Master in Actuarial and Financial Sciences

MASTER THESIS

Title: Alternative Functionals Estimation Based on Index Models and Kernel Approach

Author: Lei Dai

Advisor: Catalina Bolancé Losilla

Academic year: [2023-2024]

UNIVERSITAT DE BARCELONA
Facultat d'Economia i Empresa

Màster de Ciències Actuarials i Financeres

Faculty of Economics and Business

Universitat de Barcelona

Master thesis

Master in Actuarial and Financial Sciences

Alternative Functionals Estimation Based on Index Models and Kernel Approach

Author: Lei Dai

Advisor: Catalina Bolancé Losilla

The content of this document is the exclusive responsibility of the author, who declares that he has not committed plagiarism and that all references to other authors have been expressed in the text.

Abstract

This work explores the utilization of double cross-validation methods for determining the optimal bandwidth in kernel regression using single index-models. Kernel regression is a non-parametric technique widely employed in various fields, particularly in smoothing noisy data. The bandwidth parameter plays a crucial role in kernel regression, controlling the smoothness of the estimated function. Selecting an appropriate bandwidth is essential for achieving accurate and robust model performance. Traditional approaches to bandwidth selection often rely on heuristic methods or fixed rules, which may not be optimal for all datasets.

In this study, we investigate the use of double cross-validation techniques to systematically assess different bandwidth values and identify the one that minimizes *mean integrated squared error* (MISE). Double cross-validation offers a data-driven approach to bandwidth selection, allowing the model to adapt to the inherent complexity of the data while avoiding overfitting. We discuss the theoretical underpinnings of double cross-validation in the context of kernel regression and provide practical guidelines for its implementation.

Furthermore, we conduct empirical experiments using simulated and real-world datasets to evaluate the performance of double cross-validation-based bandwidth selection methods compared to traditional approaches. Our results demonstrate the effectiveness of double cross-validation in identifying optimal bandwidths that lead to improved predictive accuracy and generalization performance. We also discuss potential challenges and limitations associated with double cross-validation, such as computational complexity and sensitivity to data distribution.

Overall, this paper highlights the importance of rigorous model selection techniques, such as double cross-validation, in enhancing the reliability and interpretability of kernel regression models. By leveraging double cross-validation methods, practitioners can effectively tune the bandwidth parameter and construct more robust and adaptive regression models tailored to the characteristics of the underlying data.

Keywords: double cross-validation, kernel regression, single-index models, bandwidth parameters, mean integrated squared error.

Resumen

Este trabajo explora la utilización de métodos de doble validación cruzada para determinar el parametro de alizamiento óptimo en modelos de índice único de regresión con kernel. La regresión con kernel es una técnica no paramétrica ampliamente empleada en varios campos, particularmente para suavizar datos ruidosos. El parámetro de alizamiento juega un papel crucial en la regresión con kernel, controlando la suavidad de la función estimada. Seleccionar un parametro de alizamiento apropiado es esencial para lograr un rendimiento preciso y robusto del modelo. Los enfoques tradicionales para la selección del ancho de banda a menudo se basan en métodos heurísticos o reglas fijas, que pueden no ser óptimos para todos los conjuntos de datos.

En este estudio, investigamos el uso de técnicas de doble validación cruzada para evaluar sistemáticamente diferentes valores de parametro de alizamiento e identificar aquel que minimiza el *error cuadrático medio integrado* (MISE). La doble validación cruzada ofrece un enfoque basado en datos para la selección del parametro de alizamiento, permitiendo que el modelo se adapte a la complejidad inherente de los datos mientras evita el sobreajuste. Discutimos los fundamentos teóricos de la doble validación cruzada en el contexto de la regresión con kernel y proporcionamos pautas prácticas para su implementación.

Además, realizamos experimentos empíricos utilizando conjuntos de datos simulados y del mundo real para evaluar el rendimiento de los métodos de selección de parametro de alizamiento basados en doble validación cruzada en comparación con los enfoques tradicionales. Nuestros resultados demuestran la efectividad de la doble validación cruzada en la identificación de parametros de alizamiento óptimos que conducen a una mayor precisión predictiva y rendimiento de generalización. También discutimos los desafíos y limitaciones potenciales asociados con la doble validación cruzada, como la complejidad computacional y la sensibilidad a la distribución de los datos.

En general, este artículo destaca la importancia de las técnicas rigurosas de selección de modelos, como la doble validación cruzada, en la mejora de la fiabilidad e interpretabilidad de los modelos de regresión con kernel. Al aprovechar los métodos de doble validación cruzada, los practicantes pueden ajustar eficazmente el parámetro de ancho de banda y construir modelos de regresión más robustos y adaptativos, adaptados a las características de los datos subyacentes.

Palabras clave: doble validación cruzada, regresión Kernel, modelos de índice, parámetros de alizamiento, error cuadrático medio integrado.

Dedication

In the dedication of this thesis, I wish to express my deepest gratitude and appreciation to the individuals whose unwavering support and encouragement have been instrumental throughout this academic journey.

First and foremost, I dedicate this thesis to my beloved family. To my parents, whose boundless love, sacrifices, and tireless encouragement have been the foundation upon which I have built my dreams. Your unwavering belief in my abilities has been a constant source of strength, motivating me to overcome challenges and strive for excellence. Your presence in my life has brought immeasurable richness and warmth, and I am forever grateful for your love and support.

I extend my heartfelt gratitude to my advisor, Catalina Bolance, whose guidance, mentorship, and wisdom have been invaluable throughout this research journey. Your expertise, encouragement, and unwavering belief in my potential have empowered me to explore new frontiers, overcome obstacles, and push the boundaries of knowledge. Your dedication to fostering my intellectual growth and nurturing my academic curiosity have left an indelible mark on my professional development, and I am profoundly grateful for your steadfast support and mentorship.

I also wish to acknowledge the support and encouragement of my friends and colleagues. To those who have shared in the triumphs and challenges of this academic pursuit, your camaraderie, companionship, and shared experiences have enriched my journey in countless ways. Whether through late-night study sessions, lively discussions, or moments of celebration, your presence has made this journey all the more memorable and meaningful.

Finally, I dedicate this thesis to all those whose paths have intersected with mine, leaving an indelible mark on my journey. To the teachers, mentors, and role models who have inspired me to reach for the stars and pursue my passions with courage and determination. To the countless individuals who have offered their support, encouragement, and guidance along the way, your kindness, generosity, and belief in my potential have fueled my ambition and propelled me forward, even in the face of adversity.

In conclusion, this thesis is dedicated to each and every individual who has played a part, big or small, in shaping my academic and personal journey. Your support, encouragement, and unwavering belief in my abilities have been the driving force behind my success, and for that, I am eternally grateful.

Contents

Contents **4**

1 Introduction **6**

2 Functional Kernel Estimation **7**
2.1 Kernel estimator of probability density function 7
2.2 Kernel estimator of cumulative distribution function 8
2.3 Bivariate pdf/cdf and conditional pdf/cdf 9
2.4 Nadaraya-Whatson non-paremetric regression 10
2.5 Selection of bandwidth for kernel estimator 11

3 Single-Index model **13**
3.1 Definition 13
3.2 Approaches for estimating variable parameters 13
3.2.1 Minimising the sum of squared errors 13
3.2.2 Maximising the log-likelihood 14
3.3 Approach for estimating smooth parameter 15
3.3.1 Minimizing the Sum of Squared Errors 15

4 Simulations **16**
4.1 Description of Simulation Data 16
4.2 Candidate models 16
4.3 Comparison of the results 18
4.3.1 Comparison of ISE 18
4.3.2 Illustration of conditional mean prediction 21

5 Application **24**
5.1 Description of the dataset 24
5.2 Model results 25
5.3 Marginal Effect analysis 26

6 Conclusion **32**

References **33**

A Appendix **34**
A.1 Comparison table of ISE of simulations 34
A.2 Codes in R for the simulations 34
A.3 Real data from a Spanish insurance company used in application 34
A.4 Codes in R for estimate the mean cost and marginal effect analysis 34

1 Introduction

Estimating the cost of claims is a fundamental task for insurance companies, crucial for setting premiums and managing financial risk. Traditional methods often rely on simplistic assumptions of average costs, which may overlook important nuances in the underlying data distribution. To address this limitation, there's a growing interest in more sophisticated statistical models that can better capture the complexities of claim costs.

