mathematics

MDPI

Article

A Sarmanov Distribution with Beta Marginals: An Application to Motor Insurance Pricing

Catalina Bolancé †, Montserrat Guillen *,† and Albert Pitarque †

Department Econometrics, Riskcenter-IREA, Universitat de Barcelona, E08034 Barcelona, Spain; bolance@ub.edu (C.B.); albertpitarque@ub.edu (A.P.)

* Correspondence: mguillen@ub.edu; Tel.: +34-934-037-039

† These authors contributed equally to this work.

Received: 26 October 2020; Accepted: 9 November 2020; Published: 13 November 2020

Abstract: Background: The Beta distribution is useful for fitting variables that measure a probability or a relative frequency. Methods: We propose a Sarmanov distribution with Beta marginals specified as generalised linear models. We analyse its theoretical properties and its dependence limits. Results: We use a real motor insurance sample of drivers and analyse the percentage of kilometres driven above the posted speed limit and the percentage of kilometres driven at night, together with some additional covariates. We fit a Beta model for the marginals of the bivariate Sarmanov distribution. Conclusions: We find negative dependence in the high quantiles indicating that excess speed and night-time driving are not uniformly correlated.

Keywords: beta regression; dependence; bivariate Sarmanov distribution; estimation; telematics; insurance

1. Introduction

We analyse a bivariate model based on the Sarmanov distribution with marginal Beta distributions. These marginals are specified based on a generalized linear model (Beta-GLM) or Beta regression as defined by Ferrari and Cribari-Neto [1]. The objective is to fit data defined in the $(0, 1)$ interval.

Many authors have analysed bivariate Beta distributions (see, for example, [2–5]). However, these distributions pose several difficult challenges: their generalization to higher dimensions and their specification as a generalized linear model are not straightforward. The Sarmanov distribution provides a way to address these challenges.

Originally, the Sarmanov distribution in its bivariate form was introduced by Sarmanov [6], its multivariate version was suggested by Lee [7] and was generalized by Bairamov et al. [8]. Its use to model the bivariate behaviour of random variables with a marginal $Beta(\alpha, \beta)$ distribution was proposed by Gupta and Wong [3]. These authors defined the five parameter bivariate Beta distribution from what is known as Morgenstern's distribution [9] with marginal Beta, which is a particular case of the Sarmanov distribution.

The bivariate Sarmanov distribution is characterized by its flexibility in the marginal distributions and, furthermore, given that its functional form establishes that the marginals are clearly separated from the dependency model, the specification in terms of a bivariate generalized linear model turns out to be natural. Generalizing from two dimensions to higher dimensions is simple—(see [10] for an example of a trivariate Sarmanov distribution specified as a generalized linear model with Negative Binomial marginals).

In this work, we show an application of the bivariate Sarmanov distribution with Beta marginals generalised linear model to predict two of the most relevant telematics variables in motor insurance [11]. Telematics variables are obtained from GPS/inertial devices installed in vehicles and they provide an

abundant source of information to motor insurers. In our case study, a bivariate model is specified, for the proportion of kilometres driven above the posted speed limit and the proportion of kilometres driven at night. These two variables seem to be related, but researchers have not yet been able to find a good way to understand their association. The explanatory variables are the characteristics of the insured policyholder and the vehicle. The database used in our application has already been analysed in various works published in statistical, transport and risk analysis journals (see [11–17]). In all previous studies, the two telematics variables that we analyse here were used as predictors of the accident rate, and they were assumed to be uncorrelated.

In Section 2, the new bivariate Sarmanov model is specified and the particular case with marginal Beta-GLM with a domain in the $(0,1)$ interval is analysed; the estimation method is also discussed. The results of our case study are shown in Section 3. Finally, Section 4 contains the conclusions.

2. The Models

Let (Y_1, Y_2) be a bivariate random vector that, for convenience, is defined in $(0,1)^2$. Its distribution depends on a set of k quantitative or binary covariates, whose values are represented by the vector $x_j = \left(x_{1j}, ..., x_{kj}\right)'$, $j = 1,2$, where $x_{1j} = 1$ is a constant term. The relationship between Y_j and the covariates is given by the linear predictor $x_j'\beta^j$, where $\beta^j = \left(\beta_1^j, ..., \beta_k^j\right)'$, $j = 1,2$, are vectors of parameters to be estimated. To simplify the notation, the covariates are assumed to be the same for $j = 1$ and $j = 2$, and so the vector of explanatory variables is denoted as x. The bivariate probability density function (pdf) associated with the Sarmanov distribution is:

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2|x'\beta^1, x'\beta^2) &= f_1(y_1|x'\beta^1) f_2(y_2|x'\beta^2) \\ &\times \left[1 + \omega\phi_1(y_1|x'\beta^1)\phi_2(y_2|x'\beta^2)\right],\ y_1, y_2 \in (0,1) \end{aligned} \tag{1}$$

where ω is the dependence parameter and ϕ_j, $j = 1,2$, are bounded kernel functions. For the function defined in (1) to be a pdf, the following conditions must hold:

$$\int_0^1 \phi_j(y_j|x'\beta^j) f_j(y_j|x'\beta^j)\, dy_j = 0, j = 1,2 \tag{2}$$

and

$$1 + \omega\phi_1(y_1|x\beta^1)\phi_2(y_2|x\beta^2) \geq 0,\ \forall\,(y_1, y_2) \in (0,1)^2. \tag{3}$$

