0.946
industry18	-0.0088	-0.0088	0.974	0.0069	0.585	0.0069	0.589
industry19	-0.0201	-0.0199	0.965	-0.0136	0.917	-0.0135	0.914
industry20	-0.1055	-0.1053	0.961	-0.1036	0.960	-0.1036	0.960
industry21	-0.0906	-0.0904	0.899	-0.0888	0.871	-0.0888	0.880
industry22	-0.1134	-0.1135	0.960	-0.1163	0.954	-0.1163	0.951
industry23	-0.0527	-0.0525	0.960	-0.0534	0.965	-0.0535	0.967
industry24	-0.1616	-0.1615	0.960	-0.1607	0.965	-0.1607	0.963
industry25	-0.2182	-0.2178	0.958	-0.2141	0.947	-0.2140	0.947
industry26	-0.0579	-0.0581	0.962	-0.0563	0.954	-0.0563	0.956
industry27	-0.0477	-0.0477	0.947	-0.0509	0.937	-0.0508	0.944
industry28	-0.0904	-0.0899	0.962	-0.0877	0.961	-0.0876	0.956
industry29	-0.0702	-0.0703	0.956	-0.0663	0.940	-0.0663	0.934
industry30	-0.0690	-0.0688	0.959	-0.0786	0.903	-0.0784	0.899
industry31	-0.0657	-0.0659	0.968	-0.0647	0.957	-0.0649	0.963
industry32	-0.0596	-0.0606	0.909	-0.0637	0.915	-0.0636	0.917
industry33	0.0085	0.0093	0.947	0.0079	0.945	0.0080	0.949
industry34	-0.0960	-0.0959	0.965	-0.0846	0.859	-0.0845	0.855
industry35	-0.0061	-0.0057	0.958	-0.0225	0.823	-0.0225	0.825
industry36	-0.2607	-0.2607	0.958	-0.2619	0.957	-0.2618	0.960
contract	-0.1116	-0.1119	0.946	-0.1113	0.953	-0.1113	0.951
cons	4.0396	4.0397	0.964	4.0415	0.979	4.0415	0.979

Table 9.3: Imputation using external starting values versus MI-Het

the multiple imputation approaches performing a tobit regression in the first step (MI-Hom and MI-Het). Instead of adapting starting values for $\beta^{(0)}$ and the variance $\sigma^{2(0)}$ from a tobit regression, an OLS regression is performed using the complete external data set only. Then, values for the missing wages are randomly drawn from a truncated distribution in analogy to the MI-Hom approach using these starting values

$$z_i^{(0)} \sim N_{trunc_a}(x_i'\beta_{ext}^{(0)}, \sigma_{ext}^{2(0)}) \text{ if } y_i = a \text{ for } i = 1, ..., n. \tag{9.5}$$

Now the same Markov chain Monte Carlo algorithm as in the MI-Hom approach can be performed to receive $m = 5$ complete data sets. This procedure can be performed considering heteroscedasticity in the same way. Again, we start the imputation by adapting starting values for $\beta_{ext}^{(0)}$ and $\gamma_{ext}^{(0)}$ from a GLS estimation for truncated variables. Then, we draw values for the missing wages from a truncated distribution using individual variances $\sigma_i^{2(0)} = e^{w_i'\gamma_{ext}^{(0)}}$ again like in the heteroscedastic single imputation model:

$$z_i^{(0)} \sim N_{trunc_a}(x_i'\beta_{ext}^{(0)}, \sigma_i^{2(0)}) \quad \text{where} \quad \sigma_i^{2(0)} = e^{w_i'\gamma_{ext}^{(0)}} \text{ if } y_i = a \text{ for } i = 1, ..., n. \tag{9.6}$$

Afterwards, we continue like in the MI-Het approach. This approach based on starting values from external data is also compared in a simulation study to the MI-Het approach. One additional objective of this simulation study is to assess if using a tobit model to receive starting values for the imputation procedure is an applicable approach. Using a tobit model we have to assume properties of the wage distribution like normality. Comparing starting values from a tobit estimation to starting values from an uncensored external distribution, we can assess the applicability of tobit regression in our case. The simulation procedure is adapted here as follows: The complete data set is divided again into two parts: one part serving as the complete data set (external data), the other part will be artificially censored and serves as the data set with censoring. In every iteration we draw 10 percent samples from both data sets. We run a GLS regression using the complete random sample to receive the starting values for $\beta_{ext}^{(0)}$ and $\gamma_{ext}^{(0)}$. Then we discard the complete sample and go on with the imputation of the censored sample using these starting values. Afterwards we perform the analysis step and repeat the whole procedure 1,000 times and calculate the coverage rates. As imputation and analysis model we

use again the model containing as covariates X_{large} and run OLS regression in the analysis step.
Table 9.3 shows the results of this simulation study. The results of the two approaches compared in this simulation study are surprisingly similar. The results therefore indicate that applying a tobit regression is an applicable strategy, as it leads to the same results as starting values from observed complete data.
This procedure can be performed for all years, in which the GSES was conducted. While the starting values for 2001 can be estimated using the scientific use file, for all other years they have to be calculated onsite or by remote access. Then the starting values can be transferred to the IAB without any data privacy protection restrictions.

9.4 Minimum Requirements for Imputation based on External Data

The preceding simulation studies have shown that imputation approaches for right-censored wages based on starting values from a tobit estimation lead to valid imputed data. Imputation based on external information leads to a good imputation quality as well, if the data set with censored wages is combined with complete external information or the imputation is based on starting values for the parameters from external data. In this section, it will be assessed whether a minimum of information from external data is sufficient to obtain a satisfying imputation quality. To do so, we assume that the only information that is available from external data is in the form of wage quantiles. While for the approaches discussed before, the entire external data set is necessary, quantile information can be obtained easily. It can be calculated without any knowledge about multiple imputation by the data provider and it does not depend on the specific imputation problem and model. Sometimes information on the distribution of wage quantiles (i.e the median, quartiles, and deciles) can even be found in publications of statistical offices. Therefore, it can be seen as a kind a minimum amount of external information that can easily be obtained, but it may be already sufficient to perform multiple imputation based on external information. Based on the quantile information we develop imputation approaches that need no additional information to impute the cen-

sored wages and assess the quality of these approaches performing simulation studies again.

We assume once more that all wages from the 85th percentile onwards are censored. From the external data set we obtain in a first scenario the values of all wage percentiles from the 85th to the 99th percentile (in one percent steps). In order to perform the imputation, the external information does not necessarily need to be in percentiles; any kind of quantile information about the upper part of the wage distribution would be sufficient. In the second step we fit a logit model with a dummy for censored/uncensored wage as independent variable to the censored wage data in order to estimate the propensity score. The propensity score is the probability of a person having a censored wage observation given a set of known covariates. In parallel, we apply a tobit model to receive predicted values for the censored wage observations. In the next step we distribute the persons with censored wage information to the 15 cells between the 15 percentiles. The censored observations with the lowest predicted wage (or lowest probability to have a censored wage observation) go to the cell between the 85th and the 86th percentile, the next to the cell between the 86th and the 87th and so on. Once we have filled the cells, we obtain a coarsened distribution. Afterwards, we apply an interval regression for coarsened data, which represents a generalization of the tobit model. Interval regression models can fit models for data where each observation represents interval data, left-censored data, right-censored data, or point data. In the case of the IAB Employment Sample, we find data where all observations up to ceiling a are observed and all higher observation are right-censored, i.e., lie in the range (a, ∞). On the other hand, in the estimated coarsened distribution, all observations lower than a are observed as well, but the higher observations lie in 14 smaller intervals between the percentiles and in one right-censored interval $(P89, \infty)$, where $P89$ represents the 89th percentile. As we now have estimated a coarsened complete data distribution we can apply again a two-step imputation procedure. A main advantage of this procedure is that we do not need starting values from a tobit model, but instead can directly apply an interval regression and perform random draws of the parameter ψ in the first step and random draws of Y_{mis} according to their conditional predictive distribution in the second step. We fit an interval regression for that kind of coarsened data, which can be found for example in the *intreg* command in STATA, to estimate the parameters of the wage distribution. Now, we are

able to perform draws for the parameters and afterwards random draws for the missing wages from a truncated distribution:

$$z_i|\beta,\sigma^2 \sim N_{trunc_a}(x_i'\beta,\sigma^2) \text{ if } y_i = a \text{ for } i = 1,...,n \tag{9.7}$$

The draws for the parameters and the imputation step are then repeated 5 times to receive $m = 5$ complete data sets.

In a simulation study this approach is then compared to the MI approach considering heteroscedasticity. The GSES is divided again into two parts: one part serving as the complete data set (external data), the other part will be artificially censored and serves as the data set with censoring. In every iteration we draw 10 percent samples from both data sets and calculate the percentiles in the uncensored part, which is deleted afterwards. Then, the missing wage information is imputed based on the approach described above and based on MI-Het. Finally, we repeat these steps again 1,000 times to be able to calculate coverage rates. The imputation and analysis model consists of the same variables as in the preceding simulation studies (X_{large}). The results of this simulation study can be found in Table 9.4 and can be summarized as follows:

Compared to MI-Het approach the results of the quantile-information based approach are more or less the same for most variables, except for the dummies for technical college degree (education5) and university degree (education6). For this two variables representing groups with an especially high percentage of censored observations, the coverage rates are somewhat lower. The results indicate that in the simulated case a logit model is more suitable than a tobit model to distribute the censored observations into a coarsened distribution. Whereas the results for education level 6 are in both cases around 56 percent, the coverage rate for education level 6 based on a logit model in the first step (0.691) is higher compared to the result based on a tobit model in the first step (0.455).

The imputation approach based on external quantile information could be varied and adapted in several ways. First, in the first step a probit model could be applied instead of the logit model. An additional simulation showed that this modification has little effect on the imputation results. Second, the size of the cells could be modified. If less external quantile information is available (for example only in 3 or 5 percentage point steps), the censored observations have to be assigned to larger cells. If information is available only

