SORT 48 (2) July-December 2024, 1-24 DOI: 10.57645/20.8080.02.20

Conditional likelihood based inference on single-index models for motor insurance claim severity

Catalina Bolancé[1], Ricardo Cao[2] and Montserrat Guillen[1]

Abstract

Prediction of a traffic accident cost is one of the major problems in motor insurance. To identify the factors that influence costs is one of the main challenges of actuarial modelling. Telematics data about individual driving patterns could help calculating the expected claim severity in motor insurance. We propose using single-index models to assess the marginal effects of covariates on the claim severity conditional distribution. Thus, drivers with a claim cost distribution that has a long tail can be identified. These are risky drivers, who should pay a higher insurance premium and for whom preventative actions can be designed. A new kernel approach to estimate the covariance matrix of coefficients' estimator is outlined. Its statistical properties are described and an application to an innovative data set containing information on driving styles is presented. The method provides good results when the response variable is skewed.

MSC: *62G05, 62P20, 91G70.*

Keywords: *covariance matrix of estimator, kernel estimator, marginal effects, telematics covariates, right-skewed cost variable.*

1. Introduction

We analyse costs of claims in a motor insurance data set. Because higher costs occur much less frequently than lower costs of claims, the dependent variable here is right-skewed. Specifically, we are interested in modelling the distribution of costs of claims conditional on the values of covariates that reflect driving habits. We focus on the whole conditional distribution rather than on the conditional expectation to measure the influence of covariates on different quantiles, specifically on the costly claims, i.e., the right

[1] Department of Econometrics, RISKcenter-IREA, Universitat de Barcelona (UB).

[2] Research Group MODES, Department of Mathematics, CITIC, Universidade da Coruña and ITMATI.

Received: April 2023

Accepted: January 2024

tail of the severity distribution. This problem could be addressed by quantile regression, for fixed quantile levels, but this could potentially lead to contradictory results for close quantiles. Modelling the cost of claims conditional on covariate information has remained a bottleneck for insurance companies, as a result of which average costs are used in practice worldwide. We address this problem also considering data on driving patterns and driving conditions, a type of information that is available through sensor data regularly collected by insurtech firms. Some new motor insurance rate making schemes are based on near-miss telematics information which measures the propensity of risky events that do not always lead to an accident (see Guillen et al., 2019, 2020 and Guillen, Nielsen and Pérez-Marín, 2021). Risk scores such as the ones obtained with index-models can be combined with the evaluation of near-miss information to improve the performance of predictive modelling in motor insurance pricing.

Single-index regression models are semiparametric methods for generalising linear regression. They specify the dependence between a random variable Y (here the cost of a traffic accident, or claim severity) and a d-dimensional vector X as follows (see Härdle et al., 1993):

$$Y = m\left(\theta^\top X\right) + \varepsilon, \tag{1}$$

where θ is a vector of unknown parameters, m is an unknown smooth function, and ε is a random variable with zero-mean conditional on X.

Traditional approaches for estimating the linear predictor coefficients θ and the function m are based on the conditional expectation rather than on the whole conditional distribution and, as a consequence, they are vulnerable to the presence of extremes, heavy tails or strong asymmetry, as in many applications. Our contribution is to extend the maximum likelihood estimation of (1) and, in so doing, to open the door to single-index conditional distribution modelling which has enormous potential for a range of applications.

In order to estimate the vector θ, Härdle, Hall and Ichimura (1993) proposed the direct minimisation of the residual sum of squares, so their estimator is

$$\hat{\theta} = \arg\min_\theta \sum_{i=1}^{n} \left[Y_i - \hat{m}_i\left(\theta^\top X_i\right)\right]^2,$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are iid observations of the covariates and the dependent variable and $\hat{m}_i$ indicates the leave-one-out kernel estimator of m. Alternatively, Hristache, Juditsky and Spokoiny (2001) analysed the average derivative estimator of the vector of parameters in the index model, introduced by Stoker (1986) and as subsequently employed by Powell, Stock and Stoker (1989). Hristache et al. (2001) presented the method for estimating the vector of coefficients, θ, by minimising an $M-$function, with a score function ψ, that again compares Y_i with a nonparametric estimator $\hat{m}(\cdot)$, i.e., $\arg\min_\theta \sum_{i=1}^{n} \psi\left[Y_i, \hat{m}\left(\theta^\top X_i\right)\right]$. All these methods ignore the shape of the conditional distribution because they are based on fitting the conditional expectation.

Delecroix, Härdle and Hristache (2003) investigated the pseudo-maximum likelihood estimation of θ in (1). They proposed starting from a preliminary $\sqrt{n}$-consistent

estimator and, subsequently, correcting it with the gradient and the Hessian of the log-likelihood function. They showed that the corrected estimator is efficient. Previously, Klein and Spady (1993) had analysed the maximum likelihood estimation of θ but only for a binary response dependent variable. In the context of survival data with censored observations, Strzalkowska-Kominiak and Cao (2013) investigated maximum likelihood alternatives based on the kernel estimation of the conditional distribution and showed that previous methods for censored data could be improved.

Nonparametric regression is more general than the single-index model specified in (1). Indeed, it emanates from a more general specification $Y = m(X) + \varepsilon$, where the aim is to estimate the regression curve $m(x) = E(Y|X = x)$; Härdle (1990). However, in practice, nonparametric regression presents two considerable challenges. First, estimation becomes increasingly difficult as the number of covariates rises (the curse of dimensionality). The second challenge is that any interpretation of the effects of the explanatory variables cannot be carried out directly and it is necessary to plot the different relations to explore these effects. Another alternative to the single-index model is the generalised additive model (see Hastie and Tibshirani, 1990); however, it faces the same challenges as those described for nonparametric regression.

Here, a new maximum likelihood estimator of θ in (1) is proposed, inspired by the work of Strzalkowska-Kominiak and Cao (2013) with right-censored data. As these authors proposed we use two different smoothing parameters: one associated with the distribution of Y and the other one associated with the distribution of the index $\theta^\top X$. The new theoretical results that we present in Section 2 for uncensored data do not follow directly as a particular case of Strzalkowska-Kominiak and Cao (2013), since some assumptions of the censored data case can be relaxed or dropped. In this paper, we deduce the covariance matrix that can be easily estimated using a kernel estimator. We evaluate the inference power of the statistical test for the covariate effects deduced from our maximum likelihood estimator. Details on the method, some results of the simulation study and proofs are available in the Supplementary Material.

We show the superiority of our estimator, in particular, when there are extreme values, like in our application where we observe only a few severe accidents. Additionally, we show that the results of the estimated index model are easily interpretable from different points of view, for example, for the prediction of conditional mean, quantiles and marginal effects.

We analyse a data set obtained from a specific portfolio from an insurance company in Spain. The portfolio is made up of a small group of policyholders under 35 years of age, who have underwritten a new insurance contract that requires a telematics device to be installed in their vehicle. The data set contains information on mean yearly claim cost per policy and on telematic and non-telematic characteristics. Our aim is to find the influence of telematic information on pricing compared to a traditional approach with only classical non-telematic variables. The data set is available at http://www.ub.edu/rfa/R/SORT-BCG/. We observe how the mean yearly claim cost per policy does not change with a linear index; however, the shape of the distribution depends on a linear index, something that could be considered when calculating the premium.

In a simulation study presented in Section 3, the finite-sample properties of our proposal are compared with several alternative methods for different distributions with heterogeneity in the location and in the scale parameters. We also carry out basic inference about the estimators. In addition, we evaluate how the results are affected when the covariates are correlated and binary explanatory variables are included. Note that Hall and Yao (2005) and Newey and Stoker (1993) only consider continuous covariates; indeed, not many papers to date have dealt with discrete covariates in single-index models. One exception is Horowitz and Härdle (1996), who focused on analysing a direct estimator for the effect of the discrete covariates. Elsewhere, methods such as those proposed by Härdle et al. (1993), Hristache et al. (2001) and Delecroix et al. (2003), while allowing dummy (binary) variables to be incorporated, do not consider the consequences of their inclusion.

