"Flexible maximum conditional likelihood estimation for single-index models to predict accident severity with telematics data"

Bolancé C, Cao R & Guillen M

Institut de Recerca en Economia Aplicada | *Research Institute of Applied Economics*

Abstract

Estimation in single-index models for risk assessment is developed. Statistical properties are given and an application to estimate the cost of traffic accidents in an innovative insurance data set that has information on driving style is presented. A new kernel approach for the estimator covariance matrix is provided. Both, the simulation study and the real case show that the method provides the best results when data are highly skewed and when the conditional distribution is of interest. Supplementary materials containing appendices are available online.

JEL classification: C51, C14, G22

Keywords: Insurance loss data, heavy tailed distributions, quantiles, non-parametric conditional distribution.

Catalina Bolancé: Department of Econometrics, Riskcenter-IREA, University of Barcelona, Avinguda Diagonal 690, 08034 Barcelona, Spain. E-mail: bolance@ub.edu

Ricardo Cao: Research Group MODES, Department of Mathematics, CITIC, Universidade da Coruña and ITMATI Campus de Elviña, s/n 15071 A Coruña, Spain. E-mail: rcao@udc.es

Montserrat Guillen: Department of Econometrics, Riskcenter-IREA, University of Barcelona, Avinguda Diagonal 690, 08034 Barcelona, Spain. E-mail: mguillen@ub.edu

Acknowledgements

The support received by the Ministry of Economy and Competitiveness in Grant ECO2016-76203-C2-2-P for the first and third authors is gratefully acknowledged. The research of the second author has been supported by MINECO Grants MTM2014-52876-R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigaci´on de Galicia ED431G/01), all of them through the ERDF. All authors declare no conflict of interest as no sponsor has been involved in the implementation and conclusions of the research.

1 Introduction

Single-index models are a semiparametric way to generalise linear regression. They specify the dependence between a random variable Y and a d-dimensional vector X as follows (see Härdle et al., 1993):

$$Y = m\left(\theta' X\right) + \epsilon, \tag{1}$$

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. The aim is to estimate θ and m. Single-index regression models can be extended to single-index conditional distribution models by just recalling that $P(Y \leq y|X = x) = E(\quad (Y \leq y)|X = x)$, where $\quad (\cdot)$ equals 1 if the expression is true and 0 otherwise. This is important for instance, in risk management where the conditional distribution function can be useful to evaluate certain scenarios.

The traditional approaches for estimating the linear predictor coefficients θ and the link function m are based on the conditional expectation rather than 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 and, in particular, in risk analysis. Our contribution is to extend maximum likelihood estimation of (1), and this opens the door to single-index conditional distribution modeling which has an enormous potential for applications.

In order to estimate vector θ, Härdle et al. (1993) proposed to directly minimise the residual sum of squares, so their estimator is

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

where $(X_i, Y_i)\quad i = 1, \ldots, n$ 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 et al. (2001) analysed the average derivatives estimator of the vector of parameters in the index-model, introduced by Stoker (1986) as well as Powell et al. (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. $\arg\min_\theta \sum_{i=1}^{n} \psi\left[Y_i, \hat{m}\left(\theta' X_i\right)\right]$. All these methods ignore the shape of the conditional distribution because they are based on fitting the conditional expectation.

Delecroix et al. (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 maximum likelihood estimation of θ but only for a binary response dependent variable. More recently, in the context of survival data with censored observations, Strzalkowska-Kominiak and Cao (2013) investigated maximum likelihood alternatives based on a nonparametric estimation of the conditional distribution and they showed that the existing methods for censored data could be improved.

Non-parametric regression is more general than the single-index model specified in (1). It emanates from a more general specification $Y = m(X) + \epsilon$, where the aim is to estimate the regression curve $m(x) = E(Y|X = x)$ (see Härdle, 1990). However in practice, nonparametric regression presents two important difficulties. First, estimation becomes considerably difficult as the number of covariates increases (curse of dimensionality). The second difficulty is that the interpretation of the effects of explanatory variables cannot be carried out directly because 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 Tibshiran, 1990); however, it shares the same difficulties already described for nonparametric regression.

Here, a new maximum likelihood estimator of θ in (1) is proposed inspired on Strzalkowska-Kominiak and Cao (2013) who worked with censored data. We propose to use two different smoothing parameters; one associated with the distribution of Y and another one associated with the distribution of index $\theta_0' X$. These two parameters are needed to estimate the conditional distribution. Consistency properties for the estimators are obtained.

A simulation study is carried out where 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 covariates are correlated and binary variables are included.

Note that Hall and Yao (2005) or Newey and Stoker (1993) consider continuous covariates. Not many papers deal with discrete covariates in single-index models. Horowitz and Härdle (1996) focused on analysing a direct estimator for the effect of the discrete covariates. Methods such as those proposed by Härdle et al. (1993), Hristache et al. (2001) and Delecroix et al. (2003) do not consider including categorical variables, but they allow incorporating dummy (binary) variables.

We present a real data application where we analyse the cost of claims in a motor insurance data set. This dependent variable is right-skewed. We show the interpretability of the model results.

The rest of the paper is organized as follows. Section 2 describes the proposed estimator. The simulation study is presented in Section 3. Section 4 includes the application to the motor insurance data. Finally, Section 5 summarises the conclusions. The proofs of the main results are collected in Appendix A.

2 The estimator

Let Y be the response random variable that depends on a vector of covariates $X = (X_1, ..., X_d)'$ and $f(\bullet|\mathbf{x})$ be the density function of Y given $X = \mathbf{x}$, where $\mathbf{x} = (x_1, ..., x_d)$ is a fixed vector. Moreover, let

$$\theta_0 = (\theta_{0,1}, ..., \theta_{0,d})', \quad \text{where} \quad d \geq 2,$$

be the parameter vector to be estimated with the property:

$$f(y|\mathbf{x}) = f_{\theta_0}(y|\theta_0'\mathbf{x}),$$

where $f_{\theta_0}(\bullet|\theta_0'\mathbf{x})$ is the conditional density of Y given $\theta_0'X = \theta_0'\mathbf{x}$. Furthermore, let $F(y|\mathbf{x})$ and $F_{\theta_0}(y|\theta_0'\mathbf{x})$ be the conditional distribution functions, given $X = \mathbf{x}$ and $\theta_0'X = \theta_0'\mathbf{x}$, respectively. As a consequence

$$F_{\theta_0}(y|\theta_0'\mathbf{x}) = F(y|\mathbf{x}). \tag{2}$$

For any $\theta_0 \in \mathbb{R}$ fulfilling (2) and any nonzero real number λ, then vector $\lambda\theta_0$ also fulfills (2). Consequently, infinitely multiple choices exist for the single-index parameters. The usual way to solve this identification problem is to introduce some scale constraint, for example $||\theta_0|| = 1$ or $\theta_{0,1} = 1$. In practice, this implies that the sign of the effects of the covariates on the dependent variable are not identified.

For a given $\theta = \theta_0$, using the conditional distribution function we can obtain the p-th conditional quantile: $Q_\theta(p|\theta'\mathbf{x}) = F_\theta^{-1}(p|\theta'\mathbf{x})$, i.e. $F_\theta(y_p|\theta'\mathbf{x}) = p$ where $p \in (0,1)$. As it happens for any generalized 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:

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

where:

$$\frac{\partial Q_\theta(p|\theta'\mathbf{x})}{\partial x_k} = -\frac{\left.\frac{\partial F_\theta(Q_\theta(p|\theta'\mathbf{x})|t)}{\partial t}\right|_{t=\theta'\mathbf{x}} \cdot \theta_k}{f_\theta(Q_\theta(p|\theta'\mathbf{x})|\theta'\mathbf{x})}.$$

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

Let (Y_i, X_i) $i = 1, ..., n$ be a random sample of the dependent variable and the covariates, where $X_i = (X_{i1}, ..., X_{id})$. Let K be a nonnegative kernel and h_1, h_2 two positive bandwidths. The kernel conditional density estimator is (see Bashtannyk and Hyndman, 2001)

$$\hat{f}_\theta(y|t) = \hat{f}_{\theta,h_1,h_2}(y|t) = \frac{\hat{r}(t,y)}{\hat{s}(t)}, \tag{3}$$

where

$$\hat{r}(t,y) = \hat{r}_{h_1,h_2}(t,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{t-\theta'X_i}{h_1}\right) K\left(\frac{y-Y_i}{h_2}\right)$$

and

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

It is well known that the choice of the kernel function, K, is not so relevant, but the selection of the smoothing parameters, h_1 and h_2, has a big impact in the quality of estimator (3). To find a practical solution for the choice of h_1 and h_2 we introduce these two parameters in the maximization procedure of the nonparametric estimated likelihood.

The kernel estimator of the conditional distribution function is (see Hall et al., 1999):

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

where

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

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

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

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

where

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

and

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

where K' is the first derivative of the kernel.

2.1 Semiparametric conditional likelihood

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

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

Maximizing this function is equivalent to maximizing its logarithm:

$$\tilde{l}_n(\theta) = \frac{1}{n} \log \left(\tilde{L}_n(\theta) \right) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(Y_i|\theta' X_i). \tag{4}$$

An ideal "estimator" would be the one that maximizes the theoretical log-likelihood

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

2.2 Maximum conditional likelihood estimation

In practice, f_θ (or F_θ) is not known. So we need to estimated it and plug it into (4). In addition we need to modify the estimated likelihood with a leaving-one-out procedure, in order not to pick artificially small bandwidths.

Let $\hat{f}_\theta^{-i}(Y_i|\theta' X_i)$ be the estimator defined in (3), where the sum runs over $j \neq i$. Set

$$\hat{f}_\theta^{-i}(Y_i|\theta' X_i) = \frac{\hat{r}^{-i}(\theta' X_i, Y_i)}{\hat{s}^{-i}(\theta' X_i)} \tag{5}$$

where

$$\hat{r}^{-i}(\theta' X_i, Y_i) = \frac{1}{h_1 h_2} \sum_{j=1, j\neq i}^{n} K\left(\frac{\theta' X_i - \theta' X_j}{h_1}\right) K\left(\frac{Y_i - Y_j}{h_2}\right) \tag{6}$$

and

$$\hat{s}^{-i}(\theta' X_i) = \frac{1}{h_1} \sum_{j=1, j\neq i}^{n} K\left(\frac{\theta' X_i - \theta' X_j}{h_1}\right)$$

and define the leaving-one-out estimated conditional likelihood:

$$\hat{l}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log \hat{f}_\theta^{-i}(Y_i|\theta' X_i). \tag{7}$$

The final maximum conditional likelihood estimator is defined as

$$\hat{\theta}_n = \arg\max_{\theta} \hat{l}_n(\theta). \tag{8}$$

In practice, we will maximize the likelihood function defined in (8) with respect to θ and to the two smoothing parameters h_1 and h_2.

2.3 Main results

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

$$l(\theta) = E(\tilde{l}_n(\theta)). \tag{9}$$

We start by proving that the true parameter vector, θ_0, can be characterized as the maximizer of the score function. Existence of that function is the only condition needed:

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

Theorem 1. *The true single-index parameter, θ_0 , is the maximizer of the score function, i.e.,*

$$\theta_0 = \arg\max_{\theta} l(\theta) \tag{10}$$

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

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

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

The two bandwidths h_1, h_2 should fulfill the following conditions

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

Finally, let $l^{[1]}(\theta_0) = \nabla_\theta l(\theta)|_{\theta=\theta_0}$ denote the gradient of $l(\theta)$ over θ evaluated in θ_0. Further, let $l^{[2]}(\theta)$ denote the Hessian matrix of $l(\theta)$. The following regularity conditions are also assumed

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

A6: The function $h(\mathbf{x}, y) = \frac{\partial}{\partial\theta_k}f_\theta(\theta_1'\mathbf{x}, y)_{\theta=\theta_0}$ is continuous and $\frac{\partial^2}{\partial^2\theta_k}f_\theta(\theta_0'\mathbf{x}, y)_{\theta=\theta_0}$ exists.

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

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

Lemma 1. *Under A1, A4 and A6 we have*

$$\hat{\theta}_n - \theta_0 = -\left[\hat{l}_n^{[2]}(\hat{\theta}_n^*)\right]^{-1}(\hat{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0)),$$

where $\hat{\theta}_n^*$ is between $\hat{\theta}_n$ and θ_0.

Theorem 2. *Under A1-A7, we have*

$$\hat{\theta}_n \to \theta_0 \quad \textit{in probability.}$$

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

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

where

$$\Sigma = \Sigma_2\Sigma_1\Sigma_2^t,$$

$$\Sigma_2 = \left[l^{[2]}(\theta_0)\right]^{-1}$$

and

$$\begin{aligned}\Sigma_1 &= E\Big[\nabla_\theta \log(f_\theta(Y|\theta' X))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(Y|\theta' X))_{\theta=\theta_0})^t\Big] \\ &= \int(\nabla_\theta \log(f_\theta(y|\theta'\mathbf{x}))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(y|\theta'\mathbf{x}))_{\theta=\theta_0})^t f(\mathbf{x}, y)d\mathbf{x}dz.\end{aligned}$$

All the proofs can be found in the Appendix.