		before censoring		EXT-logit		EXT-tobit		MI-Het	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0363	0.0364	0.961	0.0371	0.959	0.0371	0.956	0.0368	0.963
education3	0.0682	0.0678	0.962	0.0590	0.939	0.0574	0.937	0.0619	0.966
education4	0.0732	0.0736	0.955	0.0685	0.949	0.0673	0.932	0.0674	0.933
education5	0.1364	0.1366	0.961	0.1570	0.558	0.1568	0.555	0.1474	0.852
education6	0.1799	0.1800	0.956	0.1964	0.691	0.2038	0.455	0.1777	0.947
level2	0.0097	0.0098	0.962	0.0090	0.961	0.0090	0.961	0.0085	0.970
level3	0.0417	0.0429	0.954	0.0427	0.953	0.0427	0.954	0.0419	0.952
level4	0.0237	0.0250	0.961	0.0125	0.946	0.0120	0.939	0.0084	0.934
group2	-0.0943	-0.0944	0.959	-0.0918	0.948	-0.0918	0.940	-0.0920	0.941
group3	-0.1864	-0.1864	0.965	-0.1830	0.938	-0.1829	0.939	-0.1832	0.940
group4	-0.3095	-0.3092	0.966	-0.3055	0.932	-0.3054	0.929	-0.3065	0.940
group5	0.3854	0.3837	0.964	0.3975	0.954	0.3978	0.953	0.3852	0.959
group6	0.1372	0.1356	0.961	0.1423	0.957	0.1423	0.954	0.1471	0.952
group7	0.0442	0.0428	0.960	0.0532	0.950	0.0536	0.949	0.0581	0.943
group8	-0.1719	-0.1732	0.958	-0.1606	0.949	-0.1599	0.941	-0.1564	0.933
group9	-0.3426	-0.3442	0.971	-0.3315	0.962	-0.3307	0.959	-0.3281	0.957
age	0.0249	0.0249	0.970	0.0256	0.939	0.0256	0.941	0.0248	0.981
sqage	-0.0003	-0.0003	0.971	-0.0003	0.919	-0.0003	0.910	-0.0003	0.975
region2	0.0360	0.0360	0.956	0.0451	0.620	0.0455	0.592	0.0397	0.913
region3	0.0039	0.0039	0.968	0.0083	0.877	0.0086	0.870	0.0061	0.967
region4	0.0517	0.0517	0.959	0.0540	0.934	0.0541	0.936	0.0471	0.848
industry2	-0.0459	-0.0458	0.966	-0.0446	0.961	-0.0448	0.964	-0.0445	0.958
industry3	-0.1091	-0.1090	0.959	-0.1128	0.959	-0.1130	0.959	-0.1121	0.960
industry4	0.0088	0.0085	0.966	0.0095	0.962	0.0092	0.963	0.0106	0.960
industry5	0.0774	0.0780	0.948	0.0791	0.934	0.0760	0.935	0.0742	0.946
industry6	0.0817	0.0816	0.957	0.0843	0.940	0.0840	0.935	0.0847	0.938
industry7	0.0628	0.0628	0.967	0.0712	0.861	0.0710	0.858	0.0678	0.920
industry8	-0.0144	-0.0140	0.955	-0.0103	0.945	-0.0103	0.938	-0.0101	0.936
industry9	-0.0170	-0.0163	0.961	-0.0127	0.956	-0.0129	0.952	-0.0119	0.945
industry10	0.0214	0.0219	0.958	0.0261	0.937	0.0262	0.927	0.0267	0.918
industry11	-0.0354	-0.0349	0.950	-0.0323	0.940	-0.0323	0.935	-0.0315	0.934
industry12	-0.0030	-0.0030	0.962	0.0023	0.918	0.0021	0.917	0.0022	0.915
industry13	-0.0138	-0.0138	0.963	-0.0132	0.942	-0.0142	0.955	-0.0169	0.943
industry14	-0.0302	-0.0302	0.948	-0.0290	0.958	-0.0293	0.961	-0.0294	0.963
industry15	-0.0416	-0.0413	0.956	-0.0377	0.941	-0.0379	0.942	-0.0380	0.947
industry16	0.0388	0.0387	0.962	0.0424	0.952	0.0427	0.946	0.0419	0.956
industry17	-0.0808	-0.0803	0.966	-0.0795	0.968	-0.0795	0.966	-0.0774	0.957
industry18	-0.0088	-0.0085	0.964	0.0060	0.656	0.0069	0.616	0.0073	0.590
industry19	-0.0201	-0.0198	0.963	-0.0143	0.920	-0.0142	0.916	-0.0134	0.900
industry20	-0.1055	-0.1058	0.979	-0.1056	0.976	-0.1057	0.977	-0.1040	0.972
industry21	-0.0906	-0.0903	0.912	-0.0898	0.896	-0.0899	0.899	-0.0886	0.883
industry22	-0.1134	-0.1132	0.965	-0.1174	0.950	-0.1174	0.945	-0.1163	0.958
industry23	-0.0527	-0.0523	0.972	-0.0531	0.962	-0.0536	0.967	-0.0533	0.966
industry24	-0.1616	-0.1615	0.956	-0.1626	0.957	-0.1626	0.961	-0.1610	0.966
industry25	-0.2182	-0.2177	0.965	-0.2148	0.957	-0.2149	0.961	-0.2140	0.955
industry26	-0.0579	-0.0575	0.969	-0.0570	0.960	-0.0570	0.960	-0.0559	0.948
industry27	-0.0477	-0.0476	0.973	-0.0517	0.963	-0.0518	0.966	-0.0508	0.969
industry28	-0.0904	-0.0898	0.961	-0.0881	0.965	-0.0881	0.961	-0.0874	0.960
industry29	-0.0702	-0.0701	0.966	-0.0671	0.952	-0.0671	0.955	-0.0662	0.951
industry30	-0.0690	-0.0682	0.966	-0.0839	0.808	-0.0835	0.822	-0.0783	0.914
industry31	-0.0657	-0.0655	0.963	-0.0691	0.952	-0.0691	0.962	-0.0649	0.963
industry32	-0.0596	-0.0586	0.903	-0.0639	0.922	-0.0638	0.920	-0.0616	0.903
industry33	0.0085	0.0095	0.963	0.0126	0.919	0.0099	0.912	0.0080	0.951
industry34	-0.0960	-0.0958	0.967	-0.0943	0.955	-0.0956	0.957	-0.0846	0.865
industry35	-0.0061	-0.0058	0.961	-0.0192	0.854	-0.0203	0.833	-0.0219	0.835
industry36	-0.2607	-0.2607	0.949	-0.2632	0.955	-0.2633	0.956	-0.2618	0.952
contract	-0.1116	-0.1119	0.919	-0.1149	0.933	-0.1151	0.932	-0.1114	0.943
cons	4.0396	4.0384	0.966	4.0259	0.928	4.0248	0.919	4.0405	0.979

Table 9.4: Multiple imputation based on external quantiles

in steps of 3 percentage points, a simulation study showed that this kind of modification has little effect on the quality of the imputation results. Besides, the imputation could be done considering heteroscedasticity. So far, in the last step the missing wages were drawn from a truncated normal distribution assuming homoscedasticity of the residuals. It is certainly possible to perform this step considering heteroscedasticity. Yet, as first analyses have shown, when the approach based on quantile information is applied that generalization does not contribute to a better imputation quality. Further possible modifications concern the Bayesian draws of the parameters. It is for instance possible to perform additional draws to add noise to the estimates of the initial logit or tobit model before calculating the propensity score or the predicted wages respectively. The random draws for the missing wages could be modified as well. Instead of drawing the values from a normal distribution truncated in the range (a, ∞), more restrictions could be included in the imputation step. For example, the draws could be performed in a way that ensures that every imputed values lies in the range of the cell the observation was assigned to in the beginning.

Another possibility to perform multiple imputation based on external quantile information is to apply an univariate or unconditional approach. Here, no model has to be defined in order to distribute individuals to certain cells. The persons can be just randomly assigned to the cells. Alternatively the assignment can be done by propensity scores or predicted wages, but the way chosen has no impact on the imputation results. We just need a coarsened distribution, but this distribution may be completely independent from any covariates. Once we have assigned the observations to the coarsened distribution, we are able to estimate the uncensored mean μ and the corresponding standard deviation σ of this distribution. This can easily be performed by running an interval regression with wages as independent variable just on a constant. The estimated parameter for the intercept will then be equivalent to the mean. Afterwards the missing wages can be drawn from a distribution truncated at a with mean μ and standard deviation σ. Alternatively a rejection algorithm could be applied here as well. The results of such an imputation can be found in Table 9.5. For most variables we receive here again a coverage rate comparable to the results of the MI approach considering heteroscedasticity (MI-Het). Compared to the first approach based on external quantile information, the result concerning education level 5 is significantly improved (0.919). On the

other hand, for the group with the highest percentage of censored observations, education level 6, we find a rather disappointing result (0.154).

We can conclude that the proposed imputation approaches based on external quantile information need only external information, which can be easily obtained from various sources. Based on that minimum amount of external information we receive coverage rates, which are satisfying for most variables. However, we have to admit that the imputation quality of these approaches is somewhat lower compared to the approaches based on combining the censored data with complete external data or based on starting values from external data. Consequently there is a trade-off between imputation quality and the minimum requirement of external information that is needed to perform the imputation approach. Basically, wage quantiles as a minimum amount of external information are sufficient to impute censored wages. Yet, whenever more information is available it is preferable to apply one of the approaches that use additional information.

Having evaluated various alternatives to the imputation approaches based on starting values from a tobit model, we can conclude that the proposed MI approach considering heteroscedasticity (MI-Het) is easy to implement as it requires no external information. However, the same assumptions that are applicable for estimating a tobit model have to be presumed using MI-Het. Simulation studies have shown that in the case of the German Structure of Earnings Survey the assumptions about the error distribution of the tobit model do not influence the imputation quality. More importantly, this approach yields good imputation results, which are comparable to results of approaches that require external information, in some cases even better. Hence, for most cases it is advisable to apply the multiple imputation approach considering heteroscedasticity combined with a lognormal transformation of the wages to impute the censored wage information in the IAB Employment Sample. Accordingly, we can finally summarize that there is no external information necessary to obtain valid imputations, because the MI approach considering heteroscedasticity working without external information yields imputation results that are at least comparable to the approaches requiring external information.

		before censoring		MI-Uni		MI-Het	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0363	0.0365	0.958	0.0368	0.962	0.0369	0.966
education3	0.0682	0.0688	0.964	0.0585	0.971	0.0619	0.965
education4	0.0732	0.0738	0.953	0.0659	0.937	0.0676	0.925
education5	0.1364	0.1366	0.955	0.1277	0.919	0.1473	0.864
education6	0.1799	0.1802	0.959	0.1481	0.154	0.1777	0.943
level2	0.0097	0.0098	0.956	0.0090	0.958	0.0085	0.959
level3	0.0417	0.0426	0.953	0.0467	0.961	0.0419	0.959
level4	0.0237	0.0243	0.949	0.0143	0.950	0.0079	0.929
group2	-0.0943	-0.0945	0.962	-0.0956	0.960	-0.0921	0.951
group3	-0.1864	-0.1867	0.961	-0.1882	0.963	-0.1836	0.940
group4	-0.3095	-0.3098	0.968	-0.3125	0.965	-0.3071	0.949
group5	0.3854	0.3847	0.948	0.3642	0.922	0.3859	0.948
group6	0.1372	0.1364	0.949	0.1526	0.933	0.1475	0.937
group7	0.0442	0.0433	0.944	0.0590	0.941	0.0583	0.931
group8	-0.1719	-0.1728	0.942	-0.1647	0.950	-0.1563	0.929
group9	-0.3426	-0.3438	0.958	-0.3337	0.957	-0.3279	0.932
age	0.0249	0.0248	0.956	0.0253	0.965	0.0247	0.965
sqage	-0.0003	-0.0003	0.954	-0.0003	0.954	-0.0003	0.970
region2	0.0360	0.0360	0.956	0.0418	0.850	0.0397	0.919
region3	0.0039	0.0037	0.953	0.0063	0.932	0.0060	0.962
region4	0.0517	0.0513	0.968	0.0481	0.909	0.0469	0.831
industry2	-0.0459	-0.0464	0.946	-0.0439	0.962	-0.0452	0.951
industry3	-0.1091	-0.1096	0.957	-0.1094	0.965	-0.1125	0.946
industry4	0.0088	0.0091	0.947	0.0111	0.952	0.0114	0.942
industry5	0.0774	0.0773	0.972	0.0711	0.974	0.0736	0.951
industry6	0.0817	0.0812	0.963	0.0860	0.961	0.0841	0.957
industry7	0.0628	0.0624	0.964	0.0606	0.975	0.0672	0.945
industry8	-0.0144	-0.0142	0.956	-0.0103	0.953	-0.0104	0.942
industry9	-0.0170	-0.0168	0.960	-0.0131	0.963	-0.0126	0.954
industry10	0.0214	0.0219	0.952	0.0264	0.941	0.0268	0.919
industry11	-0.0354	-0.0347	0.953	-0.0305	0.935	-0.0313	0.929
industry12	-0.0030	-0.0029	0.963	0.0027	0.923	0.0023	0.908
industry13	-0.0138	-0.0136	0.964	-0.0173	0.977	-0.0169	0.957
industry14	-0.0302	-0.0306	0.958	-0.0293	0.973	-0.0300	0.958
industry15	-0.0416	-0.0416	0.955	-0.0372	0.971	-0.0384	0.961
industry16	0.0388	0.0384	0.949	0.0399	0.961	0.0418	0.945
industry17	-0.0808	-0.0806	0.953	-0.0769	0.950	-0.0777	0.950
industry18	-0.0088	-0.0085	0.955	0.0051	0.720	0.0070	0.594
industry19	-0.0201	-0.0200	0.966	-0.0133	0.917	-0.0136	0.908
industry20	-0.1055	-0.1053	0.955	-0.1020	0.960	-0.1037	0.957
industry21	-0.0906	-0.0908	0.920	-0.0885	0.898	-0.0892	0.899
industry22	-0.1134	-0.1135	0.958	-0.1152	0.957	-0.1167	0.942
industry23	-0.0527	-0.0526	0.955	-0.0508	0.973	-0.0535	0.951
industry24	-0.1616	-0.1614	0.963	-0.1585	0.977	-0.1612	0.965
industry25	-0.2182	-0.2184	0.956	-0.2131	0.949	-0.2148	0.948
industry26	-0.0579	-0.0580	0.956	-0.0565	0.951	-0.0563	0.948
industry27	-0.0477	-0.0474	0.962	-0.0491	0.976	-0.0506	0.961
industry28	-0.0904	-0.0903	0.960	-0.0876	0.961	-0.0879	0.956
industry29	-0.0702	-0.0702	0.969	-0.0666	0.955	-0.0662	0.944
industry30	-0.0690	-0.0686	0.954	-0.0706	0.983	-0.0786	0.904
industry31	-0.0657	-0.0654	0.959	-0.0581	0.956	-0.0642	0.955
industry32	-0.0596	-0.0601	0.905	-0.0604	0.916	-0.0633	0.911
industry33	0.0085	0.0087	0.959	0.0065	0.986	0.0072	0.951
industry34	-0.0960	-0.0962	0.963	-0.0764	0.687	-0.0850	0.867
industry35	-0.0061	-0.0064	0.963	-0.0205	0.903	-0.0228	0.803
industry36	-0.2607	-0.2598	0.933	-0.2570	0.910	-0.2610	0.937
contract	-0.1116	-0.1114	0.949	-0.1126	0.954	-0.1109	0.965
cons	4.0396	4.0409	0.958	4.0393	0.966	4.0422	0.967

Table 9.5: Univariate imputation based on external quantile information versus MI-Het

Chapter 10

Applications

The main focus of this thesis is to propose a new multiple imputation approach for right-censored wages considering heteroscedasticity. In several simulation studies this method was compared to alternative approaches and the necessity and the validity of this approach was confirmed. In the following chapter, some typical real world examples will be presented to illustrate the importance of applying imputation methods before wages in the IAB Employment Sample can be analyzed. In the first part of the chapter some basic research questions will be discussed, which can only be examined using the IABS when appropriate solutions for the problem of censoring are applied. Some studies addressing these research questions were already discussed in Chapter 5. Results based on multiply imputed wages will be compared to results based on complete and censored wages to show the utility of our new approach. In the second part, some recent studies based on the IAB Employment Sample that already applied our imputation methods for right-censored wages will be presented. Finally, guidelines for researchers interested in applying multiple imputation approaches to the IAB data are suggested.

