Available online at www.sciencedirect.com

ScienceDirect

Insurance: Mathematics and Economics 42 (2008) 817–830

INSURANCE
MATHEMATICS & ECONOMICS

www.elsevier.com/locate/ime

Heavy-tailed longitudinal data modeling using copulas

Jiafeng Sun, Edward W. Frees*, Marjorie A. Rosenberg

Department of Actuarial Science, Risk Management and Insurance, School of Business, University of Wisconsin, Madison, WI 53706, USA

Received August 2006; received in revised form November 2006; accepted 11 September 2007

Abstract

In this paper, we consider "heavy-tailed" data, that is, data where extreme values are likely to occur. Heavy-tailed data have been analyzed using flexible distributions such as the generalized beta of the second kind, the generalized gamma and the Burr. These distributions allow us to handle data with either positive or negative skewness, as well as heavy tails. Moreover, it has been shown that they can also accommodate cross-sectional regression models by allowing functions of explanatory variables to serve as distribution parameters.

The objective of this paper is to extend this literature to accommodate longitudinal data, where one observes repeated observations of cross-sectional data. Specifically, we use copulas to model the dependencies over time, and heavy-tailed regression models to represent the marginal distributions. We also introduce model exploration techniques to help us with the initial choice of the copula and a goodness-of-fit test of elliptical copulas for model validation. In a longitudinal data context, we argue that elliptical copulas will be typically preferred to the Archimedean copulas. To illustrate our methods, Wisconsin nursing homes utilization data from 1995 to 2001 are analyzed. These data exhibit long tails and negative skewness and so help us to motivate the need for our new techniques. We find that time and the nursing home facility size as measured through the number of beds and square footage are important predictors of future utilization. Moreover, using our parametric model, we provide not only point predictions but also an entire predictive distribution.

JEL classification: primary I110; secondary C200

Keywords: Healthcare costs; Predictive modeling

1. Introduction

In finance and insurance, the assumption of normality is widespread. Empirical evidence shows, however, that data often exhibit "heavy-tails", meaning extreme values in the data are more likely to occur than in normally distributed data. In finance, for example, the asset pricing theories CAPM and APT assume normally distributed asset returns. Distributions of the returns of financial assets, however, suggest heavy-tailed distributions rather than normal distributions as assumed in the pricing theories (see, for example, Rachev et al., 2005). In healthcare, heavy-tailed data are also common. For example, outcomes of interest such as the number of inpatient days or inpatient expenditures are typically right skewed and heavy-tailed due to a few yet high-cost patients (Manning et al., 2005). Actuaries can also regularly encounter heavy tails in non-life insurance (Klugman et al., 2004).

To describe the tail behavior of data, extreme-value statistics has been an area of active development recently (Beirlant et al., 2004). However, the subject of extreme-value statistics is devoted to modeling tail behavior at the expense of largely ignoring the rest of the distribution. In contrast, our interest is in the entire distribution. Moreover, in this paper we focus on regression analysis of heavy-tailed data. Regression models allow researchers to understand variables of interest in terms of other "explanatory" variables. Regression modeling in extreme-value statistics is an area that has only begun to receive serious attention (Beirlant et al., 2004).

There are three commonly used techniques in regression analysis to deal with skewness and heavy tails. The most widely used method is to take a logarithmic transformation of the dependent variable and then apply ordinary least squares (Carroll and Ruppert, 1988). A second natural approach is to use generalized linear models (GLMs), an important class

* Corresponding author.
E-mail address: jfrees@bus.wisc.edu (E.W. Frees).

0167-6687/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2007.09.009

of nonlinear regression models. In GLMs, one directly assumes a parametric distribution family for the dependent variable but then allows the mean parameter to be a function of the covariates. GLMs have been applied in insurance since early 1980s, Haberman and Renshaw (1996) reviewed GLM applications to actuarial problems. Further, in the classic work by McCullagh and Nelder (first published in 1983, second edition in 1989), many examples of insurance were also given to illustrate how to fit GLMs to different types of data. A third approach is to use parametric survival models, including location–scale and proportional hazard models (Lawless, 2003). Parametric survival analysis includes regression models to analyze censored data, but the methods can certainly be applied to complete data. For complete treatments of the parametric survival models, see Lawless (2003) and Kalbfleisch and Prentice (2002).