3 Simulation Study

Here, we summarise the results of our simulation study. The aim is to evaluate the finite sample properties of our estimator assuming, on the one hand, different shapes of the distribution of the dependent variable and, on the other, different vectors of explanatory variables. We show two types of results: those related to the properties of parameter estimator, $\hat{\theta}$, and those related to basic inference about the value of these parameters.

3.1 Properties of the estimated parameters in the linear index

We compare the variance, the bias and the Mean Square Error (MSE) of $\hat{\theta}$, using our flexible maximum conditional likelihood estimator (hereinafter FMCL) with two 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 initial parameters those obtained with the HHI method which is $\sqrt{n}$-consistent. It will be shown that the proposed FMCL estimator is the best when the conditional distribution is right-skewed and also when the tail of the conditional distribution is heavy.

We generate 500 samples of size $n = 100$, 500 and $2,000$ and calculate the bias, the standard deviation (STD) and the mean squared error (MSE) of the estimators.

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

For our two choices of symmetric distributions 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 has a heavier tail (tail type Fréchet) than the

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

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

lognormal and the Weibull (tail type Gumbel) (see Buch-Larsen et al., 2005, for a full description of the Champernowne distribution).

In our simulation study, we use different vectors of covariates X that we identify as vectors V1, V2, V3 and V4; for the first three $\theta' = (1, 1.3, 0.5)$ and for the fourth $\theta' = (1, 1.3, 0.5, 0.8)$. The values in vector V1 are generated from three uncorrelated standard normal distributions. 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 $cov(X_1, X_3) = 0.5$. The vector V4 consist of considering V1 and them a binary variable whose values are generated from a Bernoulli distribution with probability 0.4, independent of the three components of V1, i.e. (X_1, X_2, X_3).

We analyse the finite sample properties of $\hat{\theta}$ obtained with: FMCL, HHI and DHH. We first consider symmetric conditional distributions. In Tables 2, 3 and 4 the results of the standard deviation, the bias and the square root of the MSE values are shown, respectively, for symmetric conditional distribution. Here we also observe the effect on the finite sample properties of the collinearity between the explanatory variables (vectors V2 and V3). Furthermore, we use the vector V4 to explore the methods when including a binary variable in the model.

For the analysed symmetric conditional distributions and all covariate vectors we observe that when the sample size is $n = 100$, the variance of FMCL is similar or a bit greater than the variance obtained for HHI. However, when the sample size increases, the variance of FMCL decreases more quickly, turning to be much lower than that obtained for HHI. The bias of FMCL also decreases when the sample size increases slightly faster than HHI. The DHH estimator is bad when the sample size is small, when there is collinearity between covariates and for the logistic distribution.

Focusing on the results for MSE in Table 4, the values obtained when using the FMCL are much lower than those obtained for the alternative methods, except for some normal distribution cases when $n = 100$, where HHI seems to reach slightly smaller values for the MSE.

Concerning the effect of collinearity, when the covariances increase the MSE increases and it is observed that this increment is not as evident in the bias compared to the variance. The MSE associated with the binary variable is the lowest in all cases for FMCL.

Next we will analyse the finite sample properties of the methods when the distribution of Y conditional on $X = \mathbf{x}$ is right-skewed. Here we only show the results using as explanatory variables the vectors V1 and V4 (the effect of collinearity is the same as that for symmetric cases). Furthermore, for the analysed asymmetric conditional distributions in Table 1, no estimator provides good results for $n = 100$.

In Table 5 the STD and the bias obtained when the conditional distribution has a Gumbel-type tail are shown. For the lognormal conditional distribution DHH provides worse results than FMCL and HHI. Comparing, FMCL and HHI we observe that the former has lower STD

Table 2: Standard deviations (STD) of the estimators $\hat{\theta}$ using alternative methods ($\theta = 1$ is fixed) for symmetric conditional distributions.

			Normal			Logistic		
			FMCL	HHI	DHH	FMCL	HHI	DHH
	V1	θ_2	0.3242	0.2910	7.2305	0.4634	0.4508	6.2274
		θ_3	0.2627	0.2952	5.7021	0.3614	0.3538	4.8824
	V2	θ_2	0.4379	0.3039	74.6637	0.5662	0.5543	12.7825
		θ_3	0.2774	0.2719	64.2189	0.4363	0.4903	15.1101
$n = 100$	V3	θ_2	1.1217	0.3200	13.4229	0.6496	0.8206	25.6798
		θ_3	0.4170	0.3835	8.3841	0.5336	0.6844	21.1687
	V4	θ_2	0.3076	0.3662	697.9336	0.4337	0.4643	27.6909
		θ_3	0.2066	0.2363	28.9431	0.3456	0.3669	24.1494
		θ_4	0.3544	0.4020	567.1738	0.6716	0.7047	43.0292
	V1	θ_2	0.0695	0.2055	0.5848	0.0986	0.2375	32.7258
		θ_3	0.0478	0.2432	0.7261	0.0719	0.2431	25.0928
	V2	θ_2	0.0979	0.2424	7.8073	0.1465	0.2849	76.6513
		θ_3	0.0694	0.2333	14.5248	0.0988	0.3071	87.8451
$n = 500$	V3	θ_2	0.1860	0.1865	19.6869	0.2801	0.4045	62.4181
		θ_3	0.0974	0.2989	56.3495	0.1437	0.4042	62.5956
	V4	θ_2	0.0744	0.1858	17.5909	0.1115	0.3411	170.1250
		θ_3	0.0515	0.2194	10.3142	0.0745	0.3320	290.1911
		θ_4	0.0896	0.2225	22.6857	0.1310	0.3447	454.0797
	V1	θ_2	0.0251	0.1977	0.4820	0.0403	0.1774	5.4796
		θ_3	0.0176	0.2414	0.6688	0.0266	0.2498	2.3952
	V2	θ_2	0.0367	0.2936	2.5533	0.0579	0.3148	5.9826
		θ_3	0.0257	0.3174	14.0713	0.0395	0.2868	4.2200
$n = 2,000$	V3	θ_2	0.0696	0.1582	13.4307	0.1092	0.2837	6.1363
		θ_3	0.0368	0.2442	34.3213	0.0565	0.2639	2.7008
	V4	θ_2	0.0292	0.2381	7.1382	0.0418	0.2226	14.7782
		θ_3	0.0191	0.3054	10.1807	0.0265	0.2983	10.6173
		θ_4	0.0358	0.0872	12.5739	0.0502	0.1601	27.9345

Table 3: Bias of the estimators $\hat{\theta}$ using alternative methods ($\theta = 1$ is fixed) for symmetric conditional distributions.

			Normal			Logistic		
			FMCL	HHI	DHH	FMCL	HHI	DHH
$n = 100$	V1	θ_2	-0.1268	-0.0239	-0.1105	-0.1002	0.0011	-0.6112
		θ_3	-0.0220	0.0083	-0.3851	0.0331	-0.0022	-0.0793
	V2	θ_2	0.0721	-0.0226	-3.6897	-0.1391	0.0089	0.1931
		θ_3	0.0295	-0.0037	-3.2652	0.0176	0.0052	0.3888
	V3	θ_2	0.2301	-0.0293	0.1905	-0.0332	0.0139	-0.6809
		θ_3	0.0489	-0.0024	-0.7695	0.2839	-0.0217	-1.3305
	V4	θ_2	0.0427	0.0233	-28.1036	-0.1012	0.0391	-1.3741
		θ_3	0.0010	-0.0163	1.2993	0.0389	-0.0048	-1.7165
		θ_4	0.0449	-0.0007	22.5352	0.1041	0.0082	-0.0628
$n = 500$	V1	θ_2	0.0040	0.0025	-0.1127	0.0012	-0.0101	-1.3913
		θ_3	0.0019	0.0112	-0.0493	-0.0001	0.0113	0.1403
	V2	θ_2	0.0058	-0.0086	-0.0503	-0.0003	0.0017	-2.1255
		θ_3	0.0028	0.0049	0.0434	-0.0034	0.0089	-5.1738
	V3	θ_2	0.0158	-0.0020	-1.3405	0.0077	-0.0118	-15.7523
		θ_3	0.0084	0.0049	1.0568	-0.0073	0.0134	-0.3868
	V4	θ_2	-0.0013	-0.0084	0.7100	0.0060	-0.0193	-5.6416
		θ_3	-0.0003	0.0066	-0.5551	-0.0011	-0.0176	-6.9701
		θ_4	0.0051	0.0018	1.9332	0.0062	0.0120	-18.0176
$n = 2,000$	V1	θ_2	-0.0002	0.0075	-0.0157	0.0005	0.0026	-0.4411
		θ_3	0.0011	0.0084	0.0292	-0.0014	0.0017	0.0675
	V2	θ_2	0.0004	-0.0003	0.3067	0.0026	0.0089	-0.5307
		θ_3	0.0010	0.0084	-1.3214	-0.0004	-0.0208	-0.4258
	V3	θ_2	0.0030	-0.0083	-0.3096	0.0088	0.0038	-0.8406
		θ_3	0.0009	-0.0023	-0.0412	0.0007	-0.0016	-0.2357
	V4	θ_2	0.0034	-0.0167	-0.3118	0.0001	-0.0041	-0.8734
		θ_3	-0.0001	0.0066	-0.8367	-0.0003	-0.0150	0.2058
		θ_4	0.0028	0.0054	0.7564	-0.0004	-0.0124	-0.8951

Table 4: Square root of MSE the estimators $\hat{\theta}$ using alternative methods ($\theta = 1$ is fixed) for symmetric conditional distributions.

			Normal			Logistic		
			FMCL	HHI	DHH	FMCL	HHI	DHH
$n = 100$	V1	θ_2	0.3482	0.2920	7.2313	0.4741	0.4508	6.2573
		θ_3	0.2636	0.2954	5.7151	0.3629	0.3538	4.8830
	V2	θ_2	0.4438	0.3047	74.7548	0.5830	0.5544	12.7840
		θ_3	0.2789	0.2719	64.3018	0.4366	0.4903	15.1151
	V3	θ_2	1.1450	0.3214	13.4243	0.6505	0.8208	25.6888
		θ_3	0.4199	0.3835	8.4193	0.6044	0.6847	21.2105
	V4	θ_2	0.3106	0.3669	698.4992	0.4454	0.4659	27.7250
		θ_3	0.2066	0.2369	28.9722	0.3478	0.3670	24.2103
		θ_4	0.3572	0.4020	567.6214	0.6796	0.7047	43.0292
$n = 500$	V1	θ_2	0.0696	0.2055	0.5956	0.0986	0.2378	32.7554
		θ_3	0.0478	0.2434	0.7277	0.0719	0.2433	25.0932
	V2	θ_2	0.0981	0.2426	7.8075	0.1465	0.2849	76.6807
		θ_3	0.0695	0.2333	14.5249	0.0988	0.3072	87.9973
	V3	θ_2	0.1867	0.1865	19.7325	0.2802	0.4046	64.3752
		θ_3	0.0977	0.2990	56.3594	0.1439	0.4044	62.5968
	V4	θ_2	0.0744	0.1860	17.6052	0.1117	0.3416	170.2186
		θ_3	0.0515	0.2195	10.3292	0.0745	0.3324	290.2748
		θ_4	0.0897	0.2225	22.7679	0.1312	0.3449	454.4370
$n = 2,000$	V1	θ_2	0.0251	0.1978	0.4822	0.0403	0.1774	5.4973
		θ_3	0.0176	0.2416	0.6694	0.0266	0.2498	2.3962
	V2	θ_2	0.0367	0.2936	2.5717	0.0579	0.3149	6.0061
		θ_3	0.0257	0.3175	14.1332	0.0395	0.2875	4.2415
	V3	θ_2	0.0697	0.1584	13.4343	0.1095	0.2838	6.1936
		θ_3	0.0368	0.2442	34.3213	0.0565	0.2639	2.7111
	V4	θ_2	0.0294	0.2387	7.1450	0.0418	0.2227	14.8039
		θ_3	0.0191	0.3055	10.2151	0.0265	0.2987	10.6193
		θ_4	0.0359	0.0874	12.5966	0.0502	0.1606	27.9488

Table 5: Standard deviation (STD) and bias of the estimators $\hat{\theta}$ using alternative methods ($\theta_1 = 1$ is fixed) for lognormal and Weibull conditional distributions.