10.1 Typical Examples from Economic Research

The problem of censoring plays an important role whenever the wage variable is in the center of interest of a research question. Even for simple descriptive questions concerning the wage distribution, a bias due to the censoring will occur if censoring is not correctly handled. This holds also for more sophis-

ticated research questions like wage inequality, for example between men and women (the so-called gender wage gap). Evaluating the change of the wage differential between high and low income groups over several years is another question where this problem arises. The aim of the following examples is to illustrate the bias that may occur when censored wages are evaluated and to demonstrate that using multiply imputed data can avoid this serious bias in the estimation results.

10.1.1 Average Wages

Calculating average wages of different groups, e.g., education groups, is a simple but important question that might be of interest to a researcher using the IAB Employment Sample. As we already have seen, the proportion of censored wages varies from education group to education group. Therefore, average wages of different groups are biased to differing extent due to the censoring. We assume that a researcher is interested in the mean daily wage of the total population and of six education groups. To assess the bias of the censoring we use again the uncensored wage information of the 2001 German Structure of Earnings Survey. The same sample restrictions as defined for the simulation studies apply here as well (male West-German residents holding a full-time job covered by social security). First, we calculate daily wages based on the original complete data set to receive reference values. Then, we artificially censor the data set at the 2001 contribution limit for West Germany (286.03 DM, 146.24 euros) and calculate the censored average daily wages. Afterwards the censored wages are multiply imputed $m = 5$ times using the MI approach considering heteroscedasticity and the multiple imputation estimate of the average wages is calculated.

Figure 10.1 displays the corresponding results. In the lower education groups the effect of the censoring is rather negligible, which is not surprising as only 0.9 percent of wages are censored for example in the lowest education group. The average wage of persons holding a technical college (44.9 percent censoring) or university degree (54.5 percent censoring) on the other hand is seriously biased if censored wages are used. While the original average wage of persons holding a technical college degree is 134.58 euros, the censored mean of this group amounts only to 124.56 euros. After censoring and re-imputing the wages the average mean is 133.84 euros. The same situation appears for persons holding

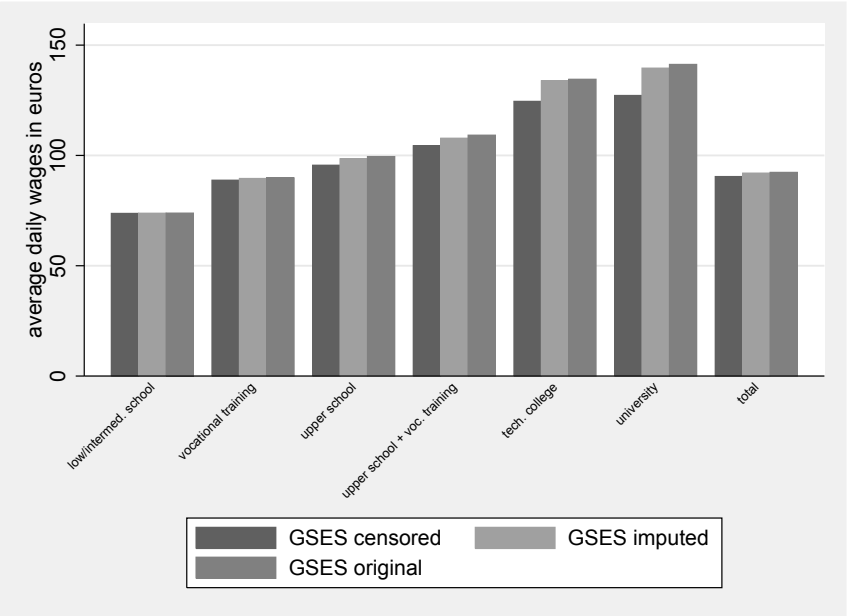

Figure 10.1: Average daily wages by education groups

an university degree. Here, the original mean wage is 141.30 euros, whereas the censored mean is only 127.23 euros and the mean after imputation 139.61 euros. Therefore, it can be concluded that using censored wage information to calculate average wage may lead to seriously biased results, while multiply imputed data allow to calculate more or less unbiased average wages. As already seen in the simulation study, this is especially apparent when highly educated groups are in the center of interest.

10.1.2 Wage Inequality

As already mentioned in the overview of studies based on the IAB Employment Sample (Section 5), a wide range of studies aims to analyze wage difference between and within certain groups over several years. Möller (2005a,b) for example, investigates the wage dispersion between employees working full-time in the lower and upper part of the wage distribution within three education groups using the regional file of the IAB Employment Sample. As a measure of wage dispersion he compares the ratio of the 90th percentile to the median and the ratio of the median to the 10th percentile in different years. The analysis is done separately for men and women and three educational groups:

	GSES censored	GSES imputed	GSES original	Reference IAB data
University or college	1.03	1.40	1.43	n.a
Vocational Training	1.54	1.54	1.54	1.59
No degree	1.40	1.40	1.40	1.42

Table 10.1: Wage inequality for men in West Germany (2001)

- Low-skilled: with no vocational degree
- Medium-skilled: with vocational degree
- High-skilled: with university or technical college degree

While the ratio of the median to the 10th percentile can be easily calculated, an important disadvantage of the IABS for this kind of analysis is that due to the censoring results for the ratio of the 90th percentile to the median can only be shown for the groups of low and medium skilled persons. In these groups the 90th wage percentile is uncensored and results can easily be calculated and reported. The 90th wage percentile of high skilled employees on the other hand is censored because almost 50 percent of wages of men in this group are censored. In this case it is impossible to obtain results without any correction for the censoring. To illustrate this problem, the study is replicated for men in West Germany in the year 2001. To do so, we again impute the daily wages $m = 5$ times. Table 10.1 shows the results concerning the ratio of the 90th percentile to the median. We can see that for the two lower education group the results based on the GSES are the same whether the original, the censored or the multiply imputed data set is used. As the censoring is less than 10 percent in these groups and hence the 90th wage percentile is uncensored, this finding is not surprising. The ratios are in general somewhat lower than the reference results of Möller (2005a,b) based on the IAB data, which may be due to minor differences in the structure of the two samples (GSES and IABS). More interesting are the results concerning the highly educated group. While we find a ratio of only 1.03 based on the censored data, which is highly biased, the result based on the multiply imputed data (1.40) is again comparable to the result based on the original data. As previously discussed a reference result from the literature is not available for this group.

	GSES censored		GSES imputed		GSES original	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Differential						
Prediction men	4.5027	0.0005	4.5220	0.0006	4.5258	0.0006
Prediction women	4.3194	0.0010	4.3244	0.0010	4.3261	0.0010
Difference	0.1833	0.0011	0.1976	0.0011	0.1997	0.0012
Decomposition						
Endowments	0.0407	0.0014	0.0432	0.0014	0.0434	0.0015
Coefficients	0.1340	0.0008	0.1412	0.0009	0.1448	0.0009
Interaction	0.0086	0.0012	0.0132	0.0012	0.0115	0.0013

Table 10.2: Blinder-Oaxaca decomposition of differences in mean wages by gender (All)

10.1.3 Blinder-Oaxaca Decomposition

Analyzing the gender wage gap is another typical research question in economics and social science research which is addressed in a wide range of studies. The counterfactual decomposition technique proposed by Blinder (1973) and Oaxaca (1973) is widely used to study outcome differences between groups, like for example differences by gender. It can be applied to study labor market outcomes by decomposing mean differences in log wages based on regression models in a counterfactual manner. The technique is called counterfactual, because it simulates a counterfactual distribution by combining data on individual characteristics from one distribution with estimated parameters from another. It represents a method that is very suitable to analyze wage differences between men and women. The procedure is known in the literature as the Blinder-Oaxaca decomposition and consists of dividing the wage differential between two groups into a part 'explained' by group differences in productivity characteristics, such as education or work experience, and a residual part that cannot be accounted for by such differences in wage determinants. This 'unexplained' part subsumes the effects of differences in unobserved variables and can often be interpreted as a measure of discrimination. For details see, e.g., Jann (2008) who provides an introduction to this method together with a STATA-ado-file, that can easily be implemented to analyze the gender wage gap.

As a further example to demonstrate the practicability of multiply imputed

	GSES censored		GSES imputed		GSES original	
	Coef.	Std. Err.	Coef.	Std. Err.	Coef.	Std. Err.
Differential						
Prediction men	4.8256	0.0010	4.9181	0.0014	4.9268	0.0015
Prediction women	4.6650	0.0028	4.6963	0.0033	4.7019	0.0034
Difference	0.1606	0.0030	0.2219	0.0036	0.2248	0.0037
Decomposition						
Endowments	0.0734	0.0028	0.0888	0.0033	0.0892	0.0034
Coefficients	0.0912	0.0022	0.1190	0.0027	0.1232	0.0028
Interaction	-0.0041	0.0019	0.0141	0.0023	0.0125	0.0024

Table 10.3: Blinder-Oaxaca decomposition of differences in mean wages by gender (University or college degree)

data, we apply a Blinder-Oaxaca decomposition to analyze differences in mean wages between men and women in West Germany. We again use the GSES 2001 and apply the same sample restrictions as before except for the restriction to male employees. As the aim is to analyze the wage gap between men and women, we need wage information on both genders. We again compare results based on the original complete data, artificially censored data and multiply imputed data. The wage imputation is performed $m = 5$ times using the MI approach considering heteroscedasticity (MI-Het) and is done separately for men and women. As determinants of wage we define here

X_{large}=*(age, age^2, 6 education categories, 4 job level categories, 9 performance groups, 4 region dummies, 36 industry dummies, contract type).*

In consequence we decompose the gender wage gap into a part that is explained by differences in these determinants of wages and a part that cannot be explained by these group differences. Certainly several further models of wage determinants could be used to analyze the gender wage gap. Here a rather simple model is chosen because the focus is just on illustrating the usefulness of the multiply imputed data. Table 10.2 shows the results for the whole sample and Table 10.3 the results for a sample restricted to highly skilled employees holding an university or college degree. These tables report the mean predictions for men and women and the difference between the predictions in the upper panel. In the lower panel of the tables this difference is decomposed into three parts. The endowments effect reflects the mean increase in

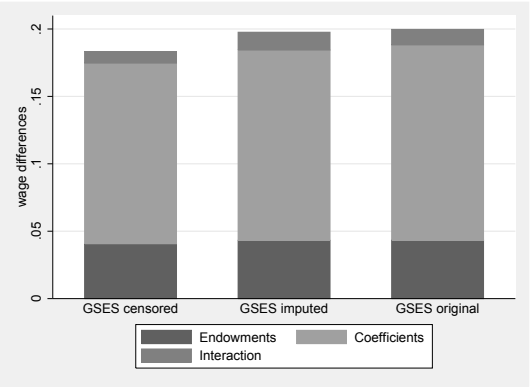

Figure 10.2: Blinder-Oaxaca decomposition results (All persons)

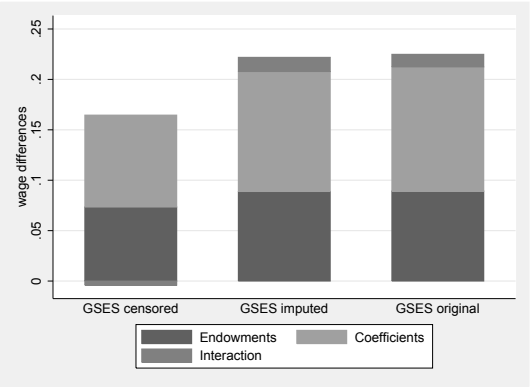

Figure 10.3: Blinder-Oaxaca decomposition results (Univ. or college degree)

women's wages if they would have had the same characteristics as men. The coefficients term indicates the change in women's wage when men's coefficients are applied to the women's characteristics and the interaction effect quantifies the simultaneous effect of differences in endowments and coefficients. Looking at the results based on the original GSES data, we find a mean difference of 0.1997 log points between the wage of men and women. Endowments account for only 0.0434 log points of this difference, while the larger part of the gap cannot be explained by individual characteristics. More interesting are the differences between the censored, multiply imputed, and original wage data. Using censored data, we find a wage gap of only 0.1833, while based on the imputed data we find a wage gap of 0.1976, which is almost identical to the result based on the original complete data. This difference is even more obvious when the sample is restricted to highly skilled employees. Here, we measure a gender wage gap of 0.2248 based on the original complete data. While this result is significantly lower, when censored data are applied (0.1605), multiple imputation yields a similar result (0.2219) compared to the original data. The decomposed effects are smaller when censored data are applied. Figure 10.2 and Figure 10.3 display the results of this example in a graphical form in order to illustrate the differences between the results based on these three different wage variables. The results confirm once more the applicability of the new MI approach to solve the problem of censored wages in an easy way.