2. Methods

Let us denote the vector of covariates $X = (X_1, \ldots, X_d)^\top$ and let $f(\cdot|\mathbf{x})$ be the density function of Y given $X = \mathbf{x}$, where $\mathbf{x} = (x_1, \ldots, x_d)$ is a fixed vector where $f(y|\mathbf{x}) = f_{\theta_0}(y|\theta_0^\top \mathbf{x})$, where $f_{\theta_0}(\cdot|\theta_0^\top \mathbf{x})$ is the conditional density of Y given $\theta_0^\top X = \theta_0^\top \mathbf{x}$ and θ_0 is the parameter vector to be estimated. Furthermore, we assume that $F(y|\mathbf{x}) = F_{\theta_0}(y|\theta_0^\top \mathbf{x})$ is its conditional cumulative distribution function. For any θ_0 and any nonzero real number λ, then vector θ_0 can be replaced by $\lambda\theta_0$. This means that the conditional distribution of the response given $X = \mathbf{x}$ only depends on this covariate vector via the linear combination $t = \theta_0^\top \mathbf{x}$. If we choose any nonzero real number λ, then, since there is a one-to-one correspondence between t and λt, it is also true that the conditional distribution only depends on the covariate vector via the linear combination $\lambda\theta_0^\top \mathbf{x}$. Consequently, infinitely multiple choices exist for the single-index parameter vector θ_0. The usual way to solve this identification problem is to introduce a scale constraint, for example $||\theta_0|| = 1$ or fixing one component of θ_0 to be equal to one. In practice, the identification problem implies that the signs of the effects of the covariates on the dependent variable are not identified but are comparable, i.e., two parameters with different sign indicate opposite effects and, if variables are measured in the same scale, then their corresponding parameter estimates can be compared directly.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of the dependent variable and the covariates, where $X_i = (X_{i1}, \ldots, X_{id})^\top$ and it is assumed that at least one covariate is continuous. Let K be a nonnegative kernel and h_1, h_2 two positive bandwidths. In line with Bashtannyk and Hyndman (2001), the kernel conditional density estimator is:

$$\hat{f}_\theta(y|t) = \frac{\hat{r}(t,y)}{\hat{s}(t)}, \tag{2}$$

where

$$\hat{s}(t) = \hat{s}_{h_1}(t) = \frac{1}{nh_1} \sum_{i=1}^{n} K\left(\frac{t - \theta^\top X_i}{h_1}\right) \tag{3}$$

and the product bivariate kernel density estimator is used for $\hat{r}(t,y)$; see Chapter 6 of Scott (2015). The product kernel is just a simple way to smooth using multiplicative weights, so:

$$\hat{r}(t,y) = \hat{r}_{h_1,h_2}(t,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) K\left(\frac{y-Y_i}{h_2}\right). \tag{4}$$

We use a Gaussian kernel, and the smoothing parameters are calculated using alternative criteria considering the estimator type, i.e., the parameter vector, the conditional density, the conditional distribution or the conditional mean.

In line with Hall, Wolff and Yao (1999), the kernel estimator of the conditional distribution function is:

$$\widehat{F}_\theta(y|t) = \frac{\widehat{R}(t,y)}{\hat{s}(t)},$$

where

$$\widehat{R}(t,y) = \widehat{R}_{h_1,h_2}(t,y) = \frac{1}{nh_1}\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) \mathbf{K}\left(\frac{y-Y_i}{h_2}\right)$$

and $\mathbf{K}$ is the kernel distribution function.

2.1. Maximum conditional likelihood estimation

If we know F_θ except for the value of the index vector θ (a highly unrealistic assumption), then we can define the following theoretical conditional likelihood function:

$$\tilde{L}_n(\theta) = \prod_{i=1}^{n} f_\theta(Y_i|\theta^\top X_i).$$

Maximising this function is equivalent to maximising its logarithm:

$$\tilde{\ell}_n(\theta) = \frac{1}{n}\log\left(\tilde{L}_n(\theta)\right) = \frac{1}{n}\sum_{i=1}^{n}\log f_\theta(Y_i|\theta^\top X_i). \tag{5}$$

Here, the ideal estimator should maximise the theoretical log-likelihood

$$\tilde{\theta}_n = \arg\max_\theta \tilde{\ell}_n(\theta).$$

In practice, f_θ (or F_θ) is unknown and so, we need to estimate it and plug it into the logarithm of the theoretical conditional likelihood function.

We propose to maximise the kernel estimation of the log-likelihood function defined in (5) with respect to θ and to the two smoothing parameters, h_1 and h_2. At this point, we note that, in the kernel estimation, when a smoothing parameter selector is obtained by optimising some criteria, such as the integrated square error or the likelihood function, which required computing a kernel estimator; using the whole observed data

set, $(X_1,Y_1),\ldots,(X_n,Y_n)$, produces undersmoothing of the optimal smoothing parameter values; see Silverman (1986). As a consequence, we need to modify the estimated likelihood with a leaving-one-out procedure so as not to pick artificially small bandwidths. Let $\hat{f}_\theta^{-i}(Y_i|\theta^\top X_i)$ be the estimator defined in (2), where the sum in (3) and (4) runs over $j \neq i$. Then, we define the leaving-one-out estimated conditional log-likelihood:

$$\hat{\ell}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log \hat{f}_\theta^{-i}(Y_i|\theta^\top X_i). \tag{6}$$

Given h_1 and h_2, the final maximum conditional likelihood estimator is defined as

$$\hat{\theta}_n = \arg\max_\theta \hat{\ell}_n(\theta).$$

The estimation procedure including the two smoothing parameters h_1 and h_2 will be described in sub-section 2.3. A similar procedure based on the leave-one-out estimator of the hazard rate model was proposed by van den Berg et al. (2021). We point out that it can be difficult to avoid local optima in the maximisation of the log-likelihood in (6). Considering the existence of local optima, in the described estimation procedure we checked how initial values for the smoothing parameters affect the final estimation. We have observed that the final estimation is practically not affected by the initial values of the covariate coefficients.

2.2. Properties

In this sub-section we study the properties of $\hat{\theta}_n$. Let the score function be defined as the expected log-likelihood:

$$\ell(\theta) = E(\tilde{\ell}_n(\theta)).$$

We start by proving that the true parameter vector, θ_0, can be characterised as the maximiser of the score function. The existence of that function is the only condition required:

A1: $E(\log f_\theta(Y_i|\theta^\top X_i)) < \infty$ for any θ

Theorem 1. *The true single-index parameter, θ_0, is the maximiser of the score function, i.e., $\theta_0 = \arg\max_\theta \ell(\theta)$.*

To establish the main results for the estimator, we need to assume some further conditions:

A2: $E(X|\theta_0^\top X, Y) = E(X|\theta_0^\top X)$

A3: $E(XX^\top) < \infty$ componentwise.

Condition A2 is a technical one needed to prove our theoretical results. It essentially means that all the information needed to predict the values of the explanatory variables

given the index and the response variable is contained just in the index. Assumption A2 also implies exogeneity of the explanatory variables, i.e., covariates are known previous to the response.

The two bandwidths h_1, h_2 should fulfill the following conditions

A4: $\sqrt{n}h_1^4 \to 0$, $\sqrt{n}h_2^2 \to 0$, $nh_1^6 \to \infty$ and $h_1, h_2 \to 0$ when $n \to \infty$.

Consider f_{θ_0} the bivariate joint density function of $(\theta_0^\top X, Y)$ and $f_{\theta_0^\top X}$ the marginal density function of $\theta_0^\top X$. Finally, let $\ell^{[1]}(\theta_0) = \nabla_\theta \ell(\theta)|_{\theta=\theta_0}$ denote the gradient of $\ell(\theta)$ over θ evaluated in θ_0. Further, let $\ell^{[2]}(\theta)$ denote the Hessian matrix of $\ell(\theta)$. The following regularity conditions are also assumed.

A5: The derivatives $\frac{\partial^j}{\partial^j u}\frac{\partial^k}{\partial^k v} f_{\theta_0}(u,v)$, $\frac{d^j}{d^j u} f_{\theta_0^\top X}(u)$ and $\frac{d^j}{d^j u} E(X|\theta_0^\top X = u)$ exist for $j = 1,2,3$ and $k = 1,2$.

