Journal of Econometrics 193 (2016) 1–16

Contents lists available at ScienceDirect

Journal of Econometrics

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

Kernel estimation of hazard functions when observations have dependent and common covariates

James Lewis Wolter *

Department of Economics, University of Oxford, United Kingdom
Oxford-Man Institute, United Kingdom

ARTICLE INFO

Article history:
Received 21 September 2014
Received in revised form
25 September 2015
Accepted 23 January 2016

JEL classification:
C13
C14
C51

Keywords:
Hazard estimation
Correlated events
Dependent covariates
Common covariates
Kernel estimation

ABSTRACT

We propose a hazard model where dependence between events is achieved by assuming dependence between covariates. This model allows for correlated variables specific to observations as well as macro variables which all observations share. This setup better fits many economic and financial applications where events are not independent. Nonparametric estimation of the hazard function is then studied. Kernel estimators proposed in Nielsen and Linton (1995) and Linton et al. (2003) are shown to have similar asymptotic properties compared with the *i.i.d.* case. Mixing conditions ensure the asymptotic results follow. These results depend on adjustments to bandwidth conditions. Simulations are conducted which verify the impact of dependence on estimators. Bandwidth selection accounting for dependence is shown to improve performance. In an empirical application, trade intensity in high-frequency financial data is estimated.

1. Introduction

Almost all research on hazard models has focused on parametric and semiparametric cases assuming *i.i.d.* observations. In econometrics, it was recognized early on that unobserved heterogeneity was an important aspect of these specifications. This lead to the mixed proportional hazard model. Lancaster (1990) and Van den Berg (2001) summarize much of the early work in this area. Hausman and Woutersen (2014a,b) are recent contributions giving up-to-date summaries of the literature.[1]

While not given as much attention, the nonparametric case has been studied in a parallel literature. Nonparametric estimation of hazard models with observed covariates is well developed in the *i.i.d.* case. In this research, kernel methods have been a frequently used approach. Much of this line of inquiry culminated in the papers of Nielsen and Linton (1995) and Linton et al. (2003) (hereafter NL and LNV). NL derived asymptotics for fully nonparametric estimators allowing exposure-time to be a covariate. In LNV, the hazard function is restricted to be additive or multiplicative, but otherwise allowed to be nonparametric. Estimators are proposed and a number of asymptotic results are derived. In particular they show that, by restricting the form of the hazard function, the rate of convergence is greatly improved. For a comprehensive overview of previous work in this area see the aforementioned papers.

In a number of economic and financial hazard situations, the assumption that the observed covariates are independent across observations is questionable or clearly false. For example, when considering mortgage default, variables such as interest rates or housing values are correlated across mortgage holders. For corporate default applications, accounting variables are correlated across members of the same industry. In addition, macroeconomic variables such as US Treasury rates or GDP growth rates can be relevant. These covariates have the same realization across observations. Their only variability is through time. If relevant covariates include macro variables, observations cannot satisfy the *i.i.d.* assumption required in the previous literature.

The most common way previous studies have allowed for dependence across observations in hazard models is with a clustering approach. Events are allowed to be correlated within groups of

* Correspondence to: Eagle House, Eagle Works, Walton Well Road, OX2 6ED Oxford, United Kingdom.

E-mail address: james.wolter@economics.ox.ac.uk.

[1] Papers include Elbers and Ridder (1982), Heckman and Singer (1984), Han (1987), Han and Hausman (1990), Honoré (1990, 1993a,b), Meyer (1990), Ridder (1990), Hahn (1994), Ishwaran (1996), Horowitz (1999), Woutersen (2000, 2002), Bijwaard and Ridder (2002, 2009), Chen (2002), Ridder and Woutersen (2003), Horowitz and Lee (2004), Frederiksen et al. (2007) and Honoré and Hu (2010).

http://dx.doi.org/10.1016/j.jeconom.2016.01.002
0304-4076/

observations, but the groups are assumed to be independent. See Martinussen and Scheike (2010) for an overview. Macro variables eliminate the possibility of a clustering approach. This is because they make it impossible to separate observations into independent groups. Relatively few hazard papers have considered situations with more extensive correlation than the clustering case. The only examples the author is aware of involve spatial correlation. See Henderson et al. (2002), Li and Ryan (2002), Banerjee et al. (2003), Banerjee and Dey (2005), Li and Lin (2005), Hennerfeind et al. (2006), Bastos and Gamerman (2006), Zhao et al. (2009), Zhao and Hanson (2011) and Lawson et al. (2014) among others. Most of these citations introduce dependence through spatially correlated unobserved heterogeneity.[2] Observed variables are assumed to be independent across observations. Li and Lin (2005) is an exception. All of these papers estimate parametric or semiparametric models.