The multiply imputed data set for this analysis was produced separately for men and women. This is an adequate approach to account for systematic differences between men and women and to produce data that allow to ana-

lyze differences between these two groups. Alternatively, the wage could be imputed in one step, whereas a dummy indicating the gender is essential in the imputation model. By doing so, the results only change to a minor extent. Then, the decomposition was repeated using a data set, which was imputed ignoring gender differences. Based on this data set, we receive a gender wage gap of 0.1966 for all employees and of 0.2164 for employees holding an university or technical college degree. This finding can be explained by the low percentage of women earning wages that are censored and the explanatory power of other variables, such as occupations and industries, which often employ a high proportion of persons of one gender.

Of course, Blinder-Oaxaca decomposition methods for tobit models can also be applied to analyze censored wage data like the IAB Employment Sample (see, e.g., Kluve and Schaffner (2007) or Bauer and Sinner (2005)). Multiple imputation on the other hand has some advantages compared to these approaches. To begin with, once the data are multiply imputed, they can be used for the analysis of various research questions. Moreover, researchers are able to apply standard techniques and do not have to familiarize themselves with multiple imputation techniques or other models for censored data as described in Chapter 4. Finally, the imputer might use more information in his imputation model which might not be available to the public due to confidentiality reasons.

10.2 First Studies Based on Imputation Approaches

This part of the chapter summarizes first studies that use one of the imputation approaches that were discussed in this thesis in order to show the growing interest in these MI approaches that solve the problem of censored wages. Again the potential of appropriate imputation approaches is illustrated.

Gartner and Rässler (2005) successfully implement the multiple imputation approach based on a tobit model assuming homoscedasticity (MI-Hom) to impute the censored wages in the IAB Employment Sample and to analyze the gender wage gap using a Juhn-Murphy-Pierce decomposition. Their main finding is that there is a general trend of the wage structure that widens the gender wage gap from 1991 to 2001 by 0.0384 log points. On the other side improvements in observed and unobserved endowments, a reduction in gender-

specific sorting and discrimination reduce the gap by 0.1122 log points.

Blien et al. (2009) analyze whether wage differences between cities and rural areas in Western Germany are due to unobserved differences in regional price levels based on the regional file of the IAB Employment Sample. Due to the censoring problem regional prices are available for only 10 percent of the regions. The same multiple imputation approach as in Gartner and Rässler (2005) is applied to be able to generate prices for all regions. The results of the study indicate that the nominal agglomeration wage differential is 2 percent, whereas the real differential is 19 percent. Controlling for the composition of the labor force and jobs, the real wage differential is 4 percent. Controlling additionally for differences in regional building land prices the agglomeration wage differential disappears.

Jensen et al. (2010) use wages imputed multiply based on the approach considering homoscedasticity to estimate earnings frontiers. In particular, individual potential incomes are estimated with stochastic earnings frontiers and overeducation is measured as the ratio between actual income and potential income. The study provides detailed evidence on the influence of experience, tenure, and education on overeducation.

Wages imputed by the single imputation approach considering heteroscedasticity are used by Brücker and Jahn (2008) to measure the wage and employment effects of migration. Here, elasticities of the wage curve for education and experience groups are identified and elasticities of substitution between different types of labor in West Germany during the period from 1980 to 2001 are estimated. As average wages in different subgroups are examined, imputation plays an important role because censoring may be higher than 50 percent in several subgroups. The authors find that the elasticity of the wage curve is particularly high for young workers and workers with an university degree, while it is low for older workers and workers with a vocational degree. The wage and employment effects of migration are found to be moderate: a 1 percent increase in the German labor force through immigration leads to an increase of the aggregate unemployment rate by less than 0.1 percentage points and reduces average wages by less than 0.1 percent.

Büttner et al. (2010) use the same single imputation approach to impute the missing wage information in the register data of the IAB (BeH). In this study they estimate the responsiveness of the occupational skill structure and occupational composition wages to the business cycle and compare the estimates

with corresponding results from a study using U.S. data (Devereux, 2002). This comparison is particularly interesting due to striking differences between U.S. and German labor market institutions. The estimates show that within occupations the skill level of new hires rises significantly in recessions and decreases in upturns. The effects for West Germany amount to about 70 percent of the corresponding U.S. results. They are, however, larger than expected given the striking institutional differences. Separate estimation of the model by establishment size groups suggests that effects are lower for small establishments, implying that a large part of the difference between both countries may be explained by a greater share of small establishments in Germany. Further differentiation of the sample into low and high wage occupations reveals that the share of unskilled is affected more strongly in low wage occupations than in high wage occupations whereas no clear pattern can be found for the high-skilled. The results regarding occupational composition wages also indicate a lower responsiveness to the business cycle than in the U.S. The estimates amount to about 30 and 40 percent of their U.S. counterparts for men and women, respectively.

As the simulation results in this thesis have confirmed the theory that in general multiple imputation is superior to single imputation and that approaches considering heteroscedasticity yield better results, we can conclude that future studies should use wage data multiply imputed using the new MI approach considering heteroscedasticity. For researchers which are interested in applying this approach to the IAB data, the next section summarizes some guidelines for the imputation of missing wages in these data.

10.3 Some Final Suggestions for Imputers

When performing imputation of wages in the IAB data, some suggestions should be considered. If one follows these suggestions, multiple imputation and especially the new approach considering heteroscedasticity are promising techniques, since they are easy to implement and offer potential for a broad range of research questions.

- First, variables to be included in the imputation model have to be chosen carefully. Variables that are good predictors of wages are needed in order to form a model that is appropriate to explain the wages. Our experience

shows that variables like education and age (or tenure) are indispensable for the imputation model.

- Besides, one has to be aware of the implications of the analysis which has to be performed using the imputed data. It is important to reflect relationships that are to be analyzed later in the analysis step. If for example an analysis on a regional level is planned, the regional structure has to be included in the imputation model. Occasionally, it happens that researchers intend to analyze, e.g., the wage returns to certain factors although these factors, possibly the establishment size, are not used in the imputation step. The same applies if differences in wages between employees with German and foreign citizenship are to be analyzed, but not considered in the imputation model. Consequently, it is advisable to include as many variables as possible in the imputation model.
- Most of the recent studies based on the IAB data focus on West Germany for several reasons. First, information for East Germany is not available for years before 1993. Second, the educational and vocational system in the former communist Eastern part differed considerably from the West German part. Moreover the productivity of East German workers may have been lower in the past as they were trained and worked with different and outdated equipment, which complicates many analyzes (see Büttner et al. (2010)). This is not only a challenge in the analysis step, but also for the imputation. A further obstacle are the contribution limits in East Germany, which are lower than in West Germany. Hence, if one is interested in wages in East and West Germany, the best strategy is to impute the wages for both parts separately.
- The imputers also have to pay attention to wage differences between groups, especially between men and women. As discussed before, there is a broad range of studies examining the gender wage gap. This gap between wages of men and women has to be reflected in the imputation model if the analyst is interested in wages of both genders. Then at least an indicator variable for the gender has to be included, even better the imputation should be performed separately
- Part-time workers have a lower monthly and daily wage than full-time workers doing the same job. As no information on hours worked is available in the IAB data, an hourly wage of part-time workers cannot be calculated. Therefore wages of part-time and full-time workers are not

to be compared. This means that part-time workers should be excluded from the sample before starting the imputation. As the proportion of part-time workers with censored wages is almost zero, imputation of wages for this group is normally not necessary anyway.

- Apprentices and marginal employed persons are not comparable to other employees. Wages of apprentices are significantly lower than the contribution limit and marginal employment by law ends with a monthly wage of 400 euros (in 2010). Hence, these groups should not be included in the imputation sample, but can be used for analyses without imputation of wages as well.

Chapter 11

Conclusion and Outlook

Top-coding or right-censoring of wages is a common problem with administrative data sets of economic interest, like the German IAB Employment Sample, which is based on the register data of the German social security system. Censoring of the wage variable is a problem which affects negatively the value of this data set. While in general, the IABS is an unique database in Germany as it covers 80 percent of the workforce and is particulary suitable to analyze a variety of research questions, the censoring hinders these possibilities seriously. Therefore, adapting and developing appropriate techniques for censored data offer new analytic potential. In the literature, there is a wide range of ways to deal with censored wage data. We suggest to use imputation approaches to estimate the missing wage information in order to offer this potential for new analyses and develop a new MI routine. The applicability of the suggested approaches is not restricted to the IAB Employment Sample, but the approaches are generally applicable to all problems of data censoring. The approaches can easily be implemented for cases of right-censoring and left-censoring.

Multiple imputation is especially useful for data sets that are to be shared by many users as it is the case with the IAB Employment Sample. The main advantages of multiple imputation are its general applicability and flexibility and the fact that it allows the data producer to create one 'adjustment' for missing data that can be used by all secondary data analysts (see, e.g., Rässler et al. (2008)). As the model used for the imputation need not to be the same as the model used in the analyses of the completed data, once the data are imputed, e.g., by an organization distributing the data, they can be used by secondary analysts to explore a wide range of models and research questions.

The job of the distributing organization then is to release already completed data sets to the researchers (or the public) or to provide imputation algorithms that are easy to implement by users without detailed knowledge about multiple imputation techniques. Ideally, these algorithms are provided as programme files for software packages that are usually used by the analysts (e.g., STATA, R, or SPSS). A great feature of the multiply imputed data is that secondary users do not have to familiarize themselves with specific techniques to analyze censored data (or incomplete data in general), but are able to perform the desired analyses using standard techniques.

There are different possibilities to impute censored wages (or other censored variables), for example using single and multiple imputation approaches which are presented here. Another important question addressed in this thesis is whether wages should be imputed considering heteroscedasticity. We know that the variance of income is smaller in lower wage categories than in higher categories. Thus we have suggested and developed a new 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 compared to conventional approaches is to perform additional draws for the parameter γ describing the heteroscedasticity in order to allow individual variances for every individual.

The simulation studies presented in this thesis show that compared to single imputation approaches and other regression-based MI approaches it is preferable to use the new multiple imputation approach considering heteroscedasticity. To begin with, we can state that MI approaches are generally superior to single imputation approaches, mainly because single imputation yields variance estimates that are biased, i.e., too small. Simulation studies have demonstrated as well that the suggested approach considering heteroscedasticity leads to better imputation results than approaches assuming homoscedasticity of the residuals. More precisely, we have seen that in case of homoscedastic residuals the same quality of imputation results can be expected compared to the conventional approach suggested by Gartner and Rässler (2005), yet if heteroscedasticity exists a simulation study shows the necessity to apply our new approach. Hence the results reveal to use the new imputation method, as this approach is more general than those based on homoscedasticity.

While these first results are based on generated data sets, the superiority of the new approach is reflected as well in a series of simulation studies using uncensored wage information from a survey (German Structure of Earnings Survey). Two MI approaches (considering heteroscedasticity vs. assuming homoscedasticity) were compared using different models and transformations. In the first step the approaches were evaluated using different transformations of the wage variable. Here, this simulation studies confirm once more the applicability of the multiple imputation approach considering heteroscedasticity. Both approaches deliver good imputation results, with some advantages for the approach considering heteroscedasticity. This result is found, also if a log or a cube root transformation is chosen. Moreover we learned that a log transformation is somewhat more suitable to impute the German wage data than a cube root transformation. Another main finding is that multiply imputed data are robust to differences between the imputer's and analyst's model. For example, once the data are imputed it does not matter if the analyst is interested in an OLS or GLS estimation.

In the same manner simulation studies have shown that imputed data are still appropriate if the analyst examines a model containing a different set of variables. There is only one small constraint to this general finding: If the analyst is interested in a model much smaller than the imputation model, there is no advantage of an imputation considering heteroscedasticity anymore. In exchange, we have seen that the heteroscedastic approach is valid, even if the imputer and analyst apply different wage transformations (i.e log and cube root transformation).

The discussed imputation approaches involve adapting starting values from a tobit estimation. To assess the validity of the suggested approach considering heteroscedasticity compared to other situations, we develop alternative approaches using uncensored wage information from a survey (GSES) instead. These alternative approaches can be distinguished by the quantity of external information required. For a first version the entire external data set is necessary. For a second version only estimation results from an OLS regression are required, while for a last version only information on quantiles is needed. Performing simulation studies, we find similar results of these approaches based on external data and the approach considering heteroscedasticity working without external information. The imputation quality of the approach based on quantile information is even somewhat lower compared to this approach. Hence,

having evaluated various alternatives to the imputation approaches based on starting values from a tobit model, we can conclude that the proposed MI approach considering heteroscedasticity (MI-Het) is easier to implement as it does not require external information and leads to an imputation quality that is at least as good and in some cases even better than approaches that require additional external information. Therefore, we can state once again that it is generally advisable to apply this approach combined with a lognormal transformation of the wages to impute the censored wage information in the IAB Employment Sample.