One such approach gaining traction is the single-index model which can be approached using kernel estimators. Motivated by its ability to flexibly capture nonlinear relationships and accommodate covariate information, this model offers a promising avenue for improving the accuracy of claim cost estimation. By incorporating covariates that reflect driving habits, such as driving patterns and conditions, the single-index model allows insurers to better understand the factors driving claim costs.

Furthermore, this model is particularly well-suited for handling right-skewed distributions, which are common in insurance data where higher claim costs occur less frequently. By focusing on the entire conditional distribution of claim costs rather than just the mean, the single-index kernel estimation model enables insurers to gain insights into the tails of the distribution, where costly claims reside. This approach aligns with the industry's increasing emphasis on understanding and mitigating tail risks.

Overall, the motivation for using the single-index model stems from the desire to improve the accuracy of claim cost estimation, better understand the factors influencing claim costs, and effectively manage financial risk for insurance companies.

In this work, we present a comprehensive exploration of estimating claim costs within the insurance industry using advanced statistical techniques. We begin by introducing functional kernel estimation, a powerful method for modeling complex relationships between covariates and claim costs. Next, we delve into the concept of the index model, highlighting its significance in capturing underlying structures within the data and its applicability to insurance risk assessment. To demonstrate the efficacy of these methodologies, we conduct simulations using various randomly generated datasets, showcasing their performance under different scenarios and data distributions. Subsequently, we transition to the application of these techniques to a real-world insurance dataset, providing insights into their practical utility and effectiveness in real-world settings. Finally, we draw conclusions based on our findings, discussing the implications for insurance companies and the broader field of risk management. Through this structured approach, we aim to offer a comprehensive understanding of functional kernel estimation, index modeling, and their application in estimating claim costs, providing valuable insights for both researchers and practitioners in the insurance industry.

2 Functional Kernel Estimation

2.1 Kernel estimator of probability density function

The kernel estimator of probability density function (PDF), denoted as $f(x)$, is a fundamental tool in statistical analysis used to estimate the underlying probability density function of a random variable based on observed data points. This method is particularly useful when the exact functional form of the distribution is unknown or complex, as it provides a flexible and non-parametric approach to density estimation (see Wand and Jones, 1995).

Let $x_1, x_2, ..., x_n$ be a sample of n independently and identically distributed data then:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right). \tag{1}$$

At its core, the kernel estimator of PDF operates by placing a kernel function, typically a smooth and symmetric function centered at each data point, and then summing these kernel functions to obtain an estimate of the density at any given point in the data space. The bandwidth parameter of the kernel function controls the smoothness of the estimated density and is a critical aspect of the estimator, influencing its bias-variance trade-off.

In Silverman (1986) the bias and the variance of this estimator are calculated:

$$\begin{aligned} bias_h(x) &= E\hat{f}(x) - f(x) \\ &= \frac{1}{2}h^2 f''(x) \int t^2 K(t)dt + \text{higher-order terms in h} \\ var\hat{f}(x) &\approx n^{-1}h^{-1} \int f(x - ht)K(t)^2 dt - n^{-1}\{f(x) + O(h^2)\}^2. \end{aligned}$$

Notice that the $bias_h$ is aymptotically proportional to h^2, so for this quantity to decrease one needs to take h to be small. However, taking h small means an increase in the leading term of the integrated variance since this quantity is proportional to $(nh)^{-1}$ and it is assumed that $\lim_{n\to\infty} nh = 0$. Therefore, as n increases h should vary in such a way that each components of the MISE becomes smaller. This is known as the *variance-bias trade-off* and is a mathematical quantification for the critical role of the bandwidth (see Wand and Jones, 1995).

This paper employs the Gaussian kernel function in the kernel estimator of probability density function (PDF). The Gaussian kernel is chosen for its computational efficiency, continuity, and flexibility in adapting to various data distributions. This choice ensures smooth estimation between observations, facilitating accurate density estimation. By utilizing the Gaussian kernel, this paper provides a practical and reliable approach for estimating probability density functions in statistical analysis.

$$K(t) = \frac{1}{\sqrt{2\pi}} exp\left(-\frac{t^2}{2}\right), \tag{2}$$

One of the key advantages of the kernel estimator of PDF is its simplicity and ease of implementation. Additionally, it can adapt well to various data distributions and is robust to outliers. However, choosing an appropriate kernel function and bandwidth parameter requires careful consideration to ensure accurate estimation.

Overall, the kernel estimator of PDF serves as a valuable tool in statistical analysis, offering a flexible and data-driven approach to estimate probability density functions, making it particularly useful in fields such as finance, epidemiology, and engineering, among others.

2.2 Kernel estimator of cumulative distribution function

The kernel estimator of cumulative distribution function (CDF), denoted as $F(x)$, is a statistical method used to estimate the underlying cumulative distribution function of a random variable based on observed data points. This technique is particularly useful when the exact form of the CDF is unknown or complex, as it provides a flexible and non-parametric approach to distribution estimation (see Bolancé et al., 2024).

$$\begin{aligned} \hat{F}(x) &= \int_{-\infty}^{x} \hat{f}(t)dt \\ &= \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x - x_i}{h}\right), \end{aligned} \tag{3}$$

Where $\Phi(x)$ is the Cumulative function of a standard normal distribution. And the kernel estimator is the generalization or a smoothed version of the empirical distribution. Instead of the indicator function, the kernel estimator uses the kernel distribution function as its weight.

In the kernel estimator of CDF, a kernel function, typically a smooth and symmetric function, is centered at each data point. The cumulative contribution of these kernel functions is then summed to obtain an estimate of the cumulative distribution function at any given point in the data space. Similar to the kernel estimator of PDF, the bandwidth parameter of the kernel function plays a crucial role in controlling the smoothness of the estimated CDF.

Nadaraya (1964) has proved that under mild conditions that $\hat{F}$ has aymptotatically the same mean and variance as $\hat{F}_n$. From (4) we can see that as $h \to 0$ and $n \to \infty$ the $E[\hat{F}(x) - F(x)] \to 0$, which implies that the kernel estimator is a consistent estimator of the CDF.

$$E\{\hat{F}(x) - F(x)\}^2 \propto F(x)\{1 - F(x)\}/n - uh/n + vh^4, \tag{4}$$

where:

$$u = f(x)\left\{3 - \int_{-3}^{3} \Phi^2(t)dt\right\}, v = \left\{\frac{1}{2}f'(x)\int_{-3}^{3} t^2 K(t)dt\right\}^2.$$

One advantage of the kernel estimator of CDF is its ability to provide estimates of percentiles and quantiles directly from the estimated cumulative distribution function. This feature makes it particularly useful for applications where understanding the distribution's tail behavior is essential, such as risk assessment in finance or reliability analysis in engineering.

Overall, the kernel estimator of CDF offers a flexible and data-driven approach to estimate cumulative distribution functions, making it a valuable tool in various statistical applications.

2.3 Bivariate pdf/cdf and conditional pdf/cdf

The kernel estimators of bivariate PDF and CDF are statistical methods used to estimate the joint distribution or cumulative distribution of two random variables based on observed data points. These estimators extend the principles of univariate kernel estimation to handle two-dimensional data, allowing for the modeling of complex relationships between two variables (see Wand and Jones, 1995). And one of the most common kernel estimators of bivariate pdf is the productive kernel estimator:

$$\hat{f}(x,y) = \hat{f}_{h_1,h_2}(x,y) = \frac{1}{nh_1h_2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right)K\left(\frac{y-y_i}{h_2}\right). \tag{5}$$

In kernel estimation of bivariate PDF, kernel functions are centered at each observed data point in the two-dimensional space, and their contributions are summed to estimate the joint probability density function across the entire domain. This approach provides a flexible and non-parametric method for capturing the joint distribution of two variables, which is especially useful when the relationship between them is not easily described by a parametric model.