STD			Lognormal			Weibull		
			FMCL	HHI	DHH	FMCL	HHI	DHH
	V1	θ_2	0.1956	0.2591	7.6728	0.8099	0.2942	12.7693
$n = 500$		θ_3	0.1222	0.2155	3.3440	0.9705	0.2643	6.4902
	V4	θ_2	0.1709	0.1858	5.1329	0.9746	0.1418	20.8384
		θ_3	0.0747	0.2111	9.0944	0.9984	0.1062	21.9198
		θ_4	0.1337	0.2225	2.8747	1.7650	0.2049	13.9250
	V1	θ_2	0.0687	0.0716	1.1335	0.2465	0.1817	14.1635
$n = 2,000$		θ_3	0.0467	0.0962	0.8517	0.1190	0.2428	12.4832
	V4	θ_2	0.0419	0.0865	2.5601	0.0697	0.3372	2.9441
		θ_3	0.0264	0.1651	1.1747	0.0437	0.3361	1.7186
		θ_4	0.0498	0.0872	3.1880	0.9616	0.3030	5.8038
Bias			Lognormal			Weibull		
			FMCL	HHI	DHH	FMCL	HHI	DHH
	V1	θ_2	0.0156	0.0090	-0.2123	-0.2700	-1.2969	-1.7129
$n = 500$		θ_3	0.0055	0.0188	0.0820	-0.1726	-0.4861	-0.8504
	V4	θ_2	0.0000	-0.0084	0.1127	-0.2437	-1.1123	-0.2708
		θ_3	-0.0027	0.0107	0.6435	0.2491	-0.4198	-1.4711
		θ_4	0.0033	0.0018	-0.0465	0.2731	-0.6416	0.0178
	V1	θ_2	0.0004	-0.0012	-0.1024	-0.0101	-1.3055	-1.7478
$n = 2,000$		θ_3	0.0034	0.0024	-0.0550	0.0033	-0.5048	0.1270
	V4	θ_2	0.0001	-0.0023	0.0009	-0.0008	-1.1559	-1.0500
		θ_3	-0.0003	0.0061	0.0204	0.0018	-0.3913	-0.2640
		θ_4	-0.0003	0.0054	-0.1637	-0.1142	-0.6191	-0.0450

than the latter. For the bias, when the sample size $n = 500$ the results are a bit better for HHI but when the sample size increases FMCL improves HHI, especially when we add the binary variable.

For the conditional Weibull distribution the worse results are obtained in the DHH column. With HHI we obtain lower values for STD, especially for smaller sample size $n = 500$; interestingly the bias of FMCL is considerably lower that for the other two methods.

The simulation results for conditional distribution with Fréchet-type tail are shown in Table 6. We know that for the two Champernowne distribution the lower the shape parameter α, the heavier the tail. The results for STD and bias clearly indicate that FMCL improves alternative

Table 6: Standard deviation (STD) and bias of the estimators $\hat{\theta}$ using alternative methods ($\theta_1 = 1$ is fixed) for Champernowne ($\alpha = 1, 2$) conditional distributions.

STD			Champernowne ($\alpha = 1$)			Champernowne ($\alpha = 2$)		
			FMCL	HHI	DHH	FMCL	HHI	DHH
	V1	θ_2	1.8842	96.6511	294.8457	0.3499	0.4030	7.2713
$n = 500$		θ_3	1.9889	39.9251	238.7030	0.1826	0.3254	4.3142
	V4	θ_2	5.3458	162.4999	167.2001	0.1877	0.6692	5.9984
		θ_3	4.0516	234.2215	241.5600	0.1176	0.5351	6.0283
		θ_4	8.5843	470.0219	474.0571	0.4181	0.6995	11.5125
	V1	θ_2	0.4903	74.2523	94.2425	0.0687	0.3990	3.1820
$n = 2,000$		θ_3	0.4143	31.0740	60.8118	0.0799	0.2741	6.3841
	V4	θ_2	0.8589	67.3781	82.8323	0.0346	0.1006	287.6499
		θ_3	0.5064	65.3621	71.4904	0.0170	0.1523	1266.5615
		θ_4	1.6937	125.9128	173.9493	0.0263	0.0741	2760.4214
Bias			Champernowne ($\alpha = 1$)			Champernowne ($\alpha = 2$)		
			FMCL	HHI	DHH	FMCL	HHI	DHH
	V1	θ_2	-0.4987	1.4114	12.0505	-0.0237	-1.3107	-1.2663
$n = 500$		θ_3	-0.2000	1.1112	-2.8499	-0.0045	-0.5067	-0.5541
	V4	θ_2	-0.9150	7.9580	8.4165	0.0116	-0.9695	-1.0882
		θ_3	-0.4476	9.8620	7.7151	0.0049	-0.4003	-0.1822
		θ_4	-0.4704	15.5404	15.1612	-0.0009	-0.5288	-1.2420
	V1	θ_2	-0.0490	-5.7725	-8.4375	-0.0046	-1.2908	-1.1824
$n = 2,000$		θ_3	-0.0176	-2.1612	0.9751	0.0094	-0.4993	-0.2626
	V4	θ_2	-0.1376	-0.6295	1.4127	0.0012	-1.1702	-14.1593
		θ_3	0.0022	-1.9380	-3.1952	0.0003	-0.4556	-57.0692
		θ_4	-0.1026	5.1837	0.3131	0.0007	-0.7016	-123.9391

methods, especially when the tail of the conditional distribution is the heaviest.

The results for the square root of the MSE that are shown in Table 7 corroborate that our proposed FMCL method improves HHI and DHH when the conditional distribution is right-skewed and this improvement is greater as the tail of the conditional distribution is heavier.

3.2 Basic inference power

Power analysis is fundamental to study whether the effect of each covariate is significantly different from zero. The null hypothesis for each parameter is $H_0 : \theta_k = 0, \ k = 1, ..., d$ and as alternative hypothesis we assume that the sign of the parameter is known, i.e. $H_1 : \theta_k >$

Table 7: Square root of MSE of the estimators $\hat{\theta}$ using alternative methods ($\theta_1 = 1$ is fixed) for lognormal, Weibull and Champernowne ($\alpha = 1, 2$) conditional distributions.

			Lognormal			Weibull		
			FMCL	HHI	DHH	FMCL	HHI	DHH
n=500	V1	θ_2	0.1962	0.2593	7.6758	0.8537	1.3298	12.8836
		θ_3	0.1224	0.2163	3.3450	0.9857	0.5533	6.5457
	V4	θ_2	0.1709	0.1860	5.1341	1.0046	1.1213	20.8402
		θ_3	0.0747	0.2114	9.1172	1.0290	0.4330	21.9691
		θ_4	0.1337	0.2225	2.8751	1.7860	0.6735	13.9250
n=2,000	V1	θ_2	0.0687	0.0716	1.1381	0.2467	1.3181	14.2709
		θ_3	0.0468	0.0963	0.8535	0.1190	0.5602	12.4839
	V4	θ_2	0.0419	0.0865	2.5601	0.0697	1.2041	3.1257
		θ_3	0.0264	0.1652	1.1749	0.0438	0.5159	1.7388
		θ_4	0.0498	0.0874	3.1922	0.9684	0.6893	5.8040
			Champernowne ($\alpha = 1$)			Champernowne ($\alpha = 2$)		
			FMCL	HHI	DHH	FMCL	HHI	DHH
n=500	V1	θ_2	1.9491	96.6614	295.0919	0.3508	1.3713	7.3807
		θ_3	1.9990	39.9406	238.7200	0.1827	0.6022	4.3497
	V4	θ_2	5.4235	162.6946	167.4118	0.1881	1.1781	6.0963
		θ_3	4.0763	234.4290	241.6832	0.1177	0.6682	6.0310
		θ_4	8.5971	470.2787	474.2994	0.4181	0.8769	11.5793
n=2,000	V1	θ_2	0.4927	74.4764	94.6194	0.0689	1.3511	3.3946
		θ_3	0.4147	31.1490	60.8196	0.0805	0.5695	6.3895
	V4	θ_2	0.8699	67.3810	82.8443	0.0346	1.1745	287.9982
		θ_3	0.5064	65.3908	71.5618	0.0170	0.4804	1267.8466
		θ_4	1.6968	126.0195	173.9495	0.0263	0.7055	2763.2023

Table 8: Power of the statistic for symmetric distributions.

	H_0	Normal			Logistic		
		n=100	n=500	n=2,000	n=100	n=500	n=2,000
V1	$\theta_2 = 0$	0.998	1.000	1.000	0.852	1.000	1.000
	$\theta_3 = 0$	0.992	1.000	1.000	0.740	1.000	1.000
V2	$\theta_2 = 0$	0.996	1.000	1.000	0.760	1.000	1.000
	$\theta_3 = 0$	0.980	1.000	1.000	0.644	1.000	1.000
V3	$\theta_2 = 0$	0.994	1.000	1.000	0.638	1.000	1.000
	$\theta_3 = 0$	0.914	1.000	1.000	0.610	1.000	1.000
V4	$\theta_2 = 0$	0.978	1.000	1.000	0.836	1.000	1.000
	$\theta_3 = 0$	0.978	1.000	1.000	0.704	1.000	1.000
	$\theta_4 = 0$	0.970	1.000	1.000	0.672	1.000	1.000
V1	$\theta_2 = \theta_3$	0.994	1.000	1.000	0.726	1.000	1.000
V2	$\theta_2 = \theta_3$	0.984	1.000	1.000	0.628	1.000	1.000
V3	$\theta_2 = \theta_3$	0.890	1.000	1.000	0.502	0.996	1.000
V4	$\theta_2 = \theta_3$	0.754	1.000	1.000	0.658	1.000	1.000

0, $k = 1, ..., d$. The statistic for the test is $Z = \frac{\hat{\theta}_j}{se(\hat{\theta}_j)}$, where se indicates the estimated standard error. The test 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 among the 500 samples obtained from each analysed conditional distribution and sample size.

Alternatively, we also analyse the power of the test where 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 $Z = \frac{\hat{\theta}_2 - \hat{\theta}_3}{se(\hat{\theta}_2 - \hat{\theta}_3)}$.

In Tables 8 and 9 the powers of the two proposed tests are shown for symmetric and skewed distributions, respectively. Both tests are at 95% confidence level. Focusing on symmetric distributions, we observe that with $n = 100$ the power is the poorest especially for a conditional logistic distribution; however, for $n = 500$ the power reaches 1 in almost all cases.

The results for skew distributions in Table 9 indicate that when $n = 500$ the power decreases considerably for Weibull and Champernowne distributions with $\alpha = 1$.

Table 9: Power of the statistic for skew distributions.

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

4 Application to the analysis of automobile accident costs

This illustration is inspired by a current problem in the analysis of data from insurance companies. The cost of an automobile accident is difficult to predict because it is linked to incidental circumstances that occur in conjunction with the collision. There are many examples of such conditions that are generally out of the control of insurers. Indeed, many circumstances cannot be predicted in advance, but may increase losses dramatically. For example, the presence of a truck, speed excess, heavy traffic, bad quality of the road at the point of the accident or adverse weather conditions. If an accident occurs when there are several passengers in the car, then there may be more victims than if the car has only the driver and no additional passengers. Consequently, the compensation cost for bodily injury is larger when more people sit on the car than when no passengers are involved.

Traditionally, insurers have identified some covariates that indirectly capture the risk of a large claim cost, such as the driving zone or the car type, but the predictive power of such covariates is rather low. In the context of actuarial statistics, it is well known that it is simpler to predict the number of accidents per year than the cost of those accidents.

In order to estimate the expected costs for one specific insurance policy, insurers usually predict the expected number of claims and multiply it by the average claim cost. This straightforward calculation is called the *pure premium*, which is smaller than the final price the customer pays because all kinds of expenses have to added to this price, namely general expenses, margins and solvency requirements, advertising, commissions and so on.

In an attempt to model the cost per accident in a simple way, insurers end up using the average claim cost for their predictions, and they do not model the cost distribution conditional on the information of the covariates. When modelling claim costs, generalised linear models are not as good in as they are for the frequency of claims, so the analysis of costs is generally done in a univariate framework and no covariate information is considered. Moreover, besides being necessary to calculate the premium that should be charged to each policyholders per year, cost distribution is also needed to estimate the heterogeneity of reserves that the insurance company must hold in order to have enough resources to pay for the compensations of the reported claims.

In this section, we analyse the cost per accident distribution conditional on the covariates and use single-index models to explain the influence of risk factors on the statistical distribution of the accident cost in a real case study. We show that single-index models provide a new tool to identify the influence of some covariates that are known to the insurance company at the beginning of the contract or during the coverage period. The proposed flexible maximum conditional likelihood method is used for estimating the parameters associated with the covariates and the smoothing parameters.