In the last part of the thesis, three examples show the applicability of the suggested approach considering heteroscedasticity not only in a simulation study, but for real world research questions as well. For these analyses uncensored wage data are used again. Descriptive wage statistics, a wage inequality analysis, and a Blinder-Oaxaca wage decomposition are taken as examples to outline that the analysis of multiply imputed data leads to results which do not significantly differ from the results based on the original complete data. These results underline once more the inherent applicability of multiple imputation to solve the problem of censored wage and to offer potential for new analyses in the IAB Employment Sample and other data sets that may have censored variables.

Having suggested several imputation approaches for censored variables and having shown the validity of these approaches in simulation studies and real world examples, still some future steps are to be performed to make multiple imputed wages accessible to researchers and still remains room for future research in this area. One important issue for future research is the adaption of appropriate models not only for cross-sectional but also for longitudinal data. Apart from this issue of future research, what are future steps to go? As already mentioned, there are basically two ways for organizations distributing data to provide access to multiply imputed data: releasing data sets with already completed wage information or releasing applicable imputation routines. To produce and distribute a multiply imputed version, e.g., of the IABS, has the inherent advantage that researchers do not have to worry about censoring and how to handle it. Therefore, it is desirable to provide a multiply imputed version of the IABS. On the other hand, many studies are not based on the 2 percent sample of the IABS, but on the entire register data (BeH) or other samples of it, sometimes in combination with data stemming from the Bene-

fit Recipient History (LeH). For these cases an already completed version of the IABS would not contribute to solve the problem of censoring. Therefore, a preferable strategy is also to improve, e.g., the STATA imputation routine which has been developed in this thesis. Enhancing this routine with a user friendly graphical user interface to allow researchers without specific knowledge about multiple imputation the use of this routine, would be very useful, but is beyond the scope of this thesis.

Having an executable STATA routine of the multiple imputation algorithm considering heteroscedasticity, like the one that is provided for the single imputation approach suggested by Gartner (2005), which is already used by researchers, wages can easily be imputed by any researcher. As analyses in this thesis have shown, researchers can expect a much better imputation quality by applying the new procedure. Thus researchers should not be deterred by the additional step of combining results from m estimations, which can be performed in the end with little additional effort. Recently published studies based on multiply imputed wages reinforce the idea that multiple imputation is a promising strategy for the future handling of the censoring. Recently received requests for advice on how to apply multiple imputation techniques in the case of censoring, indicate that there is a lot of demand for the approaches suggested.

Appendix

A.1 Additional Simulation Studies

This first part of the appendix contains the results of additional simulation studies that are described in the main part of the thesis (Chapter 8 and 9). Note that the results belong to different chapters of the thesis. Therefore, the simulations studies are based on different complete populations and different true parameters. Moreover, different imputation and analysis models are used in these simulation studies. Hence, the results in this part of the appendix are not directly comparable. The particular simulation designs are described in the corresponding sections in the main part.

		before censoring		MI homosc.		MI heterosc.	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	2.6383	2.6536	0.966	2.6160	0.964	2.6304	0.964
education3	6.8283	6.8692	0.963	4.3236	0.675	4.8369	0.675
education4	7.6188	7.6678	0.955	5.1610	0.154	5.5181	0.154
education5	15.5477	15.6366	0.961	13.2054	0.804	14.7261	0.804
education6	21.8426	21.8579	0.971	15.5703	0.004	17.9200	0.004
level2	1.3935	1.3644	0.950	1.0784	0.850	1.0927	0.850
level3	4.7039	4.6442	0.961	4.4128	0.961	4.6956	0.961
level4	4.0545	3.9673	0.953	1.0908	0.714	0.9354	0.714
group2	-8.6799	-8.6623	0.973	-8.1578	0.652	-8.1302	0.652
group3	-16.3154	-16.3087	0.961	-15.5825	0.496	-15.5373	0.496
group4	-22.9888	-22.9676	0.953	-22.5364	0.770	-22.4284	0.770
group5	46.1605	46.1996	0.951	37.3143	0.112	39.2701	0.112
group6	12.6699	12.7728	0.956	13.8099	0.912	14.1010	0.912
group7	2.5589	2.6626	0.948	5.4599	0.723	5.5946	0.723
group8	-16.6954	-16.6075	0.959	-13.5664	0.685	-13.3963	0.685
group9	-27.0961	-27.0465	0.957	-24.1514	0.744	-23.9741	0.744
age	1.9757	1.9760	0.954	1.9702	0.952	1.9691	0.952
sqage	-0.0193	-0.0193	0.954	-0.0204	0.835	-0.0202	0.835
region2	2.1436	2.1447	0.963	3.5226	0.034	3.3305	0.034
region3	-0.4471	-0.4490	0.952	0.4482	0.445	0.1876	0.445
region4	4.8344	4.8217	0.960	4.1751	0.241	4.1139	0.241
industry2	-3.5415	-3.5135	0.953	-3.1234	0.931	-3.1765	0.931
industry3	-9.3398	-9.3660	0.959	-9.2196	0.970	-9.3459	0.970
industry4	0.0721	0.0444	0.960	0.1242	0.966	0.1166	0.966
industry5	10.4967	10.5577	0.954	7.4321	0.479	7.7971	0.479
industry6	7.7718	7.7859	0.960	7.7584	0.961	7.7313	0.961
industry7	7.0124	7.0273	0.959	6.4575	0.908	6.5608	0.908
industry8	-2.3769	-2.3894	0.969	-1.4234	0.641	-1.4789	0.641
industry9	-2.5212	-2.5238	0.976	-1.3565	0.757	-1.4118	0.757
industry10	1.3130	1.3044	0.963	2.2806	0.705	2.2589	0.705
industry11	-3.9129	-3.9014	0.965	-2.8337	0.571	-2.9022	0.571
industry12	-0.5030	-0.4906	0.953	0.4092	0.630	0.3733	0.630
industry13	-0.6620	-0.6315	0.969	-1.1086	0.934	-1.1187	0.934
industry14	-2.3090	-2.2907	0.957	-2.2851	0.955	-2.2802	0.955
industry15	-3.8808	-3.9100	0.950	-2.8442	0.800	-2.9109	0.800
industry16	2.6383	2.6187	0.959	2.9684	0.942	2.9854	0.942
industry17	-7.1163	-7.1286	0.964	-6.0476	0.745	-6.1463	0.745
industry18	-2.4995	-2.5075	0.963	0.6411	0.000	0.5822	0.000
industry19	-3.1522	-3.1554	0.966	-1.7784	0.387	-1.8480	0.387
industry20	-9.5487	-9.5386	0.962	-8.5285	0.752	-8.6257	0.752
industry21	-7.9048	-7.9296	0.955	-7.0212	0.819	-7.1169	0.819
industry22	-10.4244	-10.4147	0.966	-9.8412	0.925	-10.0037	0.925
industry23	-4.4889	-4.4599	0.966	-4.4471	0.964	-4.5019	0.964
industry24	-15.0993	-15.1053	0.958	-13.4880	0.477	-13.6940	0.477
industry25	-17.1763	-17.2009	0.960	-15.9131	0.805	-16.0613	0.805
industry26	-5.7194	-5.7338	0.962	-4.9173	0.759	-5.0172	0.759
industry27	-3.9572	-3.9841	0.955	-4.1627	0.936	-4.2548	0.936
industry28	-8.5093	-8.5640	0.963	-7.2851	0.692	-7.4131	0.692
industry29	-7.5612	-7.5616	0.956	-5.9144	0.330	-6.0734	0.330
industry30	-7.1680	-7.2070	0.965	-7.0042	0.966	-7.3760	0.966
industry31	-7.4676	-7.5257	0.952	-5.5271	0.599	-5.8264	0.599
industry32	-4.1765	-4.1938	0.955	-4.0840	0.956	-4.2214	0.956
industry33	3.3915	3.3794	0.969	2.6063	0.884	2.5942	0.884
industry34	-11.6666	-11.6861	0.959	-7.4631	0.002	-7.8757	0.002
industry35	2.5266	2.5088	0.950	-1.0483	0.069	-1.0207	0.069
industry36	-17.1727	-17.1619	0.970	-16.5399	0.903	-16.7298	0.903
contract	-7.6749	-7.6837	0.966	-7.6258	0.961	-7.7394	0.961
cons	53.3483	53.3351	0.960	53.8942	0.953	53.8543	0.953

Table A.1: Simulation results based on untransformed wages (Section 8.2.2)

		before censoring		MI-Ext (het.)		MI-Het	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0363	0.0366	0.957	0.0368	0.961	0.0370	0.958
education3	0.0682	0.0690	0.960	0.0608	0.959	0.0626	0.953
education4	0.0732	0.0736	0.969	0.0684	0.955	0.0673	0.925
education5	0.1364	0.1361	0.959	0.1425	0.936	0.1473	0.854
education6	0.1799	0.1797	0.962	0.1787	0.968	0.1776	0.953
level2	0.0097	0.0096	0.960	0.0082	0.956	0.0083	0.954
level3	0.0417	0.0403	0.953	0.0386	0.956	0.0397	0.955
level4	0.0237	0.0223	0.960	0.0081	0.931	0.0059	0.917
group2	-0.0943	-0.0942	0.961	-0.0918	0.947	-0.0918	0.946
group3	-0.1864	-0.1863	0.957	-0.1831	0.934	-0.1831	0.933
group4	-0.3095	-0.3097	0.953	-0.3072	0.931	-0.3070	0.928
group5	0.3854	0.3872	0.957	0.3777	0.956	0.3883	0.952
group6	0.1372	0.1386	0.957	0.1476	0.945	0.1497	0.933
group7	0.0442	0.0458	0.958	0.0588	0.936	0.0607	0.920
group8	-0.1719	-0.1706	0.964	-0.1564	0.928	-0.1542	0.918
group9	-0.3426	-0.3408	0.957	-0.3275	0.941	-0.3251	0.933
age	0.0249	0.0248	0.963	0.0247	0.974	0.0247	0.978
sqage	-0.0003	-0.0003	0.964	-0.0003	0.975	-0.0003	0.976
region2	0.0360	0.0360	0.964	0.0386	0.957	0.0397	0.915
region3	0.0039	0.0040	0.958	0.0063	0.944	0.0063	0.958
region4	0.0517	0.0517	0.956	0.0468	0.825	0.0473	0.850
industry2	-0.0459	-0.0462	0.967	-0.0443	0.957	-0.0448	0.953
industry3	-0.1091	-0.1093	0.965	-0.1115	0.958	-0.1125	0.952
industry4	0.0088	0.0097	0.955	0.0123	0.952	0.0119	0.950
industry5	0.0774	0.0777	0.960	0.0710	0.950	0.0741	0.942
industry6	0.0817	0.0823	0.970	0.0839	0.964	0.0851	0.947
industry7	0.0628	0.0629	0.961	0.0662	0.958	0.0678	0.928
industry8	-0.0144	-0.0141	0.965	-0.0100	0.929	-0.0103	0.931
industry9	-0.0170	-0.0171	0.961	-0.0119	0.948	-0.0127	0.952
industry10	0.0214	0.0213	0.961	0.0264	0.922	0.0263	0.921
industry11	-0.0354	-0.0352	0.963	-0.0313	0.951	-0.0317	0.949
industry12	-0.0030	-0.0035	0.957	0.0013	0.933	0.0017	0.923
industry13	-0.0138	-0.0142	0.954	-0.0179	0.964	-0.0172	0.960
industry14	-0.0302	-0.0301	0.959	-0.0291	0.956	-0.0294	0.951
industry15	-0.0416	-0.0414	0.957	-0.0375	0.962	-0.0378	0.949
industry16	0.0388	0.0388	0.962	0.0416	0.965	0.0424	0.958
industry17	-0.0808	-0.0801	0.954	-0.0765	0.946	-0.0775	0.946
industry18	-0.0088	-0.0086	0.960	0.0071	0.595	0.0074	0.593
industry19	-0.0201	-0.0196	0.966	-0.0130	0.900	-0.0132	0.897
industry20	-0.1055	-0.1050	0.964	-0.1026	0.955	-0.1033	0.958
industry21	-0.0906	-0.0911	0.925	-0.0889	0.886	-0.0894	0.894
industry22	-0.1134	-0.1132	0.956	-0.1149	0.959	-0.1161	0.952
industry23	-0.0527	-0.0528	0.948	-0.0537	0.962	-0.0538	0.953
industry24	-0.1616	-0.1616	0.966	-0.1596	0.971	-0.1610	0.966
industry25	-0.2182	-0.2175	0.964	-0.2129	0.951	-0.2136	0.952
industry26	-0.0579	-0.0577	0.951	-0.0552	0.945	-0.0560	0.949
industry27	-0.0477	-0.0476	0.957	-0.0500	0.959	-0.0508	0.946
industry28	-0.0904	-0.0903	0.952	-0.0873	0.955	-0.0880	0.948
industry29	-0.0702	-0.0702	0.952	-0.0653	0.924	-0.0661	0.931
industry30	-0.0690	-0.0687	0.956	-0.0758	0.939	-0.0786	0.901
industry31	-0.0657	-0.0663	0.966	-0.0633	0.981	-0.0647	0.977
industry32	-0.0596	-0.0603	0.915	-0.0626	0.927	-0.0630	0.927
industry33	0.0085	0.0083	0.957	0.0060	0.968	0.0069	0.954
industry34	-0.0960	-0.0964	0.956	-0.0828	0.833	-0.0850	0.864
industry35	-0.0061	-0.0056	0.963	-0.0200	0.869	-0.0219	0.830
industry36	-0.2607	-0.2608	0.949	-0.2610	0.951	-0.2618	0.953
contract	-0.1116	-0.1116	0.927	-0.1108	0.930	-0.1110	0.938
cons	4.0396	4.0398	0.954	4.0417	0.968	4.0413	0.967

