Generalized Partially Linear Regression with Misclassified Data and an Application to Labour Market Transitions

Dlugosz, Stephan; Mammen, Enno; Wilke, Ralf

Document Version
Accepted author manuscript

Published in:
Computational Statistics & Data Analysis

DOI:
10.1016/j.csda.2017.01.003

Publication date:
2017

Generalized Partially Linear Regression with Misclassified Data and an Application to Labour Market Transitions

Stephan Dlugosz, Enno Mammen, and Ralf Wilke

Journal article (Accepted version)

CITE: Generalized Partially Linear Regression with Misclassified Data and an Application to Labour Market Transitions. / Dlugosz, Stephan; Mammen, Enno; Wilke, Ralf. In: *Computational Statistics & Data Analysis*, Vol. 110, 06.2017, p. 145-159.

DOI: 10.1016/j.csda.2017.01.003

Uploaded to Research@CBS: January 2018

Generalised partially linear regression with misclassified data and an application to labour market transitions

Stephan Dlugosz*, Enno Mammen†, Ralf A. Wilke‡

January 3, 2017

Abstract

Large data sets that originate from administrative or operational activity are increasingly used for statistical analysis as they often contain very precise information and a large number of observations. But there is evidence that some variables can be subject to severe misclassification or contain missing values. Given the size of the data, a flexible semiparametric misclassification model would be good choice but their use in practise is scarce. To close this gap a semiparametric model for the probability of observing labour market transitions is estimated using a sample of 20m observations from Germany. It is shown that estimated marginal effects of a number of covariates are sizeably affected by misclassification and missing values in the analysis data. The proposed generalised partially linear regression extends existing models by allowing a misclassified discrete covariate to be interacted with a nonparametric function of a continuous covariate.

Keywords: semiparametric regression, measurement error, side information

*ZEW Mannheim, L7.1, 68161 Mannheim, Germany, E–mail: stephan.dlugosz@googlemail.com

†Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany and National Research University Higher School of Economics, Russian Federation, E–mail: mammen@math.uni-heidelberg.de

‡Corresponding author: Copenhagen Business School, Department of Economics, Porcelaenshaven 16A, 2000 Frederiksberg, Denmark, and ZEW Mannheim, E–mail: rw.eco@cbs.dk

1 Introduction

The increased availability of large scale or big data opens new opportunities for the application of flexible statistical models. These data are for instance generated by public institutions through administrative processes and can comprise a country's entire population of individuals, households or firms. Other examples are internet data which are generated by user activity, or internal firm data that are generated through operational processes. While there has been tremendous progress in the development of non- and semiparametric models since the 1980s (compare for example Ruppert et al., 2003), a gap has evolved between the frontier of methodological research and what is commonly put to data in empirical research in economics and social sciences. In particular, many analysis uses parametric mean regression models or parametric logistic regression which are easy to obtain but do not exploit the richness of the data.

Empirical studies also typically assume that administrative or operational data are precise, free of errors and not subject to misclassification. While these assumptions likely hold for parts of the information they do not hold uniformly. Evidence for deficiencies in operational data has been found in financial transaction data (Chakravarty and Sarkar, 1999), public health registers (Ladouceur et al., 2007), administrative labour market registers (Johansson and Skedinger, 2009, Fitzenberger et al., 2006) and likely more. As these studies use data from different countries and continents (U.S., Sweden and Germany) and relate to different subject areas (Finance, Biometrics and Economics), a wide range of statistical applications is possibly affected by this. We claim that the existing evidence does not show the full scale of the problem due to a lack of research and knowledge about these deficiencies.

Data should be error free if they are directly resulting from operations. This could be for

example reported firm revenues or profits to tax registers or the amount of unemployment benefits paid to the jobless. However, it can also contain considerable degree of misclassification if additional information is collected and made available that is not immediately relevant for the administrative or operational processes and not checked for correctness. For instance this can be further background variables on benefit claiming individuals such as nationality or educational background. If these variables do not enter the equation for determining the level and duration of benefits entitlements, it is likely that this information is not carefully checked by the data producer. Data errors can have different natures. They can be random by accidentally entering the wrong value and not checking for correctness. Or they can be systematic if there are financial consequences for the data producer to over- or underreport certain values. In our application we focus on the educational degree in German administrative employment records, which is known to be prone to missing values and misclassification (Fitzenberger et al., 2006, Kruppe et al., 2014). Although the mechanisms behind these errors are not well researched, they are believed to be random. The affected administrative data are used in much of the academic labour market research about Germany and it serves as an important source of information for the German government and public administration. Our empirical analysis of the relevance of data quality problems in these data for estimating labour market transitions is therefore of wider academic and non-academic interest.

Once data problems are identified, there are good chances that a suitable statistical model for misclassified data or data with missing values has already been developed. Regression models with missing values are typically estimated by (mutliple) imputation methods or by maximum likelihood. See Little and Rubin (2002) for a comprehensive overview of imputa-

tion methods. Liang et al. (2004) suggest a partially linear regression model with missing values in covariates that is estimated by maximum likelihood. Other contributions have considered models with mismeasured variables. See for example Caroll et al. (2006) for a comprehensive overview. Examples of more recent works include Chen et al. (2005), Chen et al. (2008) and Yi et al. (2015) which have in common that they use the method of maximum likelihood estimation and base on the seminal work by Lee and Stepanski (1995). Messer and Natarajan (2008) and Valaste et al. (2010) study the finite sample properties of regression calibration, multiple imputation for measurement error and maximum likelihood estimation by means of simulations. Their results suggest that maximum likelihood based models are preferable as they tend to produce estimates with the smallest mean squared error, in particular if external validation data is used. Caroll et al. (2006) also link mismeasured information to a missing data problem if there is validation data available. Blackwell et al. (2015) use multiple overimputation, a variant of multiple imputation, to address measurement error and missing data simultaneously. In this paper we consider a model with a variable that is mismeasured and possesses missing values. We use external validation data and employ the method of maximum likelihood estimation. As a novelty we allow the misclassified covariate to be interacted with a nonparametric function of a continuous covariate.

We show that our proposed semiparametric generalised linear regression model can be estimated with a sample of 20m observations in a reasonable amount of time. To our knowledge similar models with or without side information have not been applied to such extensive data. Existing studies in economics that use misclassification models use less complex models and much smaller survey data (e.g. Magnac and Visser, 1999, and Hernandez and

Pudney, 2007). In our application we consider nonparametric age profiles in a labour market transition model. These age profiles are allowed to vary freely across educational degrees, where the latter are only observable with errors. We find evidence for practically relevant estimation bias in nonparametric functionals and marginal effects when misclassification is ignored.

The paper is structured as follows. Section 2 contains an informal presentation of our model. Section 3 outlines the general model and Section 4 contains the application to labour market data. Section 5 summarises the main findings.

2 Informal Presentation

We consider a regression model with dependent variable Y and covariates X and U. As a difficulty the analysis data comprises of Y and X only. U is a discrete covariate which is not observed but correlated with X. Omitting U from the model would therefore generally lead to inconsistent results. Instead of U the analysis data contains U^* which is U plus a non-classical measurement error. The measurement error is not assumed to be independent of X but conditionally independent of Y, i.e. $U^* \perp\!\!\!\perp Y|X,U$. Our model does not require that U and U^* have the same support. For example U^* can contain missing values which do not exist for U. Thus, the model does not only allow for misclassification but also for incomplete data (compare e.g. Hartley and Hocking, 1971). In addition to the analysis data, we make use of the existence of validation data for the misclassified U. The validation data contain U, U^* and $W \subseteq X$. Analysis data and validation data are independent samples of the same population but they are not linked and so small in size that we can

assume that they comprise of different population units. It is therefore possible to determine $P(U = u|U^* = u^*, W)$ with the validation data and we assume that the covariates which are in the analysis model but not in the validation data are not informative for the measurement error, i.e. $P(U = u|U^* = u^*, X) = P(U = u|U^* = u^*, W) = p_{u|u^*}$. The availability of validation data therefore allows direct estimation of the error structure. This normally leads to more precise estimation of the analysis model than if validation information was not available (Carroll et al., 2006).

We consider the generalized partial linear model (GPLM):

$$P(Y = y|X, U^*) = \sum_u f(y, \eta(X, u; \boldsymbol{\beta}), \boldsymbol{\theta}) p_{u|u^*}(X) \tag{1}$$

where f is a known density with unknown nuisance parameters $\boldsymbol{\theta}$, η is a semi-parametric regression model, and $p_{u|u^*}$ is a parametric density (for example, as for a multinomial logit model). The sum over u goes over the values on the support of U.

This model for the observed probability is a special case of a more general model and it is motivated by applying the law of total probability to the extended model

$$\begin{aligned} P(Y = y|X, U^*) &= \int P(Y = y, U = u|X, U^*) du \\ &= \int P(Y = y|X, U = u) P(U = u|X, U^*) du \end{aligned}$$

with all the densities understood as Radon-Nykodym derivatives of corresponding probability measures with respect to products of Lebesgue measures and counting measures to allow for both continuous and discrete random variables. The identification of this model is discussed e.g. in Chen et al. (2005) with and in Chen et al. (2008) without using auxiliary data.

The aim is to estimate $\boldsymbol{\theta}$, η, $p_{u|u^*}$ and $\boldsymbol{\beta}$ in model (1) on the basis of the two samples.

This can be done in one step or in two steps. In the latter case the probabilities $p_{u|u^*}$ are first estimated with the validation sample and then plugged into the model. In the second step the remaining unknowns are estimated with the analysis data. Before formally stating our general model we now sketch the simple case of a linear regression model with normal error and a dummy variable U as an illustrating example.