For given values of $x'\beta^j$, $j = 1,2$, we define:

$$m_j\left(x'\beta^j\right) = \inf_{0<y_j<1} \phi_j\left(y_j|x'\beta^j\right) \quad \text{and} \quad M_j\left(x'\beta^j\right) = \sup_{0<y_j<1} \phi_j\left(y_j|x'\beta^j\right),\ j = 1,2.$$

Taking into account the condition defined in (3), bounds can be defined for the dependency parameter ω. However, as this parameter does not depend on the linear predictor, new extreme values are defined as: $m_j^\star = \max_{\forall x'\beta^j} m_j(x'\beta^j)$ and $M_j^\star = \min_{\forall x'\beta^j} M_j(x'\beta^j)$, so that the bounds of the dependency parameter are:

$$\max\left\{-\frac{1}{m_1^\star m_2^\star}, -\frac{1}{M_1^\star M_2^\star}\right\} \leq \omega \leq \min\left\{-\frac{1}{m_1^\star M_2^\star}, -\frac{1}{M_1^\star m_2^\star}\right\}. \tag{4}$$

The previous condition holds for every vector of covariates x, which implies that the dependency parameter must be located within the narrowest bounds. In practice, we will assume that the vectors observed in the sample dataset lead to the entire domain of values of linear predictors $x'\beta^j$, $j = 1,2$. In the insurance context, where we will discuss our illustration, we assume that all possible risk profiles that can be insured by the company are already present in the portfolio.

For each vector of covariate observations x, we can also obtain the covariance between the dependent variables as:

$$cov(Y_1, Y_2) = \omega v_1(x) v_2(x), \tag{5}$$

where $v_j(x) = \int_0^1 y_j \phi_j(y_j|x\beta^j) f_j(y_j|x'\beta^j)\, dy_j,\ j = 1,2$. The correlation is obtained by dividing by the product of standard deviations.

There exist many possible specifications for the kernel functions ϕ_j, $j = 1,2$ (see [18] [for a description of kernel functions proposed in the literature). When fitting the bivariate Beta distribution without covariates, Gupta and Wong [3] propose a kernel function such as $\phi_j = 2F_j - 1$, where F_j is the cumulative distribution function (cdf). This specification has the advantage that the bounds for the dependency parameter are given by $-1 \leq \omega \leq 1$ for any vector x. However, the previous model does not allow obtaining closed expressions for some magnitudes of interest, such as the conditioned moments. In this work, we propose to use kernels $\phi_j = y_j^r - E(Y_j^r)$, where r is a value to be determined by the analyst. Next, some results obtained for the particular case of the Sarmanov distribution with marginal $Beta(\alpha, \beta)$ distribution with $r = 1$ are analyzed. These cases intuitively correspond to a situation of linear dependency, controlled by the dependence parameter ω.

2.1. The Bivariate Beta GLM Model

The pdf of a random variable Y with $Beta(\alpha, \beta)$ distribution, with parameters $\alpha,\ \beta > 0$, is:

$$f_Y(y; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} y^{\alpha-1}(1-y)^{\beta-1} = \frac{1}{B(\alpha,\beta)} y^{\alpha-1}(1-y)^{\beta-1}$$

and its cdf is:

$$F_Y(y; \alpha, \beta) = \frac{B(y, \alpha, \beta)}{B(\alpha, \beta)},$$

where $\Gamma(\cdot)$ and $B(\cdot,\cdot)$ are the Gamma and Beta functions, respectively, and $B(y,\cdot,\cdot)$ is the incomplete Beta function.

The Beta regression was proposed by Ferrari and Cribari-Neto [1], with the reparametrization $\mu = \frac{\alpha}{\alpha+\beta}$ and $\psi = \alpha + \beta$, so that:

$$f(y; \mu, \psi) = \frac{1}{B\left(\mu\psi, (1-\mu)\psi\right)} y^{\mu\psi-1}(1-y)^{(1-\mu)\psi-1},$$

where $E(Y) = \mu$, with $0 < \mu < 1$, and $V(Y) = \frac{\mu(1-\mu)}{(1+\psi)}$, with $\psi > 0$, where ψ^{-1} is the scale parameter. We note that, given the values of μ and ψ, it holds that $V(Y) < 0.25$. The specification as GLM is defined as (note that we use $\mu(x)$ to emphasize that μ depends on the linear predictor):

$$g\left[\mu(x)\right] = x'\beta,$$

where $g[\cdot]$ is a link function that can be defined in different ways, in this work, we use the logit link, $g\left[\mu(x)\right] = \log\left[\frac{\mu(x)}{1-\mu(x)}\right]$.