Table A.2: Results of a heteroscedastic imputation using external data versus MI-Het (Section 9.2)

		before censoring		MI-Ext (het.)		MI-Het	
	true β	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage	$\hat{\beta}$	Coverage
education2	0.0728	0.0730	0.967	0.0724	0.965	0.0728	0.966
education3	0.1254	0.1248	0.956	0.1193	0.970	0.1272	0.966
education4	0.1503	0.1506	0.968	0.1439	0.955	0.1467	0.957
education5	0.2973	0.2975	0.964	0.2958	0.981	0.3091	0.896
education6	0.3552	0.3557	0.963	0.3486	0.955	0.3659	0.917
level2	0.0733	0.0730	0.970	0.0713	0.946	0.0712	0.950
level3	0.3393	0.3388	0.967	0.3337	0.950	0.3372	0.957
level4	0.3030	0.3026	0.963	0.2927	0.666	0.2968	0.844
age	0.0401	0.0401	0.961	0.0395	0.950	0.0396	0.951
sqage	-0.0004	-0.0004	0.962	-0.0004	0.971	-0.0004	0.961
region2	0.0394	0.0393	0.966	0.0409	0.964	0.0418	0.945
region3	0.0168	0.0165	0.969	0.0200	0.943	0.0204	0.955
region4	0.0596	0.0594	0.956	0.0533	0.789	0.0537	0.821
industry2	-0.0588	-0.0592	0.960	-0.0570	0.958	-0.0578	0.954
industry3	-0.1231	-0.1231	0.969	-0.1243	0.975	-0.1260	0.967
industry4	0.0025	0.0032	0.955	0.0063	0.964	0.0058	0.959
industry5	0.0594	0.0596	0.966	0.0577	0.975	0.0636	0.964
industry6	0.0795	0.0794	0.967	0.0825	0.963	0.0834	0.951
industry7	0.0534	0.0534	0.957	0.0571	0.958	0.0591	0.937
industry8	-0.0343	-0.0346	0.962	-0.0308	0.954	-0.0314	0.957
industry9	-0.0283	-0.0283	0.965	-0.0232	0.954	-0.0246	0.956
industry10	0.0387	0.0389	0.970	0.0436	0.944	0.0437	0.945
industry11	-0.0374	-0.0376	0.962	-0.0340	0.957	-0.0346	0.956
industry12	0.0129	0.0126	0.948	0.0180	0.919	0.0185	0.916
industry13	0.0038	0.0042	0.956	-0.0008	0.965	0.0006	0.961
industry14	-0.0121	-0.0115	0.967	-0.0114	0.976	-0.0110	0.968
industry15	-0.0288	-0.0287	0.968	-0.0255	0.970	-0.0258	0.966
industry16	0.0594	0.0601	0.951	0.0611	0.966	0.0620	0.958
industry17	-0.0950	-0.0950	0.963	-0.0913	0.960	-0.0926	0.963
industry18	0.0189	0.0187	0.967	0.0349	0.663	0.0349	0.687
industry19	-0.0097	-0.0100	0.969	-0.0037	0.943	-0.0040	0.941
industry20	-0.0984	-0.0983	0.965	-0.0949	0.964	-0.0959	0.962
industry21	-0.0824	-0.0826	0.923	-0.0806	0.907	-0.0813	0.914
industry22	-0.1220	-0.1218	0.962	-0.1228	0.958	-0.1244	0.950
industry23	-0.0763	-0.0763	0.967	-0.0761	0.969	-0.0767	0.968
industry24	-0.2109	-0.2109	0.966	-0.2087	0.965	-0.2115	0.965
industry25	-0.2577	-0.2576	0.956	-0.2530	0.949	-0.2545	0.954
industry26	-0.0825	-0.0828	0.955	-0.0792	0.949	-0.0809	0.960
industry27	-0.0581	-0.0575	0.960	-0.0599	0.966	-0.0610	0.957
industry28	-0.1263	-0.1267	0.957	-0.1233	0.953	-0.1249	0.958
industry29	-0.0721	-0.0717	0.962	-0.0669	0.937	-0.0683	0.952
industry30	-0.0103	-0.0094	0.948	-0.0151	0.963	-0.0159	0.954
industry31	-0.0257	-0.0252	0.954	-0.0216	0.967	-0.0227	0.955
industry32	-0.0746	-0.0770	0.928	-0.0787	0.934	-0.0799	0.939
industry33	0.0058	0.0057	0.959	0.0030	0.977	0.0048	0.965
industry34	-0.1082	-0.1082	0.964	-0.0950	0.881	-0.1000	0.940
industry35	-0.0179	-0.0181	0.966	-0.0341	0.901	-0.0359	0.874
industry36	-0.3250	-0.3243	0.945	-0.3242	0.946	-0.3258	0.948
cons	3.3732	3.3730	0.961	3.3905	0.937	3.3850	0.948

Table A.3: Results of an imputation using external data versus MI-Het (only variables observed in IABS and GSES, Section 9.2)

A.2 Confidence Interval Overlap

In the literature an alternative measure to coverage rates is often applied to assess the quality of an imputation model: the confidence interval overlap. This measure plays an important role in the literature on data confidentiality, where multiple imputation is not used to solve problems of missing information, but rather to provide synthetic data that can be released to researchers or even the public without restrictions due to data protection requirements (see, e.g., Drechsler et al. (2008)). In that context the confidence interval overlap is used to determine the data utility of the synthetic data by looking at the overlap between the confidence intervals for the estimates from the original data and the confidence intervals for the estimates from the synthetic data. This measure is suggested by Karr et al. (2006). The average overlap is calculated for every estimate by:

$$J_k = \frac{1}{2}\left(\frac{U_{over,k} - L_{over,k}}{U_{orig,k} - L_{orig,k}} + \frac{U_{over,k} - L_{over,k}}{U_{syn,k} - L_{syn,k}}\right) \tag{A.1}$$

where $U_{over,k}$ and $L_{over,k}$ denote the upper and the lower bound of the overlap of the confidence intervals from the original and from the synthetic data for the estimate k, $U_{orig,k}$ and $L_{orig,k}$ denote the upper and the lower bound of the confidence interval for the estimate k from the original data, and $U_{syn,k}$ and $L_{syn,k}$ denote the upper and the lower bound of the confidence interval for the estimate k from the synthetic data. This measure can also be applied to examine the validity of imputed data. To do so, we use here the 95 percent interval to calculate the overlap. One serious drawback of this measure is that it depends to a large extent on the sample size. For large samples like the IABS or the GSES the confidence interval overlap between the confidence intervals for the estimates from the original data and the confidence intervals for the estimates from the imputed data may be rather low, even if the estimates are very similar, because the confidence intervals of the estimates are very small. Nevertheless, this approach is applied here to compare results from multiply imputed data to results from tobit model estimation and from censored data. For the analysis the GSES 2001 is used, with the sample restrictions known from the simulation studies. That means the used sample contains $N = 368,337$ persons. The multiple imputation is performed $m = 5$ times using the MI approach considering heteroscedasticity and a log transformation of the wages. To calculate the confidence interval overlap no

repetitions are necessary. This is different to the simulations studies, where we need 1,000 repetitions in order to calculate the coverage. As imputation and analysis model we basically assume again the model containing the wages in logs as dependent variable and the covariates

$$X_{large}=(age,\ age^2,\ 6\ education\ categories,\ 4\ job\ level\ categories,\ 9\ performance\ groups,\ 4\ region\ dummies,\ 36\ industry\ dummies,\ contract\ type),$$

but vary the models to illustrate the impact of an imputation model that differs from the analyst's model. Before the imputation and the application of the tobit estimation the GSES is artificially censored at the real contribution limit of 2001. To obtain the true parameters and the corresponding confidence intervals k, we run an OLS regression using the complete data set and the particular analysis model. In a first step, the imputation is performed once including dummies for education levels and once omitting these dummies. Table A.4 shows the estimates of this first example and the corresponding overlaps for the estimates. Additionally to the displayed variables 36 industry and 3 region dummies were included in the model but omitted from the table for space reasons. The average confidence interval overlap for the imputed data is 77.3 percent when education dummies are included in the imputation model and 72.1 percent when they are omitted. A considerable decrease can be observed for the estimates concerning the education dummies. For the tobit estimation we obtain an average overlap of 73.3 percent, which is somewhat, but not essentially, lower compared to imputed data. The average overlap of an OLS estimation using censored data on the other hand is considerably lower (45.6 percent) and the estimates themselves are extremely biased. Tables A.5 and A.6 show the impact of omitting variables in the imputation model that are to be analyzed in the analysis step. In the second example dummies for the establishment size are additionally included in the analysis model. This enlargement of the model has no impact on the estimation results and the overlap of the estimates concerning the firm size are higher than 90 percent. In the third example indicators of the governmental influence on the particular establishment are added. Here, the corresponding overlap is essentially lower (50.2 percent and 24.5 percent respectively). These examples illustrate that a differing analysis model not necessarily has a negative influence on the quality of the estimation results, but in some cases it may lead to seriously biased results compared to results based on the original complete data set. Summarized, the

main findings of the analyses using the measure of confidence interval overlap are: Using censored wages without any correction is not an applicable method. Multiple imputation yields the best results of the three procedures compared. But if one is only interested in estimates of linear regression, tobit estimation yields an average confidence interval overlap which is only some percentage points lower. On the other hand applying MI has several additional advantages that were already discussed (e.g., the applicability of imputed data for various purposes).

	GSES complete	GSES imputed with education		GSES imputed without education		GSES tobit		GSES censored	
	$\widehat{\beta}$	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap
education2	0.0336	0.0339	0.940	0.0338	0.960	0.0342	0.885	0.0333	0.947
education3	0.0590	0.0550	0.778	0.0521	0.629	0.0532	0.676	0.0438	0.106
education4	0.0697	0.0647	0.482	0.0626	0.265	0.0648	0.493	0.0549	0.000
education5	0.1325	0.1386	0.328	0.1264	0.324	0.1420	0.000	0.1050	0.000
education6	0.1740	0.1741	0.909	0.1504	0.000	0.1749	0.907	0.1215	0.000
level2	0.0101	0.0092	0.837	0.0092	0.834	0.0094	0.870	0.0082	0.603
level3	0.0468	0.0477	0.945	0.0477	0.948	0.0484	0.940	0.0440	0.885
level4	0.0324	0.0234	0.689	0.0256	0.767	0.0259	0.762	0.0160	0.374
group2	-0.0925	-0.0911	0.740	-0.0911	0.736	-0.0911	0.747	-0.0907	0.663
group3	-0.1851	-0.1830	0.690	-0.1830	0.691	-0.1830	0.674	-0.1824	0.565
group4	-0.3024	-0.3006	0.799	-0.3008	0.817	-0.3001	0.697	-0.3024	0.949
group5	0.3789	0.3743	0.840	0.3766	0.921	0.3814	0.906	0.3027	0.000
group6	0.1345	0.1382	0.871	0.1395	0.826	0.1374	0.890	0.1409	0.750
group7	0.0407	0.0483	0.731	0.0484	0.727	0.0464	0.784	0.0574	0.345
group8	-0.1753	-0.1660	0.680	-0.1673	0.726	-0.1682	0.740	-0.1589	0.377
group9	-0.3402	-0.3317	0.746	-0.3331	0.788	-0.3332	0.769	-0.3251	0.484
age	0.0237	0.0238	0.951	0.0238	0.950	0.0241	0.653	0.0235	0.736
sqage	-0.0002	-0.0002	0.818	-0.0002	0.815	-0.0002	0.485	-0.0002	0.716
contract	-0.1353	-0.1338	0.821	-0.1332	0.826	-0.1366	0.838	-0.1293	0.207
cons	4.0687	4.0729	0.873	4.0729	0.876	4.0669	0.936	4.0848	0.421
average overlap			0.773		0.721		0.733		0.456