In some heavy-tailed data situations, flexible positive random variable distributions are more appropriate than the distributions used with GLMs and in survival modeling. The three parameter generalized gamma (GG) and Burr distributions, as well as the four parameter generalized beta of the second kind (GB2), have been used to analyze cross-sectional data in the econometrics literature. When the data are *negatively* skewed and have a fat left tail, such as the Wisconsin nursing homes utilization data we analyze in this paper, flexible distributions are also needed. Section 2 introduces the flexible distributions that we will use in our regression modeling.

Although widely applicable, traditional regression analysis is limited in that all the observations are assumed to be statistically independent. In insurance and other fields, often the outcome of interest is measured repeatedly over time. This type of data structure is called longitudinal or panel data. In contrast to the cross-sectional data where a single outcome is observed for each subject, in a longitudinal data framework, observations of a variable of interest and a set of covariates are made repeatedly on several subjects over time. General discussions of longitudinal data and longitudinal modeling can be found in Diggle et al. (2002), Baltagi (2005) and Frees (2004).

To model heavy-tailed longitudinal data, one natural approach is to take a logarithmic (or other nonlinear) transformation of the dependent variable and then apply the usual linear models that assume that the dependent variable follows a multivariate normal distribution. As with cross-sectional modeling, this has the advantage of the ease of implementation. From a user's perspective, the main disadvantage is that one is forced to think of modeling in terms of the rescaled dependent variable which is often difficult to interpret.

Another natural approach is to use GLMs, where two classes of longitudinal models are commonly used in the literature. The first class is known as a marginal model, where the association among observations from the same subject is not of explicit research interest. The mean regression is modeled as a generalized linear regression model, separately from the association among repeated observations from each subject. Marginal models are semiparametric in that only the mean, variance and covariances among responses are specified. For long-tailed data, this could represent a serious loss of information. Moreover, moments may not even exist with long-tailed data, meaning that moment-based methods are of limited use. In marginal models, the association among repeated measures is of secondary interest compared to the regression parameters which have population average interpretations. However, for many insurance and finance applications, behavior over time is the key element of interest in the problem. Nonetheless, marginal models require fewer assumptions, and are computationally simpler, in many problems when generalized estimating equations (GEE) are employed for estimation rather than a likelihood based method. For a complete treatment of GEE, see Hardin and Hilbe (2003).

Random effects models represent a second class that uses a GLM approach. These can be motivated using a "two-stage" sampling scheme. In the first stage, n subjects are drawn randomly from the population so that certain parameters are associated with each subject. In the second stage, conditional on subject specific parameters, observations are drawn for subject i at repeated time points t. The underlying idea is that there is heterogeneity among subjects which can be modeled by a probability distribution, while the association among observations from each subject arises from the unobservable characteristics. Unlike marginal models, random effects regressions include covariate effects and within-subject association through a single equation. Regression parameters can vary across subjects; they measure the effects of explanatory variables on the response variable for each subject. Generalized linear random effects models extend the linear mixed models in that the random effects are included in the linear predictor, $g(\mu_{it}) = \mathbf{z}_{it}'\boldsymbol{\alpha}_i + \mathbf{x}_{it}'\boldsymbol{\beta}$, where g is a link function, $\mathbf{z}_{it}$ and $\mathbf{x}_{it}$ are rows in the design matrices for random and fixed effects, $\boldsymbol{\alpha}_i$ is for random effects of subject i and $\boldsymbol{\beta}$ is the parameter vector for fixed effects. The distribution of $\boldsymbol{\alpha}_i$ is usually assumed to follow a normal distribution. Given $\boldsymbol{\alpha}_i$, the responses from each subject are independent and follow a GLM.

Generalized linear random effects models are usually estimated using likelihood based methods. It is difficult to justify a particular distribution for the random effects. Moreover, maximum likelihood estimation based on the marginal distribution of the observations integrates out the random effects, which, in some cases, is numerically not feasible (Schall, 1991). Breslow and Clayton (1993) used penalized quasi-likelihood to conduct estimation. For a thorough treatment of theory and computation for the random effects models, see Pinheiro and Bates (2000).

In this paper, we introduce heavy-tailed regression models in the framework of longitudinal data using copulas. A copula is a multivariate distribution with uniform marginal distributions on the interval $(0, 1)$. As tools to construct multivariate distributions, copulas are increasingly explored in the statistics, econometrics, finance and insurance literature. Copulas separate the multivariate joint distribution into two parts: one describing the interdependency of the probabilities, the other describing the marginal distributions only. Through