Accident Analysis and Prevention 43 (2011) 1864–1871

Contents lists available at ScienceDirect

Accident Analysis and Prevention

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

On factors related to car accidents on German Autobahn connectors

Martin Garnowski[a], Hans Manner[b,*]

[a] FedEx Express Europe, Middle East, Indian Subcontinent and Africa, Cologne, Germany
[b] Department of Economic and Social Statistics, University of Cologne, Meister-Ekkehart-Str. 9, 50937 Cologne, Germany

ARTICLE INFO

Article history:
Received 7 January 2011
Received in revised form 15 April 2011
Accepted 24 April 2011

Keywords:
Highway connectors
German Autobahn
Accident causes
Negative binomial regression
Random parameters

ABSTRACT

We make an attempt to identify factors that explain accidents on German Autobahn connectors. To find these factors we perform an empirical study making use of count data models with fixed and random coefficients. The findings are based on a set of 197 ramps, which we classify into three distinct types of ramps. For these ramps, accident data is available for a period of 3 years (January 2003 until December 2005). The negative binomial model with some random coefficients proved to be an appropriate model in our cross-sectional setting for detecting factors that are related to accidents. The most significant variable is a measure of the average daily traffic. For geometric variables, not only continuous effects were found to be significant, but also threshold effects indicating the exceedance of certain values.

1. Introduction

The traffic on German highways, the so-called “Autobahn”, has been increasing drastically over the past years and is expected to grow further in the future, due to Germany’s central geographical position in Europe. The increase in traffic surpasses not only the economic growth, but also the speed of construction of roads. If the road network is not expanded significantly, the increasing number of vehicles on German Autobahns will certainly lead to an increasing number of accidents. Due to limitations in the potential expansions of the Autobahn, particularly in the short run, an important task is to identify accident factors and their influence on accident probabilities. This information could give suggestions for low-cost, short-term improvements for the prevention of accidents on existing Autobahn segments. One of the most dangerous situations for car drivers on Autobahns is the weaving out of the flow of traffic via a road connector. In the years 2003–2005 nearly 8000 accidents happened on road connectors on Autobahns in the administrative district Düsseldorf, which has an extremely dense Autobahn network and is the region we focus on in this study. Due to the safety-standards on Autobahns “only” 10 of these accidents were fatal, however, the economic damage caused by accidents is remarkable.

Several studies found that about 90% of all accidents are at least partially caused by human failure, see, e.g. Treat et al. (1977). As driver behavior is influenced by the whole environment, the aim of road construction should be to construct road sites that prevent or forgive human errors. The problem with road connectors is that each of them is constructed differently according to distinct traffic volumes or geographical constraints. The key question is which factors cause drivers to make mistakes. The aim of our study is to find a model that explains the number or the probability of accidents at various types of Autobahn connectors. This is a statistical problem. However, due to the nature of the problem, the use of standard linear regression models is inappropriate, as argued by Jovanis and Chang (1986) and Miaou and Lum (1993). The variable of interest, namely the number of accidents during a given time interval, suggests the use of count data models.

Miaou and Lum (1993), who investigated the relationship between truck accidents and roadway geometries, and Pickering et al. (1986) used the Poisson regression model to study accident data. Hauer et al. (1988), on the other hand, introduced the more appropriate negative binomial model to find that traffic flow and various road characteristics have a significant effect on the number of accidents on signalized intersections in Toronto. Another study applying the negative binomial model to determine the causes of car accidents is Shankar et al. (1995), who analyzed accidents on a section of the Interstate 90 near Seattle. Both Poisson and negative binomial models require a cross sectional setting. Chin and Quddus (2003) found that panel count data models have the advantage that they are able to deal with spatial or temporal effects in contrast to cross sectional count data models. They analyzed different types of accidents on signalized intersections in Singapore using a set of variables containing geometric variables, traffic volume variables and regulatory controls. Another paper applying panel data techniques to study accident data is Shankar et al. (1998). As accident data typically tends to have more zero-observations

* Corresponding author. Tel.: +49 221 470 4130; fax: +49 221 470 5084.
E-mail addresses: mgarnowski@fedex.com (M. Garnowski), manner@statistik.uni-koeln.de (H. Manner).

0001-4575/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2011.04.026

than are predicted by standard count data models, zero-inflated models have been introduced into traffic accident research. For example, Shankar et al. (1997) investigated accidents on arterials in Washington with two years of accident data and concluded that zero-inflated models have a great flexibility in uncovering processes affecting accident frequencies on roadway sections. For run-off-roadway accidents on a section of a highway in Washington State Lee and Mannering (2002) got promising results applying zero-inflated models in contrast to standard models. However, Washington et al. (2003) and Lord et al. (2005) provide arguments against the use of zero-inflated models in the analysis of accident data. Recently, Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009) introduced count data models with random parameters to account for unobserved heterogeneity and found that these models perform very well. For more on regression models with count data we refer to, among others, Cameron and Trivedi (1986); Lord et al. (2005); Lord and Mannering (2010) and Kibria (2006).

None of the above-mentioned studies analyzes data on highway connectors, but the statistical techniques and explanatory variables that they use are similar to the ones used here. We make an attempt to find an appropriate model for our dataset of 3 years of accidents on Autobahn connectors in the administrative district Düsseldorf (approximately a fifth of the area of North Rhine-Westphalia). In our analysis, we consider more than 60 Autobahn connectors with 197 ramps in an area of approximately 2300 km^2, using traffic data and geometric variables both in continuous form and allowing for threshold effects, which are represented by dummies indicating the exceedance of certain threshold values.