Table A.4: Comparison of confidence interval overlaps - Example 1

	GSES complete	GSES imputed		GSES tobit		GSES censored	
	$\widehat{\beta}$	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap
firm size >20	0.0298	0.0295	0.968	0.0296	0.988	0.0298	0.948
firm size >50	0.0504	0.0497	0.966	0.0498	0.972	0.0499	0.948
firm size >100	0.0758	0.0747	0.949	0.0754	0.980	0.0740	0.917
firm size >200	0.0855	0.0848	0.964	0.0857	0.989	0.0839	0.923
firm size >500	0.1058	0.1043	0.932	0.1057	0.989	0.1024	0.837
firm size >1000	0.1283	0.1270	0.943	0.1299	0.925	0.1233	0.760
firm size >2000	0.1456	0.1453	0.966	0.1495	0.820	0.1414	0.801
firm size >=2000	0.1442	0.1423	0.911	0.1471	0.862	0.1372	0.660
education2	0.0333	0.0336	0.942	0.0338	0.893	0.0330	0.947
education3	0.0542	0.0501	0.777	0.0482	0.661	0.0393	0.113
education4	0.0677	0.0626	0.475	0.0627	0.483	0.0529	0.000
education5	0.1254	0.1315	0.318	0.1349	0.000	0.0983	0.000
education6	0.1665	0.1665	0.909	0.1674	0.896	0.1144	0.000
level2	0.0200	0.0191	0.828	0.0197	0.932	0.0176	0.498
level3	0.0765	0.0772	0.942	0.0789	0.906	0.0722	0.825
level4	0.0580	0.0488	0.683	0.0521	0.784	0.0404	0.321
group2	-0.0860	-0.0846	0.740	-0.0843	0.702	-0.0844	0.714
group3	-0.1740	-0.1720	0.700	-0.1715	0.602	-0.1719	0.662
group4	-0.2880	-0.2864	0.807	-0.2851	0.620	-0.2888	0.887
group5	0.3601	0.3557	0.844	0.3619	0.934	0.2849	0.000
group6	0.1236	0.1274	0.867	0.1263	0.896	0.1304	0.731
group7	0.0234	0.0312	0.725	0.0287	0.797	0.0409	0.307
group8	-0.1874	-0.1780	0.674	-0.1804	0.743	-0.1705	0.351
group9	-0.3476	-0.3390	0.742	-0.3405	0.766	-0.3322	0.470
age	0.0231	0.0232	0.953	0.0235	0.664	0.0229	0.766
sqage	-0.0002	-0.0002	0.812	-0.0002	0.498	-0.0002	0.682
contract	-0.1403	-0.1388	0.819	-0.1417	0.819	-0.1341	0.178
cons	3.9643	3.9698	0.857	3.9608	0.901	3.9845	0.405
average overlap			0.822		0.786		0.559

Table A.5: Comparison of confidence interval overlaps - Example 2

	GSES complete	GSES imputed		GSES tobit		GSES censored	
	$\widehat{\beta}$	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap	$\widehat{\beta}$	overlap
gov2	0.0038	0.0086	0.502	0.0089	0.455	0.0097	0.346
gov3	-0.0446	-0.0394	0.245	-0.0415	0.558	-0.0332	0.000
education2	0.0336	0.0339	0.942	0.0342	0.886	0.0333	0.945
education3	0.0590	0.0550	0.781	0.0532	0.679	0.0439	0.110
education4	0.0694	0.0644	0.486	0.0645	0.494	0.0547	0.000
education5	0.1323	0.1384	0.325	0.1417	0.000	0.1048	0.000
education6	0.1746	0.1746	0.910	0.1755	0.909	0.1220	0.000
level2	0.0098	0.0090	0.841	0.0092	0.870	0.0080	0.613
level3	0.0469	0.0478	0.945	0.0485	0.937	0.0441	0.887
level4	0.0324	0.0235	0.692	0.0260	0.764	0.0161	0.378
group2	-0.0898	-0.0887	0.779	-0.0885	0.760	-0.0886	0.767
group3	-0.1842	-0.1822	0.693	-0.1821	0.667	-0.1816	0.583
group4	-0.3024	-0.3006	0.790	-0.3000	0.683	-0.3024	0.949
group5	0.3792	0.3746	0.838	0.3818	0.903	0.3030	0.000
group6	0.1348	0.1385	0.873	0.1378	0.891	0.1412	0.754
group7	0.0410	0.0486	0.733	0.0468	0.785	0.0577	0.349
group8	-0.1741	-0.1649	0.684	-0.1670	0.742	-0.1580	0.388
group9	-0.3398	-0.3314	0.746	-0.3328	0.768	-0.3248	0.485
age	0.0237	0.0237	0.951	0.0241	0.657	0.0234	0.735
sqage	-0.0002	-0.0002	0.819	-0.0002	0.491	-0.0002	0.713
contract	-0.1335	-0.1323	0.821	-0.1350	0.814	-0.1280	0.274
cons	4.0660	4.0704	0.867	4.0642	0.938	4.0827	0.401
average overlap			0.739		0.711		0.440

Table A.6: Comparison of confidence interval overlaps - Example 3

Bibliography

Achatz, J., Gartner, H., and Glück, T. (2004). How collective contracts and works councils reduce the gender wage gap. IAB Discussion Paper 2/2004, Nürnberg.

Amemiya, T. (1973). Regression analysis when the dependent variable is truncated normal. *Econometrica*, 41:997–1016.

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, 24:3–61.

Amemiya, T. (1985). *Advanced Econometrics*. Oxford: Blackwell.

An, D. and Little, R. J. A. (2007). Multiple imputation: an alternative to top coding for statistical disclosure control. *Journal of the Royal Statistical Society*, Ser. A, 170:923–940.

Autor, D. H., Katz, L. F., and Kearney, M. S. (2008). Trends in the U.S. wage inequality: Revising the Revisonists. *Review of Economics and Statistics*, 2008:300–323.

Baltagi, B. H., Blien, U., and Wolf, K. (2009). New evidence on the dynamic wage curve for Western Germany: 1980 to 2004. *Labour Economics*, 16:47–51.

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika*, 86:948–955.

Bauer, T. and Bender, S. (2001). Flexible work systems and the structure of wages: Evidence from matched employer-employee data. IZA Discussion Paper no. 353, Bonn.

Bauer, T., Bonin, H., Goette, L., and Sunde, U. (2007). Real and nominal wage rigidities and the rate of inflation: evidence from West German micro data. *The Economic Journal*, 17:F508–F529.

Bauer, T. and Sinner, M. (2005). Blinder-oaxaca decomposition for tobit models. RWI Discussion Paper No. 32.

Beck, M. and Fitzenberger, B. (2004). Changes in union membership over time: A panel analysis for West Germany. *Labour*, 18:329–362.

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.

Berg, G. D. (1998). Extending Powell's semiparametric censored estimator to include non-linear functional forms and extending Buchinsky's estimation technique. University of Colorado Discussion Paper in Economics 98-27.

Binder, J. and Schwengler, B. (2006). Korrekturverfahren zur Berechnung der Einkommen über der Beitragsbemessungsgrenze. IAB Discussion Paper 5/2006, Nürnberg.

Black, S. and Spitz-Oener, A. (2007). Explaining women's success: Technological change and the skill content of women's work. IZA Discussion Paper no. 2803, Bonn.

Blien, U., Gartner, H., Stüber, H., and Wolf, K. (2009). Regional price levels and the agglomeration wage differential in Western Germany. *Annals of Regional Science*, 43:71–88.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *The Journal of Human Resources*, 8:436–455.

Bonin, H. (2005). Wage and employment effects of immigration to Germany: Evidence from a skill group approach. IZA Discussion Paper no. 1875, Bonn.

Braun, S. and Scheffel, J. (2007). Does international outsourcing depress union wages? First evidence from Germany. SFB 649 Discussion Paper 2007-033, Berlin.

Brücker, H. and Jahn, E. J. (2008). Migration and the wage curve: A structural approach to measure the wage and employment effects of migration. IZA Discussion Paper no. 3423, Bonn.

Buchinsky, M. (1994). Changes in the U.S. wage structure 1963-1987: Application of quantile regression. *Econometrica*, 62, No.2:405–458.

Buchinsky, M. and Hahn, J. (1998). An alternative estimator for the censored quantile regression model. *Econometrica*, 66:653–671.

Burkhauser, R. V., Feng, S., Jenkins, S. P., and Larrimore, J. (2008). Estimating trends in U.S. income inequality using the CPS: The importance of controlling for censoring. CES Discussion Papers, Discussion Paper 08-25, Washington DC.

Burkhauser, R. V. and Larrimore, J. (2008). Using internal current population survey data to revaluate trends in labor earnings gaps by gender, race and educational level. CES Discussion Papers, Discussion Paper CES 08-18, Washington DC.

Büttner, T., Jacobebbinghaus, P., and Ludsteck, J. (2010). Occupational upgrading and the business cycle in West Germany. *Economics: The Open-Access, Open-Assessment E-Journal*, 4(2010-10).

Chay, K. Y. and Honoré, B. E. (1998). Estimation of semiparametric censored regression models: An application to changes in black-white earnings inequality during the 1960s. *The Journal of Human Resources*, 33:4–38.

Chay, K. Y. and Powell, J. L. (2001). Semiparametric censored regression models. *The Journal of Economic Perspectives*, 15:29–42.

Chen, S. (2010). Non-parametric identification and estimation of truncated regression models. *Review of Economic Studies*, 77:127–153.

Chen, S., Dahl, G. B., and Khan, S. (2005). Nonparametric identification and estimation of a censored location-scale regression model. *Journal of the American Statistical Association*, 100:212–221.

Chen, S. and Kahn, S. (2000). Estimating censored regression models in the presence of nonparametric multiplicative heteroskedasticity. *Journal of Econometrics*, 98:283–316.

Chernozhukov, V. and Hong, H. (2002). Three-step censored quantile regression and extramarital affairs. *Journal of the American Statistical Association*, 97:872–882.

Chib, S. (1992). Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, 51:79–99.

Czajka, J. L. and Denmead, G. (2008). Income data for policy analysis: A comparative assessment of eight surveys - final report. Mathematica Policy Research Inc., Princeton.

Deutsche Rentenversicherung (2010). Beitragsbemessungsgrenze. http://www.deutsche-rentenversicherung.de/nn_6480/SharedDocs/de/Inhalt/Servicebereich2/Lexikon/B/beitragsbemessungsgrenze.html [20.2.2010].

Deutsche Sozialversicherung (2009). German social insurance. http://www.deutsche-sozialversicherung.de/en/index.html [24.6.2009].

Devereux, P. (2002). Occupational upgrading and the business cycle. *Labour*, 14:423–452.

Drechsler, J., Bender, S., and Rässler, S. (2008). Comparing fully and partially synthetic datasets for statistical disclosure control in the German IAB Establishment Panel. *Transactions on Data Privacy*, 1:105–130.

Drews, N. (2007). Variablen der schwach anonymisierten Version der IAB-Beschäftigten- Stichprobe 1975-2004 * Handbuch-Version 1.0.1. FDZ Datenreport 3/2007, Nürnberg.

Drews, N. (2008). Das Regionalfile der IAB-Beschäftigtenstichprobe 1975-2004 * Handbuch-Version 1.0.2. FDZ Datenreport 2/2008, Nürnberg.

Dustmann, C., Ludsteck, J., and Schönberg, U. (2009). Revisiting the german wage structure. *The Quarterly Journal of Economics*, 124, No. 2:843–881.

Fair, R. (1977). A note on computation of the Tobit estimator. *Econometrica*, 45:1723–1727.

Fair, R. (1978). A theory of extramarital affairs. *Journal of Political Economy*, 86:45–61.

Fisher, L. T. (2007). Measuring the relative importance of social security benefits of the elderly. *Social Security Bulletin*, 67, No. 2:65–72.

Fitzenberger, B., Hujer, R., MaCurdy, T. E., and Schnabel, R. (2001). Testing for uniform wage trends in West Germany: A cohort analysis using quantile regression for censored data. *Empirical Economics*, 26:41–86.

Fitzenberger, B. and Kohn, K. (2006). Gleicher Lohn für gleiche Arbeit? - Zum Zusammenhang zwischen Gewerkschaftsmitgliedschaft und Lohnstruktur in Westdeutschland 1985 bis 1997. ZEW Discussion Paper no. 06-06, Mannheim.

Fitzenberger, B., Osikominu, A., and Völter, R. (2006). Imputation rules to improve the education variable in the IAB Employment Subsample. *Schmollers Jahrbuch : Journal of Applied Social Science Studies / Zeitschrift für Wirtschafts- und Sozialwissenschaften*, 126:405–436.

Fitzenberger, B. and Reize, F. (2002). Quantilsregressionen der westdeutschen Verdienste: Ein Vergleich zwischen der Gehalts- und Lohnstrukturerhebung und der IAB-Beschäftigtenstichprobe. ZEW Discussion Paper no. 02-79, Mannheim.