Similarly, kernel estimation of bivariate CDF involves summing kernel functions centered at each data point to estimate the joint cumulative distribution function. This allows for the direct estimation of probabilities associated with pairs of observations, facilitating the analysis of joint probabilities and quantiles (see Bolancé et al., 2024).

$$\begin{aligned}\hat{F}(x,y) &= \iint \hat{f}(x,y)dxdy \\ &= \frac{1}{n}\sum_{i=1}^{n} \Phi\left(\frac{x-x_i}{h_1}\right)\Phi\left(\frac{y-y_i}{h_2}\right).\end{aligned} \tag{6}$$

Kernel estimators of Conditional PDF (7) and CDF (8) extend these concepts further by estimating the conditional distribution or conditional cumulative distribution of one variable

given the value of another variable. This enables the modeling of conditional relationships between variables, providing insights into how one variable may depend on or be influenced by another.

$$\hat{f}_{Y|X}(y|x) = \frac{\hat{f}_{X,Y}(x,y)}{\hat{f}_X(x)}, \tag{7}$$

where $\hat{f}_{X,Y}(x,y)$ is estimated using (5) and $\hat{f}_X(x)$ is estimated using (1).

$$\hat{F}_{Y|X}(y|x) = \frac{\hat{F}_{X,Y}(x,y)}{\hat{f}_X(x)}, \tag{8}$$

where:

$$\hat{F}_{X,Y}(x,y) = \frac{1}{nh_1}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right)\Phi\left(\frac{y-y_i}{h_2}\right).$$

Overall, Kernel estimators of bivariate and conditional PDF/CDF offer flexible and data-driven approaches for modeling relationships between two variables, making them valuable tools in various fields such as economics, environmental science, and engineering, where understanding joint or conditional distributions is essential for decision-making and analysis.

2.4 Nadaraya-Whatson non-paremetric regression

Nadaraya-Watson nonparametric regression is a technique used for estimating the conditional expectation of a dependent variable Y given independent variable X, without assuming a specific parametric form for the relationship between the variables. It's a kernel regression method, meaning it relies on local weighted averaging of the observed data points.

In Nadaraya-Watson regression, the estimator of the conditional expectation of the dependent variable $\hat{E}(Y_i|X=x)$ given the independent variables $X=x$ is calculated as a weighted average of the observed Y values, with the weights determined by a kernel function K (see Wand and Jones, 1995):

$$\hat{m}(x) = \hat{E}(Y_i|X=x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right)}. \tag{9}$$

At the heart of Nadaraya-Watson regression is the concept of kernel smoothing. The basic idea is to estimate the conditional expectation of Y at a particular point x by giving more weight to data points that are closer to x and less weight to those that are farther away. This

weighting is accomplished through a kernel function, which assigns weights to observed data points based on their distances from x.

The formula for the Nadaraya-Watson estimator reflects this local averaging approach. For each x where estimation is desired, the estimator computes a weighted average of the observed Y values, with the weights determined by the kernel function. The bandwidth parameter h controls the width of the kernel and thus the size of the local neighborhood used for averaging. A smaller bandwidth leads to a more localized estimation, while a larger bandwidth results in a more smoothed estimate.

In Härdel (1989), it is proven that the Nadaraya-Watson estimator is an asymptotically consistent estimator of the conditional mean, as stated in Proposition 1.

Proposition 1 *Assume the stochastic design model with a one-dimensional predictor variable X and*
(A1) $\int |K(u)|du < \infty$,
(A2) $lim_{u \to \infty} uK(u) = 0$,
(A3) $EY^2 < \infty$,
(A4) $n \to \infty, h_n \to 0, nh_n \to \infty$. *Then at every point of continuity of* $m(x)$, $f(x)$ *and* $\sigma^2(x)$, *with* $f(x) > 0$,

$$\frac{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{x-x_i}{h_1}\right)} \xrightarrow{P} m(x).$$

One of the key advantages of Nadaraya-Watson regression is its ability to capture complex nonlinear relationships between variables without requiring any assumptions about the underlying distribution or functional form. This makes it particularly useful in situations where the relationship between X and Y is unknown or cannot be adequately described by a parametric model. Additionally, Nadaraya-Watson regression is robust to outliers and does not suffer from bias due to misspecification of the regression function.

2.5 Selection of bandwidth for kernel estimator

The bandwidth parameter (h) in kernel density estimation and nonparametric regression directly impacts the trade-off between model complexity and smoothness. Choosing the right bandwidth is crucial:

1. **Bias-Variance Trade-off**: It balances bias and variance in the estimator.
2. **Model Flexibility**: It determines the level of detail and smoothness in the estimated function.

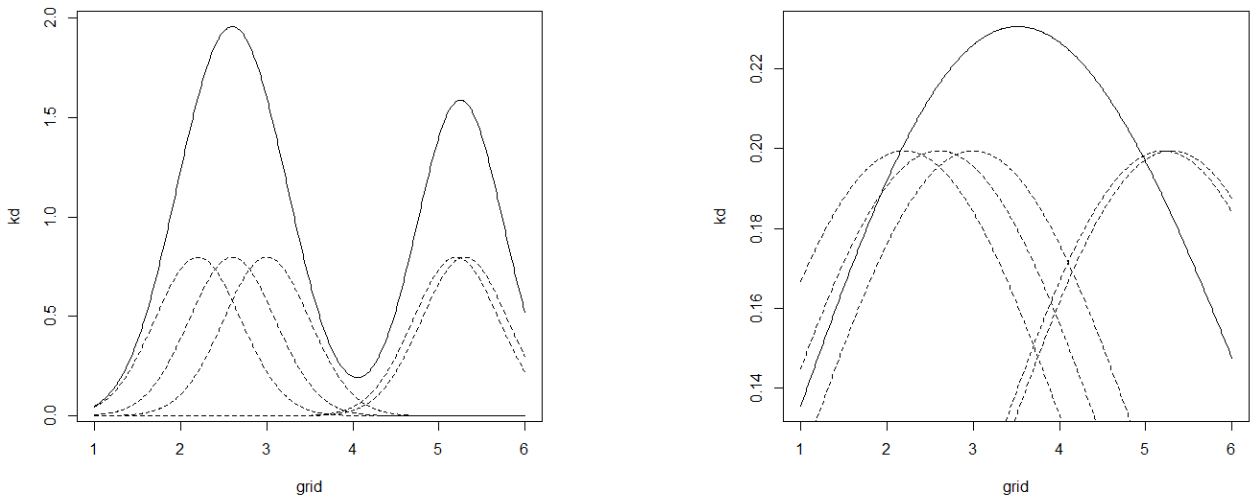

Figure 1: The illustration of how univariate kernel density estimates were constucted based on 5 observations (2.2, 2.6, 3.0, 5.2, 5.3). The left figure represents a model built with h=0.5 and the right one with h=2

3. **Interpretability**: It affects how well the estimated function captures local versus global trends.

4. **Robustness**: It influences the sensitivity of the estimator to outliers.

5. **Computational Efficiency**: It affects the computational resources required for estimation.

6. **Generalization Performance**: It impacts how well the estimator performs on unseen data.

Figure 1 shows a kernel density estimate constucted using five observations ($\vec{x} =< 2.2, 2.6, 3.0, 5.2, 5.3 >$) with the kernel chosen to be the N(0,1), which is $K(x) = \phi(x)$. It should be pointed out that only 5 observations are used here purely for clarity in illustration about how kernel method works. In practice, much more observations should be involved in the kernel estimation.

Selecting an appropriate bandwidth involves finding a balance that accurately represents the data without overfitting or oversmoothing. Techniques like cross-validation help in making this choice.

3 Single-Index model

3.1 Definition

Single-index regression models offer a semiparametric approach to extending linear regression, establishing the relationship between a random variable Y (such as the cost of a traffic accident or claim severity) and a d-dimensional vector $X = (X_1, ..., X_d)^T$. Traditionally, estimating the linear predictor coefficients $\theta = (\theta_1, ..., \theta_d)^T$ and the function m has relied on the conditional expectation, leaving models vulnerable to extremes, heavy-tailed distributions, or strong asymmetry, common in many real-world applications. Our contribution lies in extending maximum likelihood estimation to facilitate single-index conditional distribution modeling, holding significant promise across various domains.

$$Y = m(\theta^T X) + \epsilon, \tag{10}$$

were 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.