A6: The function $h(\mathbf{x}, y) = \frac{\partial}{\partial \theta_j} f_\theta(\theta^\top \mathbf{x}, y)_{\theta=\theta_0}$ is continuous and $\frac{\partial^2}{\partial^2 \theta_j} f_\theta(\theta_0^\top \mathbf{x}, y)_{\theta=\theta_0}$ exists.

A7: The Hessian matrix $\ell^{[2]}(\theta^*)$ is positive definite for θ^* belonging to a neighbourhood of θ_0.

Now we can state the first result for the proposed estimator.

Lemma 1. *Under A1, A4 and A6 we have* $\hat{\theta}_n - \theta_0 = -\left[\hat{\ell}_n^{[2]}(\hat{\theta}_n^*)\right]^{-1} (\hat{\ell}_n^{[1]}(\theta_0) - \ell^{[1]}(\theta_0))$, *where* $\hat{\theta}_n^*$ *is between* $\hat{\theta}_n$ *and* θ_0.

Theorem 2. *Under A1-A7, we have* $\hat{\theta}_n \to \theta_0$ *in probability.*

Theorem 3. *Let us assume conditions A1-A7. Then, we have*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to \mathcal{N}(0, \Sigma), \tag{7}$$

where

$$\Sigma = \Sigma_2 \Sigma_1 \Sigma_2^\top, \tag{8}$$
$$\Sigma_2 = \left[\ell^{[2]}(\theta_0)\right]^{-1}$$

and

$$\Sigma_1 = E\left[\nabla_\theta \log(f_\theta(Y|\theta^\top X))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(Y|\theta^\top X))_{\theta=\theta_0})^\top\right]$$
$$= \int (\nabla_\theta \log(f_\theta(y|\theta^\top \mathbf{x}))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(y|\theta^\top \mathbf{x}))_{\theta=\theta_0})^\top f(\mathbf{x}, y) d\mathbf{x} dy.$$

All the proofs can be found in the Supplementary Material.

The asymptotic variance-covariance matrix in (8) is different from the one obtained by Delecroix et al. (2003). These authors obtained this matrix from $\tilde{\ell}_n(\theta)$ defined in (5)

and took into account the almost sure convergence of the parameter estimator and the weak convergence of $\hat{\ell}_n(\theta)$, defined in (6), and some of its partial derivatives. Instead, to obtain the asymptotic variance-covariance matrix, we take into account that θ_0 is estimated by maximising the kernel estimator of the conditional likelihood function $\hat{\ell}_n(\theta)$ defined in (6).

2.3. Estimation procedure

To obtain $\hat{\theta}_n$, h_1 and h_2 we have used an algorithm in two steps. The first step aims to obtain $\hat{\theta}_n$ by maximising the likelihood function in (6) given fixed values for the smoothing parameters h_1 and h_2. In the second step the smoothing parameters are recalculated by maximising the same likelihood function given the values of $\hat{\theta}_n$ obtained in the previous step. Both steps are repeated until convergence. In the first step the initial values of the smoothing parameters are given by $h_1 = a\hat{\sigma}_{\theta^\top X} n^{-2/13}$ and $h_2 = a\hat{\sigma}_Y n^{-4/13}$, where $a > 0$ and $\hat{\sigma}_{\theta^\top X}$ and $\hat{\sigma}_Y$ are the empirical standard deviations (see Silverman (1986) for rule-of-thumb smoothing parameters in kernel density estimation). The sample size orders, $n^{-2/13}$ and $n^{-4/13}$, respectively for the two bandwidths, are chosen in order to fulfill the asymptotic assumptions for the bandwidths needed for Condition A4. We have observed that initial values of the smoothing parameters considerably affect the final estimation. Initially we used $a = 1$ but it is recommended to consider a grid of values around 1. The initial values of the covariate coefficients hardly affect the results, so to start the algorithm we set all these coefficients equal to 1. To maximise the likelihood function in the first step, we use the function "optim()" with the default optimization method ("Nelder-Mead") of the "stats" R package. In the second step, to recalculate the values h_1 and h_2 we also use function "optim()" but with optimization method "L-BFGS-B". We need to define limits for the smoothing parameters because it is known that $\hat{\ell}_n(\theta) \to \infty$ as $h_1, h_2 \to 0$. The limits are defined as $(c_1^{(1)}\hat{\sigma}_{\theta^\top X} n^{-2/13}, c_2^{(1)}\hat{\sigma}_{\theta^\top X} n^{-2/13})$ for h_1 and $(c_1^{(2)}\hat{\sigma}_Y n^{-4/13}, c_2^{(2)}\hat{\sigma}_Y n^{-4/13})$ for h_2, for some $c_1^{(j)} < c_2^{(j)}$, $j = 1, 2$.

Our two-step algorithm is designed to guarantee the conditions established in the theoretical properties shown in the previous sub-section. In practice, we are selecting the best estimation in a set of pre-fixed smoothing parameters which are calculated taking into account the sample size and the scale of the dependent variable and the index.

To estimate the variance-covariance matrix in (8) we calculate the corresponding derivatives of the leave-one-out kernel estimation of conditional log-likelihood defined in (6). Asymptotic normality inference, based on (7), is carried out using the estimated variance-covariance matrix, replacing theoretical derivatives by estimated ones (kernel estimator of the gradient $\nabla_\theta \log(f_\theta(y|\theta^\top \mathbf{x}))_{\theta=\theta_0}$ is direct). For kernel estimator of $\ell^{[2]}(\theta_0)$ see Lemma 9 in the Supplementary Material.

2.4. Marginal effects estimation

For a given $\theta = \theta_0$, using the conditional distribution function we can obtain the p-th conditional quantile: $Q_\theta(p|\theta^\top \mathbf{x}) = F_\theta^{-1}(p|\theta^\top \mathbf{x})$, i.e., $F_\theta(y_p|\theta^\top \mathbf{x}) = p$ where $p \in$

$(0,1)$. As in any generalised linear model, comparing marginal effects is equivalent to comparing parameters, i.e., for two covariates X_k and $X_{k'}$, with $k \neq k'$, we obtain:

$$\frac{\frac{\partial Q_\theta(p|\theta^\top \mathbf{x})}{\partial x_k}}{\frac{\partial Q_\theta(p|\theta^\top \mathbf{x})}{\partial x_{k'}}} = \frac{\theta_k}{\theta_{k'}},$$

where:

$$\frac{\partial Q_\theta(p|\theta^\top \mathbf{x})}{\partial x_k} = -\frac{\frac{\partial F_\theta(Q_\theta(p|\theta^\top \mathbf{x})|t)}{\partial t}_{t=\theta^\top \mathbf{x}} \cdot \theta_k}{f_\theta(Q_\theta(p|\theta^\top \mathbf{x})|\theta^\top \mathbf{x})}. \tag{9}$$

For estimating the marginal effects we will use kernel estimators for $f_\theta(y|\theta^\top \mathbf{x})$, $F_\theta(y|\theta^\top \mathbf{x})$ and their derivatives, as shown below.

The kernel estimator of the index marginal effects on the conditional distribution function is:

$$\frac{\partial \widehat{F}_\theta(y|t=\theta^\top \mathbf{x})}{\partial t} = \left[\frac{\hat{R}'_{h_1,h_2}(\theta^\top \mathbf{x}, y)}{\hat{s}_{h_1}(\theta^\top \mathbf{x})} - \widehat{F}_\theta(y|\theta^\top \mathbf{x})\frac{\hat{s}'_{h_1}(\theta^\top \mathbf{x})}{\hat{s}_{h_1}(\theta^\top \mathbf{x})}\right],$$

where

$$\hat{R}'_{h_1,h_2}(t,y) = \frac{1}{nh_1^2h_2}\sum_{i=1}^{n} K'\left(\frac{t-\theta^\top X_i}{h_1}\right)\mathbf{K}\left(\frac{y-Y_i}{h_2}\right)$$

and

$$\hat{s}'_{h_1}(t) = \frac{1}{nh_1^2}\sum_{i=1}^{n} K'\left(\frac{t-\theta^\top X_i}{h_1}\right),$$

where K' is the first derivative of the kernel.