To simplify the notation from now on, we eliminate the linear predictors in the conditioned part. The pdf associated with the bivariate random vector (Y_1, Y_2) with a Sarmanov distribution and Beta GLM marginals that will be called the Sarmanov-Beta-GLM is ():

$$\begin{aligned} f_{Y_1,Y_2}(y_1, y_2) &= \frac{1}{B\left(\mu_1(x)\psi_1, (1-\mu_1(x))\psi_1\right)} y_1^{\mu_1(x)\psi_1-1}(1-y_1)^{(1-\mu_1(x))\psi_1-1} \\ &\times \frac{1}{B\left(\mu_2(x)\psi_2, (1-\mu_2(x))\psi_2\right)} y_2^{\mu_2(x)\psi_2-1}(1-y_2)^{(1-\mu_2(x))\psi_2-1} \\ &\times \left[1 + \omega(y_1 - \mu_1(x))(y_2 - \mu_2(x))\right],\ y_1, y_2 \in (0,1). \end{aligned} \tag{6}$$

For the previous expression to be a pdf, the dependency parameter must be located within the bounds defined in (4), which, for the kernel functions that we propose, are:

$$\begin{aligned}&\max\left\{-\frac{1}{\max\limits_{\forall x'\beta^1}(-\mu_1(x))\max\limits_{\forall x'\beta^2}(-\mu_2(x))},-\frac{1}{\min\limits_{\forall x'\beta^1}(1-\mu_1(x))\min\limits_{\forall x'\beta^2}(1-\mu_2(x))}\right\}\\&\leq\omega\leq\\&\min\left\{-\frac{1}{\min\limits_{\forall x'\beta^1}(1-\mu_1(x))\max\limits_{\forall x'\beta^2}(-\mu_2(x))},-\frac{1}{\max\limits_{\forall x'\beta^1}(-\mu_1(x))\min\limits_{\forall x'\beta^2}(1-\mu_2(x))}\right\}.\end{aligned}\tag{7}$$

The bivariate cdf associated with a Sarmanov-Beta-GLM is obtained directly from the double integral of the bivariate pdf defined in (6):

$$\begin{aligned}F_{Y_1,Y_2}(y_1,y_2)&=\frac{B(y_1,\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}{B(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}\times\frac{B(y_2,\psi_2\mu_2(x),(1-\mu_2(x))\psi_2)}{B(\psi_2\mu_2(x),(1-\mu_2(x))\psi_2)}\\\times&\left[1+\omega\left(\frac{B(y_1,\psi_1\mu_1(x)+1,(1-\mu_1(x))\psi_1)}{B(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}-\mu_1(x)\frac{B(y_1,\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}{B(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}\right)\right.\\\times&\left.\left(\frac{B(y_2,\psi_2\mu_2(x)+1,(1-\mu_2(x))\psi_2)}{B(\psi_2\mu_2(x),(1-\mu_2(x))\psi_2)}-\mu_2(x)\frac{B(y_2,\psi_2\mu_2(x),(1-\mu_2(x))\psi_2)}{B(\psi_2\mu_2(x),(1-\mu_2(x))\psi_2)}\right)\right],\end{aligned}\tag{8}$$

where $y_1,y_2\in(0,1)$.

Proposition 1. *The conditioned pdf is:*

$$\begin{aligned}f_{Y_1|Y_2}(y_1|Y_2=y_2)&=\frac{1}{B(\mu_1(x)\psi_1,(1-\mu_1(x))\psi_1))}y_1^{\mu_1(x)\psi_1-1}(1-y_1)^{(1-\mu_1(x))\psi_1-1}\\&\times\left[1+\omega(y_1-\mu_1(x))(y_2-\mu_2(x))\right],\quad y_1,y_2\in(0,1)\end{aligned}\tag{9}$$

and similarly for $f_{Y_2|Y_1}(y_2|Y_1=y_1)$. Integrating the previous expression, the conditional cdf is obtained as

$$\begin{aligned}F_{Y_1|Y_2}(y_1|Y_2=y_2)=F_1(y_1)&\times\left[1+\omega(y_2-\mu_2(x))(1-\mu_1(x))\right]\\&-\omega(y_2-\mu_2(x))\frac{y_1(1-y_1)}{\psi_1\mu_1(x)}f_1(y_1),\ y_1,y_2\in(0,1).\end{aligned}\tag{10}$$

Proof. The conditioned pdf is obtained directly as

$$f_{Y_1|Y_2}(y_1|Y_2=y_2)=\frac{f_{Y_1,Y_2}(y_1,y_2)}{f_{Y_2}(y_2)}.$$

Integrating the result of $f_{Y_1|Y_2}(y_1|Y_2=y_2)$ in (9), we obtain:

$$\begin{aligned}&F_{Y_1|Y_2}(y_1|Y_2=y_2)=\int_0^{y_1}f_1(t)dt+\omega(y_2-\mu_2(x))\int_0^{y_1}f_1(t)(t-\mu_1(x))\,dt\\=\;&F_1(y_1)+\omega(y_2-\mu_2(x))\left[\frac{B(y_1,\psi_1\mu_1(x)+1,(1-\mu_1(x))\psi_1)}{B(\psi_1\mu_1(x),(1-\mu_1(x))\psi_1)}-\mu_1(x)F_1(y_1)\right].\end{aligned}\tag{11}$$

In addition, since

$$\begin{aligned} & \frac{B\left(y_1, \psi_1\mu_1(x)+1, (1-\mu_1(x))\psi_1\right)}{B\left(\psi_1\mu_1(x), (1-\mu_1(x))\psi_1\right)} \\ = & \frac{B\left(y_1, \psi_1\mu_1(x), (1-\mu_1(x))\psi_1\right)}{B\left(\psi_1\mu_1(x), (1-\mu_1(x))\psi_1\right)} - \frac{y_1^{\mu_1(x)\psi_1}(1-y_1)^{(1-\mu_1(x))\psi_1}}{\psi_1\mu_1(x)B\left(\psi_1\mu_1(x), (1-\mu_1(x))\psi_1\right)} \\ = & F_1(y_1) - \frac{y_1(1-y_1)}{\psi_1\mu_1(x)} f_1(y_1), \end{aligned}$$

then, by substituting the previous expression in (11), then (10) follows directly. □

The conditioned quantile is obtained from the inverse of expression (10), for which a numerical method (such as Newton's method) can be used.