Nonparametric regression offers a broader framework than the single-index model outlined in equation (10), stemming from the general specification $Y = m(X) + \epsilon$, where the objective is to estimate the regression curve $m(x) = E(Y|X = x)$, as elucidated by Härdel and Ichumura (1993). Despite its versatility, nonparametric regression encounters notable challenges in practical applications. Firstly, estimation becomes increasingly intricate with a higher number of covariates, succumbing to the curse of dimensionality. Secondly, direct interpretation of the explanatory variables' effects is impractical, necessitating the exploration of these effects through plotting various relationships.

An alternative to the single-index model is the generalized additive model, as detailed by Hastie and Hardle (1990). However, it confronts similar challenges to nonparametric regression, including the complexities associated with high-dimensional covariates and the indirect interpretation of variable effects.

3.2 Approaches for estimating variable parameters

3.2.1 Minimising the sum of squared errors

For estimating the vector θ, Härdel and Ichumura (1993) proposed directly minimizing the residual sum of squares. Their estimator is constructed using i.i.d. observations $(X_1, Y_1), ..., (X_n, Y_n)$ of the covariates and dependent variable, with $\hat{m}_{-i}$ representing the leave-one-out kernel estimator of m. Alternatively, Juditsky and Spokoiny (2001) explored the average derivative estimator of the parameter vector in the index model, initially introduced by Stoker (1986)

and further developed by Stock and Stoker (1989). Juditsky and Spokoiny (2001) presented a method for estimating the coefficient vector θ by minimizing an M-function using a score function Ψ, which compares Y_i with a nonparametric estimator $\hat{m}(\cdot)$.

Leave-One-Out Cross-Validation (LOOCV) is used for estimating model parameters because it provides an unbiased estimate of prediction error by iteratively training the model on all but one observation and then validating its prediction against the omitted observation. This method balances bias and variance, offering consistent parameter estimates as the sample size increases. LOOCV maximizes data utilization by using each observation for both training and validation, ensuring efficient estimation with robustness to the specific choice of validation sets. Overall, LOOCV is a powerful and widely used technique for parameter estimation due to its ability to provide reliable estimates of model parameters while maximizing the use of available data. That is:

$$\hat{\theta} = argmin_{\theta} \sum_{i=1}^{n} [y_i - \hat{m}_{-i}(\theta^T x_i)]^2. \tag{11}$$

In our codes, the function optim() with the method "L-BFGS-B" was used to find the parameters which minimizes the objective function.

3.2.2 Maximising the log-likelihood

To estimate the variable parameters and obtain the index values, Bolancé et al. (2024) used the method of likelihood cross-validation to estimate the vector of variable parameters θ to get the index. the estimated parameters was those which maximize the leave-one-out estimated conditional log-likelihood (see Silverman, 1986).

$$\hat{l}_n(\theta; h_1, h_2) = \frac{1}{n} \sum_{i=1}^{n} log \hat{f}_{\theta}^{-i}(Y_i|\theta^T x_i). \tag{12}$$

It should be noted that if we choose any non-zero real number λ, then, since there is a one-to-one correspondence between $\theta^T X$ and $\lambda\theta^T X$, it is also true that the conditional distribution only depends on the covariate vector via the linear combination $\lambda\theta^T X$. Consequently, infinitely multiple choices exist for the single-index parameter vector θ. The usual way to solve this identification problem is to introduce a scale constraint, for example $||\theta|| = 1$ or fixing one component of θ to one (see Bolancé et al., 2024).

The code iteratively estimates and selects the parameters θ and h_1, h_2 using the following steps:

1. Initial values: The initial values of h_1 and h_2 are determined using the variation of the samples. These initial values serve as starting points for the optimization process.

2. Optimization: The 'optim()' function is used with the "L-BFGS-B" method to find the estimate $\hat{\theta}$ that maximizes the objective function. The objective function is defined based on the specific problem being solved.

3. Fixing $\hat{\theta}$: The estimated $\hat{\theta}$ is fixed as known values, and the optimal values of h_1 and h_2 are found by maximizing the objective function. This step aims to find the best values for h_1 and h_2 given the estimated $\hat{\theta}$.

4. Repeat iterations: The first step is repeated, taking the newly selected values of h_1 and h_2 from the previous step. The process is iterated until the results converge, meaning that the parameter estimates and selected values of h_1 and h_2 stabilize.

By iteratively updating the parameter estimates and selecting the optimal values of h_1 and h_2 based on the estimated $\hat{\theta}$, the code aims to refine the model and find the best combination of parameters that maximize the objective function.

3.3 Approach for estimating smooth parameter

3.3.1 Minimizing the Sum of Squared Errors

Although in the section 3.2, when estimating the variable parameters θ, we also selected the optimal bandwidth h_1 and h_2. But according to van den Berg (2020) the optimal smoothing parameters for estimating the variables parameters is not the same ones for estimating the conditional mean, which is another functional. That is, for each different estimation use, the smooth parameters should be different.

After estimating θ using (11) or using (12), we will find another optimal smoothing parameter h_1 taking $\hat{\theta}$ as known values by minimising the sum of squared errors, which serves as a double cross validation.

$$h_1 = argmin_{h_1} \left\{ \sum_{i=1}^{n} \left(y_i - \frac{\sum_{j=1, j\neq i}^{n} K\left(\frac{t-\hat{\theta}^T x_j}{h_1}\right) y_j}{\sum_{j=1, j\neq i}^{n} K\left(\frac{t-\hat{\theta}^T x_j}{h_1}\right)} \right)^2 \right\}. \tag{13}$$

In the codes of the annex, the function optimize() was used to finde the optimal parameter h_1 which minimizes the objective function.

4 Simulations

4.1 Description of Simulation Data

Each simulation data at hand comprises 100 replications of samples with 500 observations and 2000 observations. Each observation consists of the dependent variable Y and a set of three independent variables, denoted as X. The three variables are identically and independently distributed and have a standar normal distribution.

The original index is formed as a linear combination of these independent variables X, using the original parameter vector, $\theta_0 = (1, 1.3, 0.5)^T$. This original index serves as a critical input to generate the dependent variable Y, which is produced using various conditional distributions. (See Table 1)

Through this structure, the simulation data provides a rich and diverse set of variables, allowing for a comprehensive examination of the relationships and dependencies among them. This diversity in the data is especially beneficial for understanding the impact of different distributions on the generated results.

- Lognormal Distribution: The lognormal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. It is skewed, with a heavier right tail, and is commonly used in situations where values are positively skewed and have a lower bound of zero, such as in the distribution of income or the size of populations.

- Weibull Distribution: The Weibull distribution is a flexible distribution that can assume various shapes depending on its parameters. It is often used in survival analysis and reliability engineering to model time-to-failure data. The shape of the Weibull distribution can provide insights into the nature of the failure rate.

- Log-logistic Distribution: The log-logistic function has several properties that make it useful in modeling survival data. It can handle censored data, which is data where the event of interest has not occurred for some of the observations. Additionally, it can model both increasing and decreasing hazard rates, which represent the likelihood of the event occurring at a given point in time.