The remainder of the paper is organized as follows. In the next section we describe our methodology. In Section 3 we introduce and explain our dataset, followed by the presentation of the empirical results in Section 4. Finally, Section 5 concludes our paper.

2. Methodology

As our variable of interest, the number of accidents on highway connectors, is a *count variable*, linear regression models are not an appropriate tool for our analysis. Instead, we make use of count data regression models that have been designed for the specific purpose of modeling discrete count variables. The benchmark model for count data is the Poisson regression model, which is derived from the Poisson distribution. We assume a cross sectional setting with n independent observations, the ith of which being $(y_i, \mathbf{x}_i)$, where y_i is the number of occurrences of the event of interest and $\mathbf{x}_i$ is a vector of regressors that determine the number of accidents y_i. The Poisson regression model is defined by

$$f(y_i|\mathbf{x}_i, \beta) = \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}, \qquad y_i = 0, 1, 2, \ldots \tag{2.1}$$

where $\lambda_i > 0$ is the intensity or rate parameter of observation i. Equation (2.1) measures the probability of y_i occurrences of an event during a unit of time. In this model, the mean and the variance are the same, which is called the *equidispersion*-property of the Poisson distribution. The intensity parameter λ_i is assumed to depend on the regressors through

$$\lambda_i = \exp(\mathbf{x}_i'\beta), \tag{2.2}$$

where the log-linear dependence of λ_i on $\mathbf{x}_i$ assures that the intensity parameter is always positive. It is crucial that the conditional mean equation is correctly specified and that the assumption of equidispersion is satisfied. In the case of overdispersion, maximum likelihood estimation (MLE) t-statistics are inflated, which can lead to too optimistic conclusions about the statistical significance of regressors. The assumption that y_i is Poisson distributed can be relaxed considerably as studied in Gourieroux et al. (1984a,b). Given a correctly specified mean, the pseudo MLE based on a density from the linear exponential family (LEF) is consistent. This allows the assumption of equidispersion to be relaxed either by allowing for specific variance functions or by leaving the form of the variance unspecified. In the latter case standard errors can be obtained by a robust sandwich or bootstrap estimator.

As the assumption of equidispersion is unlikely to hold in reality, a natural extension of the model is to allow for unobserved heterogeneity. Unobserved heterogeneity arises when the covariates do not account for the full amount of individual heterogeneity. An extension of the Poisson model that allows for unobserved heterogeneity and overdispersion is the negative binomial regression (NB) model. The NB model can be obtained by writing

$$\lambda_i = \exp(\mathbf{x}_i'\beta + \varepsilon_i), \tag{2.3}$$

where $\exp(\varepsilon_i)$ follows a gamma distribution with mean 1 and variance α. For this reason it is also called the Poisson-gamma model. The density of the NB distribution is given by

$$f(y_i|\mathbf{x}_i, \beta, \alpha) = \frac{\Gamma(\alpha^{-1} + y_i)}{\Gamma(\alpha^{-1})y_i!}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \lambda_i}\right)^{\alpha^{-1}}\left(\frac{\lambda_i}{\alpha^{-1} + \lambda_i}\right)^{y_i}. \tag{2.4}$$

The variance of this distribution is:

$$\mathrm{V}[y_i|\lambda_i, \alpha] = \lambda_i(1 + \alpha\lambda_i) > \lambda_i.$$

Thus, for $\alpha > 0$, this model allows for overdispersion. The NB regression is also estimated by MLE and, as it is also a member of LEF, it is robust to distributional misspecifications. However, if the model is misspecified the maximum likelihood standard errors are in general inconsistent and either robust sandwich or bootstrap standard errors should be used.

One possibility to allow for heterogeneity across observations (possibly caused by unobserved factors) is to let all or some of the parameters be random. Random parameter count data models for accident data have been proposed by Anastasopoulos and Mannering (2009) and El-Basyouny and Sayed (2009). The random parameters are written as

$$\beta_i = \beta + \varphi_i, \tag{2.5}$$

where φ_i is a random variable with density $g(\cdot)$. The most popular choice is the normal distribution with mean 0 and variance σ^2, which we also use in this paper. Conditional on the random components, the intensity parameters are given by $\lambda_i|\varphi_i = \exp(\mathbf{x}_i'\beta)$ and $\lambda_i|\varphi_i = \exp(\mathbf{x}_i'\beta + \varepsilon_i)$ for the Poisson and negative binomial regression, respectively. The log-likelihood of the random parameter model can be obtained by integrating out φ_i from the joint density of y_i and φ_i:

$$\ln \mathrm{L} = \sum_{i=1}^{n} \ln \int_{\varphi_i} g(\varphi_i) f(y_i|\varphi_i) d\varphi_i. \tag{2.6}$$

As this integral cannot be evaluated analytically and numerical integration is computational infeasible when the number of random parameters goes beyond one or two, Anastasopoulos and Mannering (2009) suggest to evaluate it by simulation. However, instead of using pseudo random numbers the integral above is evaluated using so called scrambled Halton sequences. Halton sequences are non-random sequences that cover the domain of integration more uniformly than random numbers and lead to a more precise evaluation of the integral with fewer draws. We refer to Train (1999) and Bhat (2001, 2003) for details on Halton sequences and their use in simulated maximum likelihood. Note that we treat a parameter as random when its estimated variance is significantly different from zero.

In order to decide between the competing models, it is important to test for overdispersion in the data. Besides comparing the