In this paper, we obtained the marginal effects using kernels estimators of the different functions that appear in the expression (9). The smoothing parameters of the kernel estimator of conditional density can be calculated using the sample size orders of reference rules obtained in Bashtannyk and Hyndman (2001). The kernel estimator of the conditional distribution and its derivatives are obtained directly from the estimated conditional density. Considering that in this paper the aim of estimating marginal effects is purely descriptive, we have obtained the values of smoothing parameters subjectively from graphic visualization. However, a double-cross-validation approach as suggested van den Berg et al. (2021) can be used.

2.5. Scoring rules for prediction

To evaluate the goodness of fit and the predictive capacity of the single-index model, a variety of measures is available. Gneiting and Raftery (2007) present an exhaustive review of different families of scoring rules for moments, density and distributional forecasts. We use three types of score described in Gneiting and Raftery (2007).

The predictive model choice criterion (PMCC) selects the best model based on the first two moments of the predicted values, i.e., the mean and the variance, as follows

$$PMCC = -\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{m}\left(\theta^\top X_i\right)\right]^2 - \hat{\sigma}^2\left(\theta^\top X_i\right), \tag{10}$$

where $\hat{m}\left(\theta^\top X_i\right)$ is the kernel estimator of the conditional expectation $E\left(Y_i|\theta^\top X_i\right)$ and $\hat{\sigma}^2\left(\theta^\top X_i\right)$ is estimated with the kernel estimates of both expectations as follows:

$$\hat{\sigma}^2\left(\theta^\top X_i\right) = \hat{E}\left(Y_i^2|\theta^\top X_i\right) - \left[\hat{E}\left(Y_i|\theta^\top X_i\right)\right]^2,$$

where

$$\hat{E}\left(Y_i|\theta^\top X_i\right) = \hat{m}\left(\theta^\top X_i\right) = \frac{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right)}$$

and

$$\hat{E}\left(Y_i^2|\theta^\top X_i\right) = \frac{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right) Y_i^2}{\sum_{i=1}^{n} K\left(\frac{t-\theta^\top X_i}{h_1}\right)}.$$

Here h_1 is calculated using the optimal sample size order ($n^{-1/5}$) to estimate the conditional expectation and considering the scale of the dependent variables.

The logarithmic scoring rule is calculated as

$$\hat{\ell}(\theta) = \sum_{i=1}^{n} \log\left[\hat{f}\left(Y_i|\theta^\top X_i\right)\right]. \tag{11}$$

From $\hat{l}(\theta)$ other widely used criteria such as the AIC (Akaike Information Criterion) and the BIC (Bayesian Information Criterion) can be obtained.

For the p-quantile prediction of the dependent variable, Y, the goodness of fit criterion proposed by Koenker and Bassett (1978) for quantile regression is:

$$\begin{aligned} QE_p(\theta) = \frac{1}{n} &\sum_{i=1, Y_i > \hat{Q}_\theta(p|\theta^\top X_i)}^{n} p\ \left|Y_i - \hat{Q}_\theta(p|\theta^\top X_i)\right| \\ + \frac{1}{n} &\sum_{i=1, Y_i \leq \hat{Q}_\theta(p|\theta^\top X_i)}^{n} (1-p)\ \left|Y_i - \hat{Q}_\theta(p|\theta^\top X_i)\right|\ , \end{aligned} \tag{12}$$

where $\hat{Q}_\theta(p|\theta^\top X_i)$ is the kernel conditional quantile estimator based on the kernel estimator of the conditional distribution function. For a set of probabilities $p_1, \ldots, p_k$, we define $QE = \frac{1}{k}\sum_{j=1}^{k} QE_{p_j}(\theta)$ and its corresponding weighted version, $WQE = \frac{1}{k}\sum_{j=1}^{k} p_j QE_{p_j}(\theta)$.

3. Simulation study

We carry out a simulation study, the aim being to evaluate the finite-sample properties of our estimator. The properties of the parameter estimator, $\hat{\theta}$, are summarised in the Supplementary Material and the basic inferences about the value of these parameters are presented in this section. The results are obtained using a Gaussian kernel.

We compare the variance, the bias and the mean square error (MSE) of the estimated parameters in the vector $\hat{\theta}$, using our flexible maximum conditional likelihood (FMCL) estimator and three alternatives. The first is based on fitting the single-index model to individual conditional expected values as proposed by Härdle et al. (1993) (hereinafter, HHI). The second alternative is based on Delecroix et al. (2003) (hereinafter, DHH), where we use as our initial parameters those obtained with the HHI method which are $\sqrt{n}$-consistent. The third is the direct method proposed by Hristache et al. (2001) (hereinafter, HJS).

We analyse six different conditional distributions for the dependent variable Y, two symmetric distributions (zero skewness) and four right-skewed distributions. The conditional distributions are shown in Table 1.

Table 1. *Conditional distributions for dependent variable as a function of the linear index $\theta^\top \mathbf{x}$ for the simulation study.*

Skewness	Distribution	Parameters	Density
Zero	normal	$(\mu = \theta^\top \mathbf{x}, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{\sqrt{2\pi\lvert\theta^\top \mathbf{x}\rvert^2}} \exp\left(-\frac{(y-\theta^\top \mathbf{x})^2}{2\lvert\theta^\top \mathbf{x}\rvert^2}\right)$
	logistic	$(\mu = \theta^\top \mathbf{x}, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{\lvert\theta^\top \mathbf{x}\rvert} \frac{\exp\left(\frac{(y-\theta^\top \mathbf{x})}{\lvert\theta^\top \mathbf{x}\rvert}\right)}{1+\exp\left(\frac{(y-\theta^\top \mathbf{x})}{\lvert\theta^\top \mathbf{x}\rvert}\right)}$
Positive	lognormal	$(\mu = \theta^\top \mathbf{x}, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{y\sqrt{2\pi\lvert\theta^\top \mathbf{x}\rvert^2}} \exp\left(-\frac{(ln(y)-\theta^\top \mathbf{x})^2}{2\lvert\theta^\top \mathbf{x}\rvert^2}\right)$
	Weibull	$(\alpha = 1, \sigma = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{1}{\lvert\theta^\top \mathbf{x}\rvert} \exp\left(-\frac{y}{\lvert\theta^\top \mathbf{x}\rvert}\right)$
	Champernowne	$(\alpha = 1, M = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{\lvert\theta^\top \mathbf{x}\rvert}{(y+\lvert\theta^\top \mathbf{x}\rvert)^2}$
		$(\alpha = 2, M = \lvert\theta^\top \mathbf{x}\rvert)$	$\frac{2\lvert\theta^\top \mathbf{x}\rvert^2 y}{(y^2+\lvert\theta^\top \mathbf{x}\rvert^2)^2}$

For our two choices of symmetric distribution, the logistic distribution has more kurtosis and heavier tails than the normal distribution. If we compare our selection of right-skewed distributions, we find that the Champernowne or log-logistic has a heavier tail than the lognormal and the Weibull; see Buch-Larsen et al. (2005) for a description of the Champernowne distribution.

In our simulation study, we use six vectors of covariates X that we identify as vectors V1, V2, V3, V4, V5 and V6. For the first three $\theta^\top = (1, 1.3, 0.5)$ and for the fourth $\theta^\top = (1, 1.3, 0.5, 0.8)$. The values in vector V1 are generated from three uncorrelated standard normal distributions. The vectors V2 and V3 are trivariate normal distributions with correlated marginals. For V2 the components are three standard normal distributions whose covariances are $cov(X_k, X_{k'}) = 0.3$ for $k \neq k'$ and $k, k' = 1, 2, 3$. The same holds for V3 but with covariances $cov(X_1, X_2) = cov(X_2, X_3) = 0.7$ and $cov(X_1, X_3) = 0.5$. Vector V4 consists of V1 and a binary variable whose values are generated from a Bernoulli distribution with probability 0.4, independent of the three components of V1. Furthermore, the number of categorical covariates is usually greater than one. We have carried out an alternative simulation study using two new vectors of covariates V5 and V6, with $\theta^\top = (1, 1.3, 0.5, 0.8)$. Vector V5 consists of two independent standard normal variables and two binary variables whose values are generated from two Bernoulli distributions with probabilities 0.4 and 0.7, respectively. The covariate vector V6 includes the same two binary variables, one lognormal with mean 0 and σ equal to 0.5 and one with a standard normal distribution.