Proposition 2. *The conditional expectation is:*

$$E\left(Y_1|Y_2=y_2\right) = \mu_1(x) + \omega(y_2-\mu_2(x))V\left(Y_1|x\right), \tag{12}$$

where $V\left(Y_1|x\right) = \frac{\mu_1(x)(1-\mu_1(x))}{(\psi_1+1)}$ *is the variance, which also depends on the vector of covariates. Similarly,* $E\left(Y_2|Y_1=y_1\right)$ *can be found.*

Proof. The conditional expectation is obtained directly by solving the integral:

$$\begin{aligned} E\left(Y_1|Y_2=y_2\right) &= \int_0^1 y_1 f_{Y_1|Y_2}(y_1|Y_2=y_2)\,dy_1 \\ &= \int_0^1 y_1 f_{Y_1}(y_1)\,dy_1 \times (1+\omega(y_1-\mu_1(x))(y_2-\mu_2(x))) \\ &= \int_0^1 y_1 f_{Y_1}(y_1)\,dy_1 \\ &\quad +\omega(y_2-\mu_2(x))\left(\int_0^1 y_1^2 f_{Y_1}(y_1)\,dy_1 - \mu_1(x)\int_0^1 y_1 f_{Y_1}(y_1)\,dy_1\right) \\ &= \mu_1(x) + \omega\left(y_2-\mu_2(x)\right)\left(E(Y_1^2|x) - \mu_1(x)^2\right) \\ &= \mu_1(x) + \omega\left(y_2-\mu_2(x)\right)V\left(Y_1|x\right). \end{aligned}$$

Likewise, the corresponding result is obtained for $E\left(Y_2|Y_1=y_1\right)$. □

Proposition 3. *From (5), the conditional covariance which depends on the vector of covariates x is:*

$$cov(Y_1,Y_2) = \omega V(Y_1)V(Y_2) = \omega\frac{\mu_1(x)(1-\mu_1(x))}{(\psi_1+1)}\frac{\mu_2(x)(1-\mu_2(x))}{(\psi_2+1)} \tag{13}$$

and the correlation is:

$$corr(Y_1,Y_2) = \omega\sqrt{\frac{\mu_1(x)(1-\mu_1(x))}{(\psi_1+1)}}\sqrt{\frac{\mu_2(x)(1-\mu_2(x))}{(\psi_2+1)}}. \tag{14}$$

Proof. Note that the covariance and the correlation are calculated directly if, in expression (5), we see that:

$$\begin{aligned} v_j(x) &= \int_0^1 y_j\phi_j(y_j|x\beta^j)f_j(y_j|x'\beta^j)\,dy_j \\ &= \int_0^1 y_j(y_j-\mu_j(x))f_j(y_j|x'\beta^j)\,dy_j = E(Y_j^2|x) - \mu_j(x)^2, j=1,2 \end{aligned}$$

□

The dependence parameter of the model proposed in Gupta and Wong [3], which uses kernel functions $\phi_j = 2F_j - 1$, $j = 1,2$, is located in the interval $-1 \leq \omega \leq 1$ and is the same for all x. Our proposal bounds the dependence parameter to the narrowest interval among those obtained from all x. However, the advantage of our proposal is that our model allows for obtaining closed expressions for some magnitudes of interest such as bivariate moments (covariance) and conditional moments. In the numerical analysis section, we also compare the correlations estimated from our model and that of Gupta and Wong [3].

2.2. Estimation

In practice, we start from a bivariate sample of n observations. Let us denote the sample information as (Y_{i1}, Y_{i2}), $i = 1, ..., n$, where for each i we know the values of the covariates $X_i = (X_{i1}, ..., X_{ik})'$. Our objective is to estimate the parameter vectors β^j, the scale parameters, ψ_j and $j = 1, 2$, and the dependency parameter ω, from the maximization of the logarithm of the likelihood function associated with the Sarmanov distribution:

$$\begin{aligned} l\left(\beta^1, \beta^2, \psi_1, \psi_2, \omega\right) &= \sum_{i=1}^{n} \log f_1(Y_{i1}|X_i'\beta^1) + \sum_{i=1}^{n} \log f_2(Y_{i2}|X_i'\beta^2) \\ &+ \sum_{i=1}^{n} \log\left(1 + \omega\phi_1(Y_{i1}|X_i'\beta^1)\phi_2(Y_{i2}|X_i'\beta^2)\right) \\ &= l_1(\beta^1, \psi_1) + l_2(\beta^2, \psi_2) + l_{12}(\omega, \beta^1, \beta^2, \psi_1, \psi_2), \end{aligned} \tag{15}$$

The maximization of (15) cannot be carried out directly without considering that the parametric space is restricted and, in addition, as it was shown in expression (4), the bounds of the dependence parameter are closely related to the parameters of the marginals. Thus, in the maximization process, infeasible solutions will often be reached unless a careful numerical procedure is specifically designed. One way to address these difficulties is to rely on the IFM (Inference from Margin) method that has been widely used in the estimation of copulas see [19] [for a review]. For the estimation of the Sarmanov distribution, the IFM was already used by Bolancé and Vernic [10] for the case of GLM marginals with Negative Binomial distributions.