In the linear regression model with normal error ϵ we have $\eta = \eta(x, u; \boldsymbol{\beta}) = \beta_0 + \beta_x x + \beta_u u$, $\epsilon \sim N(0, \sigma^2)$, and $\boldsymbol{\theta} = \sigma$. Suppose we have two random samples $(Y, X, U^*)_i$ for $i = 1, \ldots, n$ and $(U^*, U, W)_j$ for $j = 1, \ldots, m$. In the first step $p_{u|u^*}$ is estimated by for example a standard parametric model such as multinomial logit with the validation data to obtain $\hat{p}_{u|u^*}(X_i)$. In the second step the following log likelihood function is maximized

$$logL(\boldsymbol{\beta}, \sigma) = \sum_{i=1}^{n} \ln \left[\sum_{u=0}^{1} f_\epsilon\big(y_i, \beta_0 + x_i\beta_x + \beta_u u, \sigma\big) \cdot \hat{p}_{u|u_i^*}(x_i) \right]$$

on the grounds of the analysis data with variables Y, X and U^*.

3 The Model

$Y \in \mathbb{Y} \subset \mathbb{R}$ is a discrete or continuous outcome. $\boldsymbol{X} \in \mathbb{X} \subseteq \mathbb{R}^k$ is $1 \times k$-dimensional with discrete or continuous covariates and $Z \in \mathbb{Z} \subseteq \mathbb{R}$ is another continuous covariate. $U^* \in \mathbb{U}^*$ is one dimensional and discrete with finite number of values. $U \in \mathbb{U}$ is also one dimensional with $\mathbb{U} \subseteq \mathbb{U}^*$. U^* contains misclassified information about U. The analysis data comprises of $Y, \boldsymbol{X}, Z, U^*$ while the validation data consists of $U^*, U, \boldsymbol{W}$, where $\boldsymbol{W} \subseteq \{\boldsymbol{X}, Z\}$. The misclassification in U^* is assumed to be not related with the outcome, i.e. $U^* \perp\!\!\!\perp Y | \boldsymbol{X}, Z, U$. $\boldsymbol{\beta}$ is a $k + 1 \times 1$ vector of unknown parameters and η is a partially linear and partially unknown function with $\eta(\boldsymbol{x}, z, u) = (1, \boldsymbol{x})\boldsymbol{\beta} + \gamma_u(z)$, where γ_u are unknown but smooth

functions which are allowed to differ across values of U. Accordingly, let $\boldsymbol{\gamma}$ be the vector of functions γ_u. The analysis model can be then written as

$$P(Y = \mathbf{y}|X, Z, U^*) = \sum_u f(y, \eta(\boldsymbol{X}, Z, u; \boldsymbol{\beta}, \gamma_u), \boldsymbol{\theta}) p_{u|u^*}(\boldsymbol{X}, Z),$$

where f is a known density with unknown nuisance parameters $\boldsymbol{\theta}$, and where the conditional density $p_{u|u^*}$ is specified below.

3.1 Estimation

We assume that analysis data of size n and validation data of size m are two independent random samples from the same population. U and $\boldsymbol{W}$ are therefore observed on the same support in the two samples. The semiparameric analysis model is estimated by Smoothed Local Maximum Likelihood. The estimator is related to the approach by Severini and Wong (1992). The algorithm that we use for estimation is related Severini and Staniswalis (1994), who developed a profile likelihood estimator for GPLM models without misclassification.

In the first step the validation model $P(U = u|U^* = u^*, \boldsymbol{W})$ is estimated by parametric Maximum Likelihood such as probit or multinomial logit. The resulting estimated coefficients are then used to determine $\hat{P}(U = u|U^* = u^*, \boldsymbol{W}) = \hat{p}_{u|u^*}(\boldsymbol{x}, z)$. Whenever some components of $\{\boldsymbol{X}, Z\}$ are not available in the validation data, it is required that they are redundant in the validation model: $P(U = u|U^* = u^*, \boldsymbol{X}, Z) = P(U = u|U^* = u^*, \boldsymbol{W})$. Fitted values for the validation model are computed for all observations of the analysis data. This does not require extrapolation because U and $\boldsymbol{W}$ are observed on the same support in both samples. The fitted values are plugged into the following second stage smoothed log

likelihood

$$logL(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \int \sum_{i=1}^{n} \ln \left[\sum_{u \in \mathbb{U}} f\big(y_i, (1, \boldsymbol{x}_i)\boldsymbol{\beta} + \gamma_u(z), \boldsymbol{\theta}\big) \cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) \right] \cdot K_h(z_i - z) \mathrm{d}z, \quad (2)$$

for the calculation of an estimator for γ_u and into a parametric likelihood

$$logL(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \left[\sum_{u \in \mathbb{U}} f\big(y_i, (1, \boldsymbol{x}_i)\boldsymbol{\beta} + \gamma_u(z_i), \boldsymbol{\theta}\big) \cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) \right], \quad (3)$$

for the estimation of the parametric components. Here $K_h(\cdot)$ is a classical Kernel function which satisfies $K_h(\cdot) > 0$, $\int K_h(x)dx = 1$ and $h > 0$ is a bandwidth. This likelihood is globally maximized in $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}(\cdot)$ at a vector of functions $\mathbb{R} \to \mathbb{R}$, $z \mapsto \gamma_u(z)$ for each u. The resulting estimators are denoted $\hat{\boldsymbol{\beta}}$, $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$, where the latter is a vector whose length is determined by the number of values in $\mathbb{U}$.

One step procedure Instead of pre-estimating the misclassification probabilities with the validation data it is possible to estimate all unknown parameters in one step if the analysis data and the validation data can be jointly used. The smoothed likelihood is then formed of information from the validation and analysis data simultaneously:

$$\begin{aligned} logL(\boldsymbol{\beta}, \boldsymbol{\beta}_v, \boldsymbol{\gamma}, \boldsymbol{\theta}, \boldsymbol{\theta}_v) = \int \sum_{i=1}^{n} \ln \Bigg[\sum_{u \in \mathbb{U}} f\big(y_i, (1, \boldsymbol{x}_i)\boldsymbol{\beta} + \gamma_u(z), \boldsymbol{\theta}\big) \cdot g\big(u, (\mathbf{1}, \boldsymbol{x}_i, z_i, u_i^*)\boldsymbol{\beta}_v, \boldsymbol{\theta}_v\big) \Bigg] \\ \cdot K_h(z_i - z) + \sum_{j=1}^{m} \ln \left[g\big(u_j, (\mathbf{1}, \boldsymbol{w}_j, u_j^*)\boldsymbol{\beta}_v, \boldsymbol{\theta}_v\big) \right] \mathrm{d}z, \end{aligned}$$

where g is a known density function with unknown nuisance parameters $\boldsymbol{\theta_v}$ and $\boldsymbol{\beta}_v$ is a $(k + 3 \times 1)$ vector of unknown parameters of the validation model. For theoretical reasons the one step procedure should be more efficient than the two step procedure. Given that our analysis and validation data are held separately, we cannot estimate the model in one step. In what follows we therefore focus on the two step procedure.

Algorithm For optimizing (2) and (3) the algorithm iterates between optimizing the parametric part with parameters $\boldsymbol{\beta}, \boldsymbol{\theta}$ and the non-parametric part with the smoothed functions $\boldsymbol{\gamma}(\cdot)$, i.e. we have to solve

$$0 = \sum_{i=1}^{n} \frac{d}{d\gamma_u(z)} \ln \left[\sum_{u \in \mathbb{U}} f\big(y_i, (1, \boldsymbol{x}_i)\boldsymbol{\beta} + \gamma_u(z), \boldsymbol{\theta}\big) \cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) \right] \cdot K_h(z_i - z),$$

with respect to $\boldsymbol{\gamma}_u(z)$ and

$$0 = \sum_{i=1}^{n} \frac{d}{d(\boldsymbol{\beta}, \boldsymbol{\theta})^t} \ln \left[\sum_{u \in \mathbb{U}} f\big(y_i, (1, \boldsymbol{x}_i)\boldsymbol{\beta} + \gamma_u(z_i), \boldsymbol{\theta}\big) \cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) \right] \cdot K_h(z_i - z),$$

with respect to the coefficient vector $(\boldsymbol{\beta}, \boldsymbol{\theta})^t$.

The bandwidth selection is done by the method of cross-validation.

Inference Since the distribution of the smoothed local likelihood estimator for $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\theta})$ in (2) is difficult to derive we suggest the following bootstrap procedure for standard errors and other inference statistics. In particular, we bootstrap the analysis data $(y_i, \boldsymbol{x}_i, z_i, u_i^*)$ for model (2) by drawing n times with replacement. Instead of $\hat{p}_{u|u_i^*}$ we use for each bootstrap observation $\hat{p}^b_{u|u_i^*}(\boldsymbol{x}_i, z_i) = \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) + \phi(u_i^*, \boldsymbol{x}_i, z_i)$ where $\phi(u_i^*, \boldsymbol{x}_i, z_i)$ is a random draw from the asymptotic distribution of $\hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) - p_{u|u_i^*}(\boldsymbol{x}_i, z_i)$. Thus, we do not bootstrap the first step of the estimation procedure but use information about the asymptotic distribution of the estimated misclassification probabilities. This procedure is in particular convenient when analysis and validation data are not linked or held in separate locations as in our application.

Computation The main challenges for obtaining estimates and inference statistics stems from the large sample and the slow convergence of the Newton-Raphson-type algorithm. In our code the computation is considerably accelerated through the following measures.

