Dlugosz, Stephan; Mammen, Enno; Wilke, Ralf A.

Working Paper

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

ZEW Discussion Papers, No. 15-043

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Dlugosz, Stephan; Mammen, Enno; Wilke, Ralf A. (2015) : Generalised partially linear regression with misclassified data and an application to labour market transitions, ZEW Discussion Papers, No. 15-043, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim, https://nbn-resolving.de/urn:nbn:de:bsz:180-madoc-394880

This Version is available at:
https://hdl.handle.net/10419/112759

Discussion Paper No. 15-043

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

Stephan Dlugosz, Enno Mammen,
and Ralf A. Wilke

ZEW
Zentrum für Europäische Wirtschaftsforschung GmbH
Centre for European Economic Research

Discussion Paper No. 15-043

Generalised Partially Linear Regression With Misclassified Data and an Application to Labour Market Transitions

Stephan Dlugosz, Enno Mammen,
and Ralf A. Wilke

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

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

July 9, 2015

Abstract

We consider the semiparametric generalised linear regression model which has mainstream empirical models such as the (partially) linear mean regression, logistic and multinomial regression as special cases. As an extension to related literature we allow a misclassified covariate to be interacted with a nonparametric function of a continuous covariate. This model is tailor-made to address known data quality issues of administrative labour market data. Using a sample of 20m observations from Germany we estimate the determinants of labour market transitions and illustrate the role of considerable misclassification in the educational status on estimated transition probabilities and marginal effects.

Keywords: semiparametric regression, measurement error, side information

1 Introduction

The increased availability of large scale or big data enables empirical researchers to apply flexible statistical models which operate under mild assumptions. These data are for instance administra-

*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).

†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

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

tive data which are generated by administrative bodies through operations and can comprise the country's entire population of individuals, households or firms. Another example are internet data which are generated by user activity. While there has been tremendous progress in the development of non- and semiparametric models over the past 20 years, a gap has evolved between the frontier of methodological research and what is commonly put to data in empirical research. New research methods often suffer from practical issues such as lack of ready to use implementations and long run time which can cause user frustration. Empirical research is therefore often using much simpler methods such as the standard linear mean regression model or parametric logistic regression. These methods are simple and fast but they do not fully exploit the richness of the available data. Empirical research also often assumes that administrative data are very precise and not subject to misclassification and free of errors. But the absence of errors might only be justified for some set of variables. In particular, data should be error free if they are directly resulting from operations such as administrative activity or internet usage. However, large scale data can also contain considerable errors if the variables are not directly resulting from operations but so called variables of secondary interest. These might be supplementary background information that is entered by administrators without checking for correctness. An example is the information about the educational degree in German administrative employment records which is known to be prone to errors. This is a variable that is reported by firms to the public pension insurance without playing any role for operations. See Fitzenberger et al. (2006) or Kruppe et al. (2014) for details. On the other hand statistical regression models with errors in variables have been developed for smaller survey data. These data are known to be subject to misclassification and measurement error due to response and recall errors. Prominent examples for the application of models with measurement error include Magnac and Visser (1999) and Hernandez and Pudney (2007) among many others. But there is an increasing gap in the literature between flexible statistical models, large scale data and the presence of misclassification in variables. This paper addresses this gap by suggesting a semiparametric generalised linear regression model with a misclassified covariate. The model is purpose built for the data limitations in German administrative labour market data and it makes use of side information to estimate the extent of misclassification. We present a convenient implementation of the model and demonstrate its applicability with a sample of around 20m observations. R-code is available from the first author. Other semiparametric models with misclassified regressor - with and without side information- have to our knowledge not been applied

to large scale data. Examples of such models include Chen et al. (2005) and Chen et al. (2008) which base on the seminar work by Lee and Stepanski (1995). As another contribution we allow the misclassified covariate to be interacted with a nonparametric function of a continuous covariate. 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. Thus, in contrast to common measurement error models, our model does not simply correct regression coefficients but nonparametric age profiles.

We illustrate the practical usefulness and relevance of our model with a comprehensive application. In particular, we apply a semiparametric multinomial labour market transition model to German administrative data that are commonly used for empirical research about the German labour market. This includes academic research but also official evaluation studies of labour market reforms which are conducted on behalf of the German government. Our analysis of the amount and the relevance of data quality problems in these data are therefore of wider academic and non-academic interest. Our model differs from other contributions in economics that combine information from two datasets in order to expand the variable set. Arellano and Meghir (1992) combine information from two micro data sets, while Maddala (1971) combines time series and cross section data. We only use the validation data to incorporate information about the data quality but not to increase the number of covariates in our model. In our analysis we estimate a model which relates individual job separation probabilities to various individual level, firm-level and region-level variables. In particular we consider the probabilities of observing a transition to unemployment, another employer (locally or in another labour market region) and out of the labour force/unknown. The education variable in the analysis data is subject to considerable misclassification and has many missing values. It therefore requires special care. We use another data source as validation data for the educational degree to estimate conditional misclassification probabilities. These are then used in our analysis model with misclassification. Our analysis therefore sheds light on how the estimated effect of covariates changes when the data problems are taken into account. We find evidence for a bias in estimates when misclassification is ignored. There is no clear pattern for the direction of the bias, although it is found to be sizable for some of these variables. Our application provides detailed insights in the determinants of labour market transitions for male employees in Germany. It exceeds previous empirical research in this area by applying a multiple labour market state transition model to large scale administrative data which

are linked with regional data.