The IFM method is implemented as follows:

Inicialization. The parameters for the marginals are estimated as:

$$\left(\hat{\beta}^{1(0)}, \hat{\psi}_1^{(0)}\right) = \max_{\beta^1, \psi_1} l_1(\beta^1, \psi_1) \tag{16}$$

$$\left(\hat{\beta}^{2(0)}, \hat{\psi}_2^{(0)}\right) = \max_{\beta^2, \psi_2} l_2(\beta^2, \psi_2). \tag{17}$$

For the initial estimation, function `betareg()` of `betareg` R package is used. With these parameters of the marginals, we start the iterative process in the two steps described below.

Step 1. Given the estimated marginal parameters in iteration $m-1$ and taking into account the limits of the dependence parameter ω defined in (4), with function `optim()` and the `L-BFGS-B` method using `R`, we estimate ω from the maximization of the likelihood function given fixed values of the marginal parameters, which is:

$$\hat{\omega}^{(m)} = \max_{\omega} l_{\omega|12}\left(\omega|\hat{\beta}^{1(m-1)}, \hat{\beta}^{2(m-1)}, \hat{\psi}_1^{(m-1)}, \hat{\psi}_2^{(m-1)}\right), \tag{18}$$

where $l_{\omega|12}$ is the likelihood as a function of ω given the estimated parameters for the marginals in iteration $m-1$.

Step 2. Given the estimated dependency parameter $\hat{\omega}^{(m)}$ in step 1, the marginal parameters are re-estimated in iteration m as:

$$\left(\hat{\beta}^{1(m)}, \hat{\psi}_1^{(m)}, \hat{\beta}^{2(m)}, \hat{\psi}_2^{(m)}\right) = \max_{\beta^1,\psi_1,\beta^2,\psi_2} l_{12|\omega}(\beta^1, \psi_1, \beta^2, \psi_2|\hat{\omega}^{(m)}), \tag{19}$$

where $l_{12|\omega}$ is the likelihood as a function of the marginal parameters given the dependence parameter estimated in step 1. The above maximization is also performed with function `optim()` and the `L-BFGS-B` method of `R`.

Steps 1 and 2 described above are repeated until reaching the convergence criterion based on the differences between parameter estimates obtained in two consecutive iterations.

Remark 1. *In the initialization process, if the dependent variables contain zeros or ones, the following correction* $\tilde{Y}_j = (Y_j * (n-1) + 0.5)/n,\ j = 1,2$ *was proposed by Smithson and Verkuilen [20].*

In practice, the algorithm described above is based on the optimization of conditional likelihood functions and not on the likelihood function defined in (15). However, in the last stage, the parameters estimated with the IFM method can be used as initial parameters in the process of maximizing the full likelihood function defined in (15). For this purpose, function `optim()` and method `L-BFGS-B` of `R` are used again.

Remark 2. *To estimate the Sarmanov model proposed by Gupta and Wong [3], it is not necessary to use the two-step process, since the bounds of the dependence parameter do not depend on the parameters of the marginal distributions.*

3. Numerical Analysis

We analyse a database corresponding to a car insurance portfolio, in which part of the variables have been measured via a telematic system. The objective of our analysis is to model the joint behaviour of the percentage of kilometres driven above the posted speed limits (Y_1) and percentage of kilometres driven at night (Y_2). It is well known that both variables are related to the risk of having an accident. In Table 1, we show the main descriptive statistics of the dependent variables and the covariates used in the modelling process. For the estimation of the Sarmanov-Beta-GLM, the dependent variables have been transformed as indicated in Remark 1 in Section 2.2. Furthermore, to avoid very low coefficient values due to the scale of some covariates, variables age (X_1), age of driving license (X_2) and age of the vehicle (X_5) have been divided by 10; the vehicle power variable (X_6) is divided by 100 and the total annual distance driven in kilometres (X_7) is divided by 1000. In addition, note that, in this study, we have included a variable denoting the driver's gender (X_3) and an indicator of private garage (X_4) as covariates.

The last row of Table 1 shows the Pearson correlation between the two dependent variables. This correlation is compared with the corresponding parameter estimate obtained from the Sarmanov model with marginal Beta proposed here and with the one proposed by Gupta and Wong [3], from now on the GW model. With this objective, Table 2 shows the dependence parameters estimated with both models, and the AIC and BIC statistics without including the covariates and including them. Using expression (14) and without covariates, from the dependence parameters $\hat{\omega} = 14.883$, it can be deduced that the estimated correlation is 0.0601, which is within the confidence interval of the Pearson correlation as shown in the last row of Table 1. On the contrary, if we use the five parameter Beta distribution, the (residual) correlation that is obtained from the numerical calculation of expression (5) is practically zero. This means that the association is captured by the bivariate model. Comparing both models, with and without covariates, using the AIC and BIC statistics, the results of Table 2 show that the fit is better for the model proposed here than it is for the GW model.