Several time consuming functions are implement in C which resulted in dramatic speed increases. Cross validation and the bootstrap exploit multiple core structures by parallelization. A binning procedure as that in Fan and Marron (1994) is used and starting values of the algorithm are obtained by pre-estimating the model on a much smaller random sample. It took approximately one day to obtain the final point estimates on a mid performance server with 2 multiple core XEON CPUs and at least 64GB RAM. R-sample code for our model is available from the first author. Code for a parametric version of the model is already made available as R-package (`misclassGLM`).

3.2 Discussion of Properties

This subsection provides a discussion of the identifiability of the nonparametric functions and the validity of the bootstrap procedure. A rigorous statement of asymptotic properties and regularity conditions is given in Appendix A.I. A small simulation exercise to illustrate how our proposed misclassification model corrects estimation bias due to misclassification is given in Section S.II in the supplementary material.

Identification of the nonparametric functions $\gamma_u(\cdot)$ We start with a discussion of the model under the simplifying assumption that there are no parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and that the misclassification probabilities $p_{u|u_i^*}(\boldsymbol{x}_i, z_i)$ are known. We will skip the latter condition below but we will keep the first assumption in the following heuristic discussion and in the theoretical appendix. This simplifies the notation. Furthermore, the main aim of our discussion and theory is the study of consistency of the bootstrap. Our main point is that bootstrap of the nonparametric part works as long as the bias part of the nonparametric estimator

is asymptotically negligible. We expect that the same result applies in case that there are parametric components in the model. Because of the faster convergence of the parametric estimators they do not affect the limiting behavior of the nonparametric estimator and of its bootstrap estimator. But assuming that there are no parametric components makes the heuristic discussion and its theoretical justification more transparent. For convenience we also assume that $\mathbb{U} = \mathbb{U}^*$. Then for $u \in \mathbb{U}$ the kernel estimator $\hat{\gamma}_u(z)$ is equal to γ_u where γ_u solves:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\gamma_u} \ln \left[\sum_{u \in \mathbb{U}} f\big(y_i, \gamma_u\big) \cdot p_{u|u_i^*}(\boldsymbol{x}_i, z_i) \right] \cdot K_h(z_i - z).$$

For fixed z, we now use the notation $\hat{f}_i^u = f(y_i, \hat{\gamma}_u(z))$, $\hat{f}_{\eta,i}^u = f_\eta(y_i, \hat{\gamma}_u(z_i))$, $\bar{f}_i^u = f(y_i, \gamma_u(z))$, $\bar{f}_{\eta,i}^u = f_\eta(y_i, \gamma_u(z))$, $f_i^u = f(y_i, \gamma_u(z_i))$, $f_{\eta,i}^u = f_\eta(y_i, \gamma_u(z_i))$, $f_{\eta\eta,i}^u = f_{\eta\eta}(y_i, \gamma_u(z_i))$, and $p_i^u = p_{u|u_i^*}(\boldsymbol{x}_i, z_i)$, where $f_\eta(y, \eta)$ and $f_{\eta\eta}(y, \eta)$ are the first or second derivative of $f(y, \eta)$ with respect to η. With this notation we can rewrite the last equation:

$$0 = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_{\eta,i}^u p_i^u}{\sum_{v \in \mathbb{U}} \hat{f}_i^v p_i^v} K_h(z_i - z)$$

for $u \in \mathbb{U}$. By expansion one gets the following approximation of the right hand side of the last equation:

$$\begin{aligned}
0 \quad \approx \quad & \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\eta,i}^u p_i^u}{\sum_{v \in \mathbb{U}} f_i^v p_i^v} K_h(z_i - z) + \frac{1}{n}\sum_{i=1}^{n} \frac{(\bar{f}_{\eta,i}^u - f_{\eta,i}^u) p_i^u}{\sum_{v \in \mathbb{U}} f_i^v p_i^v} K_h(z_i - z) \\
& - \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\eta,i}^u p_i^u}{(\sum_{v \in \mathbb{U}} f_i^v p_i^v)^2} \sum_{v \in \mathbb{U}} (\bar{f}_i^v - f_i^v) p_i^v K_h(z_i - z) \\
& + \frac{1}{n}\sum_{i=1}^{n} \frac{(\hat{f}_{\eta,i}^u - \bar{f}_{\eta,i}^u) p_i^u}{\sum_{v \in \mathbb{U}} f_i^v p_i^v} K_h(z_i - z) \\
& - \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\eta,i}^u p_i^u}{(\sum_{v \in \mathbb{U}} f_i^v p_i^v)^2} \sum_{v \in \mathbb{U}} (\hat{f}_i^v - \bar{f}_i^v) p_i^v K_h(z_i - z).
\end{aligned}$$

A careful analysis shows that, under regularity conditions for bandwidth h of order $n^{-1/5}$, the error of this expansion is of order $o_P(n^{-2/5})$. The first term $S(z)$ on the right hand

side is of order $O_P(n^{-2/5})$. Note that under our conditions $E[f^u_{\eta,i}p^u_i/(\sum_{v\in\mathbb{U}} f^v_i p^v_i)|z_i] = 0$. Furthermore, one gets by common arguments of kernel smoothing theory that the second and third term is equal to $b(z)n^{-2/5} + o_P(n^{-2/5})$. For the last two terms we get that their sum is approximately equal to

$$\frac{1}{n}\sum_{i=1}^{n} \frac{f^u_{\eta\eta,i}p^u_i}{\sum_{v\in\mathbb{U}} f^v_i p^v_i} K_h(z_i - z)(\hat{\gamma}_u(z) - \gamma_u(z))$$
$$-\frac{1}{n}\sum_{i=1}^{n} \frac{f^u_{\eta,i}p^u_i}{(\sum_{v\in\mathbb{U}} f^v_i p^v_i)^2} \sum_{v\in\mathbb{U}} f^v_{\eta,i}p^v_i(\hat{\gamma}_v(z) - \gamma_v(z))K_h(z_i - z).$$

This can be written as $-\hat{M}(z)(\hat{\gamma}(z)-\gamma(z))$ with an $r\times r$ matrix $\hat{M}(z)$. Here r is the number of elements of $\mathbb{U}$. Furthermore $\hat{\gamma}(z)$ and $\gamma(z)$ are r-dimensional vectors with elements $\hat{\gamma}_u(z)$ or $\gamma_u(z)$, respectively. One can show by standard kernel smoothing theory that $\hat{M}(z) = M(z) + o_P(1)$, where $M(z)$ has (u,v)-elements

$$E\left[\frac{f^u_{\eta,i}p^u_i f^v_{\eta,i}p^v_i}{(\sum_{w\in\mathbb{U}} f^w_i p^w_i)^2}\middle| z\right] f_Z(z),$$

where f_Z is the density of Z. This matrix has full rank if there exists no values $a_u(z)$ with

$$E\left[\sum_{u\in\mathbb{U}} a_u(z)p_{u|u^*}(x,z)f_\eta(y,\gamma_u(z))\middle| z\right] = 0.$$

Suppose that this is not the case. Then, we get that the derivative of

$$E\left[\sum_{u\in\mathbb{U}} p_{u|u^*}(x,z)f(y,\gamma_u(z) + \delta a_u(z))\middle| z\right]$$

with respect to δ is equal to 0. Thus, the values of the likelihood function at the parameter value $\gamma_u(z)$ and at the value $\gamma_u(z) + \delta a_u(z)$ are negligible small for small values of δ and cannot be distinguished by finite samples. If there exists not such a function $a_u(z)$ the matrix $M(z)$ is invertible and we get that

$$\hat{\gamma}(z) - \gamma(z) = M(z)^{-1}b(z)n^{-2/5} + M(z)^{-1}S(z) + o_P(n^{-2/5}).$$

In particular, we get that the function $\gamma_u(z)$ is identifiable. The last expansion is the usual bias-variance decomposition of a kernel estimator. It can be used to determine the asymptotic distribution of $\hat{\gamma}(z)$. The asymptotic properties of $\hat{\gamma}$ are formally stated in Theorem 1 in Appendix A.I.

Consistency of the bootstrap approach We discuss again only the case that the model does not contain parametric components $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, but now we assume that the values of $p_{u|u_i^*}(\boldsymbol{x}_i, z_i)$ are not known and have been estimated in a preliminary data analysis. We suppose that in this data set the sample size is m and that $p_{u|u_i^*}(\cdot,\cdot)$ is estimated with rate $O_P(m^{-1/2})$. We assume that the first data set is independent from the second sample. By an extension of the arguments in the last paragraph one gets with $\hat{p}_i^u = \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i)$ that

$$\begin{aligned}
\hat{\gamma}(z) - \gamma(z) &= M(z)^{-1} b(z) n^{-2/5} + M(z)^{-1} S(z) \\
&+ M(z)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{f_{\eta,i}^u (\hat{p}_i^u - p_i^u)}{\sum_{v \in \mathbb{U}} f_i^v p_i^v} K_h(z_i - z) \\
&- M(z)^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{f_{\eta,i}^u p_i^u}{(\sum_{v \in \mathbb{U}} f_i^v p_i^v)^2} \sum_{v \in \mathbb{U}} f_i^v (\hat{p}_i^v - p_i^v) K_h(z_i - z) \\
&+ o_P(n^{-2/5}) + o_P(m^{-1/2}).
\end{aligned}$$