The paper is structured as follows. Section 2 contains an informal presentation of our model with the linear regression model as a motivating simple example. Section 3 outlines the general model and Section 4 contains the application to labour market data.

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). The validation data contain U, U^* and $W \subset X$. 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. $P(U = u|U^* = u^*, X) = P(U = u|U^* = u^*, W) = p_{u^*,u}$.

We consider the generalized partial linear model:

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

where f is a known density with unknown nuisance parameters $\boldsymbol{\theta}$. η and $p_{u^*,u}$ are (semi-)parametric functions and $\boldsymbol{\beta}$ is a vector of unknown parameters. 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 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 unknown quantities of the regression model 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 $\eta = \eta(x, u; \boldsymbol{\beta}) = \beta_0 + \beta_x x + \beta_u u$ and $\epsilon \sim N(0, \sigma^2)$. Then since $\boldsymbol{\theta} = \sigma$ we have

$$f_\epsilon(y, \eta, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(y-\eta)^2}{2\sigma^2}\right].$$

Suppose we have two random samples of $(Y, X, U^*)_i$ 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

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

on the grounds of the analysis data with variables Y, X and U^*. The next section considers the generalised partial linear model which includes the probit, logit and multinomial logit model for link functions as special cases.

3 The Model

$Y \in \mathbb{Y} \subset \mathbb{R}$ is a discrete or continuous outcome. $\boldsymbol{X} \in \mathbb{X} \subset \mathbb{R}^k$ is $1 \times k$-dimensional with discrete or continuous covariates and $Z \in \mathbb{Z} \subset \mathbb{R}$ 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 $\mathbb{U} \subset \mathbb{U}^*$. 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 $\boldsymbol{W} \subset \{\boldsymbol{X} \bigcup Z\}$. $U^* \perp\!\!\!\perp Y | \boldsymbol{X}, Z, U$. $\boldsymbol{\beta} = \beta_0, \beta_1, \ldots, \beta_k$ is a $k \times 1$ 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 be $\boldsymbol{\gamma}$ 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 $\boldsymbol{\theta}$.

3.1 Estimation

We assume that analysis data of size n and validation data of size m are two independent 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 $\hat{P}(U = u | U^* = u^*, \boldsymbol{W}) = \hat{p}_{u^*,u}(\boldsymbol{x}, z)$. If not all covariates of the analysis model are contained in the validation model this requires $P(U = u | U^* = u^*, \boldsymbol{X}, Z) = P(U = u | U^* = u^*, \boldsymbol{W})$ that these variables are not informative in the validation model. The fitted values are computed for all observations of the analysis data and plugged into the following second stage smoothed local log likelihood around a value of z

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

where $K_h(\cdot)$ 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 are physically available in one place. The likelihood is then formed of information from the validation and analysis data simultaneously:

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

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 practical reasons we use the two step procedure in the application, although for theoretical reasons the one step procedure should be more efficient.

Algorithm For optimizing (2) 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

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

with respect to $\boldsymbol{\gamma}_{\beta,\theta}(z)$ and

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

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

The resulting Newton-Raphson-like algorithm can be sped up by binning procedures like those in Fan and Marron (1994).

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 $\hat{p}_{u_i^*,u}$ we use for each bootstrap observation $\hat{p}^b_{u_i^*,u}(\boldsymbol{x}_i, z_i) = \hat{p}_{u_i^*,u}(\boldsymbol{x}_i, z_i) + \phi(u_i^*, \boldsymbol{x}_i, z_i)$ where $\phi(u_i^*, \boldsymbol{x}_i, z_i)$ is a random draw from the asymptotic distribution of $\hat{p}_{u_i^*,u}(\boldsymbol{x}_i, z_i) - p_{u_i^*,u}(\boldsymbol{x}_i, z_i)$. Thus we do not bootstrap the first step of the estimation procedure but use information about the asymptotic distribution of the estimated misclassification probabilities.

3.2 Discussion of Properties

This subsection provides a discussion of the identifiability of the nonparametric functions and the validity of the bootstrap procedure.

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 $p_{u_i^*,u}(\boldsymbol{x}_i, z_i)$ are known. 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:

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

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 $p_i^u = p_{u_i^*,u}(\boldsymbol{x}_i, z_i)$, 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:

$$\begin{aligned}
0 \quad \approx \quad & \frac{1}{n}\sum_{i=1}^{n} \frac{f^u_{\eta,i} p^u_i}{\sum_{v\in\mathbb{U}} f^v_i p^v_i} K_h(z_i - z) + \frac{1}{n}\sum_{i=1}^{n} \frac{(\bar{f}^u_{\eta,i} - f^u_{\eta,i}) p^u_i}{\sum_{v\in\mathbb{U}} f^v_i p^v_i} K_h(z_i - z) \\
& - \frac{1}{n}\sum_{i=1}^{n} \frac{f^u_{\eta,i} p^u_i}{(\sum_{v\in\mathbb{U}} f^v_i p^v_i)^2} \sum_{v\in\mathbb{U}} (\bar{f}^v_i - f^v_i) p^v_i K_h(z_i - z) \\
& + \frac{1}{n}\sum_{i=1}^{n} \frac{(\hat{f}^u_{\eta,i} - \bar{f}^u_{\eta,i}) p^u_i}{\sum_{v\in\mathbb{U}} f^v_i p^v_i} K_h(z_i - z) \\
& - \frac{1}{n}\sum_{i=1}^{n} \frac{f^u_{\eta,i} p^u_i}{(\sum_{v\in\mathbb{U}} f^v_i p^v_i)^2} \sum_{v\in\mathbb{U}} (\hat{f}^v_i - \bar{f}^v_i) p^v_i K_h(z_i - z).
\end{aligned}$$

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

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

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 $\hat{\gamma}(z)$.

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 $p_{u_i^*,u}(\boldsymbol{x}_i, z_i)$ 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 $p_{u_i^*,u}(\cdot,\cdot)$ 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 $\hat{p}_i^u = \hat{p}_{u_i^*,u}(\boldsymbol{x}_i, z_i)$ 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:

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

where $\hat{p}_{u^*,\cdot}(\boldsymbol{x}, z)$ and $p_{u^*,\cdot}(\boldsymbol{x}, z))$ denote the vectors with elements $\hat{p}_{u^*,v}(\boldsymbol{x}, z)$ and $p_{u^*,v}(\boldsymbol{x}, z))$ $(v \in \mathbb{U})$, 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 arise 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 bootstrap can be easily understood if one of the two rates $n^{-2/5}$ and $m^{-1/2}$ dominates the other. In that 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 $\sum_{u^* \in \mathbb{U}} \int W(z, u^*, x)(\hat{p}_{u^*,\cdot}(\boldsymbol{x}, z) - p_{u^*,\cdot}(\boldsymbol{x}, z))\ d\boldsymbol{x}$ are asymptotically independent. Thus, we get, that also in this case bootstrap gives a consistent estimate of the variance.

4 Application: Labour Market Transitions

In this section we present an application of the model of Section 3 to show its practicality and relevance for empirical research. In particular we put it to large linked administrative labour market data from Germany to estimate the probability of transitions from employment to other labour market states. A flexible semiparametric statistical model is a natural candidate for the analysis because we use a sample of more than 20m observations. Information in administrative data is known to be often accurate but a high degree of misclassification may also exist in some variables. A well known example is the education variable in German employment records which is prone of misclassification and missing values (compare Fitzenberger et al., 2006, Dlugosz, 2011). Kruppe et al. (2014) use linked information from administrative sources and an interview based survey to analyse the degree of misclassification of the educational degree in the administrative data. They use the ALWA-ADIAB survey because it comprises of validated information about the educational degree. For more details about these data see Antoni and Seth (2011). But due to the limited size of the survey, the linked admin-survey sample is not a natural candidate for

analysis data, we will, however, use it as validation data. In the first step of our application we therefore estimate misclassification probabilities on the grounds of the ALWA-ADIAB (validation data) using a restricted set of covariates $\mathbf{W}$. For the main analysis we use the IAB Employment Sample 04- Regional File (IABS) as analysis data. The IABS is a 2% random sample of employees who make payments into the social security system. It is linked administrative 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). See Drews (2008) for more details about the IABS which covers the period 1975-2004. While we consider labour market transitions in the period 1999-2002, we use the information since the year 1980 to construct a number of employment history variables on individual level. These include 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 individual level data by a number of regional indicators on county level which are provided by the German Federal Statistical Office. In our analysis model we include information about the type of the region (urban, sub-urban and rural) and the monthly unemployment rate. We have also included additional regional variables but these were eventually left out because they did not reveal additional interesting result patterns. Table 5 in the Appendix contains the covariate lists of our analysis and our validation model along with some basic descriptive statistics.

Our main analysis 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 a 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 anaylsis of labour market transitions of employees. Early analysis for Germany has used household survey data (Bergemann and Mertens, 2002, Gangl, 2003). Analysis based on large linked administrative data has been conducted mainly as employment duration analysis within a competing risks framework. Bookmann and Steffes (2005) consider transitions into unemployment, nonemployment and into new jobs but do not distinguish between local and distant new job. Dütsch and Struck (2011) mainly focus on within firm trajectories and pool all job separations into one risk. Wichert and Wilke (2012) and Westerheide and Kauermann (2014) use a similar sample as in this paper but only model transitions from employment to unemployment. All these papers do not present a satisfactory solution for dealing with misclassification in covariates.