The first contribution of this paper is to propose a specific construction of correlated random events. The proposed construction makes clear how dependence in events is related to observed covariates. This connection allows application of methods used in the *i.i.d.* case to the dependent case. In the proposed construction, dependence is assumed directly on the covariates. This in turn determines dependence between random events. The model is not spatial. Indeed, one of the main goals of this research is to incorporate macro covariates as described above. Dependence through time is an important aspect of our model not captured in spatial situations.

Compared with the *i.i.d.* case, this setup better fits many economic situations. The model is particularly well suited for analysis of credit events when observations overlap with financial crises. During these episodes, many important macroeconomic and financial variables persistently take on values rarely seen in tranquil periods. This implies these covariates are dependent through time as assumed in the proposed model. These variables are frequently used as measures of macroeconomic conditions by decision making agents. This requires they are accounted for in hazard analysis. Covariates with these properties include yield curve variables, VIX, the TED spread, Moody's corporate default spread and aggregate consumption, investment, income or GDP growth rates. In addition, empirical analysis often uses variables which are cross-sectionally correlated. This is another aspect of our model. Relevant credit situations include mortgage default, corporate default, bank lending and corporate exercise of credit lines.

Using the proposed model, asymptotic results for estimators considered in NL and LNV are derived. Dependence between random events is controlled using mixing and martingale conditions. This control is then utilized to derive asymptotics. Most of the results from NL and LNV are extended to the dependence case. Compared with the *i.i.d.* situation, the rate at which the bandwidth converges to zero must be slowed down for similar results to hold. This adjustment changes confidence intervals, uniform convergence rates and asymptotic variance estimates.[3]

With the proposed setup, many previous papers can be extended to the dependence case. In this paper we trace the consequences of dependence for kernel approaches. A similar exercise could be undertaken with much of the hazard literature. This work is also related to a large body of research on panel data with dependence. This dependence can manifest itself cross-sectionally, through time or spatially. The literature is too large to comprehensively overview here. We only mention a few areas related to this paper. The first area is panel data models with common factors. Recent citations include Pesaran (2006), Kapetanios and Pesaran (2007), Chudik et al. (2011), Pesaran and Tosetti (2011), Su and Jin (2012) and Jin and Su (2013). See Chudik and Pesaran (2013) for a survey. A second area is panel data models with spatial dependence. Recent work includes Kapoor et al. (2007), Yu et al. (2008), Lee and Yu (2010a,b, 2014), Su and Jin (2010), Parent and LeSage (2012), Su (2012), Baltagi et al. (2013, 2014) and Su and Yang (2015). A third area is panel data models with common shocks as in Kuersteiner and Prucha (2013) who build on the results of Andrews (2005). A final area is dependent discrete choice. See Robinson (1982), Poirier and Ruud (1988), De Jong and Woutersen (2011) and Hahn and Kuersteiner (2011). Many of these papers assume more complicated forms of dependence than the present work. Ideas from this literature can likely be used to extend our hazard results in various directions. We leave these extensions to future research.

Another application of our results is estimation of trade intensity in high-frequency financial data. Several papers have shown that an asset's quotes can partially explain trading rates (see, for example, Hall and Hautsch (2007)). It is also possible information from related assets influences this rate. A commonly used variable is depth imbalance. In an empirical application of our results, the hazard rate of trade arrival for silver futures is estimated. The depth imbalance of both gold and silver futures are used as covariates. We show both these covariates have substantial autocorrelation. Therefore, this application fits into the framework described in this paper. It is not obvious a priori which shape the hazard function will take. This makes kernel estimation an attractive approach.

The remainder of the paper is organized as follows. In Section 2, we propose the construction of random events that is used throughout the paper. In Section 3, nonparametric asymptotic results related to NL are derived assuming covariates are dependent. Section 3.1 presents results related to LNV where the functional form of the hazard is restricted to be additive or multiplicative. Section 4 conducts a number of simulations which examine the effect of dependence on finite sample properties of the relevant estimators. Section 5 presents our application to trade intensity. Section 6 concludes. All proofs are presented in the Online Supplement (see Appendix B).

2. Model of random events

This section describes the random events used throughout the sequel. Special attention is given to incorporating macro covariates across observations. First, the covariate processes assigned to observations are presented. Then these processes are used to construct random times. The hazard function arises naturally in the setup. The construction is presented without censoring. General forms of censoring can be added with few complications.[4] The notation of NL and LNV is used where possible.

2.1. Sampling

Observations are indexed by $i \in \mathbb{N}$. Each observation i has a deterministic constant $G^i \in [0, \infty)$ which is the calendar-time it starts to be at risk of default. These represent, for example, the date a mortgage is signed. Each observation is at risk over a fixed time

[2] Unobserved heterogeneity is referred to as frailty in many of these papers.

[3] This paper considers the local constant estimation case. There is a related literature for local linear estimation. See Nielsen (1998), Nielsen and Tanggaard (2001), Mammen et al. (2011) and Gámiz et al. (2013). Similarly to NL and LNV, the local linear approach has not been modified to account for dependence. This line of research would be an extension of the results that are presented below.

[4] See the Online Supplement for more details (see Appendix B).