Table 1. Definition of variables and descriptive statistics: mean, standard deviation (STD), minimum (Min) and Maximum (Max). The last row shows the linear correlation between dependent variables and a confidence interval at the 95% level.

Variable	Description	Mean	STD	Min	Max
Y_1	Percentage of kilometres driven above the speed limit	0.063	0.068	0.000	0.704
Y_2	Percentage of kilometres driven at night	0.069	0.064	0.000	1.000
X_1	Age of the driver	27.565	3.094	19.849	36.904
X_2	Age if driver License	7.174	3.053	1.810	15.910
X_3	Gender (=1 Men, =0 Women)	0.489	0.500	0.000	1.000
X_4	Night parking (=1 yes, 0=no)	0.774	0.418	0.000	1.000
X_5	Age of the vehicle	8.749	4.174	1.938	20.468
X_6	Power of the vehicle in Horse Power (HP)	97.226	27.772	12.000	500.000
X_7	Total Km	7159.510	4191.753	1.590	50,035.560
ρ	Pearson correlation between dependent variables (CI)		0.070 (0.057,0.082)		

Table 2. Estimated dependence from Sarmanov-Beta models and goodness of fit criteria.

		$\hat{\omega}$ (p-Value)		AIC	BIC
Proposed Model	No covariates	14.883	(<0.001)	−171,282.2	−171,241.5
	With all covariates	2.388	(0.055)	−177,508.8	−177,354.4
GW Model	No Covariates	0.002	(0.346)	−171,165.4	−171,124.8
	With all covariates	0.002	(0.356)	−177,497.2	−177,342.8

Table 3 shows the results of our Sarmanov-Beta-GLM using different vectors of covariates. Model I includes all the explanatory variables, among which we have the age (X_1), the age of the driving license (X_2) and the total distance driven annually (X_7), these three variables are associated with driving experience. To analyze the robustness of the results, in Model II, age (X_1) is eliminated, and, in addition, in Model III, the age of a driver's license (X_2) is also eliminated. The results of Model I show that the effect of age is negative on both Y_1 and Y_2 that the effect of the driver's license age is positive on Y_1 and negative on Y_2 and the effect of total distance, X_7, is positive on both dependent variables. By eliminating age (X_1) in Model II, the signs of the parameters associated with X_2 and X_7 are maintained, although the value is smaller in the case of X_2 and remains practically the same for X_7. After eliminating variables X_1 and X_2, we see that the effect of the total annual distance driven remains practically the same. If we observe the effects of the rest of covariates, these are practically the same in models I, II, and III. A man driver (X_3) with a powerful vehicle (X_6) would have larger Y_1 and Y_2 than the rest, all other characteristics being the same. However, using parking at night (X_4) has a positive effect on the percentage of speeding distance (Y_1) and a negative effect on the percentage of night-time driving (Y_2); the opposite happens with the age of the vehicle (X_5). The effect of X_5 indicates that, when the vehicle is older, drivers tends to diminish the percent of speed driving, while night-time driving is larger.

To visualize the dependence between Y_1 and Y_2 in different quantiles, the following three examples of insured drivers are graphically analysed:

- **Profile 1** corresponds to a 27-year-old man, who drives about 7000 kilometres per year, with a 7-year-old driving license, with parking, with a vehicle of about 8 years and 100 HP.
- **Profile 2** corresponds to a 20-year-old man, who drives about 4000 kilometres per year, with a 2-year-old driving license, with parking, with a vehicle of about 2 years and 75 HP.
- **Profile 3** corresponds to a 36-year-old man, who drives about 10,000 kilometres per year, with a 15-year-old driving license, without parking, with a vehicle of about 15 years and 200 HP.

Table 3. Parameter estimates (p-values) for the Sarmanov-Beta models and goodness of fit statistics.

	Model I		**Model II**		**Model III**	
	Y1	**Y2**	**Y1**	**Y2**	**Y1**	**Y2**
Cons.	−3.055 (<0.001)	−2.556 (<0.001)	−3.819 (<0.001)	−2.975 (<0.001)	−3.796 (<0.001)	−3.061 (<0.001)
X_1	−0.339 (<0.001)	−0.185 (<0.001)	-		-	
X_2	0.294 (<0.001)	−0.052 (0.018)	0.048 (0.002)	−0.187 (<0.001)	-	
X_3	0.097 (<0.001)	0.274 (<0.001)	0.107 (<0.001)	0.281 (<0.001)	0.109 (<0.001)	0.274 (<0.001)
X_4	0.108 (<0.001)	−0.031 (0.007)	0.107 (<0.001)	−0.031 (0.007)	0.107 (<0.001)	−0.031 (0.007)
X_5	−0.043 (0.001)	0.055 (<0.001)	−0.043 (0.001)	0.055 (<0.001)	−0.043 (0.001)	0.055 (<0.001)
X_6	0.653 (<0.001)	0.077 (<0.001)	0.654 (<0.001)	0.079 (<0.001)	0.664 (<0.001)	0.038 (0.027)
X_7	0.045 (<0.001)	0.035 (<0.001)	0.046 (<0.001)	0.035 (<0.001)	0.046 (<0.001)	0.035 (<0.001)
ϕ_1	18.480 (<0.001)		18.300 (<0.001)		18.294 (<0.001)	
ϕ_2	14.823 (<0.001)		14.782 (<0.001)		14.703 (<0.001)	
ω	2.388 (0.055)		2.325 (0.059)		2.214 (0.060)	
AIC	−177,508.8		−177,238.5		−177,113.5	
BIC	−177,354.4		−177,100.3		−176,991.6	