One can show that up to order $o_P(m^{-1/2})$, the last two terms are equal to their conditional expectation given the first data set. This gives with a matrix valued function W:

$$\begin{aligned}
\hat{\gamma}(z) - \gamma(z) &= M(z)^{-1} b(z) n^{-2/5} + M(z)^{-1} S(z) \\
&+ \sum_{u^* \in \mathbb{U}} \int W(z, u^*, x) (\hat{p}_{\cdot|u^*}(\boldsymbol{x}, z) - p_{\cdot|u^*}(\boldsymbol{x}, z))\, d\boldsymbol{x} \\
&+ o_P(n^{-2/5}) + o_P(m^{-1/2}),
\end{aligned}$$

where $\hat{p}_{\cdot|u^*}(\boldsymbol{x}, z)$ and $p_{\cdot|u^*}(\boldsymbol{x}, z))$ denote the vectors with elements $\hat{p}_{v|u^*}(\boldsymbol{x}, z)$ and $p_{v|u^*}(\boldsymbol{x}, z))$ $(v \in \mathbb{U})$, respectively. The stochastic behaviour of $\hat{\gamma}(z)$ is driven by the second term or by

the third or by both terms, depending on the relation between the rate of convergence for the two sequences $n^{-2/5}$ and $m^{-1/2}$. The most complex situation arises if $n^{-2/5}$ and $m^{-1/2}$ are of the same order. Then $\hat{\gamma}(z) - \gamma(z)$ can be decomposed into three components: a deterministic bias term and two independent stochastic terms, where one comes from the first estimation step and the other arises in the second step. The performance of the bootstrap can be easily understood if one of the two rates $n^{-2/5}$ and $m^{-1/2}$ dominates the other. In this case, in the real world and in the bootstrap world the estimation error of the step with faster rate is negligible. If $n^{-2/5} << m^{-1/2}$ this gives consistency of the bootstrap. If $m^{-1/2} << n^{-2/5}$ the bootstrap distribution is asymptotically equal to the limiting distribution of $M(z)^{-1}S(z)$, thus it gives a consistent estimate of the variance of $\hat{\gamma}(z) - \gamma(z)$ but the bias estimate is asymptotically equal to zero. This can be understood as for related bootstrap methods in standard kernel estimation problems with one estimation step. If $m^{-1/2}$ and $n^{-2/5}$ are of the same order we get that also in the bootstrap world the bootstrap analogues of $M(z)^{-1}S(z)$ and of $\sum_{u^* \in \mathbb{U}} \int W(z, u^*, x)(\hat{p}_{\cdot|u^*}(\boldsymbol{x}, z) - p_{\cdot|u^*}(\boldsymbol{x}, z))\, d\boldsymbol{x}$ are asymptotically independent. Thus, we get, that also in this case the bootstrap gives a consistent estimate of the variance. The validity of our bootstrap procedure is formally established by Theorem 2 in Appendix A.I.

4 Application: Labour Market Transitions

In this section we apply the model of Section 3 by estimating the probability of transitions from employment to other labour market states. A flexible semiparametric statistical model is a natural candidate for the analysis because the data consists of more than 20m

observations. The data contains administrative records which are generated by the German statutory social insurance and by the German Federal Employment Agency through operational processes. Some of the variables result from operational activity directly, such as periods of unemployment benefit claims. Others, such as employment period and salaries, are used for determining pension entitlements. While these variables are believed to be very precise, the data also contains background variables on individuals which are not used in processes. They are only collected and added as supplementary information for statistical analysis and few systematic efforts were made to check for their correctness. A well known example is the education variable in German employment records which is prone to misclassification and missing values (compare Fitzenberger et al., 2006, Dlugosz, 2011, and Kruppe et al., 2014).

In addition to the administrative records we have access to a smaller survey sample that is linked to administrative records. This survey, the ALWA-ADIAB contains precise information about the educational background (Antoni and Seth, 2011). We use this survey as our validation data. While it is linked to administrative records, the survey information is not linked to the main analysis data that we use. Both data sets are held in separate environments and cannot be linked by us for data security reasons.

As analysis data we use the IAB Employment Sample 04- Regional File (IABS). The IABS is a 2% random sample of employees who make payments into the social security system in the period 1975-2004 (Drews, 2008). It is daily spell data comprising start and end dates of employment records and unemployment benefit claim spells. The data also comprise of a number of variables on individual level such as salary, gender, nationality and job characteristics. It also contains information about the employer such as business

sector and geographic location (county). While we consider labour market transitions in the period 1999-2002, we use the information since the year 1980 to construct a number of individual employment history variables such as labour market experience, tenure, previous job changes and past unemployment experiences among other things. We focus on West-Germany and only consider employment with contributions to the public social insurance (thus our analysis excludes minor employment, life-time civil servants and self-employed). Due to the availability of information about the geographic location of the workplace we enrich the analysis data by a number of regional indicators on county level which are provided by the German Federal Statistical Office. In particular, we include information about the type of the region (urban, sub-urban and rural) and the monthly unemployment rate. Table 4 in the Appendix contains the covariate lists of the analysis and the validation model along with some basic descriptive statistics.

Our analysis model relates probabilities for labour market transitions of male full-time employees to a larger set of variables on individual, firm and regional level. In particular, we estimate the probability for an employee in month t to be in one of the following labour market states in month $t+1$:

0: continue employment with existing employer

1: local employer change (same labour market region)

2: distant employer change (different labour market region)

3: unemployment (claiming unemployment benefits)

4: unknown (out of the labour force, not observed in the data)

Our analysis model is a Multinomial Logit Model with base outcome 0. There is a wealth of empirical literature about the empirical analysis of labour market transitions of employees in Germany. Examples include Bergemann and Mertens (2002), Gangl (2003), Bookmann and Steffes (2005), Dütsch and Struck (2011), Wichert and Wilke (2012) and Westerheide and Kauermann (2014). All these papers do not present a satisfactory solution for dealing with misclassification.

U^* is the education variable in the administrative employment records (BeH, Beschäftigtenhistorie). U is the survey based education variable of the ALWA-ADIAB. This variable is only available for the survey population. These variables contain information about the educational degree but not about the years of education. Following Wichert and Wilke (2012) we do not use the raw education variables but consider three groups of educational levels. In particular, $U \in \{$higher education [HE], vocational training [VT], no degree [ND]$\}$. HE corresponds to tertiary education (university or polytechnic/applied university) and VT corresponds to completion of vocational education combined with an apprenticeship. ND is if neither of the former two is applicable. U^* also takes on missing values [NA]. The raw variable has more values but we group them because some of them correspond to almost identical educational backgrounds. The grouping into broader categories eliminates a larger number of inconsistencies in the data without deleting analysis relevant information. U, the validation variable, takes on the same values as U^* but there are no missing values. This is because we have dropped the affected observations (about 1%).

Table 1 reports misclassification probabilities for the education information in the validation data. Findings of previous research are confirmed that there is a sizeable amount of misclassification (compare Kruppe et al., 2014). The observed education information in the

analysis data is incorrect in around every other observation for individuals with "no degree" or "higher education". In order to address missing values and obvious data inconsistencies in this variable, many empirical research applies a heuristic imputation technique (Fitzenberger et al., 2006). In order to obtain some insights how well these imputations work, we apply the IP1 procedure and recompute the misclassification probabilities for the corrected version of U^*. Table 2 confirms that the IP1 correction generally reduces misclassification for VT and HE but fails to do so for ND. It is apparent that ND and HE are often reported as VT in U^*. This wipes out a considerable amount of variation in this variable (provided that it contains ordered information). It is also apparent that IP1 mainly eliminates missing values in U^*. Although, still containing considerable misclassification, the IP1 corrected variable is better than the non-imputed version. We only report results for the corrected variable in our following analysis because we observed that the precision of the misclassification model was higher when there are fewer missing values in the analysis data.

Given that U contains ordered information, we estimate an Ordered Probit Model for $P(U|U^*, \boldsymbol{W})$ as validation model. $\boldsymbol{W}$ contains Z and some components of $\boldsymbol{X}$. The full list of covariates is given in Table 4 in the appendix. Both validation and analysis data are randomly drawn from the population. Given their sizes we do not expect that a notable share of individuals is in both samples and therefore the assumption of independent samples appears innocent. The estimation results and computed estimated marginal effects for this model are given in Table S2 in the supplementary material. U^* and a number of individual background variables are found to sizably affect the estimated probability of observing the true value of education (U). Based on this model we compute $\hat{P}(u_i|u_i^*, \boldsymbol{w}_i)$ which are the estimated probabilities of observing the true value of education for all observations in

Table 1: Misclassification matrix for U^* (grouped raw data), validation sample.

Grouped education	ALWA-ADIAB (U)		
BeH (U^*)	ND	VT	HE
NA	13.47	12.70	11.75
ND	54.26	6.88	2.16
VT	31.98	78.73	34.23
HE	.30	1.70	51.86
Total	100.00	100.00	100.00

Table 2: Misclassification matrix for U^* (grouped, IP1 corrected data), validation sample.

Grouped IP1	ALWA-ADIAB (U)		
BeH (U^*)	ND	VT	HE
NA	0.99	0.40	0.78
ND	53.27	3.59	1.26
VT	45.35	90.96	33.15
HE	.40	5.05	64.81
Total	100.00	100.00	100.00

Table 3: Sample average of $\hat{P}(u_i|u_i^*, \boldsymbol{w}_i)$ tabulated by U^*.

	U			
U^*	ND	VT	HE	Total
NA	4.54	83.33	12.13	100
ND	62.54	37.37	0.09	100
VT	5.70	86.83	7.47	100
HE	0.00	19.53	80.47	100

our validation sample. Table 3 reports the sample average of $\hat{P}(u_i|u_i^*, \boldsymbol{w}_i)$ for all values of U and U^*. It is apparent that also conditional probabilities point to the presence of considerable data errors. Using these results we also compute fitted values $\hat{P}(u_i|u_i^*, \boldsymbol{w}_i)$ for all observations in the analysis data to be used in the second step.