The aim of our analysis is to obtain a comprehensive understanding of the determinants of labour market transitions on individual level by estimating the monthly probability of being in one of the above labour market states. We therefore avoid strong identifying interdependence assumptions of classical competing risks duration models and all our covariates are allowed to vary over time. In contrast to the previous studies based on German administrative data we put a misclassification model to data to address the bias of results incurred by the measurement error in the education variable. Following Wichert and Wilke (2012) we only consider three distinct grouped values of the education variable as the misclassification in the raw education information appears to be to a larger extent due to having very similar values. In particular, $U \in \{$higher education [HE], vocational training [VT], no degree [ND]$\}$. The education variable in the analysis data (U^*) can also take on missing values [NA]. There are no missing values about the educational degree in our validation sample because we have dropped the affected observations (about 1%). Given the small number of cases we do not expected that this affects our results. In the first stage of the analysis we compare the education information in the employment records (BeH, Beschäftigtenhistorie) in the administrative data with the information in the ALWA-ADIAB survey data for the validation sample. We do so by using U^* directly constructed from the education variable and a corrected version of U^*. The latter is obtained by applying the IP1 imputation of Fitzenberger et al. (2006) which overwrites missing values and apparent inconsistent information

using the individual employment history. This correction is commonly applied in academic research which uses these data but it is not clear how much of the misclassification is eliminated by this imputation. Tables 1 and 2 report misclassification probabilities of the education information in the analysis data. It confirms that there is substantial misclassification in the grouped education variable of the administrative employment records. The observed education information in the anaylsis data is incorrect in around every other observation if the true level of education is "no degree" or "higher education" (compare Table 1). Table 2 confirms that the IP1 correction reduces misclassification for the higher two categories but fails to do so for the lowest. It is apparent that no degree and higher educational degrees are often reported as vocational training in the employment records, which wipes out a considerable amount of variation in this variable. Thus, estimated effects of education in labour market studies based on these data are likely under estimating the true effect. Although, still containing considerable misclassification, the IP1 corrected variable is better than the uncorrected version and for this reason we only report results for the former in what follows.

In order to obtain estimates for $P(U|U^*, \boldsymbol{W})$ we estimate an Ordered Probit Model as validation model as values of U are ordered. The covariate list for this model can be found in Table 5. The number of observations in our validation sample is 22,974. 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 we can assume independence between them. The estimation results and computed estimated marginal effects for this model are given in Table 6 in the Appendix. 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 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 data errors. It is therefore likely that ignoring these errors will lead to bias in empirical results. On the grounds of the parameter estimates for the validation model we compute conditional probabilities $\hat{P}(U_i|U_i^*, \boldsymbol{W}_i)$ for all observations in the analysis data which are then plugged into our misclassification regression model.

We use a partially linear Multinomial Logit Model (PLM) for our analysis of the probability of

Table 1: Misclassification matrix for the education variable (uncorrected) in the administrative employment records.

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 the education variable (IP1) in the administrative employment records.

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 $\hat{P}(U_i|U_i^*, \boldsymbol{W}_i)$; this is average estimated probability of the correct value of education given the IP1 corrected grouped education in the administrative employment records and a number of control variables ($\boldsymbol{W}$).

	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

transiting into one of the labour market states:

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

for $j = 1, \ldots, 4$ and $\gamma_{uj}(z)$ is a nonparametric age (z) profile which differs across educational degrees (u) and labour market state j. This model is used for the density in the log-likelihood (2) which is then maximised in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Since the estimated coefficients are only limited informative due to the nonlinearities, we construct marginal effects. This is the partial derivative of the response probability in response to a covariate change in case of a continuous regressor. In case of a binary covariate we take the difference in response probabilities evaluated for the two values of the covariate. In all cases the other covariates are set to their sample averages. In what follows we briefly discuss the main result patterns and compare whether the estimates change when the misclassification of the education variable is being taken into account by the model (misPLM).

Estimated marginal effects for covariates $\mathbf{x}$ are reported in Table 4. The table also contains baseline transition probabilities ($\hat{P}(j|\bar{u}, \bar{x}, \bar{z})$, i.e. at the sample means of covariates) to provide a reference for the estimated marginal effects. Both baseline probabilities and marginal effects appear to be small in terms of size. This is because we consider monthly transition probabilities out of existing jobs. These are rather small given that the vast majority of individuals simply continues in the current job. Looking deeper it becomes clear that a number of marginal effects are quite sizable relative to the baseline probability. For instance, having a low paid job increases the probability of a local job change by around 50% (0.3/0.64). Estimated marginal effects are often larger when misclassification has been taken into account but not always (e.g. the marginal effect of past job changes on local employer changes increases) and in some cases there are even changes in the direction of the effect (e.g. the negative effect of past unemployment periods on local employer changes becomes positive in misPLM). We also observe very different roles of the covariates for different labour market states. While having had past recalls to the former employer only increases the probability of entering unemployment, past distant job changes decrease the probability of a future local job change and unemployment but increase the probability of future distant job changes. The latter increases by almost one half when the misclassification in the education variable has been taken into account. Thus, we find evidence for sizable bias of estimated marginal effects for variables which are not subject to misclassification.