We generate 500 samples of size $n = 100$, 500 and $2,000$ and calculate the bias, the standard deviation (STD) and the MSE of the estimators using each method, FMCL, HHI, DHH and HJS. The results of the simulation study show that the proposed FMCL estimator is the most suitable when the conditional distribution is right-skewed and also when the tail of the conditional distribution is heavy. Moreover, the FMCL is more robust to multicollinearity and to the presence of binary and asymmetric covariates.

3.1. Basic inference

Power analysis of hypothesis tests is fundamental to determining whether the effect of a covariate is significantly different from zero. The null hypothesis for each parameter is $H_0: \ \theta_k = 0, \ k = 1, \ldots, d$ and as an alternative hypothesis we assume that the sign of the parameter is known, i.e., $H_1: \ \theta_k > 0, \ k = 1, \ldots, d$. The statistic test is $Z = \hat{\theta}_j / se(\hat{\theta}_j)$, where *se* indicates the standard error. The statistic Z asymptotically follows a $N(0,1)$ distribution. To obtain the power of the test we calculate the proportion of times that we reject the null hypothesis in the 500 samples obtained from each analysed conditional distribution and sample size. Alternatively, we also analyse the power of the test when the null hypothesis is $H_0: \ \theta_2 = \theta_3$ and the alternative hypothesis $H_1: \ \theta_2 > \theta_3$. Again, we know that the alternative hypothesis is true. The statistic for this test is $Z = (\hat{\theta}_2 - \hat{\theta}_3) / se(\hat{\theta}_2 - \hat{\theta}_3)$.

Table 2. *Power of the test for skewed distributions. The values are calculated using the 500 samples for each skewed distribution in Table 1.*

		Lognormal		Weibull		Champernowne $\alpha=1$		Champernowne $\alpha=2$	
	H_0	$n=500$	$n=2,000$	$n=500$	$n=2,000$	$n=500$	$n=2,000$	$n=500$	$n=2,000$
V1	$\theta_2=0$	1.000	1.000	0.864	0.996	0.722	0.970	0.984	0.998
	$\theta_3=0$	1.000	1.000	0.876	0.998	0.702	0.972	0.992	1.000
V4	$\theta_2=0$	1.000	1.000	0.856	1.000	0.636	0.908	1.000	1.000
	$\theta_3=0$	1.000	1.000	0.828	1.000	0.622	0.902	1.000	1.000
	$\theta_4=0$	1.000	1.000	0.770	0.984	0.584	0.862	0.996	1.000
V1	$\theta_2=\theta_3$	1.000	1.000	0.882	0.996	0.730	0.976	0.988	1.000
V4	$\theta_2=\theta_3$	1.000	1.000	0.662	1.000	0.598	0.880	0.998	1.000

Table 3. *Percent of no-rejection of null hypothesis. The values are calculated using the 200 samples for each distribution in Table 1.*

		Normal		Logistic		Lognormal	
	H_0	$n=500$	$n=2000$	$n=500$	$n=2000$	$n=500$	$n=2000$
V1	$\theta_4=0$	0.848	0.955	0.942	0.985	0.696	0.850
V4	$\theta_5=0$	0.828	0.980	0.992	0.980	0.345	0.890
		Weibull		Champernowne $\alpha=1$		Champernowne $\alpha=2$	
	H_0	$n=500$	$n=2000$	$n=500$	$n=2000$	$n=500$	$n=2000$
V1	$\theta_4=0$	0.850	0.965	0.530	0.570	0.752	0.720
V4	$\theta_5=0$	0.924	0.965	0.478	0.795	0.720	0.770

The results for symmetric distributions have a power about 100% for almost all tests when $n \geq 500$, these results are shown in the Supplementary Material. Here we focus on the results for the power of tests for skewed distributions.

Table 2 shows the powers of the two tests proposed for skewed distributions. Both tests are at the 95% confidence level. These results indicate that when $n = 500$ the power decreases considerably for the Weibull and the Champernowne distribution with $\alpha = 1$, compared to a larger sample size, $n = 2,000$.

To analyse the percent of times the null hypothesis that the parameter is equal to zero is not rejected, we have designed an alternative reduced simulation study that consists of adding a new covariate with associated parameter equal zero in the estimation procedure; this implies to re-estimate the parameters. To reduce the computation time, instead of 500 replicates, we use 200 replicates of sizes $n = 500$ and $n = 2,000$. The null hypothesis is $H_0 : \theta_j = 0$, $j = 4,5$, and the results of the percent of no-rejection of the null hypothesis, for the models described in Table 1 and using extended covariate vectors, are shown in Table 3. For $n = 500$ the results for skewed distributions are poorer than those obtained for a symmetric distributions. For $n = 2,000$, in general, the results improve compared to a smaller sample size, except for the Champernowne distribution, which is heavy tailed. These results suggest that if the dependent variable is asymmetric, a transformation to achieve a symmetric distribution should be suitable.

4. Data analysis and model estimations of automobile claim costs

In this section we analyse the effect of risk factors on the distribution of the cost per automobile claim in a real case study. We show that single-index models constitute a new tool for identifying the influence of some of those covariates that are known to the insurer at the beginning of the contract or during the coverage period. We estimate the single-index model coefficients with the FMLC method. The results are obtained using a Gaussian kernel. Some parametric models based on Weibull, gamma, log-normal and log-logistic distributions, which are not reported here, produced poor fits. Furthermore, significant effects of the covariates were not found.

We analyse a data set obtained from a Spanish insurance company. The original portfolio consists of policyholders between age 18 and 35, who underwrote a motor insurance policy and accepted a telematics engine that allows the company to gather data on the policyholder's driving behaviour. In the available data set, all claims are settled. In the original data set, a few claims result from no fault agreements between insurers, in these cases the amount recorded is equal to the legally established cost. Claims regulated by a no-fault agreement were excluded from our analysis. Hence, our data are not censored. Those in the no-fault agreement had to be removed because there was no information on the true cost of the claim, which could be lower or higher than the amount established by the agreement. To estimate the proposed single-index model, we have selected a sample of $n = 489$ car insurance policyholders who reported at least one claim in 2011. Furthermore, we have also selected another sample of 100 policyholders to carry out a predictive analysis. The claims correspond to third-party liability accident. For each policyholder in the sample, the total incurred losses and the number of claims along the year is known, the ratio between both values is equal to the yearly mean claim cost per policy. The cost refers to incurred and paid losses.

For each policyholder, we have information about the following covariates (labels in parentheses): cost per policyholder in thousands of euros (cost), age in years (age), number of years holding a driving licence (agelic), age of car in years (agecar), a binary indicator equal to 1 if car is parked in a garage overnight and 0 otherwise (parking), annual distance driven in thousands of kilometres (tkm), percentage of kilometres driven at night (nightkm), percentage of kilometres driven on urban roads (urbankm) and percentage of kilometres driven above the speed limit (speedkm). These data correspond to a sample of insureds for whom the company collected driving behaviour information employing a telematics device installed in their vehicle. Thus, "tkm", "urbankm", "nightkm" and "speedkm" correspond to the so-called "telematics covariates" that capture policyholders' driving style and driving patterns. We do not include the gender variable in the model because European Union regulations prohibit discrimination between men and women in the field of insurance premiums; for more information on these data, see Guillen et al. (2019).

Table 4 shows our descriptive statistics for the cost per policyholder variable in the original scale, transformed into logarithmic form (log(cost)), and information on the

covariates. We show that our data set contains one extreme observation for the response variable corresponding to a claim that exceeded €130,000 (natural logarithm close to 5).

Table 4. *Descriptive statistics of the variables in the claim costs dataset.*

	Mean	Std.	Min.	Q25	Median	Q75	Max.
cost	1.810	6.191	0.018	0.417	0.818	1.878	130.870
log(cost)	-0.145	1.128	-4.031	-0.874	-0.201	0.630	4.874
age	27.009	3.246	20.586	24.496	26.820	29.886	34.067
agelic	6.429	2.833	2.001	4.337	5.864	7.992	14.686
agecar	8.916	4.162	2.111	5.777	7.943	11.370	20.468
parking	0.763	0.426	0.000	1.000	1.000	1.000	1.000
tkm	8.356	4.530	1.220	5.174	7.549	10.635	35.105
nightkm	7.514	6.504	0.044	2.979	5.841	9.954	42.830
urbankm	27.127	14.163	3.810	16.565	24.401	35.245	80.659
speedkm	7.203	7.100	0.122	2.286	4.969	9.403	48.002

Q25 and Q75 are the first and third quartiles.

log(•) denotes natural logarithm.