We analyse a data set from a Spanish insurance company that contains a sample of 974 policyholders with car insurance that have submitted at least one claim in the analysed period of one year. The data were collected in 2011. These claims correspond to accidents with third party liability.

Our aim is to analyse the conditional distribution of the annual cost per claim based on the

Table 10: Descriptive statistics of the variables in the claim costs dataset.

	Mean	Std.	Min.	Q25	Median	Q75	Max.
cost	1.657	4.579	0.018	0.418	0.817	1.874	130.870
Log(cost)	-0.149	1.102	-4.031	-0.873	-0.202	0.628	4.874
age	26.875	3.160	20.060	24.465	26.667	29.495	34.067
agelic	6.340	2.843	1.859	4.305	5.743	7.977	14.754
agecar	8.752	4.133	1.944	5.615	7.775	11.358	20.468
parking	0.760	0.427	0.000	1.000	1.000	1.000	1.000
tkm	8.116	4.461	0.560	5.063	7.320	10.570	42.022
nightkm	7.538	6.305	0.000	2.979	5.884	10.380	42.830
urbankm	28.214	14.463	2.686	17.263	26.070	36.582	94.998
speedkm	6.886	6.774	0.000	2.109	4.698	9.041	48.002

Q25 and Q75 are the first and third quartiles.

characteristics of the insured. For each policyholder we have information about the following covariates (labels between parenthesis): cost per claim in thousands of euros (cost), age in years (age), age of driving licence in years (agelic), age of the car in years (agecar), A binary indicator which equals 1 if car is parked in the garage at night or 0 otherwise (parking), annual kilometers driven (tkm) in thousands, percentage of kilometers driven at night (nightkm), percentage of kilometers on urban roads (urbankm) and percentage of kilometers above the speed limit (speedkm). These data correspond to a sample of insured customers for whom the company collected information on the driving behaviour with a telematics device installed in their vehicle. Therefore, the information on total distance driven during one year, the percentage of distance in urban versus non urban, the percentage of distance driven in nighttime hours compared to daytime, and the percentage of distance driven above the speed limit, correspond to the so-called "telematics covariates" that capture the driving style and patterns of the policy holder. We do not include the gender variable in the model because the European regulation prohibits discrimination between men and women in insurance premiums (see, Guillen et al., 2018, for more information on the data).

In Table 10 we present descriptive statistics of the variable cost per claim in the original scale, transformed into logarithm (Log(cost)), and information on the covariates.

The index-models that we estimate in this section are fitted using the “Log(cost)” as dependent variable. In Figure 1 we show the dispersion plots of the dependent variable “Log(cost)” versus each covariate. We observe that is difficult to find a clear association pattern between the claim cost and the covariates. The mean of the dependent variable “Log(cost)” seems to remain constant for the different values of the covariates.

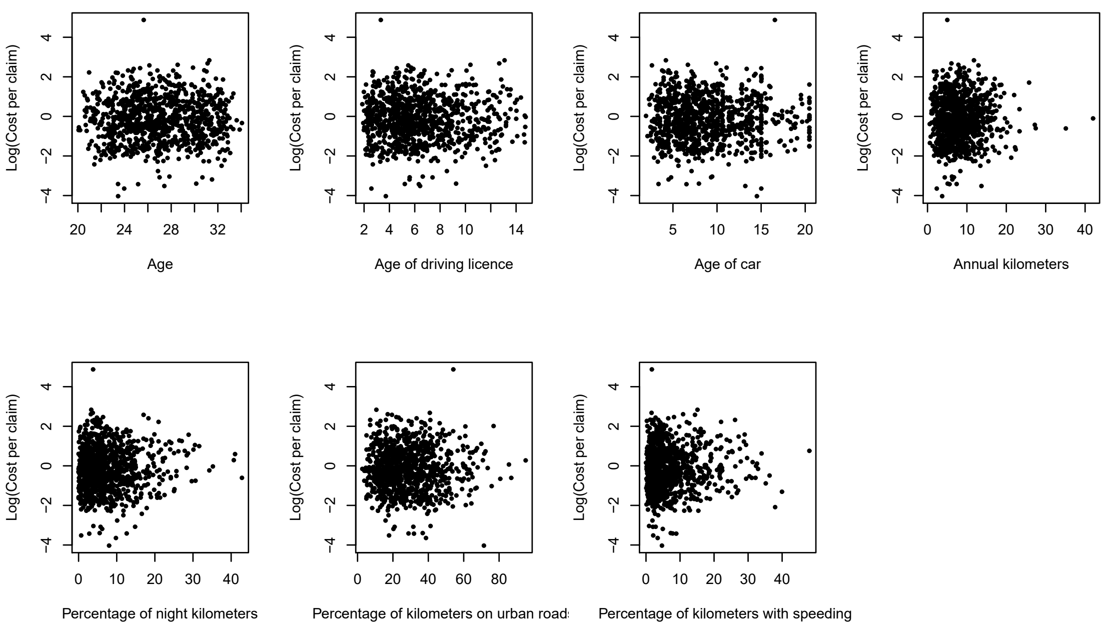

Figure 1: Dispersion plots of the log-transformed cost per claim versus the covariates.

In Table 11 we show the results of the estimated parameters ($\hat{\theta}$) of the single-index models obtained using different covariate vectors: all explanatory variables, only telematics variables and only traditional rating factors, i.e. non-telematics covariates. We set “speedkm” as the variable with the constrained coefficient $\theta_1 = 1$ and, for the model with non-telematics variables we use “age”. This is convenient because the nature of the covariates makes interpretation straightforward in this context. The reason to fix the effect of the speed variable is that we believe that high speed levels result in higher risk of severe accidents, which in turn is more costly than a minor accident. For each estimated parameter $\hat{\theta}_j,\ j = 2, ..., 8$ we test its

Table 11: Estimated parameters and their significance for the single index-model in the accident cost data set.

	Model		
	All variables	Only telematics	Only non-telematics
speedkm	1.000	1.000	–
age	-0.021	–	1.000
agelic	0.005	–	0.115**
agecar	-0.011	–	-0.031**
parking	-0.019	–	-0.276**
tkm	-0.017*	-0.013**	–
nightkm	0.130**	0.130**	–
urbankm	0.051**	0.052**	–
Significant at 5% level * and at 1% level **			

individual significance. As it is defined at the bottom of Table 11, we use asterisks to indicate if the estimated coefficient is significantly different from zero.

When looking at the results on Table 11 we observe that the effect of telematics variables does not change if we compare the models with all variables and only telematics variables. This indicates that in a single-index model for the cost per claim, what matters most is the driving pattern, however, all the no-telematics variables have a significant effect when telematics variables are excluded from the model.

As we show in Section 2, given that we assume $\theta_1 = 1$, even if the signs of the coefficients of the explanatory variables are not identified, we can analyse the relation between these effects; for example, on the one hand, in Table 11 we observe that "tkm" has an opposite effect to "speedkm", i.e. we interpret that excess speed can be compensated for by driving experience, which means total distance driven. On the other hand, the coefficients of "nightkm" and "urbankm" have the same sign as the coefficient of "speedkm". So, if higher percentage of excess speed implies higher value of the index, then the same happens when the night time driving and/or the urban driving increase.

To analyse the results with more detail, we use different plots that are shown in the original scale of the cost per accident rather than the logarithm. In Figures 2, 3 and 4 we plot the index versus the fitted mean with each model, the median and p-th quantiles with $p = 0.90$, $p = 0.95$ and $p = 0.99$. The curve for the mean is estimated using the Nadaraya-Watson estimator of the regression function between the dependent variable and the estimated linear index. The median and the quantiles are estimated from the inverse of the estimated conditional distribution function. The smoothing parameters are calculated specifically for each estimated curve. The main result is that cost distribution conditional on the value of the index is not constant. This is an evidence that there are some combinations of the covariates that lead to a conditional distribution of the cost with longer tails than others. This feature is not captured by the mean curve, which is flat, thus showing that the use of a single-index model prediction can be helpful to insurance companies to set up wider margins corresponding to the values of the predicted in the intervals where the conditional distribution has a remarkable heavy tail.

When comparing the plots of the models with all the variables (Figure 2) and with only the traditional rating variables (Figure 4), we observe the benefits of including "telematics"' regressors that measure the driving patterns. By doing so, the intervals of the index that correspond to a conditional distribution that has longer tails are easily identified and, as a consequence, in those cases the insurance company should expect a slight increase of the median cost and a large increase of the upper quantiles. In the model with telematics information. Special attention should be given to those policyholders for whom the index is slightly above 5 or between 15 and 20 (Figure 3).

In this dataset, there is one extreme observation for the response variable, which corresponds to an accident claim that exceeded 130 thousand euros. The results without this extreme cost are shown in the Appendix. In that case, for the model with all variables, "age" and "agecar" are significantly different from zero in the model with all variables, but the rest of conclusions are stable, which means that the method is quite robust under the presence of an outlier observation in the dependent variable.

An additional interest of the results in this case study is that the single-index value provides a one-dimensional summary of the characteristics that discriminate the policyholders in terms of the conditional cost distribution.

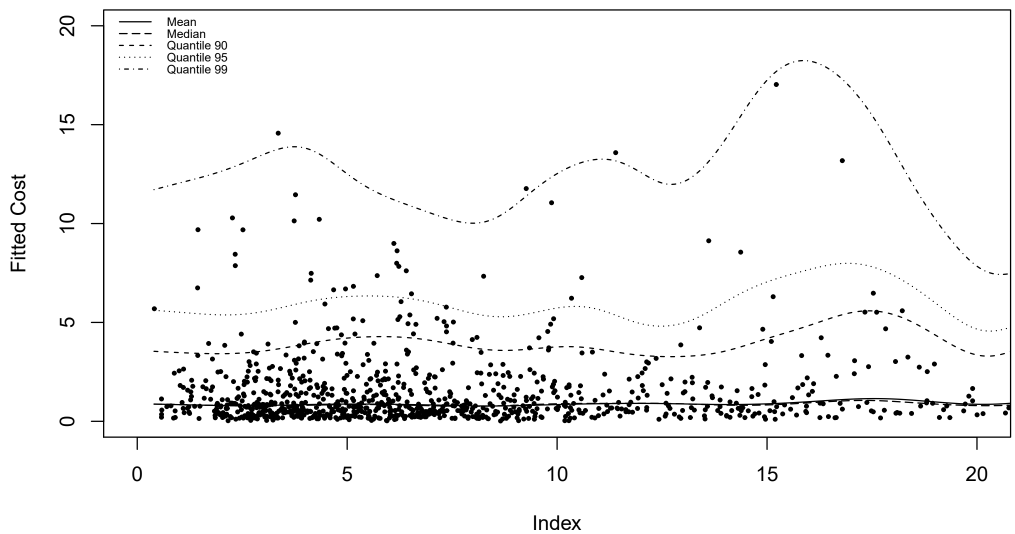

Figure 2: Fitted values of the conditional mean and quantiles with all variables.

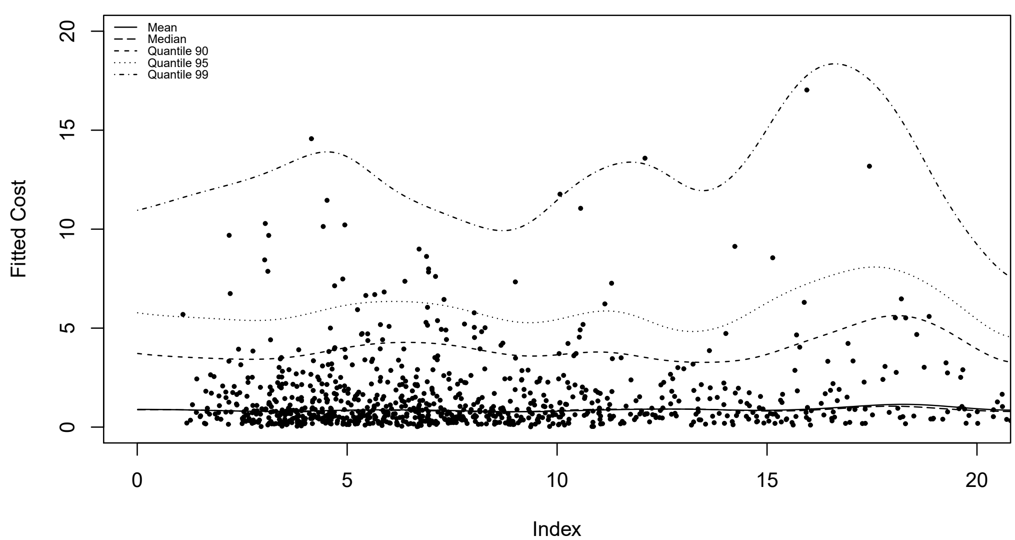

Figure 3: Fitted values of the conditional mean and quantiles with telematics variables.