The results in Table 4 also reveal a number of interesting result patterns related to the subject

Table 4: Marginal effects for changes in X (analysis model, in %-points).

	local employer change				distant employer change				unemployment			
	PLM		misPLM		PLM		misPLM		PLM		misPLM	
Variable	ME	SE	ME	SE	ME	SE	ME	SE	ME	SE	ME	SE
Work History												
Job changes	0.17	0.01	0.26	0.01	0.01	0.00	0.01	0.01	-0.01	0.00	0.01	0.00
Out of labour force periods	0.04	0.00	0.04	0.00	0.00	0.00	0.03	0.00	0.02	0.00	0.02	0.00
Distant job changes	-0.06	0.01	-0.08	0.00	0.22	0.01	0.30	0.02	0.00	0.00	-0.01	0.00
Unemployment periods	-0.05	0.00	0.03	0.01	-0.02	0.00	-0.03	0.00	0.31	0.01	0.26	0.01
Recalls to p. employer (=1)	-0.07	0.01	-0.09	0.01	-0.01	0.00	-0.02	0.00	0.13	0.00	0.13	0.01
Tenure 1-4 months	-0.23	0.01	-0.24	0.01	-0.09	0.00	-0.12	0.01	-0.03	0.00	-0.01	0.00
Tenure 5-11 months	-0.34	0.01	-0.35	0.01	-0.10	0.00	-0.13	0.01	-0.01	0.00	0.01	0.00
Tenure 12-23 months	-0.42	0.00	-0.44	0.01	-0.12	0.00	-0.15	0.01	-0.15	0.00	-0.11	0.00
Tenure 2-<4 years	-0.49	0.01	-0.51	0.01	-0.13	0.00	-0.16	0.01	-0.21	0.00	-0.15	0.01
Tenure 4-<8 years	-0.59	0.01	-0.61	0.01	-0.14	0.00	-0.18	0.01	-0.28	0.00	-0.23	0.01
Tenure 8-<15 years	-0.66	0.00	-0.68	0.01	-0.15	0.00	-0.20	0.01	-0.35	0.00	-0.29	0.01
Tenure ≥15 years	-0.68	0.01	-0.70	0.01	-0.17	0.00	-0.22	0.01	-0.35	0.00	-0.28	0.01
Add. Experience 6-11 months	-0.09	0.01	-0.14	0.01	-0.04	0.00	-0.05	0.01	-0.07	0.00	-0.07	0.00
Add. Experience 12-23 months	-0.09	0.01	-0.16	0.01	-0.04	0.00	-0.05	0.01	-0.09	0.00	-0.08	0.00
Add. Experience 2-<4 years	-0.10	0.01	-0.17	0.01	-0.07	0.00	-0.09	0.01	-0.12	0.00	-0.11	0.01
Add. Experience 4-<8 years	-0.15	0.01	-0.25	0.01	-0.08	0.00	-0.10	0.01	-0.13	0.00	-0.12	0.01
Add. Experience 8-<15 years	-0.12	0.01	-0.23	0.01	-0.10	0.00	-0.12	0.01	-0.15	0.01	-0.14	0.01
Add. Experience ≥15 years	-0.16	0.01	-0.25	0.01	-0.10	0.00	-0.12	0.01	-0.17	0.00	-0.16	0.01
Job characteristics												
Seasonal job type	-0.23	0.01	-0.19	0.01	-0.02	0.00	-0.05	0.00	0.07	0.00	0.06	0.00
White collar	-0.21	0.00	-0.19	0.01	0.00	0.00	-0.05	0.00	-0.09	0.00	-0.05	0.00
Vocational trainee	-0.43	0.00	-0.42	0.01	-0.03	0.00	-0.04	0.01	-0.17	0.00	-0.14	0.01
Part-time	-0.23	0.01	-0.22	0.01	-0.01	0.00	-0.03	0.00	-0.14	0.00	-0.11	0.00
Low wage	0.33	0.01	0.30	0.01	-0.03	0.00	-0.05	0.01	0.20	0.00	0.16	0.01
Immigration status												
Immigrantion background	-0.08	0.01	-0.09	0.01	-0.03	0.00	-0.05	0.00	0.02	0.00	-0.02	0.00
Immigration status unknown	-0.14	0.00	-0.15	0.01	0.09	0.00	0.11	0.01	0.00	0.00	0.00	0.00
Calender time												
January	-0.33	0.01	-0.32	0.01	-0.06	0.00	-0.06	0.01	0.11	0.01	0.03	0.01
February	-0.32	0.01	-0.32	0.01	-0.06	0.00	-0.07	0.01	-0.06	0.00	-0.09	0.01
March	-0.12	0.01	-0.11	0.01	0.00	0.00	0.02	0.01	-0.02	0.01	-0.09	0.01
April	-0.30	0.01	-0.29	0.01	-0.05	0.00	-0.06	0.01	-0.12	0.00	-0.13	0.01
May	-0.30	0.01	-0.29	0.01	-0.05	0.00	-0.06	0.01	-0.12	0.00	-0.13	0.01
July	-0.18	0.01	-0.16	0.01	-0.05	0.00	-0.05	0.01	-0.08	0.00	-0.10	0.01
August	-0.16	0.01	-0.14	0.01	-0.04	0.00	-0.05	0.01	-0.10	0.00	-0.11	0.01
September	-0.13	0.01	-0.11	0.01	0.00	0.01	0.02	0.01	-0.08	0.00	-0.10	0.01
October	-0.30	0.01	-0.29	0.01	-0.05	0.00	-0.06	0.01	-0.07	0.00	-0.09	0.01