For the analysis model we specify the probability of transiting into one of the labour market states as a partially linear Multinomial Logit Model (PLM):

$$P(Y = j|U, \mathbf{X}, Z) = \frac{\exp((1, \mathbf{x})\boldsymbol{\beta}_j + \gamma_{uj}(z))}{1 + \sum_{h=1}^{4} \exp((1, \mathbf{x})\boldsymbol{\beta}_h + \gamma_{uh}(z))}$$

for $j = 1, \ldots, 4$ and $\gamma_{uj}(z)$ is a nonparametric age (z) profile which is allow to differ across educational degree (u) and labour market state (j). This model is used for the density in the log-likelihoods (2) and (3). One global bandwidth is used for the estimation of all $\gamma_{uj}(z)$, that is obtained by the method of cross validation. Due to the nonlinearity of the model the estimated coefficients are not directly informative. We therefore construct marginal effects of covariate changes on the response probability, holding all other covariates constant at their sample averages.

Estimated marginal effects for our misclassification model (misPLM) and a model that ignores the data errors (PLM) are given in the supplementary material to this paper (Table

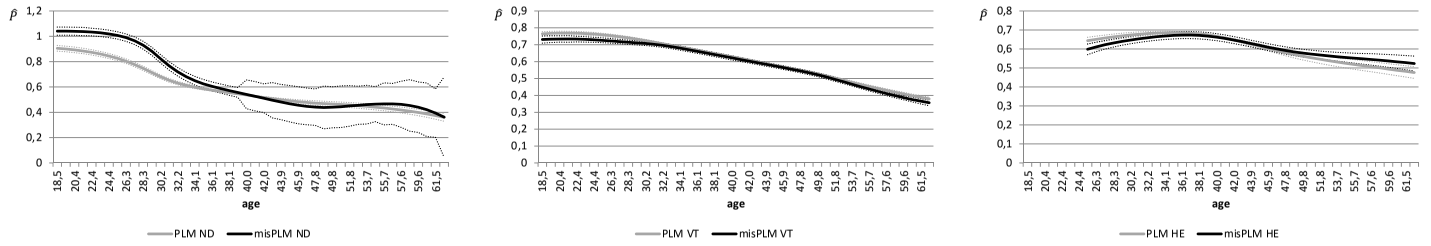
S3), which also contains a detailed discussion of the results (Section S.III). We only describe here briefly the main result pattern related to accounting for the misclassification. There is not a uniform pattern for changes in estimated marginal effects when misclassification has been taken into account. They often increase in size (attenuation bias pattern), but there are also a number of size decreases and even changes in the direction (e.g. the effect of past unemployment periods on local employer changes is negative in PLM and positive in misPLM). Table S3 also provides evidence for the effect of some covariates strongly differing across destination states. This highlights the importance of using a multiple state transition model. In general it is somewhat surprising that some of the estimated effects change sizeably due to misclassification in another variable. In order to explore a possible pattern in the direction of the bias, we have also estimated the marginal effects conditional on the different levels of education. These results are given in Table S4 in the supplementary material. Unfortunately, it is also difficult to determine a systematic bias pattern in these results except maybe that marginal effects for individuals with HE tend to have the largest bias.

Figure 1 shows estimated transition probabilities $\hat{P}(j|U, \boldsymbol{X}, Z)$ in age and by educational degree with their 90% bootstrap confidence intervals. While estimated age profiles for PLM and misPLM often posses a similar shape, they are also significant differences. In some cases the misPLM model produces higher estimated transition probabilities than the PLM and in some cases they are lower. There is no clear pattern in the direction of the bias and the probability functions sometimes cross as age increases. For VT, the value of U that occurs most often in the data and has the fewest errors, the two profiles are very similar. For the other two categories, there are several interesting and important differences. For example

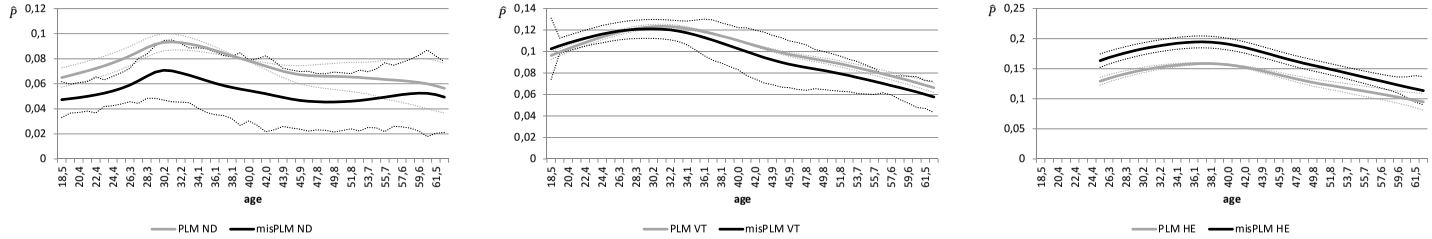
Figure 1: Estimated transition probabilities (in %) in age by education (analysis model).

No Degree | Vocational Training | Higher Education

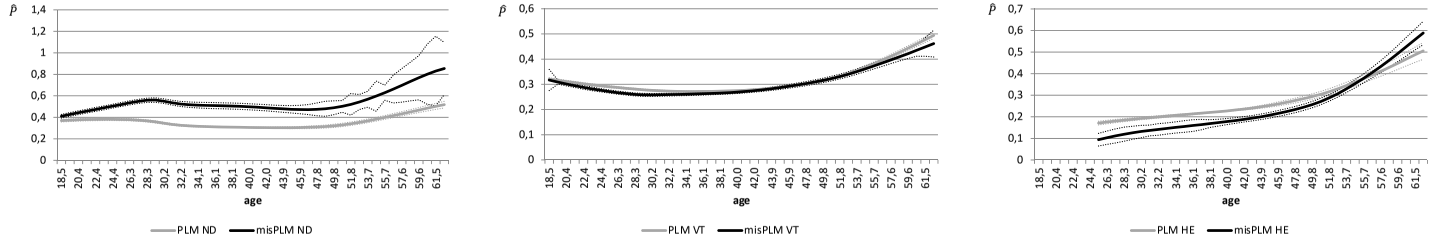
(a) Local employer change



(b) Distant employer change



(c) Transition to unemployment



for ND, the probability of redundancy is estimated to increase by up to one half (50%) when misclassification is taken into account. Similarly for HE the estimated probability for making a distant employer change increases by 20-30% for younger ages in misPLM. While the probability for redundancy decreases by around 60% when misclassification has been accounted for. These changes also lead to different conclusions about the relevance of the educational degree for labour market dynamics. While for instance the results for the PLM suggest that men with ND and VT have similar probabilities of redundancy for higher ages, this complectly changes when the misPLM is applied. Here, the probability of redundancy is estimated to be almost twice as high for ND than for VT, suggesting that the misclassification mixes up the role of the education background for old-age unemployment. The width of the confidence intervals varies strongly for the different functions. This is because of the different number of observed transitions in the relevant age-education combinations. For example for the destination state "distant job change" we obtain the narrowest confidence bands for the group with higher education degree because there are more observed transitions for this group than for the other two education groups.

The economic content of these results is as follows. The probability for a local employer change is estimated to generally non-increase in age, except for men aged less than 35 with higher education degree. When comparing a younger employed with an older employed, the decrease in local job mobility is estimated to be smallest for HE and largest for ND. Men without educational degree have the highest local job transition probabilities in younger ages (in their twenties) but the lowest for higher ages (aged > 40). Men with higher educational degree have the lowest local job probabilities for younger ages (< 30) but the highest for higher ages (> 50). Distant employer changes are estimated to be most likely for men

with higher education degree and least likely for those without any degree. The estimated probability functions increase in younger ages for those with completed vocational training and higher education degree but decrease in age after they have reached their maximum somewhere in the 30s. For those without degree this pattern is least pronounced.

The probability of entering unemployment increases in age for all education groups (with the exception of young ages for men with completed vocational training). The increase in redundancy risk accelerates with age and is strongest for ages > 50. This well known pattern is related to early retirement schemes that use unemployment benefits as a bridge for the time gap between leaving employment and entry into old age pension (compare Wichert and Wilke, 2012, or Westerheide and Kauermann, 2014). Given that our model controls for tenure and additional labour market experience, these results confirm that there is a strong age discriminating pattern. Due to strong dismissal protection laws it is common in the German labour market that the employer negotiates a comprehensive early retirement package with its older employees in exchange of that they accept the redundancy.

5 Summary and Conclusions

We have presented a generalised semiparametric regression model with a misclassified covariate and put this model to extensive administrative labour market data. Our application proves that a flexible, nonlinear misclassification model is operational even with sample sizes in the 2 digit millions. Although, a number of computational challenges had to be resolved, including a substantial amount of code was implemented in `C` and parallel computing was used as much as possible.

Despite the large amount of misclassification, a number of main result patterns remain after accounting for misclassification. This is good news as the data we use are the main source of information for empirical labour market research about Germany. However, there is also evidence of some results being misleadingly biased when the misclassification is ignored. It is therefore preferable in an application to use a model that accounts for misclassification. Somewhat surprisingly we obtain evidence that not only results for the misclassified variable can be sizeably biased but also the marginal effects for other (presumably correct) variables in the model. Given that there is no common attenuation bias pattern in non-linear models with non classical measurement error, it is difficult to anticipate the direction and the size of the bias.

Future research should extend the model to errors in several covariates as it is unlikely that only one variable is affected by misclassification. It would be also of interest to examine the quality of similar data from other countries. From a more practical perspective of users of the German data it would be interesting to conduct a comparative analysis of using the original education variable or the IP1 corrected version each with or without the misclassification model. This would reveal to what extent direct data correction rules, which do not require validation data, contribute to reducing inconsistencies. We experimented with this in early stages and found that a combination of the data correction rule and the misclassification model produced the best results.