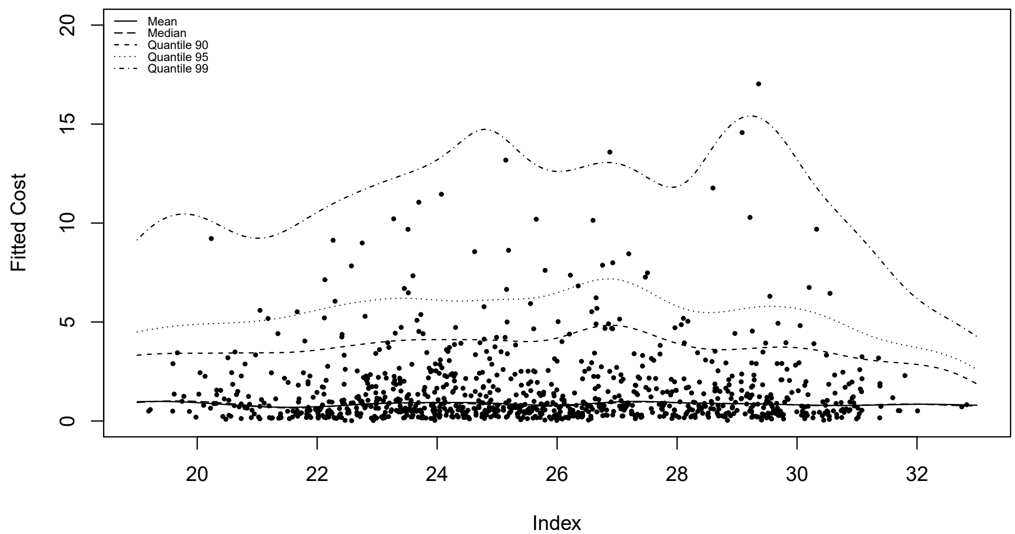

Figure 4: Fitted values of the conditional mean and quantiles with non-telematics variables.

5 Conclusions

The continuous evolution of big-data confronts data analysts with the challenge of studying phenomena much more complex to represent than in the past; for example, the existence of extremes or the possibility of having heterogeneity not only in the mean of a statistical distribution of interest, but also in its whole domain.

A limitation associated with the traditional approach in generalised linear modeling 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. As opposed to this principle, our method provides a full specification of the conditional distribution, while preserving the nature of the single-index abd flexibility.

In many contexts, heterogeneity might be associated with the shape of the distribution and not so much with the location. This is precisely the example application shown in the case study section. An index-model allows to analyse all the parts of the motor insurance claim cost distribution: namely, the mathematical expectation, the median, the quantiles, and so on.

We have developed an estimator for the single-index model based on maximising the estimated conditional likelihood. We have used this approach to estimate the conditional distribution and, in particular, its quantiles. This is fundamental in data analysis, given that in some applications the focus on the mean has lost interest in front of other characteristics of the distribution.

From the expression of the marginal effect of a covariate on a given quantile, we analyse a way to interpret the estimated parameters of the index. Since this is not the aim of our analysis, we have not done a full analysis of the marginal effect estimator, but we highlight its importance.

Our main theoretical results prove the asymptotic properties of the estimator for vector of parameters in the single-index model and provide an expression for its covariance matrix. With a simulation study we prove the power of the inference using the kernel estimator of the covariance matrix. These results are fundamental in situations where the analyst does not have any prior knowledge that can identify the variables that actually cause changes in the distribution of the dependent variable.

A simulation study shows how our method improves with respecto to the finite sample properties of some known alternative methods, especially when the conditional distribution is skewed and has a long right tail. This type of distribution is frequent in economic variables that measure revenues and expenses. The proposed estimator improves considerably the analysed alternatives, showing greater robustness in the presence of extreme values. However, for this shape of the distribution, in small sample the results are still not good and, therefore, alternative procedures for improving finite sample properties should be studied.

In the application, the observed characteristics of the insured drivers can be used to understand the distribution of the cost of claims. Second, if these models are implemented in practice, they will allow insurers to combine the cost per claim distribution with the predicted expectation on the number of claims, which is currently the baseline for premium calculation because it depends on covariates such as age, year of driving license, power of the vehicle, age of the vehicle, and so on. Moreover, when driving behaviour information is included in the model, such as distance driven or driving habits, our approach offers an opportunity to identify the vales of the single-index that correspond to a long-tailed cost distribution, and therefore to find situations where the probability of observing a large claim increases.

Acknowledgements

The support received by the Ministry of Economy and Competitiveness in Grant ECO2016-76203-C2-2-P for the first and third authors is gratefully acknowledged. The research of the second author has been supported by MINECO Grants MTM2014-52876-R and MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the ERDF. All authors declare no conflict of interest as no sponsor has been involved in the implementation and conclusions of the research.

Supplementary Materials

Appendix A Proofs of Theorems 1 - 3. (PDF file)

Appendix B Results without an extreme cost value. (PDF file)

References

Bashtannyk, D.M., 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., Bolancé, C., 2005. Kernel density estimation for heavy-tailed distributions using the Champernowne transformation. Statistics 39, 503–518.

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

Guillen, M., Nielsen, J.P., Ayuso, M., Pérez-Marín, A.M., 2018. The use of telematics devices to improve automobile insurance rates. Risk Analysis , Accepted (in press).

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

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

Härdle, W., 1990. Applied Nonparametric Regression. Cambridge University Press, UK.

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

Hastie, T.J., Tibshiran, R., 1990. Generalized Additive Models. Chapman & Hall/CRC, London.

Horowitz, J.L., 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., Spokoiny, V., 2001. Direct estimation of the index coefficient in a single-index model. Annals of Statistics 29, 595–623.

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

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

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

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

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

Appendix A Proofs of Theorems 1 - 3. (PDF file)

Proof of Theorem 1.

Let us consider θ_0 implicitely defined by (2) and fix some other θ. Then, using Jensen's inequality, (4) and (9),

$$\begin{aligned} l(\theta) - l(\theta_0) &= E(\tilde{l}_n(\theta_0)) - E(\tilde{l}_n(\theta)) = E\left(\log\left(\frac{f_\theta(Y|\theta' X)}{f_{\theta_0}(Y|\theta_0' X)} \right) \right) \\ &\leq \log\left(E\left(\frac{f_\theta(Y|\theta' X)}{f_{\theta_0}(Y|\theta_0' X)} \right) \right) = \log\left(\int\int \frac{f_\theta(y|\theta'\mathbf{x})}{f_{\theta_0}(y|\theta_0'\mathbf{x})} f_{\theta_0}(y|\theta_0'\mathbf{x}) f_X(\mathbf{x}) d\mathbf{x} dy \right) \\ &= \log\left(\int \left(\int f_\theta(y|\theta'\mathbf{x}) dy \right) f_X(\mathbf{x}) d\mathbf{x} \right) = \log 1 = 0. \end{aligned}$$

So $l(\theta) \leq l(\theta_0)\ \forall\theta$, which completes the proof. ⊠

Proof of Lemma 1. Using (10) and (8) we have

$$l^{[1]}(\theta_0) = 0 \ \text{ and } \ \hat{l}_n^{[1]}(\hat{\theta}_n) = 0.$$

Now a Taylor expansion gives

$$l^{[1]}(\theta_0) = 0 = \hat{l}_n^{[1]}(\hat{\theta}_n) = \hat{l}_n^{[1]}(\theta_0) + \hat{l}_n^{[2]}(\hat{\theta}_n^*)(\hat{\theta}_n - \theta_0),$$

where $\hat{\theta}_n^*$ is between $\hat{\theta}_n$ and θ_0. This completes the proof. ⊠

The proof of Theorem 2 is postponed. Let us first state and prove an auxiliary lemma.

Lemma 2. *Under the conditions in Theorem 3 we have*

$$\tilde{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0) \xrightarrow{P} 0$$

and

$$\sqrt{n}(\tilde{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0)) \to \mathcal{N}(0, \Sigma_1)$$

Proof. According to (4) and since $l^{[1]}(\theta_0) = E(\tilde{l}_n^{[1]}(\theta_0)) = 0$, we have

$$\sqrt{n}(\tilde{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{f_\theta^{[1]}(Y_i|\theta' X_i)_{\theta=\theta_0}}{f_{\theta_0}(Y_i|\theta_0' X_i)} \right].$$

Then, $\tilde{l}_n^{[1]}(\theta) - l^{[1]}(\theta)$ is a sum of i.i.d. random vectors and the Law of Large Numbers gives the convergence to zero in probability. Now the Central Limit Theorem leads to

$$\sqrt{n}(\tilde{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0)) \to \mathcal{N}(0, \Sigma_1)$$

The matrix Σ_1 is easily computed:

$$\begin{aligned}\Sigma_1 &= E\Bigg[\left(\frac{f_\theta^{[1]}(Y_1|\theta' X_1)_{\theta=\theta_0}}{f_{\theta_0}(Y_1|\theta_0' X_1)}\right)\left(\frac{f_\theta^{[1]}(Y_1|\theta' X_1)_{\theta=\theta_0}}{f_{\theta_0}(Y_1|\theta' X_1)}\right)^t\Bigg]\\ &= \int (\nabla_\theta \log(f_\theta(z|\theta'\mathbf{x}))_{\theta=\theta_0})(\nabla_\theta \log(f_\theta(z|\theta'\mathbf{x}))_{\theta=\theta_0})^t f(\mathbf{x},z)d\mathbf{x}dz.\end{aligned}$$

Observe that the last equation is a consequence of Lemma 4 below. ☒

Proof of Theorem 3. In view of Lemma 1, the term

$$\hat{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0) = \alpha_n(\theta_0) + \beta_n(\theta_0),$$

has to be studied, where

$$\alpha_n(\theta_0) = \hat{l}_n^{[1]}(\theta_0) - \tilde{l}_n^{[1]}(\theta_0) \tag{11}$$

$$\beta_n(\theta_0) = \tilde{l}_n^{[1]}(\theta_0) - l^{[1]}(\theta_0).$$

To deal with $\beta_n(\theta_0)$, we use Lemma 2, to prove

$$\sqrt{n}\beta_n(\theta_0) \to \mathcal{N}(0, \Sigma_1)$$

and

$$\beta_n(\theta_0) \to 0 \ \text{ in probability.}$$

Concerning $\alpha_n(\theta)$, using (4), (7) and (5), we have

$$\begin{aligned}&\hat{l}_n(\theta_0) - \tilde{l}_n(\theta_0)\\ &= \frac{1}{n}\sum_{i=1}^n \Big[(\log(\hat{r}^{-i}(\theta_0' X_i, Y_i)) - \log(f_{\theta_0}(\theta_0' X_i, Y_i))) + (\log(f_{\theta_0' X}(\theta_0' X_i)) - \log(\hat{s}^{-i}(\theta_0' X_i)))\Big],\end{aligned}$$

which, in view of (11), implies

$$\begin{aligned}&\hat{l}_n^{[1]}(\theta_0) - \tilde{l}_n^{[1]}(\theta_0)\\ &= \frac{1}{n}\sum_{i=1}^n \Bigg[\left(\frac{\hat{r}^{-i,[1]}(\theta_0' X_i, Y_i)}{\hat{r}^{-i}(\theta_0' X_i, Y_i)} - \frac{f_{\theta_0}^{[1]}(\theta_0' X_i, Y_i)}{f_{\theta_0}(\theta_0' X_i, Y_i)}\right) + \left(\frac{f_{\theta_0' X}^{[1]}(\theta_0' X_i)}{f_{\theta_0' X}(\theta_0' X_i)} - \frac{\hat{s}^{-i,[1]}(\theta_0' X_i)}{\hat{s}^{-i}(\theta_0' X_i)}\right)\Bigg].\end{aligned}$$