November	-0.39	0.01	-0.38	0.01	-0.06	0.00	-0.07	0.01	-0.01	0.01	-0.06	0.00
December	0.73	0.02	0.85	0.03	0.19	0.01	0.28	0.02	0.27	0.01	0.13	0.01
Year 1999	-0.10	0.01	-0.10	0.01	-0.06	0.00	-0.07	0.01	-0.10	0.00	-0.09	0.00
Year 2000	-0.02	0.01	-0.03	0.01	-0.04	0.00	-0.05	0.00	-0.09	0.00	-0.08	0.00
Year 2002	-0.17	0.01	-0.18	0.01	-0.05	0.00	-0.07	0.01	0.00	0.00	0.00	0.00
Business Sector												
Commodities	-0.47	0.01	-0.41	0.01	0.30	0.05	0.01	0.02	-0.05	0.01	0.13	0.02
Manufacturing (machines)	-0.49	0.01	-0.42	0.01	0.28	0.04	0.02	0.02	-0.09	0.01	0.07	0.01
Manufacturing (vehicles)	-0.45	0.01	-0.38	0.01	0.31	0.04	0.03	0.02	-0.09	0.01	0.06	0.01
Manufacturing (consumption)	-0.39	0.01	-0.32	0.01	0.32	0.05	0.02	0.02	0.00	0.01	0.21	0.02
Food production	-0.41	0.01	-0.34	0.01	0.67	0.10	0.13	0.04	-0.03	0.01	0.17	0.02
Construction	-0.34	0.01	-0.27	0.01	0.62	0.09	0.16	0.03	0.10	0.01	0.38	0.03
Finishing trade	-0.32	0.01	-0.25	0.01	0.67	0.09	0.17	0.04	0.05	0.01	0.27	0.02
Whole sale	-0.40	0.01	-0.33	0.01	0.56	0.08	0.13	0.03	-0.01	0.01	0.18	0.02
Retail	-0.39	0.01	-0.31	0.01	0.56	0.08	0.12	0.03	-0.04	0.01	0.14	0.02
Transport and Communication	-0.40	0.01	-0.26	0.01	0.51	0.07	0.08	0.03	-0.08	0.01	0.09	0.01
Services (business)	-0.35	0.02	-0.25	0.02	0.50	0.06	0.09	0.03	-0.07	0.01	0.10	0.01
Services (private)	-0.42	0.01	-0.36	0.01	0.49	0.07	0.07	0.03	-0.08	0.01	0.10	0.02
Services (care and health)	-0.39	0.01	-0.32	0.01	0.49	0.07	0.10	0.03	-0.01	0.01	0.17	0.02
Services (other public)	-0.40	0.01	-0.34	0.01	0.56	0.08	0.11	0.03	-0.04	0.01	0.13	0.02
Public institutions	-0.48	0.01	-0.43	0.01	0.72	0.10	0.19	0.04	-0.07	0.01	0.24	0.03
Region characteristics												
urban	0.04	0.00	0.04	0.01	-0.01	0.00	0.00	0.00	-0.06	0.00	-0.05	0.00
rural	-0.12	0.01	-0.11	0.01	-0.01	0.00	-0.01	0.00	0.03	0.00	0.03	0.00
Unemployment rate 4-<5%	0.45	0.06	0.17	0.04	-0.07	0.01	0.03	0.02	-0.17	0.01	-0.18	0.01
Unemployment rate 5-<6%	0.46	0.06	0.19	0.04	-0.09	0.01	-0.02	0.01	-0.19	0.01	-0.20	0.01
Unemployment rate 6-<7%	0.46	0.06	0.18	0.04	-0.09	0.01	-0.01	0.01	-0.19	0.01	-0.20	0.01
Unemployment rate 7-<8%	0.29	0.05	0.06	0.04	-0.10	0.01	-0.01	0.01	-0.19	0.01	-0.21	0.01
Unemployment rate 8-<9%	0.33	0.05	0.08	0.04	-0.09	0.01	-0.01	0.01	-0.16	0.01	-0.18	0.01
Unemployment rate 9-<10%	0.42	0.06	0.15	0.04	-0.09	0.01	-0.01	0.01	-0.16	0.01	-0.18	0.01
Unemployment rate 10-<11%	0.40	0.06	0.13	0.04	-0.09	0.01	-0.01	0.01	-0.15	0.01	-0.17	0.01
Unemployment rate 11-<12%	0.41	0.06	0.13	0.04	-0.07	0.01	0.05	0.02	-0.14	0.01	-0.17	0.01
Unemployment rate 12-<13%	0.37	0.06	0.11	0.04	-0.07	0.01	0.06	0.02	-0.14	0.01	-0.16	0.01
Unemployment rate 13-<14%	0.42	0.06	0.14	0.05	-0.06	0.01	0.08	0.03	-0.12	0.01	-0.15	0.01
Unemployment rate 14-<15%	0.37	0.06	0.10	0.05	-0.05	0.01	0.10	0.03	-0.12	0.01	-0.14	0.01
Unemployment rate 15-<16%	0.34	0.05	0.08	0.06	-0.06	0.01	0.08	0.03	-0.12	0.01	-0.14	0.01
Unemployment rate 16-<17%	1.03	0.09	0.63	0.11	0.03	0.02	0.40	0.08	0.04	0.02	-0.06	0.01
Unemployment rate 17-<18%	1.13	0.10	0.65	0.20	0.26	0.05	1.40	0.25	-0.06	0.01	-0.12	0.02
Unemployment rate 18-<19%	1.35	0.11	0.86	0.20	0.03	0.02	0.45	0.12	0.00	0.02	-0.08	0.02
Unemployment rate 19-<20%	8.07	0.49	5.68	0.77	0.53	0.09	2.70	0.68	0.85	0.07	0.34	0.10
base probability	0.64		0.63		0.11		0.10		0.27		0.27	

