I N S T I T U T   D E

S T A T I S T I Q U E

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

D I S C U S S I O N

P A P E R

0123

SEMIPARAMETRIC ESTIMATION IN SINGLE INDEX POISSON REGRESSION A PRACTICAL APPROACH

D. CLIMOV, M. DELECROIX and L. SIMAR

http://www.stat.ucl.ac.be

Semiparametric Estimation in Single Index Poisson Regression: A Practical Approach

Daniela Climov*
Institut de Statistique
Université Catholique de Louvain

Michel Delecroix‡
CREST-ENSAI
Rennes

Léopold Simar *,‡
Institut de Statistique
Université Catholique de Louvain

June 22, 2001

Abstract

In a single index Poisson regression model with unknown link function, the index parameter can be root-n consistently estimated by the method of pseudo maximumum likelihood. In this paper, we study, by simulation arguments, the practical validity of the asymptotic behavior of the pseudo maximum likelihood index estimator and of some associated cross-validation bandwidths. A robust practical rule for implementing the pseudo maximum likelihood estimation method is suggested, which uses the bootstrap for estimating the variance of the index estimator and a variant of bagging for numerically stabilizing its variance. Our method gives reasonable results even for moderate sized samples thus it can be used for doing statistical inference in practical situations. The procedure is illustrated through a real data example.

Keywords: Single index models, Poisson regression, kernel estimation, bandwidth selection, bootstrap

---

*Constructive comments on previous versions from Jeffrey D. Hart and Research support from "Projet d'Actions de Recherche Concertées" (No. 98/03–217) from the Belgian Government are acknowledged.

‡Visiting Institut für Statistik und Ökonometrie, Humboldt-Universität zu Berlin with research support carried out within Sonderforschungbereich 373.

1 Introduction

We address the problem of estimating the direction parameter and the regression function in a Poisson single index model. The observed data $(X_i, Y_i) \in I\!R^k \times I\!N$, for $i = 1, \ldots, n$ are independent and the conditional distribution of Y_i given the vector of explicative variables X_i is Poisson with parameter depending on X_i.

We moreover assume that we have a Single Index Model (SIM), defined by the following condition:

$$\exists \beta_0 \in I\!R^k : \mathrm{E}[Y_i|X_i] = \mathrm{E}[Y_i|\beta_0 X_i], \tag{1.1}$$

where $\beta_0 x$ is the usual scalar product of two vectors from $I\!R^k$. Note that, if we denote by $R(\cdot)$ the regression function of Y_i on X_i, condition (1.1) is equivalent to:

$$R(x) = \mathrm{E}[Y_i|X_i = x] = g_{\beta_0}(\beta_0 x), \tag{1.2}$$

where g_β, for $\beta \in I\!R^k$ is defined as:

$$g_\beta(z) = \mathrm{E}[Y_i|\beta X_i = z]. \tag{1.3}$$

The unknown function $g_{\beta_0}(\cdot)$ is usually called the "link" function of the SIM. Since g_{β_0} is identified up to a multiplicative constant, we choose, as usual (see, e.g, Sherman, 1994, Härdle, Hall, Ichimura, 1993), to fix the first component of β_0 to 1. Another solution would be to fix the norm of β_0.

This kind of models has been extensively used in the literature in actuarial sciences, in biometrics or in econometrics, but with a fixed link function in the framework of General Linear Models (GLM, see McCullagh and Nelder, 1989). Here we focus on the problem of estimating simultaneously the link and the parameters β in the case of a Poisson Single Index regression model:

$$Y_i|X_i = x \sim Po(g_{\beta_0}(\beta_0 x)) \tag{1.4}$$

One of the most attractive approaches for estimating this kind of models is based on M-estimation methods. Under only the condition (1.1), a consistent estimator $\hat{\beta}_n$ is defined by maximizing with respect to β the empirical mean of some objective function Ψ:

$$\hat{\beta}_n = \arg\max_\beta \frac{1}{n} \sum_{i=1}^n \Psi(Y_i, \hat{g}_{\beta,h_n}(\beta X_i)), \tag{1.5}$$

where $\hat{g}_{\beta,h_n}$ is a nonparametric estimator of the function g_β and h_n is a series of bandwidths used in the nonparametric estimator, which tends to zero at some appropriate rate as $n \to \infty$.

Usually, the Nadaraya-Watson leave-one-out estimator of $g_\beta(\beta X_i)$, is used in (1.5). It is defined as:

$$\hat{g}_{\beta,h_n}^{(-i)}(\beta X_i) = \frac{\sum_{j\neq i} Y_j \, K_{h_n}(\beta X_i - \beta X_j)}{\sum_{j\neq i} K_{h_n}(\beta X_i - \beta X_j)}, \tag{1.6}$$

where $K_{h_n}(x) = h_n^{-1} K(x/h_n)$ and K is a fixed kernel function (typically a symmetric probability function).

Many objective functions Ψ can be chosen and under general regularity conditions (Sherman (1994), Delecroix and Hristache (1999)), it is easy to prove that $\hat{\beta}_n$ achieves the root-n consistency. The idea is based on the fact that, since $\hat{g}_{\beta,h_n}$ converges to g_β as $n \to \infty$, we have

$$\hat{\beta}_n \longrightarrow \arg\max_\beta \mathrm{E}\left[\Psi(Y_i, g_\beta(\beta X_i)\right], \tag{1.7}$$

at the usual root-n rate of convergence. The remaining point is then to analyze under which conditions the limiting term on the right hand side of (1.7) is equal to the true unknown β_0.