The single-index models that we estimate in this section are fitted using "log(cost)" as the dependent variable. Table 5 shows the results of the estimated parameters ($\hat{\theta}$) of the single-index models when using our FMCL method and three different covariate vectors, that is, all the explanatory variables, only the telematics variables and only the traditional rating factors, i.e., the non-telematics covariates. Note that the smoothing parameters h_1 and h_2 obtained for each estimated parameter vector are the same. This is just a coincidence which does not occur for other analyses. We establish "speedkm" as the variable with the constrained coefficient $\theta_1 = 1$ while for the model with the non-telematics variables we use "age". This is convenient because the nature of these covariates makes interpretation straightforward in this context. The reason we opt to fix the effect of the speed variable is that we believe that high speeds result in a greater risk of being involved in severe accidents, which in turn are more costly than minor accidents. For each estimated parameter $\hat{\theta}_j,\ j = 2,\ldots,8$, we test individual significance.

We also estimated the parameters with methods HHI and DHH, that are shown in the Supplementary Material. These results indicate that the estimated parameters with the HHI method have larger standard errors than those obtained with our method. In general, HHI and DHH are more sensitive than FMCL to the selection of the covariate vector.

Table 5 shows that the sign of the effect of the telematics variables is unchanged when comparing the model with all the variables and that one using only the telematics variables. This indicates that, in the single-index model of the logarithm of the cost per policyholder, driving patterns matter. For example, in the model with all variables, as the effect of "speedkm" is the reference and the effects of "age" and "agelic" are positive and

we conclude that the longer the driving experience is, the greater the risk is; however, driving experience is associated with "tkm', the effect of which is negative.

As shown above in Section 2, given that we assume $\theta_1 = 1$, even when the signs of the coefficients of the explanatory variables are not identified, we are still able to analyse the relation between these effects. For example, in Table 5, we observe, on the one hand, that "tkm" has an opposite effect to "speedkm", i.e., excess speed can be offset by the amount of time spent driving (measured here in terms of total distance driven). On the other hand, the coefficients of "nightkm" and "urbankm" present the same sign as the "speedkm" coefficient. Thus, if a higher percentage of driving at speeds above the limit implies higher values of the index, then the same is true as night-time and/or urban driving increase.

Table 5. *Estimated parameters and their significance (p-value in parentheses) for the single-index model in the claim cost data set.*

	Model		
	All variables	Only telematics	Only non-telematics
speedkm	1.000	1.000	–
age	0.153 (<0.0001)	–	1.000
agelic	0.097 (0.0034)	–	-0.246 (<0.0001)
agecar	-0.107 (<0.0001)	–	0.074 (<0.0001)
parking	-0.162 (0.2570)	–	-0.655 (<0.0001)
tkm	-0.044 (0.0004)	-0.423 (<0.0001)	–
nightkm	0.117 (<0.0001)	0.089 (0.0005)	–
urbankm	0.141 (<0.0001)	0.080 (<0.0001)	–
$h_1 = 0.3857$ and $h_2 = 0.1488$			
The first coefficient of each model is fixed and equal to 1.000.			
The computational times are 4.28 minutes with all variables,			
35.27 seconds with telematics and 21.49 seconds with no telematics			

The values of the index do not have a direct interpretation. These values allow us to analyse how the shape of the conditional distribution and the marginal effects change. To analyse these results in greater detail, we use plots, shown now in the original scale of the cost per claim as opposed to their log-transformation. In Figures 1 and 2, we plot the index against the fitted mean of the model with all variables and with non-telematics variables only, the median and p-th quantiles with $p = 0.90$, $p = 0.95$ and $p = 0.99$ (the plot with only telematics variables is similar to Figure 1). The mean curve is estimated using the Nadaraya-Watson estimator of the regression function between the dependent variable and the estimated linear index. The median and the higher quantiles are estimated from the inverse of the estimated conditional distribution function. The smoothing parameters are calculated specifically for each estimated curve, i.e., for the kernel regression the order is $n^{-1/5}$ and for the quantile it is $n^{-1/3}$. The main result is that

the cost distribution conditional on the value of the index is not constant. Furthermore quantiles are not monotonic in the index. This is evidence that some combinations of the covariates lead to a conditional cost distribution with a longer tail than others. For example, Figure 1 shows that when the index takes values around 22.5 and around 31 the conditional distribution has a heavier tail than for the rest of the index domain. We have calculated the mean of the covariates for the policyholders with index values between 22 and 23 and the results indicate that these individuals tend to use night parking and the means for the telematics covariates (tkm, nightkm, urbankm and speedkm) are higher than the means for the whole sample. A second group of policyholders with heavier tail takes index values around 31. The means of the covariates for policyholders with index values between 30 and 32 indicate that these individuals also use parking and drive more than 20% of total kilometres above the speed limit. These features are not captured by the mean curve, which is flat; thus, we can conclude that using a single-index conditional distribution model prediction is helpful to insurance companies when setting up wider margins that correspond to the values of those predicted in the intervals where the conditional distribution presents a remarkable heavy tail.

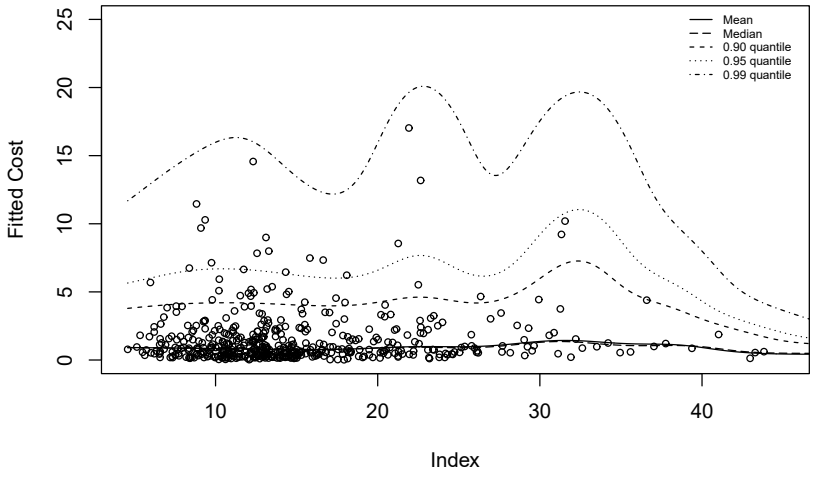

Figure 1. *Fitted values of the conditional mean (solid line) and quantiles (dotted and dashed lines) with all covariables in the model.*

When comparing the plot of the model with all the variables (Figure 1) with the plot of that with only the traditional rating variables (Figure 2), the benefits of including the telematics regressors become evident. By doing so, the intervals of the index corresponding to a conditional distribution with a longer tail are easily identified and, as a result, in such cases the insurance company estimates a slight increase in the median cost and a marked increase in the upper quantiles.

The single-index value provides a one-dimensional summary of the characteristics that discriminate between the policyholders in terms of the conditional cost distribution.