content. For local employer changes we find that past job mobility and having a low paid job strongly enhances the prospects of future local job changes. Long tenure, a lot of additional labour market experience, part timers, vocational trainees and seasonal job types are estimated to have a considerably lower probability of making a local employer change. The most likely month for a local job change to take place is December and men working in agriculture are estimated to have the highest probability for locally changing employer. Local job changes take place least likely in rural areas and more likely the higher the regional unemployment rate is.

We find distant job changes to take place more likely in December and when the individual had already past distant job changes, is not working in agriculture, or is located in a region with higher unemployment rate. Men with long tenure and a lot of additional labour market experience are estimated to have considerably lower probabilities of making a distant job change. The marginal effects for distant job change are often very large compared to the baseline probability, which is only 0.1%. For example the existence of past distant employer changes is related with a 200% higher probability of observing a future distant job change. The marginal effects for the business sectors decrease strongly (from about 0.5-0.6 to around 0.1-0.2) when misclassification of the education variable has been taken into account.

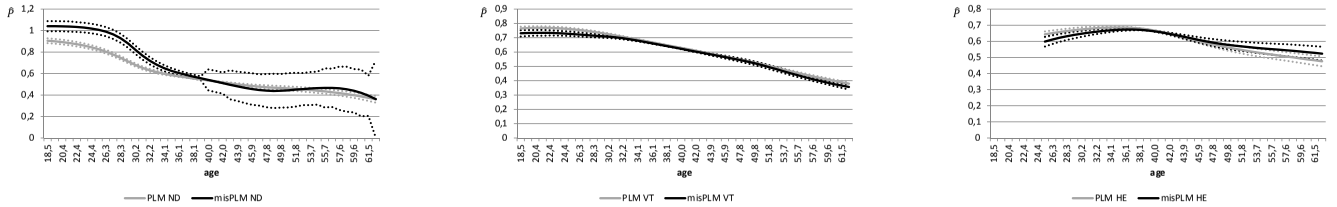
The probability of entering unemployment is considerably higher when the individual had been unemployed in the past or had been previously recalled to the same employer. Transitions take mainly place at calender year change and are much less likely for men with long tenure or a lot of additional labour market experience. Entries into unemployment are more likely in rural areas and less likely in urban areas and they are generally less likely in regions with lower unemployment rate. The patterns for the marginal effects for entering unemployment are often similar to what has been estimated by Wichert and Wilke (2012) and Westerheide and Kauermann (2014). When comparing the results for the PLM with misPLM we find sizable differences for the end of year effect (December and January) when misclassification in education is taken into consideration. The direction of the marginal effects for most business sectors changes, which suggests that agriculture is not among the business sectors with the highest lay-off probability but the one with the lowest - notice, after having controlled for seasonality patterns in the job type and employment history.

Figure 1 shows estimated transition probabilities $\hat{P}(j|u,x,z)$ as functions in age by educational degree with their 90% bootstrap confidence intervals. When comparing PLM with misPLM estimates it becomes apparent that the general shape of these age profiles are often similar. Ac-

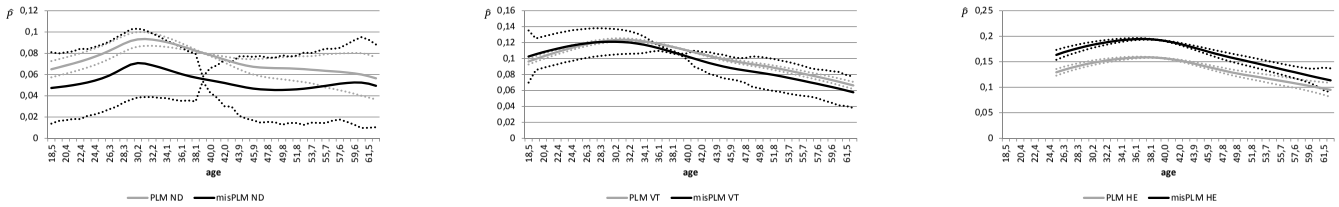
Figure 1: Estimated transition probabilities in age by education (analysis model).