Acknowledgements

Financial support by the German Research Foundation (DFG) through research grants FI692/9-2 and Research Training Group RTG 1953 is gratefully acknowledged. Research of the second author was carried out within the framework of a subsidy granted to the HSE by the Government of the Russian Federation for the implementation of the Global Competitiveness Program. The empirical research uses the IABS-04 and ALWA-ADIAB which have been provided by the Research Data Centre of the Institute for Employment Research (IAB-FDZ).

References

[1] Antoni, M. and Seth, S. (2011): ALWA-ADIAB- Linked individual Survey and Administrative Data for Substantive and Methodological Research. FDZ Methodenreport 12/2011. Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

[2] Bergemann, A. and Mertens, A. (2002): Job Stability Trends, Layoffs and Quits - An Empirical Analysis for West Germany. 10th International Conference on Panel Data, Berlin, July 5-6, 2002 C1-4, International Conferences on Panel Data.

[3] Blackwell, M.; Honaker, J. and King, G. (2015): A Unified Approach to Measurement Error and Missing Data: Overview and Applications. Sociological Methods & Research. To appear.

[4] Bookmann, B. and Steffes, S. (2005): Individual and Plant-level Determinants of Job Durations in Germany. ZEW Discussion Paper 05-89, ZEW Mannheim.

[5] Carroll, R.J.; Ruppert, D.; Stefanski, L.A. and Crainiceanu, C.M. (2006): Measurement Error in Nonlinear Models. 2nd Edt., Monographs on Statistics and Applied Probability No. 105, Chapman & Hall.

[6] Chakravarty, S. and Sarkar, A. (1999): Liquidity in U.S. Fixed Income Markets: A Comparison of the Bid-Ask Spread in Corporate, Government and Municipal Bond Markets. FRB of New York Staff Report, No. 73.

[7] Chen, X.; Hong, H. and Tamer, E. (2005): Measurement Error Models with Auxiliary Data. The Review of Economic Studies. 72, 343–366.

[8] Chen, X.; Hu, Y. and Lewbel, A. (2008): Nonparametric identification of regression models containing a misclassified dichotomous regressor without instruments. Economics Letters, 100, 381–384.

[9] Dlugosz, S. (2011): Combined Stochastic and Rule-based Approach to Improve Regression Models with Mismeasured Monotonic Covariates Without Side Information. ZEW Discussion Paper No. 11-013, Mannheim.

[10] Drews,N. (2008): Das Regionalfile der IAB-Beschäftigtenstichprobe 1975-2004. FDZ Methodenreport 02/2008. Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg.

[11] Dütsch, M. and Struck, O. (2011): Individual, Firm-specific and Regional Effects on Internal Employment Trajectories in Germany. University of Bamberg, Chair of Labour Studies, Working Paper No.5.

[12] Fan, J. and Marron, S. (1994): Fast Implementations of Nonparametric Curve Estimators. Journal of Computational and Graphical Statistics, 3(1), 35–56.

[13] Fitzenberger, B.; Osikominu, A. and Völter, R. (2006): Imputation rules to improve the education variable in the IAB employment subsample. Journal of Applied Social Science Studies (Schmollers Jahrbuch), 126(3), 405-436, 2006.

[14] Gangl, M. (2003): Unemployment Dynamics in the United States and West Germany. Heidelberg: Physica.

[15] Hartley, H.O. and Hocking, R.R. (1971): The Analysis of Incomplete Data. Biometrics. 27(4), 783–823.

[16] Hernandez, M. and Pudney, S. (2007): Measurement error in models of welfare participation. Journal of Public Economics. 91(1-2), 327–341.

[17] Johansson, P. and Skedinger, P. (2009): Misreporting in register data on disability status: evidence from the Swedish Public Employment Service. Empirical Economics, 37, 411–434.

[18] Kruppe, T.; Matthes, B. and Unger, S. (2014): Effectiveness of data correction rules in process-produced data: The case of educational attainment. IAB-Discussion Paper, 15/2014, Nürnberg.

[19] Ladouceur, M.; Rahme, E.; Pineau, C.A. and Joseph, L. (2007): Robustness of Prevalence Estimates Derived from Misclassified Data from Administrative Databases. Biometrics, 63, 272-279.

[20] Liang, H.; Wang, S.; Robin, J.M. and Carroll, R.J. (2004): Estimation in Partially Linear Models With Missing Covariates, Journal of the American Statistical Association, 99(466), 357–367.

[21] Little, R.J.A. and Rubin, D.B. (2002): Statistical Analysis with Missing Data. 2nd Edt., John Wiley & Sons.

[22] Lee, L. and Sepanski, J.H. (1995): Estimation of Linear and Nonlinear Errors-in-Variables Models Using Validation Data. Journal of the American Statistical Association, 90, 130–140.

[23] Magnac, T. and Visser, M. (1999): Transition Models With Measurement Errors. The Review of Economics and Statistics, 81(3), 466–474.

[24] Messer, K. and Natarayan, L. (2008): Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. Stat Med., 27(30), 6332-6350.

[25] Ruppert, D.; Wang, M.P and Carroll, R.J. (2003): Semiparametric Regression. Cambridge University Press, Cambridge.

[26] Severini, T.A. and Staniswalis, J.G. (1994): Quasi-likelihood estimation in semiparametric models, Journal of the American Statistical Association, 89, 501–511.

[27] Severini, T.A. and Wong, W.H. (1992): Profile Likelihood and Conditionally Parametric Models. Annals of Statistics, 20(4), 1768–1802.

[28] Valaste, M.; Lehtonen, R. and Vehkalahti, K. (2010): Multiple imputation for measurement error correction in survey data. Unpublished manuscript.

[29] Westerheide, N. and Kauermann, G. (2014): Unemployed in Germany: Factors Influencing the Risk of Losing the Job. Research in World Economy, 5(2), 43–55.

[30] Wichert, L. and Wilke, R.A. (2012): Which factors safeguard employment? An analysis with misclassified German register data. Journal of the Royal Statistical Society A, 175, 135-151.

[31] Yi, G.Y.; Yanyuan, M.; Spiegelman, D. and Carroll, R.J. (2015): Functional and Structural Methods With Mixed Measurement Error and Misclassification in Covariates. Journal of the American Statistical Association, 110(510), 681–696.

Appendix

A.I: Asymptotic Properties of the Kernel Estimator and the Bootstrap

This appendix complements the heuristic discussion of the statistical properties of the kernel estimator $\hat{\gamma}_u(z)$ in Section 3.2 by deriving and proving rigorous asymptotic statements. In particular, we establish the asymptotic distribution of the kernel estimator $\hat{\gamma}_u(z)$ and discuss consistency of the bootstrap procedure.

As in Section 3.2 we make the simplifying assumption that there are no parameters $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ and that $\mathbb{U} = \mathbb{U}^*$. Then in our model we observe i.i.d. tuples $(y_i, \boldsymbol{x}_i, z_i, u_i^*)$ $(i = 1, ..., n)$, where the conditional density of y_i given $(\boldsymbol{x}_i, z_i, u_i^*)$ is

$$\sum_{u \in \mathbb{U}} f\big(y_i, \gamma_u(z_i)\big) \cdot p_{u|u_i^*}(\boldsymbol{x}_i, z_i).$$

Here, $f\big(y_i, \gamma_u(z_i)\big)$ is the conditional density of y_i, given $(\boldsymbol{x}_i, z_i, u_i)$ and $p_{u|u_i^*}(\boldsymbol{x}_i, z_i)$ is the conditional probability of u_i, given $(\boldsymbol{x}_i, z_i, u_i^*)$. As in our general model $f\big(y_i, \gamma_u\big)$ does not depend on z_i and because of our simplifying assumption that there is no parameter β it also does not depend on $\boldsymbol{x}_i$. The conditional probability $p_{u|u_i^*}(\boldsymbol{x}_i, z_i)$ is unknown but an estimate $\hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i)$ of it is available that is based on an additional validation sample. We assume that the validation sample is independent of the sample $(y_i, \boldsymbol{x}_i, z_i, u_i^*)$ $(i = 1, ..., n)$.

Our kernel estimator $\hat{\gamma}(z) = (\hat{\gamma}_u(z))_{u \in \mathbb{U}}$ is equal to $\gamma = (\gamma_u)_{u \in \mathbb{U}}$ where γ maximizes

$$\frac{1}{n} \sum_{i=1}^{n} \ln \left[\sum_{u \in \mathbb{U}} f\big(y_i, \gamma_u\big) \cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i, z_i) \right] \cdot K_h(z_i - z).$$

We will discuss the asymptotic distribution of $\hat{\gamma}$ at a fixed point $z = z_0$. We also write $\hat{\gamma} = (\hat{\gamma}_u)_{u \in \mathbb{U}}$ for $\hat{\gamma}(z_0)$. The true function γ will be denoted by $\gamma_0 = (\gamma_{0,u})_{u \in \mathbb{U}}$.

The estimator $\hat{\gamma}$ is defined by

$$G_n(\hat{\gamma}) = 0,$$

where

$$\begin{aligned} G_n(\gamma) &= \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\gamma} \ln\left[\sum_{u\in\mathbb{U}} f\big(y_i,\gamma_u\big)\cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i,z_i)\right]\cdot K_h(z_i-z_0) \\ &= \left(\frac{1}{n}\sum_{i=1}^{n} \frac{f_\eta\big(y_i,\gamma_u\big)\cdot \hat{p}_{u|u_i^*}(\boldsymbol{x}_i,z_i)}{\sum_{v\in\mathbb{U}} f\big(y_i,\gamma_v\big)\cdot \hat{p}_{v|u_i^*}(\boldsymbol{x}_i,z_i)}\cdot K_h(z_i-z_0)\right)_{u\in\mathbb{U}} \end{aligned}$$

with $K_h(\zeta) = h^{-1}K(h^{-1}\zeta)$.