4.2 Candidate models

Three alternatives of estimating θ and h are proposed in this paper which combines different methods of estimating θ and h. For comparing the performance of the different alternatives it

Distribution	Parameters	Density	Mean
Lognormal	$(\mu = \theta^T x, \sigma = 0.5)$ $(\mu = \theta^T x, \sigma = 2)$ $(\mu = \theta^T x, \sigma = 5)$	$\frac{1}{y\sqrt{2\pi * \sigma^2}} exp\left(-\frac{(ln(y)-\mu)^2}{2*\sigma^2}\right)$	$e^{\mu+\frac{1}{2}\sigma^2}$
Weibull	$(k = 1, \lambda = (\theta^T x)^2)$ $(k = 2, \lambda = (\theta^T x)^2)$ $(k = 3, \lambda = (\theta^T x)^2)$	$\frac{k}{\lambda}\left(\frac{y}{\lambda}\right)^{k-1} e^{-(y/\lambda)^k}$	$\lambda\Gamma(1+\frac{1}{k})$
Log-logistic	$(\beta = 1.5, \alpha = (\theta^T x)^2)$ $(\beta = 2, \alpha = (\theta^T x)^2)$ $(\beta = 3, \alpha = (\theta^T x)^2)$	$\frac{(\beta/\alpha)(y/\alpha)^{\beta-1}}{(1+(y/\alpha)^\beta)^2}$	$\frac{\alpha\pi/\beta}{sin(\pi/\beta)}$

Table 1: Conditional distributions for dependent variable as a function of index for the simulation study

is introduced in this paper the concept of Integrated Squared Error of expected values, which is defined as follow:

$$ISE := \int \{E[Y|t] - \hat{m}(t)\}^2 dt, \tag{14}$$

with $t = \hat{\theta}^T x_i$.

Unlike the sum of squared errors, the ISE measures magnitude of prediction error along all the field of the index. So this will measure the global performance of the predictor.

The three alternatives are described as follows:

- Alternative 1: To estimate θ and h simultaneously by minimising the Leave One Out Sum of Squared Errors, which is $(\hat{\theta}, h_1) = argmin_{(\theta,h_1)} \sum_{i=1}^{n} [y_i - m_{-i}(\theta^T x_i)]$, where $m(\cdot)$ is defined through the formula (9) (see Härdel, 1989);

- Alternative 2: At first to estimate θ and h_1 simultaneously by minimising the Leave-One-Out Sum of Squares Errors (11). But taking into account that the optimal smooth parameters h_1 don´t have to be the optimal ones to predict the conditional mean of the response variable. For this reason, we´ll take the estimated $\hat{\theta}$ as known parameters and obtain the optimal smooth parameters by minimising (11) again;

- Alternative 3: At first we will estimate θ and h_1, h_2 simultaneously by maximizing the Leave-One-Out Log-Likelihood (12). Samely, taking into account that the optimal h_1 for estimating the conditional mean value of the response variable. For this reason we will take the estimated $\hat{\theta}$ as known parameters and obtain the optimal smooth parameter h_1 by minimising the Leave-One-Out Sum of Squared Errors (11).

4.3 Comparison of the results

4.3.1 Comparison of ISE

Table 2 presents the comparison of the Integrated Squared Errors (ISE) of 18 samples, which are generated randomly with three different distributions. The samples are further generated by changing the shape parameters, which control the variance and tail behavior of the distribution. The comparison results are presented as ratios to the ISE of alternative 1.

To analyze how the sample size affects the performance of the different alternatives, samples of size 500 and 2000 are simulated for each distribution described above.

As the parameters increase, the variance of the distribution grows exponentially, and so does the ISE. For simplicity of illustration, the values in the table are presented as relative values to the ISE of the first alternative. Original ISE values can be found in the Appendix.

From the Table 2, it is evident that alternative 2 slightly improves upon alternative 1 in all simulations, while the improvement of alternative 3 is larger. In most cases, the ISE using alternative 3 is less than half of alternative 1.

Furthermore, the sample size has an impact on the model's performance. With a larger sample size, the improvement of alternative 3 becomes even bigger compared to alternative 1. It is worth noting that as the shape parameters of each distribution are changed to make the distribution heavier in the right tail, the superiority of alternative 3 over alternatives 1 and 2 becomes more pronounced.

In summary, both the choice of method and the sample size play a crucial role in the model's performance. Alternative 3 consistently outperforms the other alternatives, and a larger sample size further enhances its superiority. Additionally, modifying the shape parameters of the distributions reinforces the strength of alternative 3 compared to alternatives 1 and 2.

		Sample size					
		n=500			n=2000		
Distribution	Parameters	Alt. 1	Alt. 2	Alt. 3	Alt. 1	Alt. 2	Alt. 3
Log Normal	$\mu = \theta^T x$, $\sigma = 0.5$	1	0.970	0.412	1	0.974	0.038
	$\mu = \theta^T x$, $\sigma = 2$	1	0.992	0.116	1	0.974	0.141
	$\mu = \theta^T x$, $\sigma = 5$	1	1.029	0.641	1	1.008	0.993
Weibull	$k = 1$, $\lambda = (\theta^T x)^2$	1	0.832	0.484	1	1.003	0.216
	$k = 2$, $\lambda = (\theta^T x)^2$	1	0.995	0.377	1	0.984	0.386
	$k = 3$, $\lambda = (\theta^T x)^2$	1	1.000	0.307	1	1.013	0.343
Log-logistic	$\beta = 1.5$, $\alpha = (\theta^T x)^2$	1	0.991	1.111	1	0.606	0.114
	$\beta = 2$, $\alpha = (\theta^T x)^2$	1	0.954	0.408	1	0.979	0.201
	$\beta = 3$, $\alpha = (\theta^T x)^2$	1	0.999	0.413	1	0.871	0.106

Table 2: Comparison of the ISE using different methods for estimating θ and h in different simulations with sample size 500 and 2000. The values are ratio values to the ISE of Alternative 1

4.3.2 Illustration of conditional mean prediction

For a more visual comparison of the precision of mean values using the three alternatives for estimating variable parameters and selecting optimal smoothing parameters, we provide a sample of size 2000. This sample follows the same conditions of the simulated index and conditional distribution as described in the previous section. Additionally, in each plot, the theoretical mean values are represented using solid lines.

- Log-Normal: From Figure 2, it can be observed that the first two alternatives yield almost

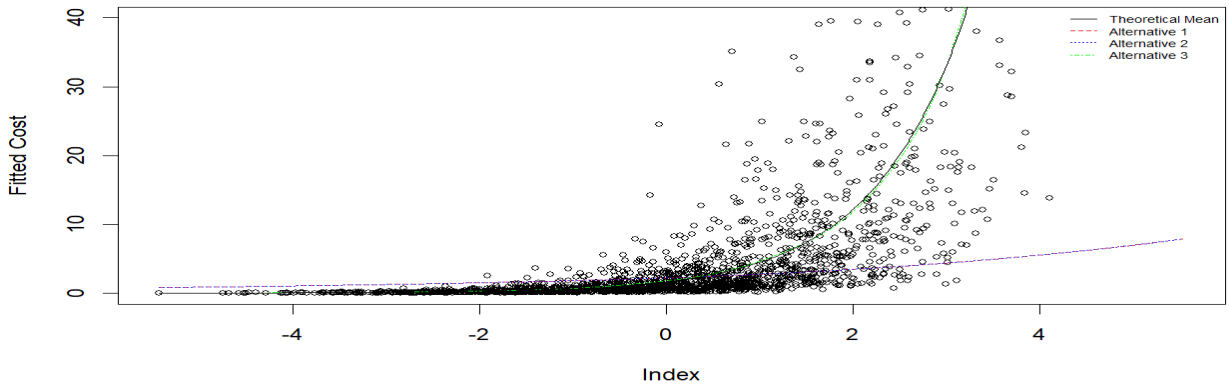

Figure 2: Comparison of the estimated conditional means using three alternatives to estimate the parameters and theoretical means. The theoretical distribution is a log-normal distribution with $\mu = \theta^T x$ and $\sigma = 1$ using a sample with 2000 observations

identical estimations. Initially, all three alternatives estimate the mean values very close to the theoretical means. However, as the index increases, indicating a heavier tail, the first two alternatives consistently underestimate the theoretical values. In contrast, the third alternative continues to estimate the mean value accurately. Thus, the third alternative proves to be much more robust than the first two.

- Weibull: In the case of a Weibull conditional distribution (Figure 3), all three alternatives perform very similarly and are very close to the theoretical value. Notably, at the right tail of the index, the third alternative outperforms the first two, while the first two demonstrate better performance at the left tail. However, it is evident that all alternatives tend to underestimate the mean value at the left tail and overestimate it at the right tail of the index.