No Degree | Vocational Training | Higher Education

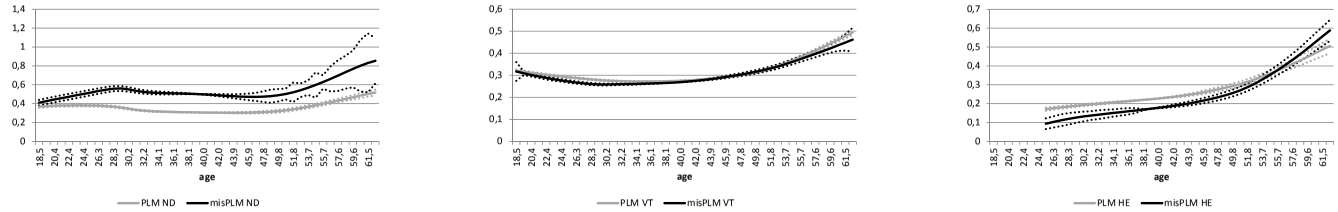
(a) Local employer change



(b) Distant employer change



(c) Transition to unemployment



counting for misclassification therefore does not completely alter estimates. However, there are some changes and in some cases the estimates are statistically different. In some cases the mis-PLM model produces higher estimated probabilities than the PLM and in some cases they are lower. The two lines sometimes cross, which means there is no clear pattern in the direction of the bias. We make the following observations with regard to the subject content.

The estimated probability for a local employer change generally non-increases in age except for men aged less than 35 with higher education degree. The overall decrease is also less pronounced for the latter group, in particular at higher ages. Men without educational degree have the highest transition probabilities in younger ages (less than 30) but the lowest for higher ages (aged > 40). Men with higher educational degree have the lowest probabilities for younger ages (< 30) but the highest for higher ages (> 50).

Distant employer changes are most likely for men with higher education degree and lowest for those without any degree. The estimated probability functions increase for those with completed vocational training and higher education degree for younger ages (< 30 and < 37, respectively) and thereafter they fall. For those without completed degree there is no systematic fall.

The probability of entering unemployment in contrast increases in age for all education groups, where the increase is most pronounced for ages > 50. These pattern are related to early retirement schemes which used unemployment benefits as a bridge between employment and some other form of compensation. The probability of entering unemployment is only decreasing in age for men with completed vocational training at younger ages (< 32). Given that we control for tenure and additional labour market experience in our model, these results suggest a strong age discriminating pattern. But it would be misleading to speak only about compulsory redundancies as many of the terminations of employment contracts have been agreed by the older employees after negotiating a comprehensive early retirement package.

To summarise, our application has revealed a number of interesting results on the determinants of labour market transitions of male employees. Despite the large amount of misclassification in the education variable, the main result patterns do not change when applying our misclassification model. Somewhat surprisingly we observe that estimated covariate effects of the variables without misclassification are more affected by the misclassification than the interacted age-education profiles.

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] Arellano, M. and Meghir, C. (1992): Female Labour Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets. The Review of Economic Studies. 59, 537–557.

[3] 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.

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

[6] 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

[7] 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.

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

[9] 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.

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

[11] 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.

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

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

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

[15] 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.

[16] Maddala, G.S. (1971): The Likelihood Approach to Pooling Cross Section and Time Series Data. Econometrica, 39, 939–953.

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

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

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

[20] 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.

[21] 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.

Appendix

Table 5: 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 (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	
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	✓

Continued on next page

Table 5 – continued from previous page

Variable	Sample Average	Validation Model
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	
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	

Continued on next page

Table 5 – continued from previous page

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

Table 6: Estimation results of an Ordered Probit Model for the probability of the true value of education (U) given the IP1 corrected value U^* in the administrative employment records (BeH) and a number of control variables $\boldsymbol{W}$.

	Model: $P(U\|U^*, \boldsymbol{W})$				
	coef.		marginal effect†		
$U^*, \boldsymbol{W}$	(standard error)		U =ND	U =VT	U =HE
U^*, ref: ND					
NA	1.684***	(0.566)	−0.025	−0.525	0.550
VT	1.292***	(0.128)	−0.143	−0.017	0.160
HE	3.367***	(0.161)	−0.088	−0.811	0.899
Immigration Background ref: German					
Yes	−0.146	(0.120)	0.01	0.013	−0.023
Missing value	0.278	(0.209)	−0.013	−0.043	0.056
individual background					
age	0.016***	(0.005)	−0.001	−0.002	0.003
part time	0.387***	(0.104)	−0.019	−0.056	0.075
daily salary (in €)	0.064***	(0.011)	−0.004	−0.007	0.011
white collar	0.562***	(0.061)	−0.032	−0.068	0.100
vocational training	−0.315***	(0.114)	0.024	0.021	−0.045
business sector, ref: others					
construction	−0.055	(0.123)	0.003	0.006	−0.009
trade	−0.060	(0.082)	0.004	0.006	−0.010
calendar time, ref: year 2001					
year 1999	−0.023	(0.061)	0.002	0.002	−0.004
year 2000	0.040*	(0.024)	−0.002	−0.005	0.007
year 2002	0.001	(0.024)	−0.0001	−0.0001	0.0002
Log. Pseudo-Likelihood	-10,372.759				
Pseudo R^2	0.4353				
Number of observations	22,974				

Note: †: Evaluated at the sample mean of the other covariates.

*p<0.1; **p<0.05; ***p<0.01

2 coefficients for cut points not reported.