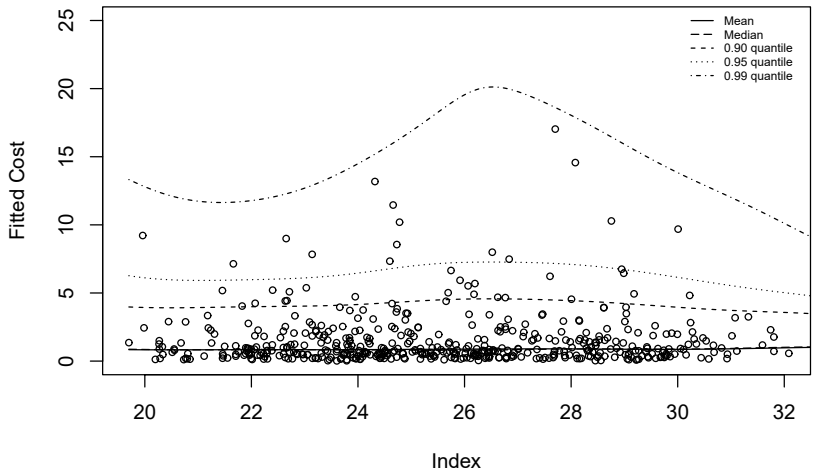

Figure 2. *Fitted values of the conditional mean (solid line) and quantiles (dotted and dashed lines) with only the non-telematics covariables in the model.*

4.1. Marginal effects on extreme quantiles

The study of marginal effects on extreme quantiles of the cost per policyholder, obtained from the derivative of the estimated inverse conditional distribution function, provides us with information about the changes of the risk of high losses when explanatory variables increase or decrease. Furthermore, we analysed to what extent the effects of the variables are different in the extremes and in the central values of the variable cost per policyholder. The results of the marginal effects have been obtained from the kernel estimates described in Subsection 2.4, using the significant parameters. These results are purely descriptive and show the flexibility of our proposal.

As we explained in Section 2, for a given vector of values of the covariates $\mathbf{x} = (x_1,\ldots,x_d)$, we estimated the marginal effects along a grid of values of the covariate x_k. We focused this analysis on telematics variables and studied some examples for certain policyholders. Specifically, once the grid for each telematics variable was fixed, we estimated the marginal effects for the younger and the older individuals, when the rest of the variables take their minimum values. Figures 3 and 4 show the marginal effects for variables "speedkm", and "nightkm", respectively, similar results for "tkm" and "ubankm" are shown in Supplementary Material. In general, we observe that marginal effects of telematics variables are different for the median and for the 0.95 quantile of the cost per policyholder. Furthermore, the marginal effects for the younger and older policyholders exchange their position when they are calculated at the median or when they are calculated at the 0.95 quantile. For example, in Figure 3 we see that the impact on the severity of a claim of exceeding approximately 10% or more kilometers over the posted speed limit is higher for older than for younger drivers at the 95% quantile (right plot) while it is lower at the median (left plot). Note that a negative marginal effect is possible because drivers that exceed speed limits by more that 10% could be more skilled than the rest

and for them the median cost may be lower (left plot) while the extreme quantile may be higher (right plot).

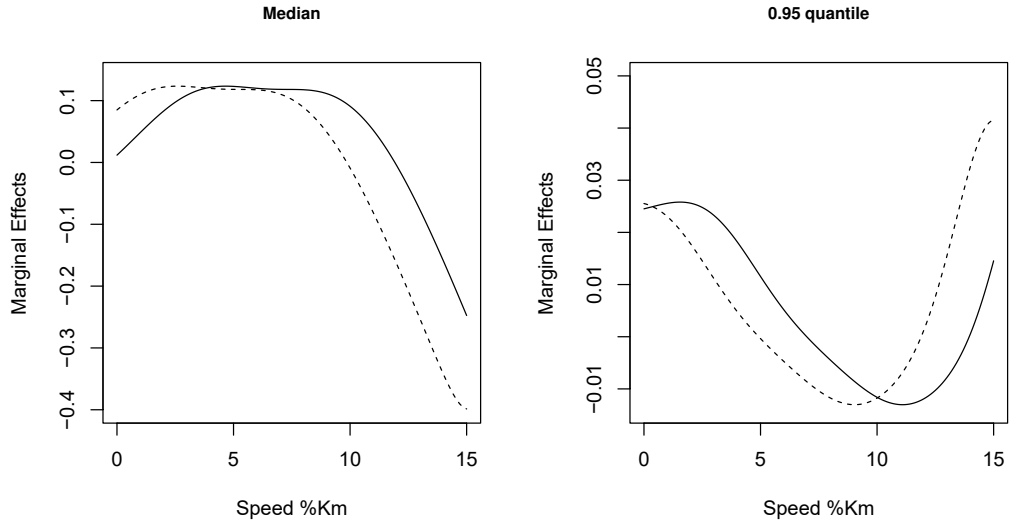

Figure 3. *Marginal effects on the median (left plot) and on the 0.95 quantile (right plot) of the cost per policyholder vs the percentage of kilometers with excess speed. Younger policyholder (solid line) and older policyholder (dashed line). The rest of covariates take their minimum values.*

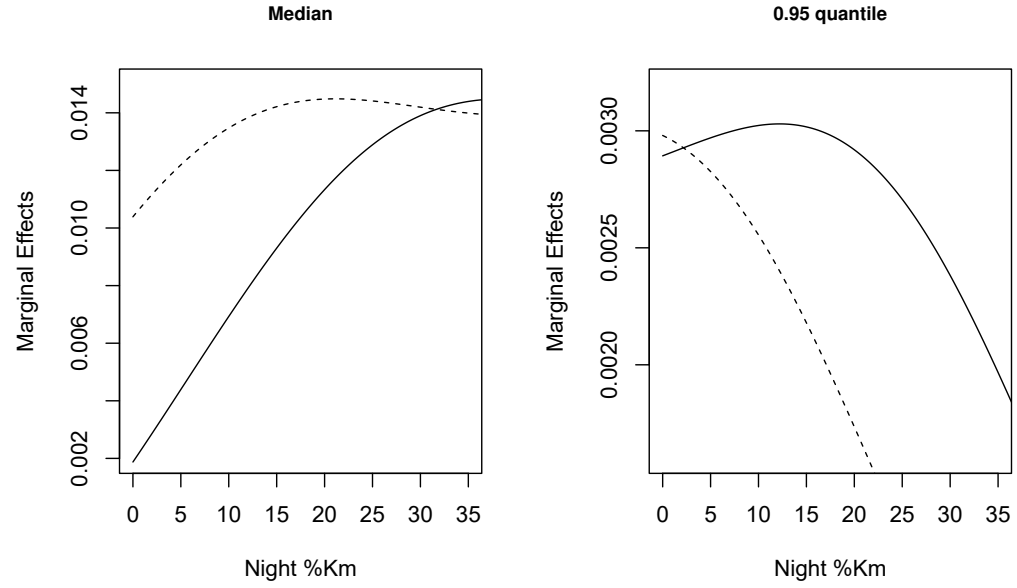

Figure 4. *Marginal effects on the median (left plot) and on the 0.95 quantile (right plot) of the cost per policyholder vs the percentage of kilometers at night. Younger policyholder (solid line) and older policyholder (dashed line). The rest of covariates take their minimum values.*

4.2. Predictive analysis of automobile red claim costs

To analyse the predictive capacity of the estimated single-index models with the three estimators FMCL, HHI and DHH, we used a sample of 100 cases, which were not included in the sample used to estimate the models in the previous subsection. The results using HHI and DHH are practically the same, only some differences are observed from the fourth decimal, for this reason we only analyse the results for HHI. In Table 6 we present the descriptive statistics of this new sample.

Table 6. *Descriptive statistics of the variables in the claim costs sample for prediction analysis.*

	Mean	Std.	Min.	Q25	Median	Q75	Max.
cost	1.557	2.080	0.030	0.438	0.820	1.880	13.584
Log(cost)	-0.138	1.107	-3.518	-0.827	-0.199	0.631	2.609
age	26.438	2.763	20.594	24.410	26.418	28.461	32.769
agelic	5.952	2.670	1.859	3.908	5.458	7.592	14.628
agecar	8.349	3.875	2.283	5.175	7.943	10.665	20.468
parking	0.790	0.409	0.000	1.000	1.000	1.000	1.000
tkm	7.644	4.006	0.560	5.018	7.267	9.876	23.336
nightkm	8.022	6.794	0.462	3.235	5.889	12.115	40.694
urbankm	29.232	14.522	10.266	17.386	27.811	36.246	85.553
speedkm	6.638	6.369	0.155	1.911	4.582	9.103	29.420

Q25 and Q75 are the first and third quartiles.