In our notation we will use the following convention. Suppose that a random variable $r = (s,t)$ has a continuous component s with values in $\mathbb{R}^d$ and a discrete component t with values in a finite set J and suppose that r has a density ϕ with respect to the product of the Lebesgue measure on $\mathbb{R}^d$ and the counting measure on J, that is $\mathbb{E}[h(s,t)] = \sum_{t\in J}\int_{\mathbb{R}^d} h(s,t)\phi(s,t)\mathrm{d}s$. In this case we also write $\int h(s,t)\phi(s,t)\mathrm{d}s\ \mathrm{d}t$ instead of $\mathbb{E}[h(s,t)]$.

In our theory we will make use of the following assumptions.

(A1) For $u\in\mathbb{U}$ the functions $\gamma_{0,u}$ are twice continuously differentiable in a neighborhood of the point z_0. The point z_0 lies in the interior of the support of z_i. The density $g(u^*,\boldsymbol{x},z)$ of $(u_i^*,\boldsymbol{x}_i,z_i)$, with respect to products of Lebesgue measures and counting measures, is differentiable with respect to z in a neighborhood of z_0. The derivative is continuous in z in a neighborhood of z_0, uniformly in z and over $\boldsymbol{x}\in\mathbb{X}$ and $u^*\in\mathbb{U}$. The transition density $p_{u^*,u}(\boldsymbol{x},z)$ is differentiable with respect to z in a neighborhood of z_0. The derivative is continuous in z in the neighborhood of z_0, uniformly in $z\in\mathbb{Z}$ and over $\boldsymbol{x}\in\mathbb{X}$ and $u,u^*\in\mathbb{U}$.

(A2) The density $f(y,\eta)$ is three times differentiable with respect to η for all y and for all $\eta\in I$. Here I is an interval that is chosen such that $\gamma_{0,u}(z)\in I$ for all $u\in\mathbb{U}$ and for

all z in a neighborhood of z_0. For some functions ρ_1,... ρ_4 it holds for all $\eta, \eta' \in I$ that

$$\frac{f(y,\eta)}{f(y,\eta')} \leq \rho_1(y), \qquad \left|\frac{f_\eta(y,\eta)}{f(y,\eta)}\right| \leq \rho_2(y),$$
$$\left|\frac{f_{\eta\eta}(y,\eta)}{f(y,\eta)}\right| \leq \rho_3(y), \qquad \left|\frac{f_{\eta\eta\eta}(y,\eta)}{f(y,\eta)}\right| \leq \rho_4(y),$$

where the functions ρ_1,... ρ_4 for some $0 < C < \infty$ fulfill for z in a neighborhood of z_0

$$\mathbb{E}[\rho_1^8(y)|z] \leq C, \qquad \mathbb{E}[\rho_2^4(y)|z] \leq C,$$
$$E[\ \rho_3^4(y)|z] \leq C, \qquad E[\rho_4^2(y)|z] \leq C.$$

(A3) The matrix M is invertible where M is defined by its elements

$$M_{u,w} = \int \frac{f_\eta(y,\gamma_{0,u}(z_0))p_{u|u^*}(\boldsymbol{x},z_0)f_\eta(y,\gamma_{0,w}(z_0))p_{w|u^*}(\boldsymbol{x},z_0)}{\sum_{v\in\mathbb{U}} f_\eta(y,\gamma_{0,v}(z_0)p_{v|u^*}(\boldsymbol{x},z_0)} \times g(u^*,x,z_0)\mathrm{d}y\ \mathrm{d}u^*\ \mathrm{d}\boldsymbol{x},$$

$(u, w \in \mathbb{U})$.

(A4) It holds that

$$\mathbb{E}\left[\sup_{|\eta'-\eta|\leq\delta} |f_\eta(y,\eta)|\right] < \infty, \qquad \mathbb{E}\left[\sup_{|\eta'-\eta|\leq\delta} |f_{\eta\eta}(y,\eta)|\right] < \infty.$$

(A5) Put

$$S_{1,n} = \int Q(u^*,\boldsymbol{x})\left(\hat{p}_{w|u^*}(\boldsymbol{x},z_0) - p_{w|u^*}(\boldsymbol{x},z_0)\right)_{w\in\mathbb{U}} \mathrm{d}u^*\ \mathrm{d}\boldsymbol{x},$$

where the matrix valued function $Q(u^*, \boldsymbol{x})$ has elements

$$Q(u^*,\boldsymbol{x})_{v,w} = \int \frac{f_\eta(y,\gamma_{0,v}(z_0))p_{v|u^*}(\boldsymbol{x},z_0)f(y,\gamma_{0,w}(z_0))}{\sum_{u\in\mathbb{U}} f_\eta(y,\gamma_{0,u}(z_0)(z)p_{u|u^*}(\boldsymbol{x},z_0)} \times g(u^*,x,z_0)\mathrm{d}y.$$

It holds that

$$\sqrt{m}S_{1,n} \to \mathrm{N}(0,\Sigma_1),$$

in distribution, for a symmetric positive semidefinite matrix Σ_1. Furthermore it holds that

$$\sqrt{m}\sup\Big\{\big|\hat{p}_{u|u^*}(\boldsymbol{x},z)-p_{u|u^*}(\boldsymbol{x},z)\big| : u|u^*\in\mathbb{U},\boldsymbol{x}\in\mathbb{X},z\in\mathbb{Z},\\ |z-z_0|\le h\Big\}=O_P(1),$$

$$\sqrt{m}\sup\Big\{\big|\hat{p}_{u|u^*}(\boldsymbol{x},z)-\hat{p}_{u|u^*}(\boldsymbol{x},z_0)\big| : u|u^*\in\mathbb{U},\boldsymbol{x}\in\mathbb{X},z\in\mathbb{Z},\\ |z-z_0|\le h\Big\}\to 0,$$

in probability.

(A6) The kernel K is a symmetric probability density function with compact support, $[-1,1]$ say. For the bandwidth h is holds that $nh\to\infty$ and $h=O(n^{-1/5})$.

We briefly comment on our assumptions. Assumption (A1) is a standard smoothness condition. Condition (A2) is used to get limit results for some terms using the theorem of dominated convergence. Assumption (A3) is essential for identification of γ_0. Assumption (A4) is used to allow for interchanging the order of integrating and taking derivatives. This allows to show that some expectations are equal to zero. The same argument is used in asymptotics for parametric maximum likelihood theory. Assumptions (A5) can be easily checked under mild regularity conditions if a parametric estimator is used in the validation step. Condition (A6) is a standard assumption used in kernel smoothing.

Theorem 1 *Make the assumptions (A1)–(A6). It holds that*

$$\hat{\gamma}-\gamma_0(z_0)=-h^2\int\zeta^2K(\zeta)\mathrm{d}\zeta\ M^{-1}b-M^{-1}S_{1,n}-M^{-1}S_{2,n}+o_P(m^{-1/2}+(nh)^{-1/2}),$$

where for $u \in \mathbb{U}$ *and for* $i = 1, ..., n$

$$\begin{aligned} S_{2,n,u} &= \frac{1}{n}\sum_{i=1}^{n} e_{i,u} K_h(z_i - z_0), \\ e_{i,u} &= \frac{f_\eta(y_i, \gamma_{0,u}(z_0))}{\sum_{v\in\mathbb{U}} f(y_i, \gamma_{0,v}(z_i)) p_{v|u_i^*}(\boldsymbol{x}_i, z_i)} p_{u|u_i^*}(\boldsymbol{x}_i, z_i). \end{aligned}$$

Furthermore, the bias vector b *has elements*

$$\begin{aligned} b_u = \int \sum_{w\in\mathbb{U}} & \frac{f_\eta(y, \gamma_{0,u}(z_0)) p_{u|u^*}(\boldsymbol{x}, z_0)}{\sum_{v\in\mathbb{U}} f(y, \gamma_{0,v}(z_0) p_{v|u^*}(\boldsymbol{x}, z_0)} g(u^*, x, z_0) \\ & \times \Bigg[f_\eta(y, \gamma_{0,w}(z_0)) \Bigg\{ \gamma'_{0,w}(z_0) \times \Bigg(\frac{\partial_z p_{w|u^*}(\boldsymbol{x}, z_0)}{p_{w|u^*}(\boldsymbol{x}, z_0)} \\ & + \frac{\partial_z p_{u|u^*}(\boldsymbol{x}, z_0)}{p_{u|u^*}(\boldsymbol{x}, z_0)} + \frac{\partial_z g(u^*, x, z_0)}{g(u^*, x, z_0)} - \frac{\sum_{v\in\mathbb{U}} f(y, \gamma_{0,v}(z_0) \partial_z p_{v|u^*}(\boldsymbol{x}, z_0)}{\sum_{v\in\mathbb{U}} f(y, \gamma_{0,v}(z_0) p_{v|u^*}(\boldsymbol{x}, z_0)} \Bigg) \\ & + \frac{1}{2}\gamma''_{0,w}(z_0) \Bigg\} + \frac{1}{2} f_{\eta\eta}(y, \gamma_{0,w}(z_0)) \gamma'_{0,w}(z_0)^2 \Bigg] \mathrm{d}y \ \mathrm{d}u^* \ \mathrm{d}\boldsymbol{x}. \end{aligned}$$

Here $\partial_z p_{w|u^*}$ *and* $\partial_z g$ *denote the partial derivative of* $p_{w|u^*}$ *or of* g*, respectively, with respect to* z*. Furthermore, it holds that*

$$\Sigma_n^{-1/2} \left(\hat{\gamma} - \gamma_0(z_0) - h^2 \int \zeta^2 K(\zeta) \mathrm{d}\zeta \ M^{-1} b \right)$$

converges to a normal distribution with identity covariance matrix. Here the matrix Σ_n *is equal to*

$$\frac{1}{m} M^{-1} \Sigma_1 M^{-1} + \frac{1}{nh} M^{-1}.$$

The limiting covariance matrix Σ_n depends on the ratio of m and nh. If nh/m converges to zero the asymptotic covariance is asymptotically related only to the estimation error in the validation step. If this ratio converges to infinity the covariance is driven by the stochastic errors of the second step. If m and nh are of the same order, the asymptotic variance has terms coming from both estimation steps. The proof of Theorem 1 can be found in Section S.I of the supplementary material.