Now using

$$\frac{1}{\hat{r}^{-i}} = \frac{1}{f_{\theta_0}} + \frac{f_{\theta_0} - \hat{r}^{-i}}{f_{\theta_0}\hat{r}^{-i}}, \quad \frac{1}{\hat{s}^{-i}} = \frac{1}{f_{\theta_0' X}} + \frac{f_{\theta_0' X} - \hat{s}^{-i}}{f_{\theta_0' X}\hat{s}^{-i}}$$

we obtain

$$\hat{l}_n^{[1]}(\theta_0) - \tilde{l}_n^{[1]}(\theta_0) = \sum_{k=1}^{8} A_{kn} + o_P(n^{-1/2}), \tag{12}$$

where

$$\begin{aligned}
A_{1n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{r}^{-i,[1]}(\theta_0'X_i, Y_i) - f_{\theta_0}^{[1]}(\theta_0'X_i, Y_i)}{f_{\theta_0}(\theta_0'X_i, Y_i)} \\
A_{2n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\theta_0}^{[1]}(\theta_0'X_i, Y_i)}{f_{\theta_0}^2(\theta_0'X_i, Y_i)}(f_{\theta_0}(\theta_0'X_i, Y_i) - \hat{r}^{-i}(\theta_0'X_i, Y_i)) \\
A_{3n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\theta_0'X}^{[1]}(\theta_0'X_i) - \hat{s}^{-i,[1]}(\theta_0'X_i)}{f_{\theta_0'X}(\theta_0'X_i)} \\
A_{4n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{f_{\theta_0'X}^{[1]}(\theta_0'X_i)}{f_{\theta_0'X}^2(\theta_0'X_i)}(\hat{s}^{-i}(\theta_0'X_i) - f_{\theta_0'X}(\theta_0'X_i)) \\
A_{5n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(\hat{r}^{-i,[1]}(\theta_0'X_i, Y_i) - f_{\theta_0}^{[1]}(\theta_0'X_i, Y_i))(f_{\theta_0}(\theta_0'X_i, Y_i) - \hat{r}^{-i}(\theta_0'X_i, Y_i))}{f_{\theta_0}^2(\theta_0'X_i, Y_i)} \\
A_{6n} &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log(\hat{r}^{-i}(\theta'X_i, Y_i))|_{\theta=\theta_0} \frac{(f_{\theta_0}(\theta_0'X_i, Y_i) - \hat{r}^{-i}(\theta_0'X_i, Y_i))^2}{f_{\theta_0}^2(\theta_0'X_i, Y_i)} \\
A_{7n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(f_{\theta_0'X}^{[1]}(\theta_0'X_i) - \hat{s}^{-i,[1]}(\theta_0'X_i))(\hat{s}^{-i}(\theta_0'X_i) - f_{\theta_0'X}(\theta_0'X_i))}{f_{\theta_0'X}^2(\theta_0'X_i)} \\
A_{8n} &= \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \log(\hat{s}^{-i}(\theta'X_i))|_{\theta=\theta_0} \frac{(f_{\theta_0'X}(\theta_0'X_i) - \hat{s}^{-i}(\theta_0'X_i))^2}{f_{\theta_0'X}^2(\theta_0'X_i)}
\end{aligned}$$

To deal with each of these terms separately, let us define the following functions:

$$\begin{aligned}
\tilde{r}(\theta'\mathbf{x}, y) &= \frac{1}{h_1 h_2}\int K\left(\frac{\theta'(\mathbf{x}-\mathbf{u})}{h_1}\right) K\left(\frac{y-v}{h_2}\right) f(\mathbf{u}, v)d\mathbf{u}dv, \qquad (13) \\
\tilde{s}(\theta'\mathbf{x}) &= \frac{1}{h_1}\int K\left(\frac{\theta'(\mathbf{x}-\mathbf{u})}{h_1}\right) f(\mathbf{u}, v)d\mathbf{u}dv
\end{aligned}$$

For instance, to deal with A_{1n} we telescope using $\tilde{r}^{[1]}(\theta_0'X_i, Y_i)$ to obtain $A_{1n} = B_{1n} + C_{1n}$, where

$$\begin{aligned}
B_{1n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{r}^{-i,[1]}(\theta_0'X_i, Y_i) - \tilde{r}^{[1]}(\theta_0'X_i, Y_i)}{f_{\theta_0}(\theta_0'X_i, Y_i)} \qquad (14) \\
C_{1n} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\tilde{r}^{[1]}(\theta_0'X_i, Y_i) - f_{\theta}^{[1]}(\theta_0'X_i, Y_i)}{f_{\theta_0}(\theta_0'X_i, Y_i)} \qquad (15)
\end{aligned}$$

and $A_{2n} = B_{2n} + C_{2n}$, where

$$B_{2n} = \frac{1}{n}\sum_{i=1}^{n} \frac{f^{[1]}_{\theta_0}(\theta_0' X_i, Y_i)}{f^2_{\theta_0}(\theta_0' X_i, Y_i)} (\tilde{r}^{-i}(\theta_0' X_i, Y_i) - \hat{r}(\theta_0' X_i, Y_i)) \tag{16}$$

$$C_{2n} = \frac{1}{n}\sum_{i=1}^{n} \frac{f^{[1]}_{\theta_0}(\theta_0' X_i, Y_i)}{f^2_{\theta_0}(\theta_0' X_i, Y_i)} (f_{\theta_0}(\theta_0' X_i, Y_i) - \tilde{r}(\theta_0' X_i, Y_i)).$$

In a parallel way $A_{in} = B_{in} + C_{in}$ for $i = 3, 4$. Just defining $B_n = B_{1n} + B_{2n} + B_{3n} + B_{4n}$, $C_n = C_{1n} + C_{2n} + C_{3n} + C_{4n}$ and $D_n = A_{5n} + A_{6n} + A_{7n} + A_{8n}$ and using equation (12) we have

$$\alpha_n(\theta_0) = B_n + C_n + D_n + o_P(n^{-1/2}).$$

Now using Lemmas 5, 8 and 9 imply that

$$\sqrt{n} C_n = o_P(1), \quad \sqrt{n} B_n = o_P(1), \quad \sqrt{n} D_n = o_P(1).$$

Finally, Lemma 10 can be used to prove $\hat{l}^{[2]}_n(\theta) \to l^{[2]}(\theta)$. This completes the proof. ⊠

Next we consider the gradients of f_θ. We have following results:

Lemma 3. *Under A6, we have*

a)

$$\frac{\partial}{\partial \theta_k} f_\theta(\theta' \mathbf{x}, y)_{\theta=\theta_0} = \frac{\partial}{\partial t}\left(f_{\theta_0}(t, y)\left[\mathbf{x}_k - E(X_k|\theta_0' X = t, Y = y)\right]\right)_{t=\theta_0'\mathbf{x}}.$$

b)

$$\frac{\partial}{\partial \theta_k} f_{\theta' X}(\theta' \mathbf{x})_{\theta=\theta_0} = \frac{\partial}{\partial t}\left(f_{\theta_0' X}(t)\left[\mathbf{x}_k - E(X_k|\theta_0' X = t)\right]\right)_{t=\theta_0'\mathbf{x}}$$

Proof. We will prove only part a). The proof of b) is very similar. Let us consider $\theta_0 = (\theta_{01}, ..., \theta_{0d})'$ and $\theta^{(k)}_{0,\delta} = \theta_0 + (0, ..., \delta, ..., 0)'$, where the δ is in position k. Then, for every $\mathbf{x} = (x_1, ..., x_d)'$ and $y \in \mathbb{R}$, we have

$$\frac{\partial}{\partial \theta_k} f_\theta(\theta' \mathbf{x}, y)_{\theta=\theta_0} = \beta_1 + \beta_2 + \beta_3, \tag{17}$$

where

$$\beta_1 = \lim_{\delta\to 0} \frac{f_{\theta^{(k)}_{0,\delta}}(\theta_0'\mathbf{x}, y) - f_{\theta_0}(\theta_0'\mathbf{x}, y)}{\delta}$$

$$\beta_2 = \lim_{\delta\to 0} \frac{f_{\theta_0}(\theta^{(k)'}_{0,\delta}\mathbf{x}, y) - f_{\theta_0}(\theta_0'\mathbf{x}, y)}{\delta} \tag{18}$$

$$\beta_3 = \lim_{\delta\to 0} \left[\frac{f_{\theta^{(k)'}_{0,\delta}}(\theta^{(k)'}_{0,\delta}\mathbf{x}, y) - f_{\theta_0}(\theta^{(k)'}_{0,\delta}\mathbf{x}, y)}{\delta} - \frac{f_{\theta^{(k)}_{0,\delta}}(\theta_0'\mathbf{x}, y) - f_{\theta_0}(\theta_0'\mathbf{x}, y)}{\delta} \right]. \tag{19}$$

Using two Taylor expansions, the term in (19) becomes

$$\begin{aligned}\beta_3 &= \lim_{\delta\to 0}\left[\frac{\partial}{\partial\theta_k} f_\theta(\theta_{0,\delta}^{(k)'}\mathbf{x}, y) - \frac{\partial}{\partial\theta_k} f_\theta(\theta_0'\mathbf{x}, y)\right] \\ &+ \lim_{\delta\to 0}\frac{\delta}{2}\left[\frac{\partial^2}{\partial\theta_k^2} f_\theta(\theta_{0,\delta}^{(k)'}\mathbf{x}, y)_{\theta=\tilde\theta} - \frac{\partial^2}{\partial\theta_k^2} f_\theta(\theta_0'\mathbf{x}, y)_{\theta=\tilde{\tilde\theta}}\right] = 0\end{aligned}$$

where $\tilde\theta = \theta_0 + (0, ..., \tilde\delta, ..., 0)'$, $\tilde{\tilde\theta} = \theta_0 + (0, ..., \tilde{\tilde\delta}, ..., 0)'$ and $\tilde\delta$ and $\tilde{\tilde\delta}$ are intermediate points between 0 and δ and the last step comes from the continuity of the first and second partial derivatives (see Condition A6).

A direct inspection of (18) leads to

$$\beta_2 = \frac{\partial}{\partial\theta_k} f_{\theta_0}(\theta'\mathbf{x}, y)_{\theta=\theta_0} = x_k\frac{\partial}{\partial t} f_{\theta_0}(t, y)_{t=\theta_0'\mathbf{x}}.$$

On the other hand

$$\beta_1 = \frac{\partial}{\partial\theta_k} f_\theta(\theta_0'\mathbf{x}, y)_{\theta=\theta_0} = \frac{\partial}{\partial u}\left(\frac{\partial}{\partial v}\left(\frac{\partial}{\partial\theta_k} P(\theta'X \le u, Y \le v)_{\theta=\theta_0}\right)_{v=y}\right)_{u=\theta_0'\mathbf{x}}. \tag{20}$$

Let $f^{X_k,\theta_0'X,Y}(x_k, u, \tilde y)$ denote the density of $(X_k, \theta_0'X, Y)$. Now using standard algebra for the inner partial derivative in (20), we obtain

$$\begin{aligned}&\frac{\partial}{\partial\theta_k} P(\theta'X \le u, Y \le v)|_{\theta=\theta_0} = \lim_{\delta\to 0}\frac{P(\theta_0'X + \delta X_k \le u, Y \le v) - P(\theta_0'X \le u, Y \le v)}{\delta} \\ &= \int_{-\infty}^{v}\int_{-\infty}^{\infty}\lim_{\delta\to 0}\frac{1}{\delta}\left[\int_{-\infty}^{u-\delta z_1} f^{X_k,\theta_0'X,Y}(z_1, z_2, z_3)dz_2 - \int_{-\infty}^{u} f^{X_k,\theta_0'X,Y}(z_1, z_2, z_3)dz_2\right] dz_1dz_3 \\ &= -\int_{-\infty}^{v}\int_{-\infty}^{\infty} z_1 f^{X_k,\theta_0'X,Y}(z_1, u, z_3)dz_1dz_3. \end{aligned}\tag{21}$$

Now using (21) in (20) gives

$$\beta_1 = -\frac{\partial}{\partial u}\left(f_{\theta_0}(u, y)E(X_k|\theta_0'X = u, Y = y)\right)_{t=\theta_0'\mathbf{x}} \tag{22}$$

Using (19), (18) and (22) in (17) the proof of a) is concluded. ⊠

Lemma 4. *Under A2, we have*

a)

$$\nabla_\theta f_\theta(y|\theta'\mathbf{x})_{\theta=\theta_0} = \left[\mathbf{x} - E(X|\theta_0'X = \theta_0'\mathbf{x})\right]\frac{\partial}{\partial t} f_{\theta_0}(y|t)_{t=\theta_0'\mathbf{x}}.$$

b)

$$E(\nabla_\theta[\log f_\theta(Y_i|\theta' X_i)]_{\theta=\theta_0}|Y_i, \theta_0' X_i) = 0$$

Proof. Part a) is an immediate consequence of Lemma 3. For part b), A2 implies

$$E\Bigg([X_i - E(X|\theta_0' X = \theta_0' X_i)]|Y_i, \theta_0' X_i\Bigg) = E(X|\theta_0' X = \theta_0' X_i) - E(X|\theta_0' X = \theta_0' X_i) = 0.$$

This completes the proof. ⊠

Lemma 5. *Under A2, A4 and A5,* $\sqrt{n} C_n = O_P(\sqrt{n} h_1^4 + \sqrt{n} h_2^2) = o_P(1)$.