Profile 1 represents the average insured individual of the portfolio; Profile 2 is a younger man driver, less experienced than Profile 1 and with a newer and less powerful vehicle; finally, Profile 3 is an older man driver, more experienced than Profile 1 and an older and more powerful vehicle. Figure 1 represents different quantiles of the variable kilometres driven above the speed limit (Y_1) in the y-axis given the values of the percentage of kilometres driven at night (Y_2) for Profile 1 in the x-axis. Quantiles have been obtained from the expression (10). Note that, if the dependence parameter was zero, all the curves would remain constant. The adjusted dependence structure results in the represented conditional quantiles having a negative nonlinear relationship and, furthermore, the curves for the different quantile levels are non-parallel. Figure 1 indicates that, for Profile 1, the higher the percentage of kilometres driven at night (Y_2), the greater the caution in driving and, therefore, the lower the percentage of distance driven above the speed limits (Y_1). The same quantiles at 75% (plot on the left) and 95% (plot on the right) confidence levels are represented in Figure 2. These plots show that the curves are non-parallel and that Profile 3 is the most risky, followed by Profiles 1 and 2.

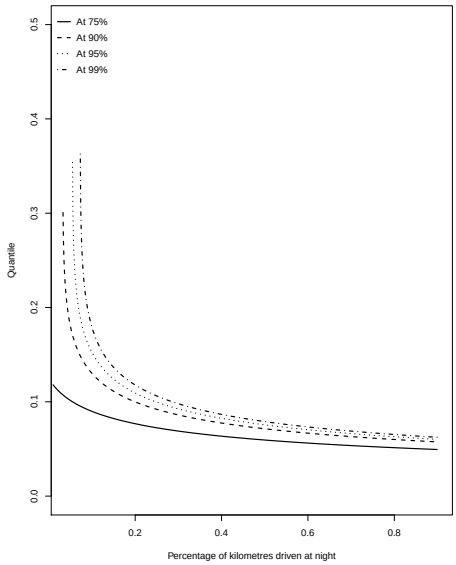

Figure 1. Quantiles of percentage of kilometres driven over the speed limit (Y_1) in the y-axis for Profile 1 given the values of percentage of kilometres driven at night (Y_2) in the x-axis.

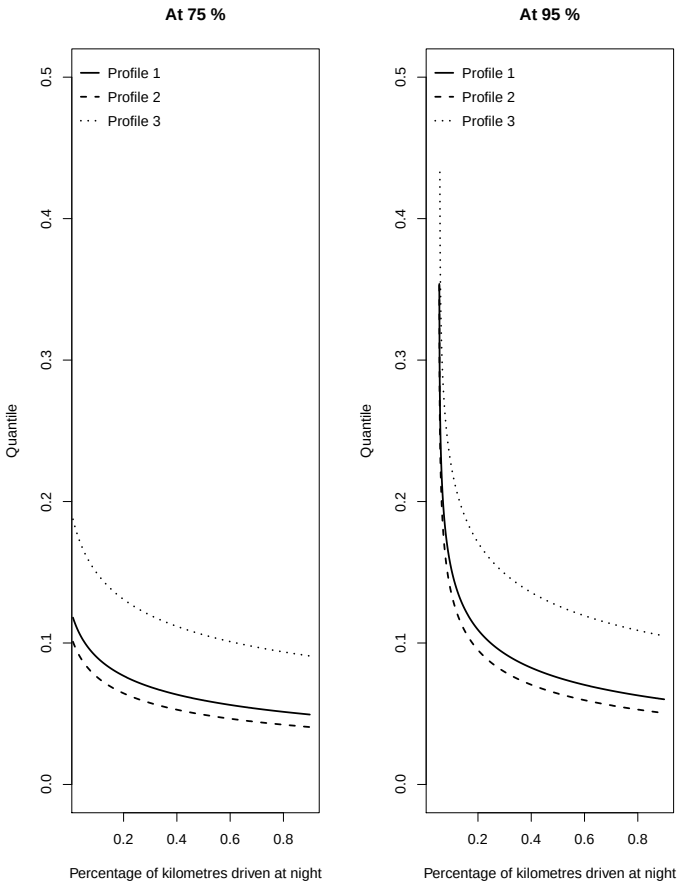

Figure 2. Quantiles of percentage of kilometres driven over the speed limit (Y_1) for each driver profile given the values of percentage of kilometres driven at night (Y_2), (**left**) 75% level and (**right**) 95% level.

4. Conclusions

We have developed a bivariate model based on the Sarmanov distribution with marginal Beta GLM which has allowed us to model two important variables in modern motor insurance telematics databases. Our model is an alternative to a proposal previously made by Gupta and Wong [3] based on what is known as Morgenstern's distribution, which is a particular case of the Sarmanov distribution. Our proposal allows for obtaining closed expressions for some magnitudes of interest, such as the bivariate cdf and conditioned moments, covariance and correlation, which are fundamental in risk analysis. We have shown that our Sarmanov-Beta-GLM model presents better fits than previous proposals also based on the Sarmanov distribution.