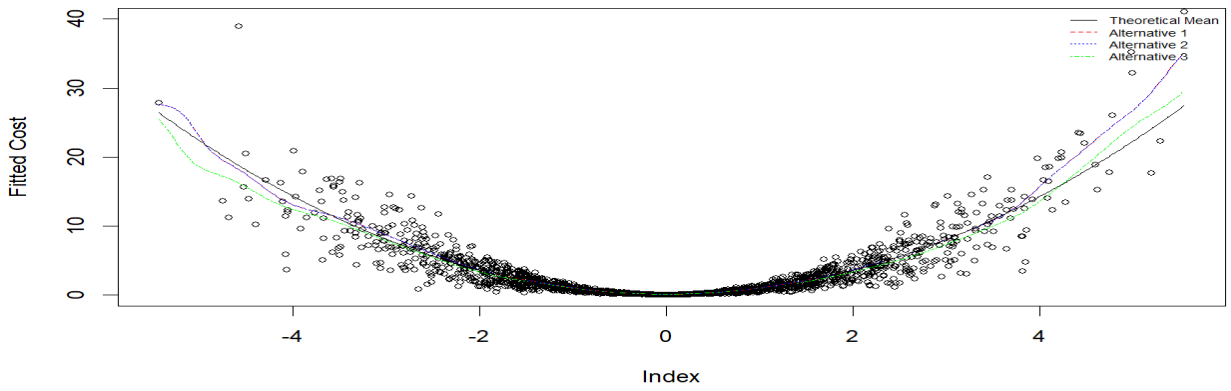

Figure 3: Comparison of the estimated conditional means using three alternatives to estimate the parameters and theoretical means. The theoretical distribution is a Weibull distribution with $\lambda = (\theta^T x)^2$ and $k = 3$ using a sample with 2000 observations

- Log-logistic: Figure 4 illustrates that when the conditional distribution is a log-logistic, there is a severe problem of overfitting with the first alternative. However, after reselecting the optimal bandwidths by minimizing the Sum of Squared Errors (SSE) function, the second alternative provides a much more robust estimation. Notably, the estimated values of alternatives 2 and 3 are very close, with the third alternative performing slightly better than the second one.

As conclusion of this section, the comparison reveals that the third alternative, which involves first estimating the variable parameters by maximizing the log-likelihood and subsequently selecting the optimal smooth parameters by minimizing the sum of squared errors of the predicted conditional mean values, stands out due to its precision and robustness in estimation especially when the conditional distribution presents heavy tails.

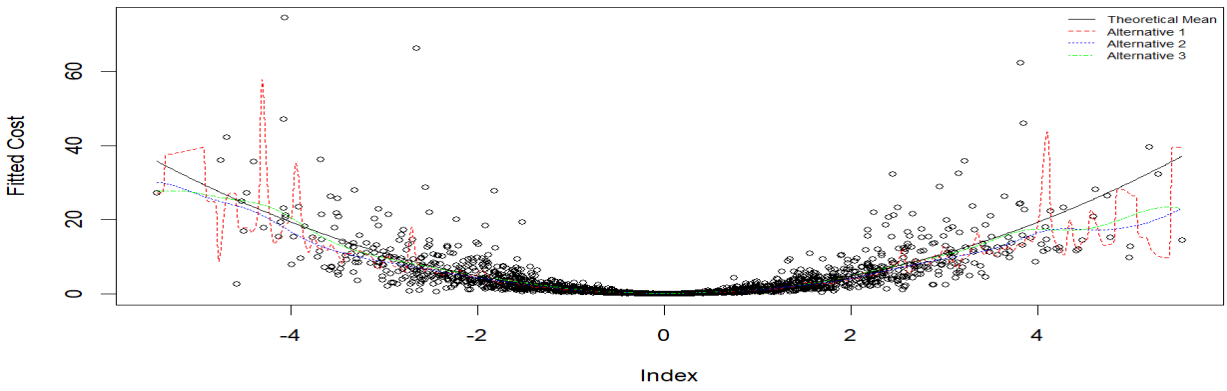

Figure 4: Comparison of the estimated conditional means using three alternatives to estimate the parameters and theoretical means. The theoretical distribution is a Log-logistic distribution with $\alpha = (\theta^T x)^2$ and $\beta = 3$ using a sample with 2000 observations

5 Application

5.1 Description of the dataset

The dataset used in the application is the same one used by Bolancé et al. (2024). Our analysis focuses on a dataset sourced from a Spanish insurance company´s portfolio, specifically pertaining to policyholders aged between 18 and 35. These policyholders, all of whom had underwritten a motor insurance policy, accepted a telematics engine enabling the company to collect data on their driving behavior. The dataset only includes settled claims.

A few claims in the original dataset arose from no-fault agreements between insurers, where the recorded amount aligns with the legally established cost. However, these claims do not provide information on the true cost of the claim, which could potentially be lower or higher than the agreed-upon amount. Therefore, to maintain the integrity of our analysis, we excluded these claims from our dataset. Consequently, our data are not censored.

Our primary analysis focuses on a sample of 489 car insurance policyholders who reported at least one claim in 2011. These claims correspond to third-party liability accidents. For each policyholder, we have data on the total incurred losses and the number of claims throughout the year. The ratio of these two values provides the yearly mean claim cost per policy, which refers to incurred and paid losses.

The dataset under examination provides information on several covariates for each policyholder. These include (the descriptive statistics are presented with Table 3):

1. Cost per policyholder in thousands of euros (cost)
2. Age in years (age)
3. Number of years holding a driving licence (agelic)
4. Age of car in years (agecar)
5. A binary indicator equal to 1 if the car is parked in a garage overnight and 0 otherwise (parking)
6. Annual distance driven in thousands of kilometres (tkm)
7. Percentage of kilometres driven at night (nightkm)
8. Percentage of kilometres driven on urban roads (urbankm)
9. Percentage of kilometres driven above the speed limit (speedkm)

This information was collected using a telematics device installed in the policyholders' vehicles, providing insights into their driving style and patterns. Therefore, "tkm", "urbankm", "nightkm", and "speedkm" are referred to as "telematics covariates".

It is worth noting that the gender variable is not included in the model, in compliance with European Union regulations that prohibit discrimination between men and women in the context of insurance premiums (See Guillen M. and Pérez-Marín, 2019). Our analysis is, thus, focused on exploring the impact of the aforementioned covariates on insurance claims, in light of the available telematics data.

5.2 Model results

Based on the simulations, the best double cross-validation approach to predict the conditional mean cost given the characteristic parameters is to first maximize the log-likelihood and then minimize the sum of squared errors. And in the model the coefficient of the variable "speedkm" was constrained to one in order to address the issue of identification. It makes sense since it is a straightforward intuition that the speed variable will contribute to the claim cost. The estimated variable parameters and their significance hypothesis test are given in the Table 4 and the following plot below (Figure 5) illustrates the prediction results of the insurance company.

By examining the p-values of the hypothesis tests for the different variable parameters, we can see that all variables, except for whether the car is parked in a garage overnight or not, affect the prediction of the expected cost in a statistically significant way.

Furthermore, by observing the signs of the estimated parameters, we can see that the variables—percentage of kilometers driven above the speed limit (speedkm), age in years (age), number of years holding a driving license (agelic), percentage of kilometers driven at night (nightkm), and percentage of kilometers driven on urban roads (urbankm)—influence the expected costs in a similar manner.

According to common knowledge, when drivers frequently exceed the speed limit, the expected costs are higher. Therefore, we can also conclude that higher percentages of night and urban drives lead to higher expected costs. This is logical since visibility is reduced at night, and urban traffic is usually more complex.

One less intuitive finding is that both the age of the driver and the number of years holding a license affect costs similarly. This can be interpreted as more experienced drivers potentially becoming overconfident and less cautious. Additionally, the age of the car plays a role: people driving older cars may be more careful, aware of potential issues with their vehicles.