We now come to a discussion of the properties of our bootstrap procedure. Our bootstrap method makes use of bootstrap samples $(y_i^b, \boldsymbol{x}_i^b, z_i^b, u_i^{*b})$ $(i = 1, ..., n)$ drawn independently with replacement from $(y_i^b, \boldsymbol{x}_i^b, z_i^b, u_i^{*b})$ $(i = 1, ..., n)$. Furthermore, independently

from the bootstrap sample one generates independent $\phi_{i,u}(u^*, \boldsymbol{x}, z)$ $(i = 1, ..., n)$ such that $\sqrt{m}\phi_{i,u}(u^*, \boldsymbol{x}, z)$ is distributed according to the asymptotic distribution of $\sqrt{m}(\hat{p}_{u|u^*}(\boldsymbol{x}, z)$. This is formalised in Assumption (A7) below. Define $\hat{p}^b_{u|u^*_i}(\boldsymbol{x}_i, z_i) = \hat{p}_{u|u^*_i}(\boldsymbol{x}_i, z_i) + \phi_{i,u}(u^*_i, \boldsymbol{x}_i, z_i)$. The bootstrap estimator $\hat{\gamma}^b$ of $\gamma_0(z_0)$ is such that it solves $G^b_n(\hat{\gamma}^b) = 0$, where

$$G^b_n(\gamma) \quad = \quad \left(\frac{1}{n} \sum_{i=1}^{n} \frac{f_\eta\big(y^b_i, \gamma_u\big) \cdot \hat{p}^b_{u|u^{*b}_i}(\boldsymbol{x}^b_i, z^b_i)}{\sum_{v \in \mathbb{U}} f\big(y^b_i, \gamma_v\big) \cdot \hat{p}^b_{v|u^{*b}_i}(\boldsymbol{x}^b_i, z^b_i)} \cdot K_h(z^b_i - z_0) \right)_{u \in \mathbb{U}}.$$

We consider the asymptotic behaviour of the bootstrap procedure under the following additional assumption.

(A7) Put

$$\begin{aligned} S^b_{1,n} &= \int Q(u^*, \boldsymbol{x}) \left(\hat{p}^b_{w|u^*}(\boldsymbol{x}, z_0) - \hat{p}_{w|u^*}(\boldsymbol{x}, z_0) \right)_{w \in \mathbb{U}} \mathrm{d}u^* \ \mathrm{d}\boldsymbol{x} \\ &= \int Q(u^*, \boldsymbol{x}) \left(\phi_{i,w}(u^*, \boldsymbol{x}, z_0) \right)_{w \in \mathbb{U}} \mathrm{d}u^* \ \mathrm{d}\boldsymbol{x}, \end{aligned}$$

where the matrix valued function $Q(u^*, \boldsymbol{x})$ has been defined in (A5). It holds that $\sqrt{m}S^b_{1,n}$ has a normal distribution $\mathrm{N}(0, \Sigma_1)$ with covariance matrix Σ_1 defined in (A5), and that

$$\begin{aligned} &\sqrt{m} \sup \Big\{ |\phi_{i,w}(u^*, \boldsymbol{x}, z)| : w | u^* \in \mathbb{U}, \boldsymbol{x} \in \mathbb{X}, z \in \mathbb{Z}, \\ &\qquad |z - z_0| \leq h \Big\} = O_P(1), \\ &\sqrt{m} \sup \Big\{ |\phi_{i,w}(u^*, \boldsymbol{x}, z_0)| : w | u^* \in \mathbb{U}, \boldsymbol{x} \in \mathbb{X}, z \in \mathbb{Z}, \\ &\qquad |z - z_0| \leq h \Big\} \to 0, \end{aligned}$$

in probability.

Theorem 2 *Make the assumptions (A1)–(A7). It holds that*

$$\hat{\gamma}^b - \hat{\gamma} = -M^{-1} S^b_{1,n} - M^{-1} S^b_{2,n} + o_P(m^{-1/2} + (nh)^{-1/2}),$$

where for $u \in \mathbb{U}$ *and for* $i = 1, ..., n$

$$\begin{aligned} S^b_{2,n,u} &= \frac{1}{n}\sum_{i=1}^{n} e^b_{i,u} K_h(z^b_i - z_0), \\ e^b_{i,u} &= \frac{f_\eta(y^b_i, \hat{\gamma}_u)}{\sum_{v\in\mathbb{U}} f(y^b_i, \hat{\gamma}_v)\hat{p}_{v|u^{*b}_i}(\boldsymbol{x}^b_i, z_i)} \hat{p}_{u|u^{*b}_i}(\boldsymbol{x}_i, z_i). \end{aligned}$$

Furthermore, it holds that, conditionally given the two samples,

$$\Sigma_n^{-1/2}\left(\hat{\gamma}^b - \hat{\gamma}\right)$$

converges weakly to a normal distribution with identity covariance matrix.

The theorem shows that the bootstrap is consistent if $h^2 \ll m^{-1/2} + (nh)^{-1/2}$. This is the case for under smoothing, $h^2 \ll (nh)^{-1/2}$, or if nh/m converges to zero. Otherwise, if h^2, $m^{-1/2}$ and $(nh)^{-1/2}$ are of the same order, the bootstrap does not take care of the bias term and is inconsistent. Nevertheless, it gives a consistent estimate of the variance of the estimator. The proof of Theorem 2 can be found in Section S.I of the supplementary material.

5.1 A.II: Tables

Table 4: Covariate lists for the analysis model $(U^*, \boldsymbol{X}, Z)$ and the validation model $(U^*, \boldsymbol{W})$.

Variable	Sample Average	Validation Model
U^*: Educational Degree (IP1 corrected) (ref: vocational training)		
Missing value	0.01	✓
No degree	0.14	✓
Higher education degree	0.11	✓
Demographics		
Age (Z)	38.70	✓
Work History		
Job changes (=1)	0.60	

Continued on next page

Table 4 – continued from previous page

Variable	Sample Average	Validation Model
Out of labour force periods (=1)	0.40	
Distant job changes (=1)	0.14	
Unemployment periods (=1)	0.38	
Recalls to pre-unemployment employer (=1)	0.10	
Tenure 1-4 months	0.07	
Tenure 5-11 months	0.11	
Tenure 12-23 months	0.14	
Tenure 2-<4 years	0.16	
Tenure 4-<8 years	0.17	
Tenure 8-<15 years	0.17	
Tenure $\geq$ 15 years	0.14	
Additional Experience 6-11 months	0.03	
Additional Experience 12-23 months	0.05	
Additional Experience 2-<4 years	0.12	
Additional Experience 4-<8 years	0.19	
Additional Experience 8-<15 years	0.21	
Additional Experience $\geq$15 years	0.11	
Job Characteristics		
Seasonal job type (=1)	0.15	
White collar (=1)	0.40	✓
Vocational trainee (=1)	0.06	✓
Part-time (=1)	0.16	✓
Low wage (lowest 20% of full-time wages) (=1)	0.36	✓
Immigration Background (ref: German)		
Yes	0.11	✓
Missing value	0.03	✓
Calendar Time (ref: June 2001)		
January	0.08	
February	0.08	
March	0.08	
April	0.08	
May	0.08	
July	0.08	
August	0.08	
September	0.08	
October	0.08	
November	0.08	

Continued on next page

Table 4 – continued from previous page

Variable	Sample Average	Validation Model
December	0.08	
Year 1999	0.24	✓
Year 2000	0.25	✓
Year 2002	0.25	✓
Business Sector (ref: agriculture)		
Commodities	0.06	
Manufacturing (machines)	0.09	
Manufacturing (vehicles)	0.08	
Manufacturing (consumption goods)	0.05	
Food production	0.03	
Construction	0.04	✓
Finishing trade	0.03	✓
Whole sale	0.06	✓
Retail	0.08	✓
Transport and communication	0.05	
Services (business)	0.15	
Services (private)	0.05	
Services (care and health)	0.11	
Services (other public)	0.06	
Public institutions	0.06	
Region Characteristics (ref: suburban, unemp. rate <4%)		
urban	0.56	
rural	0.10	
Unemployment rate 4-<5%	0.06	
Unemployment rate 5-<6%	0.11	
Unemployment rate 6-<7%	0.13	
Unemployment rate 7-<8%	0.16	
Unemployment rate 8-<9%	0.12	
Unemployment rate 9-<10%	0.10	
Unemployment rate 10-<11%	0.11	
Unemployment rate 11-<12%	0.08	
Unemployment rate 12-<13%	0.06	
Unemployment rate 13-<14%	0.04	
Unemployment rate 14-<15%	0.02	
Unemployment rate 15-<16%	0.01	
Unemployment rate 16-<17%	0.01	
Unemployment rate 17-<18%	0.00	
Unemployment rate 18-<19%	0.00	

Continued on next page

Table 4 – continued from previous page

Variable	Sample Average	Validation Model
Unemployment rate 19-20%	0.00	
Observations	20,660,311	22,974