The results of our case study have shown that, for a specific example, although the dependence parameter is positive, which directly implies that, in the mean, the relationship between the conditioned mean and the values of the variable that conditions is positive, the conditional quantiles show that the relationship between the conditioned quantile, and the value of the conditioning variable may be negative for high quantile levels, a result that is consistent with the expected behaviour of drivers.

Author Contributions: Conceptualization, C.B. and M.G.; methodology, C.B.; software, A.P.; validation, C.B., M.G., and A.P.; formal analysis, A.P.; investigation, A.P.; resources, M.G.; data curation, M.G.; writing—original draft preparation, C.B.; writing—review and editing, M.G. and A.P.; visualization, C.B.; supervision, C.B.; project administration, M.G.; funding acquisition, M.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Spanish Ministry of Science and Innovation grant PID2019–105986GB-C21, Fundación BBVA Research on Big Data and ICREA Academia.

Acknowledgments: We thank seminar participants and members of the Riskcenter, Universitat de Barcelona.

Conflicts of Interest: The authors declare no conflict of interest.The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Ferrari, S.; Cribari-Neto, F. Beta regression for modelling rates and proportions. *J. Appl. Stat.* **2004**, *31*, 799–815. [CrossRef]
2. Arnold, B.; Ng, H. Flexible bivariate Beta distributions. *J. Multivar. Anal.* **2011**, *102*, 1194–1202. [CrossRef]
3. Gupta, A.; Wong, C. On three and five parameter bivariate beta distributions. *Metrika* **1985**, *32*, 85–91. [CrossRef]
4. Olkin, I.; Liu, R. A bivariate beta distribution. *Stat. Probab. Lett.* **2003**, *62*, 407–412. [CrossRef]
5. Olkin, I.; Trikalinos, T. Constructions for a bivariate beta distribution. *Stat. Probab. Lett.* **2015**, *96*, 54–60. [CrossRef]
6. Sarmanov, O. Generalized normal correlation and two-dimensional frechet classes. *Doclady Soviet Math.* **1966**, *168*, 596–599.
7. Lee, M. Properties and applications of the sarmanov family of bivariate distributions. *Commun. Stat. Theory Methods* **1996**, *25*, 1207–1222.
8. Bairamov, I.; Altinsoy, B.; Kerns, G. On generalized Sarmanov bivariate distributions. *TWMS J. Appl. Eng. Math.***2011**, *1*, 86–97.
9. Morgenstern, D. Einfache beispiele zweidimen-sionaler verteilungen. *Mitteilingsblatt Math. Stat.* **1956**, *8*, 234–235.
10. Bolancé, C.; Vernic, R. Multivariate count data generalized linear models: Three approaches based on the sarmanovdistribution. *Insur. Math. Econ.* **2019**, *85*, 89–103. [CrossRef]
11. Guillen, M.; Nielsen, J.; Ayuso, M.; Pérez-Marín, A. The use of telematics devices to improve automobile insurance rates. *Risk Anal.* **2019**, *39*, 662–672. [CrossRef] [PubMed]
12. Ayuso, M.; Guillen, M.; Nielsen, J. Improving automobile insurance ratemaking using telematics: Incorporating mileage and driver behaviour data. *Transportation* **2019**, *46*, 735–752. [CrossRef]
13. Pérez-Marín, A.; Guillen, M. Semi-autonomous vehicles: Usage-based data evidences of what could be expected from eliminating speed limit violations. *Accid. Anal. Prev.* **2019**, *123*, 99–106. [CrossRef] [PubMed]
14. Pérez-Marín, A.; Ayuso, M.; Guillen, M. Do young insured drivers slow down after suffering an accident? *Transp. Res. Part F Psychol. Behav.* **2019**, *62*, 690–699. [CrossRef]
15. Pérez-Marin, A.; Guillen, M.; Alcañiz, M.; Bermúdez, L. Quantile regression with telematics information to assess the risk of driving above the posted speed limit. *Risks* **2019**, *7*, 80. [CrossRef]
16. Pesantez-Narvaez, J.; Guillen, M.; Alcañiz, M. Predicting motor insurance claims using telematics data-xgboost versus logistic regression. *Risks* 2019, 7, 70. [CrossRef]
17. Sun, S.; Bi, J.; Guillen, M.; Pérez-Marín, A. Assessing driving risk using internet of vehicles data: An analysis based on generalized linear models. *Sensors* **2020**, *20*, 2712. [CrossRef] [PubMed]
18. Bahraoui, Z.; Bolancé, C.; Pelican, E.; Vernic, R. On the bivariate Sarmanov distribution and copula. An application on insurance data using truncated marginal distributions. *Stat. Oper. Res. Trans. SORT* **2015**, *39*, 209–230.
19. Joe, H.; Xu, J. The estimation method of inference functions for margins for multivariate models. *Open Collect.* **1996**. [CrossRef]
20. Smithson, M.; Verkuilen, J. A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychol. Methods* **2006**, *11*, 54–71. [CrossRef] [PubMed]