From the Figure 5, we can draw the following conclusions. When the index of the insured is below 25, the expected mean cost remains relatively stable, although there is a slight tendency for the fitted cost to drop as the index increases from 5 to 15, then rise as the index continues

to increase from 15 to 25. Generally, the estimated cost in this range falls within the interval of [9000, 10000]. However, when the index exceeds 25, the expected cost increases much more rapidly until it reaches 33. According to the graph, the cost starts to decline again after this point, but the results in the tails are not credible due to the lack of data in those regions.

5.3 Marginal Effect analysis

Because our model combines both parametric and non-parametric elements, analyzing the marginal effect of the variables by simply looking at their parameters is very complicated. Therefore, we need to create a grid of the variables for a more detailed analysis. Here is how we do it:

First, we choose a variable to analyze its marginal effect. For all other variables, we set them at their minimum values and keep them unchanged. Additionally, we want to analyze the marginal effect of this variable across different groups. In this case, we distinguish between younger and older age groups. In the graphics, the continuous line represents the younger group, while the dotted line represents the older group. The pre-set values of the variables are "agelic" = 2.001; "agecar" = 2.11; "parking" = 1; "tkm" =1,219.77; "nightkm" = 4.38%; "urbankm" = 3.81%; "speedkm" = 12.23%; "age" = 20.59 for younger groups and "age" = 34.07 for older groups.

We analyze the marginal effect of the following four variables:

1. Percentage of kilometers driven above the speed limit (speedkm)
2. Annual distance driven in thousands of kilometers
3. Percentage of kilometers driven at night
4. Percentage of kilometers driven on urban roads

- For the variable "speedkm", we observe similar marginal effects for both age groups. Initially, the cost decreases as the "speedkm" increases, but then it starts to increase beyond a certain point. Notably, the turning point for the older group occurs earlier than for the younger group. The point can be interpreted as follows: When the "speedkm" is below a certain level, a higher value of "speedkm" indicates more experience and can result in a reduction in cost. However, once this variable reaches a limit beyond which experience cannot control it, the "speedkm" will start to contribute to the cost.

- The marginal effect of "tkm" on the cost is relatively small. We observe that for both age groups, the cost increases as "tkm" increases, but the rate of increase is not significant. Additionally, it's notable that the cost level for the younger group is consistently higher than for the older group.
- The variable "nkm" exhibits a converse marginal effect compared to "tkm". In other words, the cost level decreases as "nkm" increases. Additionally, the rate of decrease in cost with respect to the increment of "nkm" is much larger compared to "tkm", but it is still true that the cost level for the younger group is consistently higher than for the older group.
- The variable "ukm" exhibits a very similar marginal effect to the variable "nkm". As "ukm" increases, the cost decreases, and this variable corresponds to an even higher decreasing rate compared to "nkm". Additionally, the cost level for the younger group is consistently higher than for the older group. However, it's interesting to note that the two groups seem to converge at very high levels of "ukm".

	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

Table 3: Descriptive statisitcs of the variables in the claim costs dataset

Variables	Coefficients	STE	Z	p-value
speedkm	1.000	-	-	-
age	0.153	0.039	3.910	***
agelic	0.097	0.036	2.706	***
agecar	-0.107	0.012	-9.016	***
parking	-0.162	0.248	-0.653	
tkm	-0.044	0.013	-3.367	***
nightkm	0.117	0.006	20.315	***
urbankm	0.141	0.005	27.061	***

Table 4: Estimated variable parameters and the selected smooth parameter is h= 3.27 (*** means p-value ≤ 0.005; **means p-value ≤ 0.01; * means p-value ≤ 0.5)

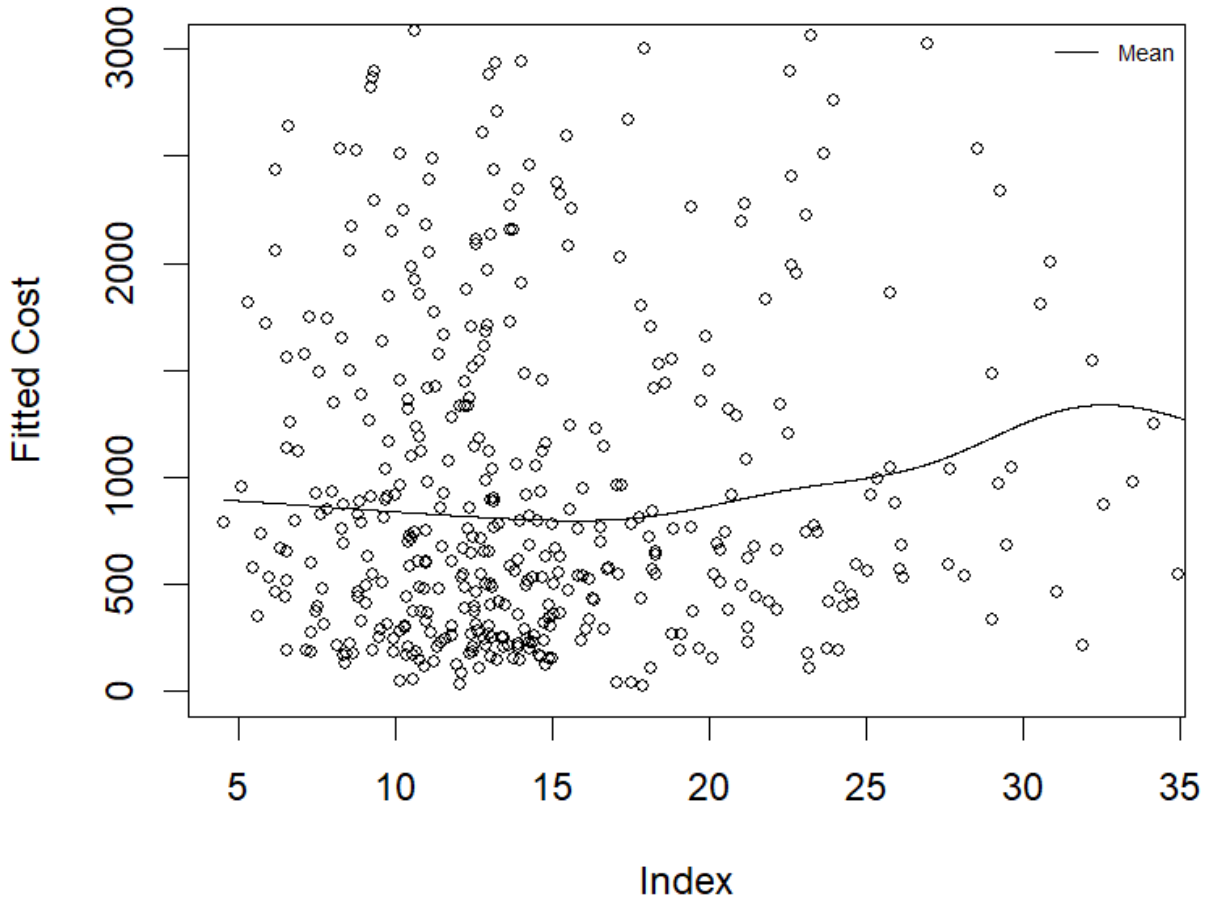

Figure 5: Prediction of conditional mean cost by combing Log-loglikelihood and MSE methods to estimate the variable parameters and smooth parameters respectively in a Single-index model