Proof. First, using several changes of variables, we have

$$\begin{aligned}
\tilde{r}(\theta_0' X_i, Y_i) 1_{\{Y_i \le a_n\}} &= \frac{1}{h_1 h_2} \int\int K\left(\frac{\theta_0' X_i - u}{h_1}\right) K\left(\frac{Y_i - v}{h_2}\right) f_{\theta_0}(u, v) du dv \\
&= \int\int K(z_1) K(z_2) f_{\theta_0}(\theta_0' X_i - h_1 z_1, Y_i - h_2 z_2) dz_1 dz_2
\end{aligned}$$

and

$$\begin{aligned}
&\tilde{r}^{[1]}(\theta_0' X_i, Y_i) 1_{\{Y_i \le a_n\}} \\
= &\ \frac{1}{h_1^2 h_2} \int\int (X_i - E(X|\theta_0' X = u)) K'\left(\frac{\theta_0' X_i - u}{h_1}\right) K\left(\frac{Y_i - v}{h_2}\right) f_{\theta_0}(u, v) du dv \\
= &\ \frac{1}{h_1} \int (X_i - E(X|\theta_0' X = \theta_0' X_i - h_1 z_1)) K'(z_1) K(z_2) f_{\theta_0}(\theta_0' X_i - h_1 z_1, Y_i - h_2 z_2) dz_1 dz_2.
\end{aligned}$$

Using A5, the function $f_\theta(u, v)$ is three times differentiable. Now applying a Taylor expansion and using $\int z^a K(z) dz = 0$ for a odd, $\int K(z) dz = 1$ and $d_K = \int z^2 K(z) dz < \infty$, together with $\int z^b K'(z) dz = 0$ for b even, $\int z K'(z) dz = -1$ and $\int z^3 K'(z) dz = -3 \int z^2 K(z) dz$ we obtain

$$\begin{aligned}
\tilde{r}(\theta_0' X_i, Y_i) &= f_{\theta_0}(\theta_0' X_i, Y_i) + \frac{d_K h_1^2}{2} \frac{\partial^2}{\partial^2 u} f_{\theta_0}(u, Y_i)_{u=\theta_0' X_i} \\
&\quad + \frac{d_K h_2^2}{2} \frac{\partial^2}{\partial^2 v} \left(f_{\theta_0}(\theta_0' X_i, v) 1_{\{Y_i \le a_n\}}\right)_{v=Y_i} + O_P(h_1^4 + h_1^2 h_2^2 + h_2^4). \qquad (23)
\end{aligned}$$

Similarly, Lemma 3 implies that

$$\begin{aligned}
\tilde{r}^{[1]}(\theta_0' X_i, Y_i) &= f_{\theta_0}^{[1]}(\theta_0' X_i, Y_i) + \frac{d_K h_1^2}{2} \frac{\partial^3}{\partial^3 u} \left[(X_i - E(X|\theta_0' X = u)) f_{\theta_0}(u, Y_i)\right]_{u=\theta_0' X_i} \\
&+ \frac{d_K h_2^2}{2} \frac{\partial}{\partial u} \frac{\partial^2}{\partial^2 v} \left[(X_i - E(X|\theta_0' X = u)) f_{\theta_0}(u, v)\right]_{u=\theta_0' X_i, v=Y_i} \\
&+ O_P(h_1^4 + h_1^2 h_2^2 + h_2^4). \qquad (24)
\end{aligned}$$

Now starting from (15), using (23) and (24) and repeating similar steps for $\tilde{s}(\theta_0' X_i)$, we obtain

$$C_n = \tilde{C}_{1n} + \tilde{C}_{2n} + O_P(h_1^4 + h_2^2), \tag{25}$$

where

$$\tilde{C}_{1n} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_K h_1^2}{2} \tilde{C}_{1i},$$

$$\tilde{C}_{2n} = \frac{1}{n}\sum_{i=1}^{n} \frac{d_K h_1^2}{2} \tilde{C}_{2i},$$

$$\begin{aligned}\tilde{C}_{1i} &= \frac{1}{f_{\theta_0}(\theta_0' X_i, Y_i)} \frac{d^3}{du^3}\Big([X_i - E(X|\theta_0' X = u)] f_{\theta_0}(u, Y_i)\Big)_{u=\theta_0' X_i} \\ &\quad - \frac{1}{f_{\theta_0' X}(\theta_0' X_i)} \frac{d^3}{du^3}\Big([X_i - E(X|\theta_0' X = u)] f_{\theta_0' X}(u)\Big)_{u=\theta_0' X_i}\end{aligned}$$

and

$$\tilde{C}_{2i} = -\frac{\nabla_\theta f_\theta(\theta' X_i, Y_i)_{\theta=\theta_0}}{f_{\theta_0}^2(Y_i, \theta_0' X_i)} \frac{d^2}{du^2} f_{\theta_0}(u, Y_i)|_{u=\theta_0' X_i} + \frac{\nabla_\theta f_{\theta' X}(\theta' X_i)_{\theta=\theta_0}}{f_{\theta_0' X}^2(\theta_0' X_i)} f''_{\theta_0' X}(\theta_0' X_i).$$

We will show that $E(\tilde{C}_{kn}) = 0$ and $Var(\tilde{C}_{kn}) = o(n^{-1})$ for $k = 1, 2$. Since $f_{\theta_0}(z|\theta_0'\mathbf{x}) = f(z|\mathbf{x})$, we have that $f(\mathbf{x}, z) = f_{\theta_0}(\theta_0'\mathbf{x}, z) f_X(\mathbf{x}) / f_{\theta_0' X}(\theta_0'\mathbf{x})$. This equation and condition A2 can be used to obtain:

$$\begin{aligned}E\left(\tilde{C}_{1i}\right) &= \int \frac{1}{f_{\theta_0}(\theta_0'\mathbf{x}, z)} \frac{d^3}{du^3}\Big([\mathbf{x} - E(X|\theta_0' X = u)] f_{\theta_0}(u, y)\Big)_{u=\theta_0'\mathbf{x}} f(y|\mathbf{x}) f_X(\mathbf{x}) d\mathbf{x} dy \\ &\quad - \int \frac{1}{f_{\theta_0' X}(\theta_0'\mathbf{x})} \frac{d^3}{du^3}\Big([\mathbf{x} - E(X|\theta_0' X = u)] f_{\theta_0' X}(u)\Big)_{u=\theta_0'\mathbf{x}} f_X(\mathbf{x}) d\mathbf{x} = 0.\end{aligned}$$

Moreover, the $\tilde{C}_{1i}$ are iid and, since $h_1^2 \to 0$, $Var(\sqrt{n}\tilde{C}_{1n}) \to 0$. Hence $\sqrt{n}\tilde{C}_{1n} \to 0$ in probability. Similar arguments can be used to conclude $\sqrt{n}\tilde{C}_{2n} \to 0$. Now using (25), we have $\sqrt{n} C_n = O_P(h_1^2 + \sqrt{n}(h_1^4 + h_2^2)) = o_P(1)$,which completes the proof. ⊠

Lemma 6. *Under the conditions in Theorem 3 we have* $B_{1n} = o_P(n^{-1/2})$.

Proof. Starting from (14) and using (13) and (6) we obtain

$$\begin{aligned}
B_{1n} &= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{f_{\theta_0}(\theta_0' X_i, Y_i)}\left[\frac{1}{h_1^2 h_2(n-1)}\sum_{j\neq i}(X_i - X_j)K'\left(\frac{\theta_0' X_i - \theta_0' X_j}{h_1}\right)K\left(\frac{Y_i - Y_j}{h_2}\right)\right.\\
&- \left.\frac{1}{h_1^2 h_2}\int\int (X_i - \mathbf{u})K'\left(\frac{\theta_0' X_i - \theta_0'\mathbf{u}}{h_1}\right)K\left(\frac{Y_i - v}{h_2}\right)f(\mathbf{u}, v)\right]d\mathbf{u}dv\\
&= \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n} H_{ij},
\end{aligned}$$

where $K'(t)$ denotes the derivative of K with respect to t and

$$\begin{aligned}
H_{ij} &= \frac{1}{h_1^2 h_2}\frac{1}{f_{\theta_0}(\theta_0' X_i, Y_i)}\left[(X_i - X_j)K'\left(\frac{\theta_0' X_i - \theta_0' X_j}{h_1}\right)K\left(\frac{Y_i - Y_j}{h_2}\right)\right.\\
&\left. - \int\int (X_i - \mathbf{u})K'\left(\frac{\theta_0' X_i - \theta_0'\mathbf{u}}{h_1}\right)K\left(\frac{Y_i - v}{h_2}\right)f(\mathbf{u}, v)d\mathbf{u}dv\right]
\end{aligned}$$

Let us define

$$\bar{H}_j = E(H_{ij}|X_j, Y_j, \delta_j).$$

Then $B_{1n} = B_{1n1} + B_{1n2}$, where

$$B_{1n1} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}(H_{ij} - \bar{H}_j) \tag{26}$$

$$B_{1n2} = \frac{1}{n}\sum_{j=1}^{n}\bar{H}_j. \tag{27}$$

For the term in (26) it is evident that

$$E(H_{ij} - \bar{H}_j) = E(H_{ij} - \bar{H}_j|X_i, Y_i) = E(H_{ij} - \bar{H}_j|X_j, Y_j) = 0,$$

so $E(B_{1n1}) = 0$. On the other hand, long but straightforward calculations can be performed to compute the variance of $B_{1n1}^{(m)}$, the m-th component of the random vector B_{1n1}, which results:

$$\begin{aligned}
Var(B_{1n1}^{(m)}) &= E({B_{1n1}^{(m)}}^2) = \frac{1}{n^2(n-1)^2}\sum_{i=1}^{n}\sum_{j=1, j\neq i}^{n}\sum_{k=1}^{n}\sum_{l=1, l\neq k}^{n} E((H_{ij}^{(m)} - \bar{H}_j^{(m)})(H_{kl}^{(m)} - \bar{H}_l^{(m)}))\\
&= \frac{1}{n(n-1)}[E[(H_{12}^{(m)} - \bar{H}_2^{(m)})^2)] + E[(H_{12}^{(m)} - \bar{H}_2^{(m)})(H_{21}^{(m)} - \bar{H}_1^{(m)})]].
\end{aligned}$$

Using standard algebra, the last two expectations can be proven to be of order $O(h_1^{-3}h_2^{-1})$. As a consequence $Var(B_{1n1}^{(m)}) = O(n^{-2}h_1^{-3}h_2^{-1})$. Now Condition A4 implies $B_{1n1}^{(m)} = o_P(n^{-1/2})$. The term in (27) can be handled in a similar way to prove $E(B_{1n2}) = 0$ and $Var(B_{1n2}^{(m)}) = O(n^{-1}h_1)$. Thus A4 implies $B_{1n2}^{(m)} = o_P(n^{-1/2})$. This concludes the proof. ⊠

Now we state a similar lemma.

Lemma 7. *Under the conditions in Theorem 3 we have* $B_{2n} = o_P(n^{-1/2})$, $B_{3n} = o_P(n^{-1/2})$ *and* $B_{4n} = o_P(n^{-1/2})$.

Proof. Given the similarities between the term B_{1n} in (14) and the term B_{2n} in (16), and also the terms B_{3n} and B_{4n}, the proof goes parallel to the proof of Lemma 6, obtaining $B_{2n} = B_{2n1}+B_{2n2}$, $B_{3n} = B_{3n1}+B_{3n2}$, $B_{4n} = B_{4n1}+B_{4n2}$, where $E[B_{jnk}] = 0$, for $j = 2, 3, 4$ and $k = 1, 2$ and $Var(B_{2n1}) = O(n^{-2}h_1^{-1}h_2^{-1})$, $Var(B_{2n2}) = O(n^{-1}h_1^2)$, $Var(B_{3n1}) = O(n^{-2}h_1^{-3})$, $Var(B_{2n2}) = O(n^{-1}h_1)$, $Var(B_{4n1}) = O(n^{-2}h_1^{-1})$, $Var(B_{4n2}) = O(n^{-1}h_1^2)$. Now assumption A4 implies that the orders in the six variance terms are all $o(n^{-1})$, which concludes the proof. ⊠

Lemma 8. *Under the conditions in Theorem 3,* $B_n = o_P(n^{-1/2})$.

Proof. The proof goes through using the definition of $B_n = B_{1n} + B_{2n} + B_{3n} + B_{4n}$ and Lemmas 6 and 7. ⊠

Lemma 9. *Under the conditions in Theorem 3,* $D_n = o_P(n^{-1/2})$.