Table 7. *Criteria for the scoring rules for prediction. QE and WQE are multiplied by* 10*.*

Methods	FMCL	HHI
MSE	1.303	1.514
PMCC	-2.529	-2.662
$\hat{l}(\theta)$	-149.592	-153.595
AIC	313.184	321.191
BIC	331.420	339.427
$10 \times QE$	11.495	11.531
$10 \times WQE$	8.388	8.436

For each of the indices that were estimated with the alternative methods, all the criteria defined in Section 2.5 are calculated and presented in Table 7. In addition, the mean squared error (MSE) associated with the first order predicted conditional moment corresponding to the first sum of the PMCC criterion is also calculated. All these criteria are obtained with the estimated parameters of the single-index models that include all the covariates. These parameters are applied to the sample described in Table 6 to obtain the values of the index, i.e., the parametric part of the single-index model. To obtain the nonparametric estimation of the conditional functions used in the evaluated criteria, we

need to calculate the appropriate degree of smoothing in each case, which depends on the type of function, on the sample size and on the scale of the variable. So, the optimal bandwidth for the kernel estimation of the density function is of order $n^{-1/5}$ and for the distribution function and quantile it is of order $n^{-1/3}$.

Scoring rules are shown in Table 7, QE and WQE are calculated for a sequence of values for p between 0.5 and 0.999 in intervals of 0.001 units. The results show that the best fit is provided by the FMCL method for prediction in all cases.

5. Conclusion

The method proposed herein provides a full specification of the conditional distribution, while preserving the flexible nature of the single-index. Contrary to this principle, one limitation of the traditional approach to generalised linear modelling is the fact that the linear predictor is linked to the mean which, in general, is related to the location parameter of a given distribution that is assumed to be true.

In many contexts, heterogeneity is likely to be more closely associated with the shape of the distribution and not so much with location. This is precisely the case of the application presented as a case study herein. The use of a single-index model allows us to analyse all the components of the motor insurance claims cost distribution: that is, its mathematical expectation, its median, its quantiles and the marginal effects of the covariates at their different values.

Here, we have developed an estimator for the conditional distribution single-index model based on maximisation of the estimated conditional likelihood. We have used this approach to estimate the conditional distribution and, more specifically, its quantiles. This, today, is fundamental in data analysis, given that in certain applications a knowledge of the mean is not as interesting as a knowledge of other characteristics of the distribution. In our application, the estimation of the probability of a severe accident given some covariates, i.e., a cost larger than a fixed value, is a measure of the risk of driving unsafely.

From the expression of the marginal effect of a covariate on a given quantile, we have developed a way to interpret the estimated parameters of the index. Furthermore, we can also interpret the specific marginal effects for each insured individual.

Our main theoretical results demonstrate the asymptotic properties of the estimator for a vector of parameters in a conditional distribution single-index model and provide an expression for its covariance matrix. Likewise, the simulation study conducted herein demonstrates the power of the inference using the kernel estimator of the covariance matrix. These results are fundamental in situations in which the analyst does not have any prior knowledge for identifying the variables that are actually responsible for changes in the distribution of the dependent variable. The estimation of the variance-covariance matrix considering the possible censored data in line with what is described by Laudagé, Desmettre and Wenzel (2019) is a future goal.

Moreover, the simulation study shows how our method is, in fact, an improvement with respect to the finite-sample properties of certain known alternative methods, especially when the conditional distribution is skewed and has a long right tail. This is frequent in economic variables measuring revenues and expenses. The estimator proposed is a considerable improvement on the alternatives analysed, showing robustness in the presence of extreme values. However, for a distribution of this shape using a sample of small size, the results are still not especially good, but they can be improved with the use of a logarithmic transformation.

In the application described here, the observed characteristics of the insured drivers can be usefully employed to understand the distribution of claims cost. Additionally, if single-index models were implemented in practice, they would enable insurers to combine the cost per policyholder distribution with predictions about the expected number of claims, which is currently the baseline for premium calculation dependent on such covariates as age, number of years holding a driving licence, power of the vehicle, age of the vehicle, and so on. Moreover, when we include driving behaviour information in the model (that is, variables such as distance driven and a range of driving habits), our approach allows us to identify the values of the single-index that correspond to a long-tailed cost distribution and, therefore, to detect situations in which the probability of observing a large claim increases. In addition, our proposal presents better predictive scores and, therefore, more adjusted predictions than other existing alternatives.

SUPPLEMENTARY MATERIAL

SM: The file contains: 1. The proofs of the theoretical results in Section 2.2. 2. The results of simulation study related to the properties (bias, variance and MSE) of the alternative methods and the inference power of FMLC for symmetric distributions. 3. The results of application using HHI and HHD methods and additional plots marginal effects using FMLC method and are available in http://www.ub.edu/rfa/R/SORT-BCG/.

DS: Data set and R program used in the illustration of FMCL method in Section 4 are available in https://data.mendeley.com/datasets/py3kb2hn2b/1 and http://www.ub.edu/rfa/R/SORT-BCG/

Acknowledgements

This article is part of the I+D+i projects PID2019-105986GB-C21 and grant TED2021-130187B-I00, financed by MCIN/ AEI/10.13039/501100011033. MG thanks ICREA Academia.

References

Bashtannyk, D. M. and Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation. *Computational Statistics & Data Analysis*, 36, 503–518.

Buch-Larsen, T., Guillen, M., Nielsen, J. P. and Bolancé, C. (2005). Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. *Statistics*, 39, 503–518.

Delecroix, M., Härdle, W. and Hristache, M. (2003). Efficient estimation in conditional single-index regression. *Journal of Multivariate Analysis*, 86, 213–226.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association*, 102, 359–378.

Guillen, M., Nielsen, J. P., Ayuso, M. and Pérez-Marín, A. M. (2019). The use of telematics devices to improve automobile insurance rates. *Risk Analysis*, 39, 662–672.

Guillen, M., Nielsen, J. P. and Pérez-Marín, A. M. (2021). Near-miss telematics in motor insurance. *Journal of Risk and Insurance*, 88, 569–589.

Guillen, M., Nielsen, J. P., Pérez-Marín, A. M. and Elpidorou, V. (2020). Can automobile insurance telematics predict the risk of near-miss events? *North American Actuarial Journal*, 24, 141–152.

Hall, P., Wolff, R. C. L. and Yao, Q. (1999). Methods for estimating a conditional distribution function. *Journal of the American Statistical Association*, 94, 154–163.

Hall, P. and Yao, Q. (2005). Approximating conditional distribution function using dimension reduction. *The Annals of Statistics*, 33, 1404–1421.

Härdle, W. (1990). *Applied Nonparametric Regression*. UK: Cambridge University Press.

Härdle, W., Hall, P. and Ichimura, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics*, 21, 157–178.

Hastie, T. J. and Tibshirani, R. (1990). *Generalized Additive Models*. London: Chapman & Hall/CRC.

Horowitz, J. L. and Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association*, 91, 1632–1640.

Hristache, M., Juditsky, A. and Spokoiny, V. (2001). Direct estimation of the index coefficient in a single-index model. *Annals of Statistics*, 29, 595–623.

Klein, R. W. and Spady, R. H. (1993). Efficient semiparametric estimator for binary response models. *Econometrica*, 61, 387–421.

Koenker, R. and Bassett, G. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Laudagé, C., Desmettre, S. and Wenzel, J. (2019). Severity modeling of extreme insurance claims for tariffication. *Insurance: Mathematics and Economics*, 88, 77–92.

Newey, W. K. and Stoker, T. M. (1993). Efficient of weighted average derivatives estimators and index models. *Econometrica*, 61, 1199–1223.

Powell, J. L., Stock, J. H. and Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, 57, 1403–1430.

Scott, D. (2015). *Multivariate Density Estimation: Theory, Practice, and Visualization*. New Jersey: John Wiley & Sons.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall/CRC.

Stoker, T. M. (1986). Consistent estimation of scaled coefficients. *Econometrica*, 54, 1461–1481.

Strzalkowska-Kominiak, E. and Cao, R. (2013). Maximum likelihood estimation for conditional distribution single-index models under censoring. *Journal of Multivariate Analysis*, 114, 74–98.

van den Berg, G. J., Janys, L., Mammen, E. and Nielsen, J. P. (2021). A general semi-parametric approach to inference with marker-dependent hazard rate models. *Journal of Econometrics*, 221, 43–67.