Computational Statistics and Data Analysis 110 (2017) 145–159

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

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

Stephan Dlugosz [a], Enno Mammen [b], Ralf A. Wilke [c,d,a,*]

[a] ZEW Mannheim, L7.1, 68161 Mannheim, Germany

[b] Institute for Applied Mathematics, Heidelberg University, Im Neuenheimer Feld 294, 69120 Heidelberg, Germany

[c] Copenhagen Business School, Department of Economics, Porcelaenshaven 16A, 2000 Frederiksberg, Denmark

[d] University of Strasbourg, France

HIGHLIGHTS

- A semiparametric generalized partially linear model with a misclassified covariate is proposed.
- Validation data is used to address misclassification and missing values.
- The model is applied to estimate the determinants of labour market transitions in Germany.
- There is no clear bias pattern for estimated partial effects.

ARTICLE INFO

Article history:
Received 1 December 2015
Received in revised form 13 January 2017
Accepted 18 January 2017
Available online 27 January 2017

Keywords:
Semiparametric regression
Measurement error
Side information

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 20 m 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 generalized partially linear regression extends existing models by allowing a misclassified discrete covariate to be interacted with a nonparametric function of a continuous covariate.

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

* Corresponding author at: Copenhagen Business School, Department of Economics, Porcelaenshaven 16A, 2000 Frederiksberg, Denmark.
E-mail addresses: stephan.dlugosz@googlemail.com (S. Dlugosz), mammen@math.uni-heidelberg.de (E. Mammen), rw.eco@cbs.dk (R.A. Wilke).

http://dx.doi.org/10.1016/j.csda.2017.01.003
0167-9473/

economics and social sciences. In particular, many analyses 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 (US, 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 (multiple) imputation methods or by maximum likelihood. See Little and Rubin (2002) for a comprehensive overview of imputation 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 Carroll et al. (2006) for a comprehensive overview. Examples of more recent works include Chen et al. (2005, 2008) and Yi et al. (2015) which have in common that they use the method of maximum likelihood estimation and based on the seminal work by Lee and Sepanski (1995). Messer and Natarayan (2008) and Valaste et al. (unpublished manuscript) 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. Carroll 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 generalized linear regression model can be estimated with a sample of 20 m 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 summarizes 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