Proof. Each of the terms A_{5n}, A_{6n}, A_{7n}, A_{8n} can be bounded using Cauchy-Schwartz inequality. Let us consider A_{5n}:

$$|A_{5n}| \leq A_{5n1}^{1/2} A_{5n2}^{1/2}, \tag{28}$$

where

$$\begin{aligned} A_{5n1} &= \frac{1}{n}\sum_{i=1}^{n} \frac{(\hat{r}^{-i,[1]}(\theta_0' X_i, Y_i) - f_{\theta_0}^{[1]}(\theta_0' X_i, Y_i))^2}{f_{\theta_0}^2(\theta_0' X_i, Y_i)} \\ A_{5n2} &= \frac{1}{n}\sum_{i=1}^{n} \frac{((f_{\theta_0}(\theta_0' X_i, Y_i) - \hat{r}^{-i}(\theta_0' X_i, Y_i))^2}{f_{\theta_0}^2(\theta_0' X_i, Y_i)} \end{aligned}$$

These two terms can be expanded as

$$A_{5n1} = A_{5n11} + A_{5n12} + A_{5n13}, \tag{29}$$

$$A_{5n2} = A_{5n21} + A_{5n22} + A_{5n23}, \tag{30}$$

where

$$\begin{aligned}
A_{5n11} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\hat{r}^{-i,[1]}(\theta_0'X_i,Y_i)-\tilde{r}^{[1]}(\theta_0'X_i,Y_i))^2}{f_{\theta_0}^2(\theta_0'X_i,Y_i)}\\
A_{5n12} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\tilde{r}^{[1]}(\theta_0'X_i,Y_i)-f_{\theta_0}^{[1]}(\theta_0'X_i,Y_i))^2}{f_{\theta_0}^2(\theta_0'X_i,Y_i)}\\
A_{5n13} &= \frac{2}{n}\sum_{i=1}^{n}\frac{(\hat{r}^{-i,[1]}(\theta_0'X_i,Y_i)-\tilde{r}^{[1]}(\theta_0'X_i,Y_i))(\tilde{r}^{[1]}(\theta_0'X_i,Y_i)-f_{\theta_0}^{[1]}(\theta_0'X_i,Y_i))}{f_{\theta_0}^2(\theta_0'X_i,Y_i)}\\
A_{5n21} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(f_{\theta_0}(\theta_0'X_i,Y_i)-\tilde{r}(\theta_0'X_i,Y_i))^2}{f_{\theta_0}^2(\theta_0'X_i,Y_i)}\\
A_{5n22} &= \frac{1}{n}\sum_{i=1}^{n}\frac{(\tilde{r}(\theta_0'X_i,Y_i)-\hat{r}^{-i}(\theta_0'X_i,Y_i))^2}{f_{\theta_0}^2(\theta_0'X_i,Y_i)}\\
A_{5n23} &= \frac{2}{n}\sum_{i=1}^{n}\frac{(f_{\theta_0}(\theta_0'X_i,Y_i)-\tilde{r}(\theta_0'X_i,Y_i))(\tilde{r}(\theta_0'X_i,Y_i)-\hat{r}^{-i}(\theta_0'X_i,Y_i))}{f_{\theta_0}^2(\theta_0'X_i,Y_i)}.
\end{aligned}$$

The terms A_{5n11}, A_{5n12}, A_{5n21} and A_{5n22} can be treated similarly to B_{1n}, C_{1n}, B_{2n}, C_{2n}, respectively. On the other hand, A_{5n13} (respectively A_{5n23}) can be bounded, using Cauchy-Schwartz inequality, in terms of A_{5n11} and A_{5n12} (respectively A_{5n21} and A_{5n22}). All in all results in $A_{5njk}=o_P(n^{-1/2})$ for $j=1,2$ and $k=1,2$. In view of (29) (30) we have $A_{5n1}=o_P(n^{-1/2})$ and $A_{5n2}=o_P(n^{-1/2})$, which, using (28), gives $A_{5n}=o_P(n^{-1/2}$.

Following parallel arguments, straightforward but tedious algebra can be used to prove that $A_{6n}=o_P(n^{-1/2})$, $A_{7n}=o_P(n^{-1/2})$and $A_{8n}=o_P(n^{-1/2})$. Since $D_n=A_{5n}+A_{6n}+A_{7n}+A_{8n}$, this concludes the proof. ⊠

Lemma 10. *Under the conditions in Theorem 3,* $\hat{l}_n^{[2]}(\theta)\to l^{[2]}(\theta)$.

Proof. We have

$$\begin{aligned}
\hat{l}_n^{[2]}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\Bigg(\frac{\hat{r}^{-i,[2]}(\theta'X_i,Y_i)}{\hat{r}^{-i}(\theta'X_i,Y_i)}-\frac{\hat{r}^{-i,[1]}(\theta'X_i,Y_i)(\hat{r}^{-i,[1]}(\theta'X_i,Y_i))^t}{\hat{r}^{-i}(\theta'X_i,Y_i)^2}-\frac{\hat{s}^{-i,[2]}(\theta'X_i)}{\hat{s}^{-i}(\theta'X_i)}\\
&+ \frac{\hat{s}^{-i,[1]}(\theta'X_i)(\hat{s}^{-i,[1]}(\theta'X_i))^t}{\hat{s}^{-i}(\theta'X_i)^2}\Bigg).
\end{aligned}$$

Since

$$\hat{r}^{-i,[2]}(y,\theta'x)=\frac{1}{nh_1^3h_2}\sum_{j=1}^{n}(x-X_j)(x-X_j)^tK''\left(\frac{\theta'x-\theta'X_j}{h_1}\right)K\left(\frac{y-Y_j}{h_2}\right).$$

it is easy to show, that

$$\hat{r}^{-i,[2]}(\theta'x,y)\to\frac{\partial^2}{\partial u^2}\left\{f_\theta(u,y)E((\mathbf{x}-X)(\mathbf{x}-X)^t|\theta'X=u)\right\}_{u=\theta'\mathbf{x}}\quad\text{in probability.}$$

Moreover, using Lemma 3, it can be proven that

$$\frac{\partial^2}{\partial u^2}\left\{f_\theta(u,u)E((\mathbf{x}-X)(\mathbf{x}-X)^t|\theta' X = u)\right\}_{u=\theta' x} = f_\theta^{[2]}(\theta'\mathbf{x}, y).$$

Similarly, it can be shown that $\hat{s}^{[2]}(\theta'\mathbf{x}) \to f_{\theta' X}^{[2]}(\theta'\mathbf{x})$ in probability. This completes the proof. ⊠

Proof of Theorem 2. In view of Lemmas 1, 2 and 10, it remains to show that $\sum_{k=1}^{8} A_{kn} \xrightarrow{P} 0$. But this can be proven following the lines of Lemmas 5, 8 and 9 but even simpler, since only convergence in probability to zero (and no rate) is required. ⊠

Appendix B Results without an extreme cost value. (PDF file)

Table 12: Estimated parameters and their significance for the single index-model in the accident cost data set (without one outlier extreme cost).

	Model		
	All variables	Only telematics	Only non-telematics
speedkm	1.000	–	–
age	-0.093*	–	1.000
agelic	-0.014	–	0.122**
agecar	0.069**	–	-0.039**
parking	0.078	–	-0.171**
tkm	-0.049**	-0.013**	–
nightkm	0.100**	0.130**	–
urbankm	0.030**	0.052**	–
Significant at 5% level * and at 1% level **			

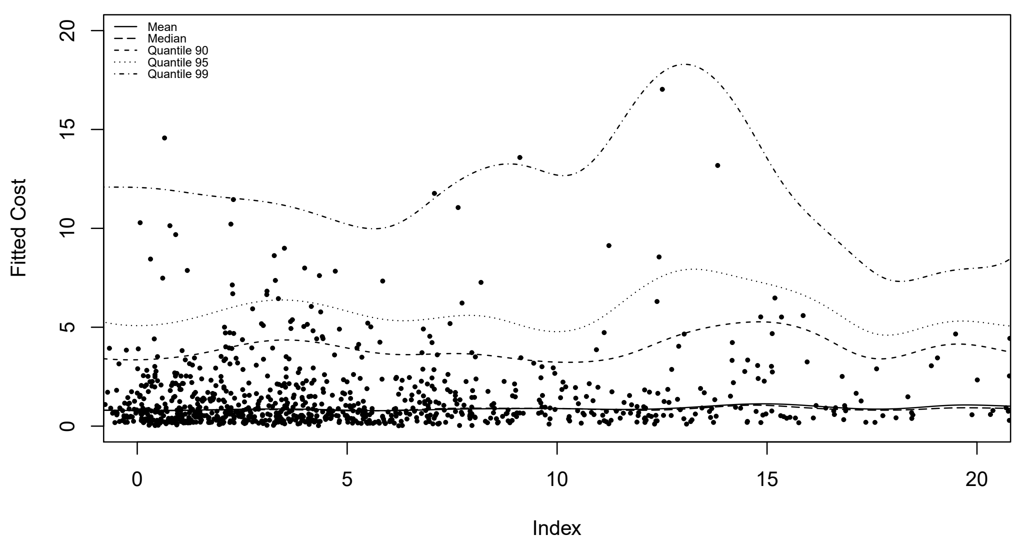

Figure 5: Fitted values of the conditional mean and quantiles with all variables in the accident cost data set without outlier extreme cost.

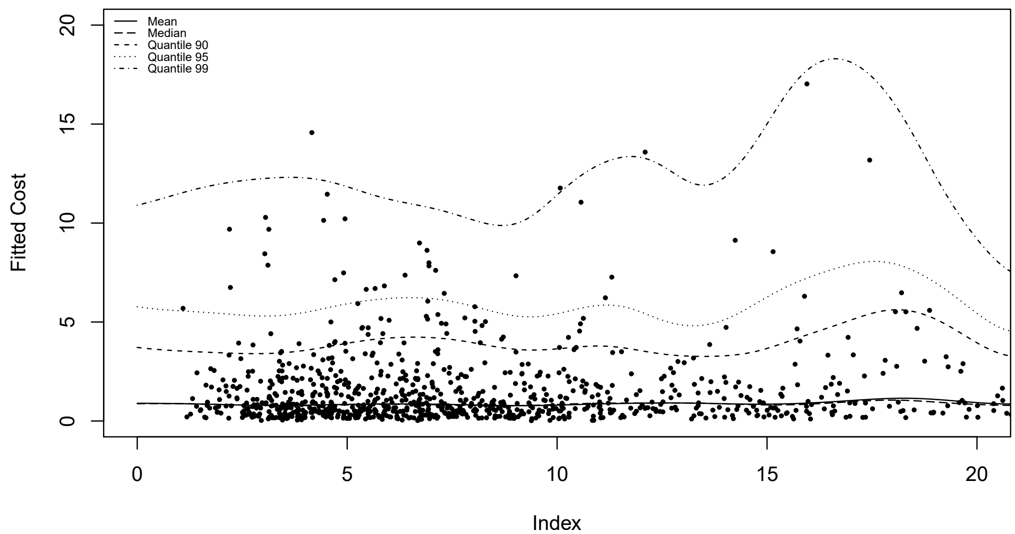

Figure 6: Fitted values of the conditional mean and quantiles with telematics variables in the accident cost data set without outlier extreme cost.

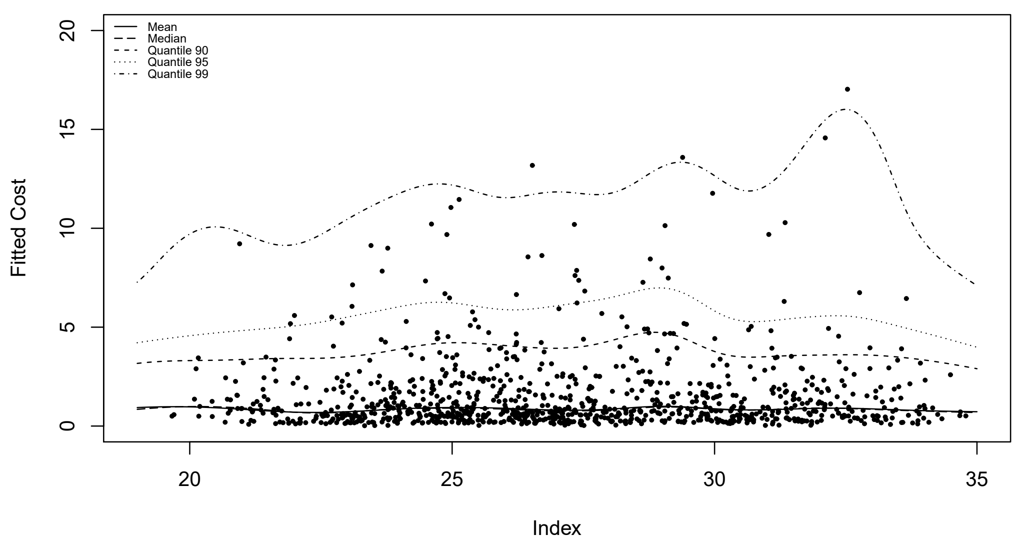

Figure 7: Fitted values of the conditional mean and quantiles with non-telematics variables in the accident cost data set without outlier extreme cost.

Institut de Recerca en Economia Aplicada Regional i Pública
Research Institute of Applied Economics
Universitat de Barcelona
Av. Diagonal, 690 • 08034 Barcelona