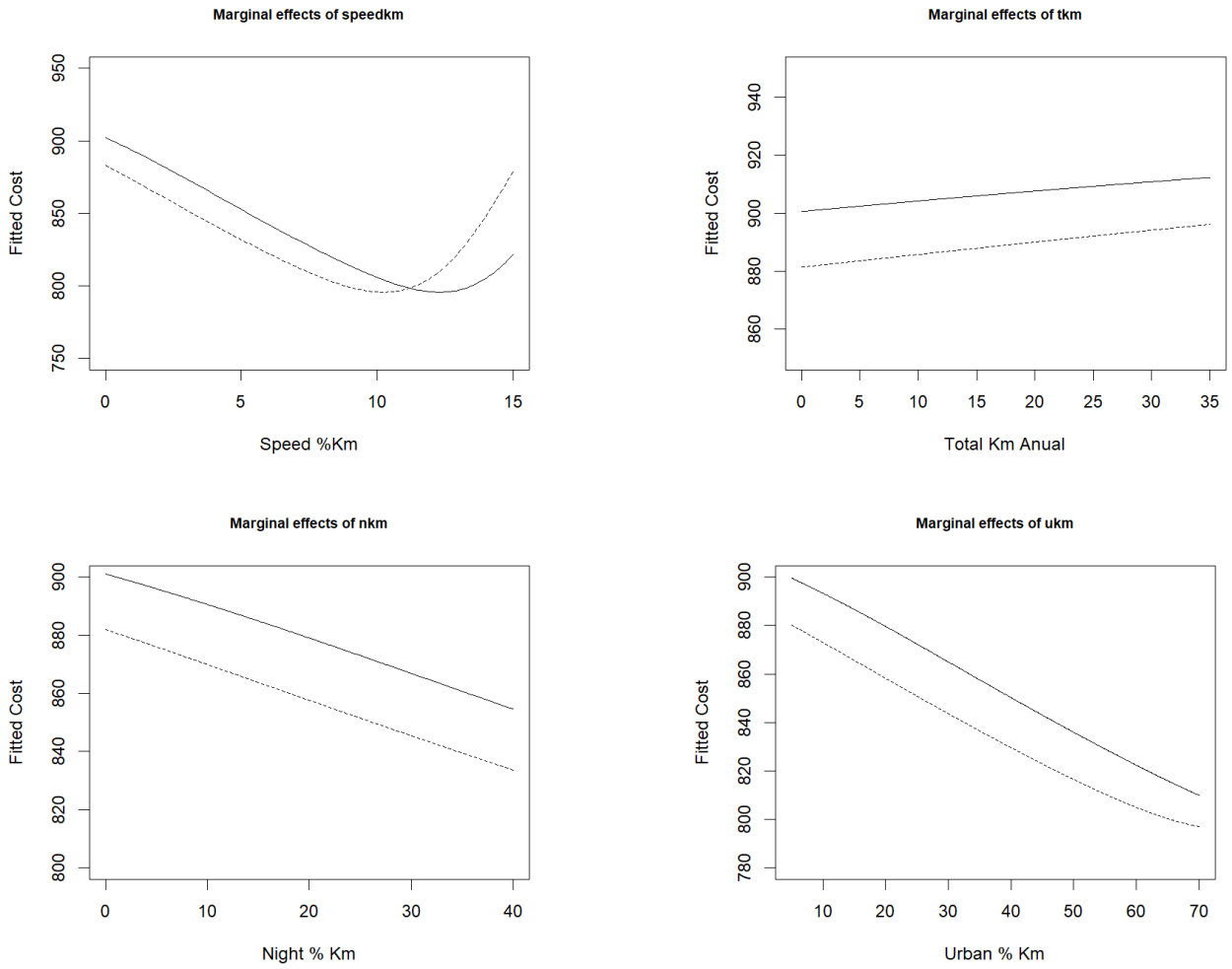

Figure 6: The marginal effects of the variables: speedkm(top left), tkm(top right), nkm(bottom left) and ukm(bottom righjt). Continuous lines: Younger group; Dotted lines: Older group

6 Conclusion

Compared to the traditional models the Single-Index model has some irreplaceble advantages. First of all, The Kernel Single-Index Model (KSIM) offers a high degree of flexibility in capturing nonlinear relationships between variables, making it well-suited for modeling complex data patterns that may not be adequately captured by linear models. This flexibility allows KSIM to handle a wide range of data types and distributions, providing robust modeling capabilities for diverse datasets. Further more KSIM is robust to outliers and noise in the data, as it relies on a single index rather than individual data points. This robustness ensures that the model can effectively capture the underlying structure of the data while minimizing the influence of outliers, resulting in more reliable and stable predictions. Additionally, KSIM is a nonparametric method that does not impose specific assumptions about the underlying data distribution. This flexibility allows KSIM to accommodate a wide range of data types and distributions, making it suitable for various applications across different domains.

Despite its strengths, The performance of KSIM can be sensitive to the choice of kernel function, bandwidth parameters, and other tuning parameters. This paper provides possible ways to estimate the variable parameters and find the potential optimal bandwidths for estimating the conditional mean value.

Here are some proposed improvements for the paper:

- Expanding Statistical Analysis: Explore methods to derive additional statistics from the conditional function, such as quantiles and expected values in the tail. This would require finding optimal bandwidths for each statistic, which could enhance the richness of the analysis and provide deeper insights into the data distribution.

- Optimizing Computational Efficiency: Implement optimizations in the code to reduce execution time, particularly as the number of observations increases. Since kernel approaches involve all observations in the estimation process, computational complexity grows exponentially with the dataset size. Optimizing the code can help mitigate this issue and improve overall efficiency.

- Theoretical Marginal Effects: Conduct theoretical analysis to calculate the marginal effects of each variable. While the paper analyzes empirical marginal effects based on observed data, theoretical calculations could provide additional insights and complement the empirical findings. This theoretical exploration could offer a deeper understanding of the underlying relationships between variables and help validate empirical results.

References

van den Berg, G.J., 2020. A general semiparametric approach to inference with marker-dependent hazard rate models. Journal of Econometrics , 43–67.

Bolancé, C., Cao, R., Guillen, M., 2024. Conditional likelihood based inference on sigle index-models for motor insurance claim severity. Universitat de Barcelona .

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

Hastie, T., Hardle, W., 1990. Generalised additive models. London: Chapman Hall/CRC .

Härdel, W.H., 1989. Applied nonparametric regression. Cambridge university press .

Härdel, W.H., Ichumura, H., 1993. Optimal smoothing in sigle-index models. Annals of statistics , 157–178.

Juditsky, H.M.A., Spokoiny, V., 2001. Direct estimation of the index coefficient in a sigle-index model. Annals of Statistics , 595–623.

Nadaraya, E.A., 1964. Some new estimates for distribution functions. Theory Prob. Applic. , 497–500.

Silverman, B., 1986. Density estimation for statistics and data anaysis. Chapman and Hall , 34–93.

Stock, P.J.J., Stoker, T., 1989. Semiparametric estimation of index coefficients. Econometrica , 1403–1430.

Stoker, T., 1986. Consistent extimation of scaled coefficients. Econometrica , 1461–1481.

Wand, M., Jones, M., 1995. Kernel smoothing. Chapman and Hall, 2-6 Boundary Row, London SE1 8HN, UK , 10–110.

A Appendix

A.1 Comparison table of ISE of simulations

A.2 Codes in R for the simulations

A.3 Real data from a Spanish insurance company used in application

A.4 Codes in R for estimate the mean cost and marginal effect analysis

		Sample size					
		n=500			n=2000		
Distribution	Parameters	Alt. 1	Alt. 2	Alt. 3	Alt. 1	Alt. 2	Alt. 3
Log Normal	$\mu = \theta^T x$, $\sigma = 0.5$	59.96	58.18	24.68	176.86	172.32	6.80
	$\mu = \theta^T x$, $\sigma = 2$	6,228.32	6,181.37	723.20	2,154.14	2,098.32	303.03
	$\mu = \theta^T x$, $\sigma = 5$	76,158,854	78,368,800	48,832,289	25,783,384	25,987,540	25,596,413
Weibull	$k = 1$, $\lambda = (\theta^T x)^2$	23.06	19.18	11.16	11.34	11.37	2.45
	$k = 2$, $\lambda = (\theta^T x)^2$	7.33	7.29	2.76	2.09	2.06	0.81
	$k = 3$, $\lambda = (\theta^T x)^2$	4.47	4.48	1.37	1.35	1.36	0.46
Log Logistic	$\beta = 1.5$, $\alpha = (\theta^T x)^2$	61.03	60.50	67.81	154.57	93.69	17.66
	$\beta = 2$, $\alpha = (\theta^T x)^2$	38.94	37.16	15.90	25.75	25.20	5.19
	$\beta = 3$, $\alpha = (\theta^T x)^2$	14.38	14.37	5.94	16.06	13.99	1.70

Table 5: Comparison of the ISE using different methods for estimating θ and h in different simulations with sample size 500 and 2000.