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.

Gartner, H. and Stephan, G. (2004). Bonus oder Bias? - Mechanismen geschlechtsspezifischer Entlohnung. IAB Discussion Paper 7/2004, Nürnberg.

Gelman, A., Carlin, J. B., Stern, H., and Rubin, D. B. (2003). *Bayesian Data Analysis*. Boca Raton: Chapman & Hall/CRC, 2nd edition.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierachical models*. New York: Cambridge University Press.

Geman, S. and Geman, D. (1984). Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IIEE Trans. Pattern Analysis and Machine Intelligence*, 6:721–741.

Greene, W. H. (2008). *Econometric Analysis*. Upper Saddle River: Prentice Hall, 6th edition.

Haisken-DeNew, J. P. and Frick, J. R. (2005). DTC - Desktop companion to the German socio-economic panel. http://www.diw.de/documents/dokumentenarchiv/17/38951/dtc.354256.pdf [9.7.2009].

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109.

Haupt, H. and Ludsteck, J. (2007). An empirical test of the Reder hypothesis. University of Munich, Department of Economics, Discussion Paper 2007-11, Munich.

Headey, B. and Holst, E. (2008). A quarter century of change: Results from the German socio-economic panel. SOEP Wave Report, 1-2008, Berlin.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, No.1:153–161.

Heeringa, S. G., Little, R. J. A., and Raghunathan, T. E. (2002). Multivariate imputation of coarsened survey data on household wealth. In R.M, G., Dillman, D., Eltinge, J., and Little, J., editors, *Survey Nonresponse*. New York: Wiley.

Heining, J. (2010). The Research Data Centre of the German Federal Employment Agency: Data supply and demand between 2004 and 2009. *Journal for Labour Market Research*, 42:337–350.

Heinze, A. (2009). Earnings of men and women in firms with a female dominated workforce - what drives the impact of sex segregation on wages? ZEW Discussion Paper no. 09-12, Mannheim.

Heinze, A. and Wolf, E. (2006). Gender earnings gap in German firms: The impact of firm characteristics and institutions. ZEW Discussion Paper no. 06-020, Mannheim.

Heinze, A. and Wolf, E. (2007). How to limit discrimination? - analyzing the effects of innovative workplace practices on intra- firm gender wage gaps using linked employer-employee data. ZEW Discussion Paper no. 07-077, Mannheim.

Heitjan, D. F. (1994). Ignorability in general incomplete-data models. *Biometrika*, 81:701–708.

Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *Journal of the American Statistical Association*, 85:304–314.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics*, 19:2244–2253.

Helsel, D. R. (1990). Less than obvious: Statistical treatment of data below detection limit. *Environmental Science and Technology*, 24:1767–1774.

Hirsch, B., Schank, T., and Schnabel, C. (2006). Gender differences in labor supply to monopsonistic firms: An empirical analysis using linked employer-employee data from Germany. IZA Discussion Paper no. 2443, Bonn.

Hirsch, B., Schnabel, C., and Schank, T. (2008). Differences in labor supply to monopsonistic firms and the gender pay gap: An empirical analysis using linked employer-employee data from Germany. Laser Discussion Papers - Paper No. 25, Nürnberg.

Hofer, H. and Weber, A. (2002). Wage mobility in Austria 1986 - 1996. *Labour Economics*, 9:563–577.

Honoré, B. E. (1992). Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects. *Econometrica*, 60:533–565.

Honoré, B. E. and Powell, J. L. (1994). Pairwise difference estimators for censored and truncated regression models. *Journal of Econometrics*, 64:241–278.

Hopke, P. K., Liu, C., and Rubin, D. B. (2001). Multiple imputation for multivariate data with missing and below-threshold measurements: Time-series concentrations of pollutants in the arctic. *Biometrics*, 57:22–33.

Hübler, O. (2003). Geschlechtsspezifische Lohnunterschiede. Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 4/2003, Nürnberg.

Humer, B., Wuellrich, J.-P., and Zweimüller, J. (2007). Integrating severely disabled individuals into the labour market: The Austrian case. IZA Discussion Paper no. 2639, Bonn.

Jann, B. (2008). The Blinder-Oaxaca decomposition for linear regression models. *The Stata Journal*, 8:453–479.

Jenkins, S. P., Burkhauser, R. V., Feng, S., and Larrimore, J. (2009). Measuring inequality using censored data: A multiple imputation approach. DIW Discussion Paper No. 866, Berlin.

Jensen, U., Gartner, H., and Rässler, S. (2010). Estimating German overqualification with stochastic earnings frontiers. *Advances in Statistical Analysis*, 94:33–51.

Juhn, C., Murphy, K. M., and Pierce, B. (1993). Wage inequality and the rise in returns to skill. *Journal of Political Economy*, 101:410–442.

Jurajda, S. and Harmgart, H. (2004). When are 'female' occupations paying more? IZA Discussion Paper no. 985, Bonn.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481.

Karr, A. F., Kohen, C. N., Oganian, A., Reiter, J. P., and Sanil, A. P. (2006). A framework for evaluating the utility of data altered to protect confidentiality. *The American Statistician*, 60:224–232.

Katz, L. F. and Murphy, K. M. (1992). Changes in relative wages, 1963-1987: Supply and demand factors. *Quarterly Journal of Economics*, 107:35–78.

Khan, S. and Powell, J. (2001). Two-step estimation of semiparametric censored regression models. *Journal of Econometrics*, 103:73–110.

Kluve, J. and Schaffner, S. (2007). Gender wage differentials and the occupational injury risk. Ruhr Economic Paper 28, Essen.

Knoppik, C. and Beissinger, T. (2003). How rigid are nominal wages? Evidence and implications for Germany. *The Scandinavian Journal of Economics*, 105:619–641.

Koenker, R. and Basset, G. S. (1978). Regression quantiles. *Econometrica*, 46:33–50.

Kohn, K. (2006). Rising wage dispersion, after all! The German wage structure at the turn of the century. ZEW Discussion Paper no. 06-031, Mannheim.

Larrimore, J., Burkhauser, R. V., Feng, S., and Zayatz, L. (2008). Consistent cell means for topcoded incomes in the public use march CPS (1976-2007). NBER Working Papers, Working Paper 13941, Cambridge.

Lazear, E. P. and Shaw, K. (2007). Wage structure, raises and mobility: International comparisons of the structure of wages within and across firms. NBER Working paper series, Working paper 13654, Cambridge.

Lehmer, F. and Ludsteck, J. (2008). The returns to job mobility and inter-regional migration. IAB Discussion Paper 6/2008, Nürnberg.

Lehmer, F. and Möller, J. (2008). Group-specific effects of inter-regional mobility on earnings - A microdata analysis for Germany. *Regional Studies*, 42:657–673.

Lehmer, F. and Möller, J. (2009). Interrelations between the urban wage premium and firm-size wage differentials: a microdata cohort analysis for Germany. *Annals of Regional Science*, 41:375–400.

Lemieux, T. (2006). Increasing residual wage inequality. Composition effects, noisy data, or rising demand for skill? *The American Economic Review*, 96:461–498.

Lewbel, A. and Linton, O. B. (2002). Nonparametric censored and truncated regression. *Journal of Econometrics*, 97:145–177.

Lewis-Beck, M. S., Bryman, A., and Futing Liao, T., editors (2004). *The SAGE Encyclopedia of Social Science Research Methods*, volume 3. Thousands Oaks: SAGE Publications Inc.

Li, Q. and Racine, J. S. (2007). *Nonparametric Econometrics: Theory and Praxis*. Princeton: Princeton University Press.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley, 1st edition.

Little, R. J. A. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*. Hoboken: Wiley, 2nd edition.

Ludsteck, J. (2008). Wage cyclicality and the wage curve under the microscope. IAB Discussion Paper 11/2008, Nürnberg.

Machado, J. A. F. and Mata, J. (2005). Counterfactual decomposition of changes in wage distributions using quantile regression. *Journal of Applied Econometrics*, 20:445–465.

Maddala, G. S. (2001). *Introduction to Econometrics*. New York: Wiley, 3rd edition.

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

Merz, J., Hirschel, D., and Zwick, M. (2005). Struktur und Verteilung hoher Einkommen-Mikroanalysen auf Basis der Einkommensteuerstatistik. *Beitrag zum zweiten Armuts- und Reichtumsbericht 2004 der Bundesregierung*.

Metropolis, N. and Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, 49:335–341.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, 66:281–302.

Möller, J. (2005a). Die Entwicklung der Lohnspreizung in West- und Ostdeutschland. In Bellmann, L., Hübler, O., Meyer, W., and Stephan, G., editors, *Institutionen, Löhne und Beschäftigung*. Nürnberg: IAB.

Möller, J. (2005b). Lohnungleichheit in West- und Ostdeutschland im Vergleich zu den USA. ZEW Working Paper, www.zew.de/de/publikationen/dfgflex/paperMoeller4.pdf [10.9.2009].

Newey, W. K. (1991). Efficient estimation of tobit models under conditional symmetry. In Barnett, W., Powell, J. L., and Tauchen, G., editors, *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. New York: Cambridge University Press.

Newton, E. and Rudel, R. (2007). Estimating correlation with multiply censored data arising from the adjustment of singly censored data. *Environmental Science and Technology*, 41:221–228.

Oaxaca, R. (1973). Male-female wage differentials in urban labor markets. *International Economic Review*, 12:693–709.

Paarsch, H. (1984). A Monte Carlo comparison of estimators for censored regression models. *Journal of Econometrics*, 24:197–213.

Pan, W. (2000). A multiple imputation approach to cox regression with interval-censored data. *Biometrics*, 56:199–203.

Powell, J. L. (1984). Least absolute deviations estimation for the censored regression model. *Journal of Econometrics*, 25, No. 3:303–325.

Powell, J. L. (1986). Symmetrically trimmed least squares estimation for Tobit models. *Econometrica*, 54:1435–1460.

Powell, J. L. (1994). Estimation of semiparametric models. In Engle, R. F. and McFadden, D. L., editors, *Handbook of Econometrics, Volume VI*. Amsterdam: North Holland.

Prais, S. and Houthakker, H. (1955). *The Analysis of Family Budgets*. New York: Cambridge University Press.

Raghunathan, T. E., Solenberger, P. W., and van Hoewyk, J. (2002). IVEware: Imputation and variance estimation software - User guide. Survey Methodology Program, Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor.

Rässler, S., Rubin, D. B., and Zell, E. R. (2008). Incomplete data in epidemology and medical statistics. *Handbook of Statistics*, 27:569–601.

Reiter, J. P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation. *Journal of the American Statistical Association*, 102:1462–1471.

Robins, P. K. (1985). A comparison of the labor supply findings from the four negative income tax experiments. *The Journal of Human Resources*, 20:567–582.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63:538–543.

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*, 1978:20–40.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business and Economic Statistics*, 4:87–95.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 1st edition.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91:473–489.

Rubin, D. B. (2004a). The design of a general and flexible system for handling nonresponse in sample surveys. *The American Statistician*, 58:298–302.

Rubin, D. B. (2004b). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley, 2nd edition.

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.

Rubin, D. B. and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. *Statistics in Medicine*, 10:585–598.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

Schank, T., Schnabel, C., and Wagner, J. (2004). Exporting firms do not pay higher wages, ceteris paribus. First evidence from linked employer-employee data. IZA Discussion Paper no. 1185, Bonn.

Schönberg, U. (2004). Wage growth due to human capital accumulation and job search: The United States versus West-Germany. Society for Economic Dynamics, 2004 Meeting Papers, Storrs.

Schönberg, U. (2009). Does the IAB Employment Sample reliably identify maternity leave taking? *Journal for Labour Market Research*, 42:49–70.

Schwartz, J. E. (1985). The utility of the cube root of income. *Journal of Official Statistics*, 1:5–19.

Schwarz, N. (2001). The German microcensus. *Schmollers Jahrbuch*, 121:649–654.

Statistische Ämter des Bundes und der Länder (2009). Datenangebot - Lohn- und Einkommensteuerstatistik. http://www.forschungsdatenzentrum.de/bestand/lest/index.asp [9.7.2009].

Stevens, K. (2007). Adult labour market outcomes: the role of economic conditions at entry into the labour market. IZA Conference Paper, Bonn.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, 82:528–540.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36.

Wei, G. C. G. and Tanner, M. A. (1990). Posterior computations for censored regression data. *Journal of the American Statistical Association*, 85:829–839.

Wei, G. C. G. and Tanner, M. A. (1991). Applications of multiple imputation to the analysis of censored regression data. *Biometrics*, 47:1297–1309.

Wooldridge, J. (2002). *Economic Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Zweimüller, J., Winter-Ebmer, R., Lalive, R., Kuhn, A., Wuellrich, J.-P., Ruf, O., and Büchi, S. (2009). Austrian social security database. The Austrian Center for Labor Economics and the Analysis of the Welfare State, Working Paper No. 0903, Linz.