Delecroix and Hristache (1999) have obtained the following general result: for *any* distribution for the vector (X_i, Y_i), the single index model assumption (1.1), implies that

$$\beta_0 = \arg\max_\beta \mathrm{E}\left[\Psi(Y_i, g_\beta(\beta X_i))\right]$$

if and only if the objective function Ψ is the log of some linear exponential density, that is:

$$\Psi(y, m) = \log f(y, m) = A(m) + B(y) + C(m)y, \tag{1.8}$$

where A and C are twice continuously differentiable and m is the mean of the distribution whose density is $f(y, m)$.

In our problem here (Poisson regression), it is obviously the case if we choose as objective function Ψ the maximum likelihood function. One can find in Delecroix and Hristache (1999), asymptotic efficiency arguments justifying this particular choice. In this case, the function Ψ used in the maximization problem (1.5) turns out to be:

$$\Psi(Y_i, \hat{g}_{\beta,h_n}^{(-i)}(\beta X_i)) = Y_i \log(\hat{g}_{\beta,h_n}^{(-i)}(\beta X_i)) - \hat{g}_{\beta,h_n}^{(-i)}(\beta X_i). \tag{1.9}$$

The estimation procedure is also called "pseudo maximum likelihood" since the true unknown link function g_{β_0} is replaced by some appropriate estimator of g_β. From now on, in this paper, we will keep this particular choice of objective function.

Once β_0 has been consistently estimated by solving (1.5), the regression function $R(x) = \mathrm{E}(Y|X = x)$ can be estimated, in a second stage, from the nonparametric regression of Y_i on the estimated index $\hat{\beta}_n X_i$, using the Nadaraya-Watson estimator, which has the same form as in (1.6), except that here, the ith observation is included in the sum:

$$\begin{aligned}\hat{R}_n(x) &= \hat{g}_{\hat{\beta}_n,h'_n}(\hat{\beta}_n x) \\ &= \frac{\sum_{i=1}^n Y_i \, K_{h'_n}(\hat{\beta}_n X_i - \hat{\beta}_n x)}{\sum_{i=1}^n K_{h'_n}(\hat{\beta}_n X_i - \hat{\beta}_n x)},\end{aligned} \tag{1.10}$$

where h'_n is another series of bandwidths converging to zero such as $\hat{R}_n(x)$ converges at the optimal rate of convergence in nonparametric regression, *i.e.*, $h'_n \approx n^{-1/5}$.

This two-steps approach of the M-estimation method presents the inconvenient that two series of bandwidths h_n and h'_n need to be chosen in advance. Practitionners know how sensitive are the results, in nonparametric regression, to the particular choice of the bandwidths and that the existing theoretical asymptotic formulae do not help too much.

An alternative approach to this bandwidth selection problem is proposed by Härdle, Hall, Ichimura (1993), generalized by Delecroix, Hristache, Patilea (1999). They suggest an *one-step* method for selecting the same bandwidth for the root-n estimation of β_0 and for the kernel estimation of the function R, by optimizing the objective function Ψ simultaneously with respect to β and to h. The resulting bandwidth is again of order $n^{-1/5}$. Thus, the one-step M-estimation is defined as follows:

$$(\hat{\beta}_n, \hat{h}_n) = \arg\max_{h,\beta} \frac{1}{n} \sum_{i=1}^n \Psi[Y_i, \hat{g}_{\beta,h}^{(-i)}(\beta X_i)] \tag{1.11}$$

and the estimator of the regression fonction is then given by (1.10) with the bandwidth $\hat{h}_n$:

$$\hat{R}_n(x) = \hat{g}_{\hat{\beta}_n,\hat{h}_n}(\hat{\beta}_n x). \tag{1.12}$$

This approach seems to be attractive and coherent: the procedure gives in fact, for any given value of β, the optimal value for h by a cross-validation criterion on Ψ, and the optimal value of β can be derived.

The asymptotics of the above estimators is well-known (see *e.g.* Ichimura, 1993, Sherman, 1994, Delecroix, Hristache, 1999, Delecroix, Hristache, Patilea, 1999). In both cases (one-step and two steps), under technical conditions on the joint distribution of (X_i, Y_i), on the smoothness of the functions g_β, on the kernel K and on the bandwidths, one gets the almost everywhere root-n convergence of $\hat{\beta}_n$ to β_0.

In our case of Poisson regression model, this result particularizes as follows:

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma), \tag{1.13}$$

where

$$\Sigma^{-1} = \mathrm{E}\left\{ \frac{1}{g_{\beta_0}(\beta_0 X_i)} \left.\frac{\partial g_\beta(\beta X_i)}{\partial \beta}\right|_{\beta=\beta_0} \left.\frac{\partial g_\beta(\beta X_i)}{\partial \beta^T}\right|_{\beta=\beta_0} \right\}. \tag{1.14}$$

As a consequence, $\hat{\beta}_n$ is asymptotically efficient, since it reaches the asymptotic semiparametric efficiency bound (see Newey, 1990, for a general formulation of the semiparametric efficiency bounds in single index models).

To the best of our knowledge, the only existing result on the asymptotic behavior of estimators of the regression function $R(x)$ in this framework, concerns the one-step method. Delecroix, Hristache, Patilea (1999) show that $\hat{R}_n(x)$, as defined in (1.12), shares the following property:

$$\sqrt{n\hat{h}_n}\left(\hat{g}_{\hat{\beta}_n,\hat{h}_n}(\hat{\beta}_n x) - g_{\beta_0}(\beta_0 x) - \hat{h}_n^2\omega(\beta_0 x)\right) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, V(\beta_0 x)\right) \tag{1.15}$$

where

$$\begin{aligned} \omega(z) &= (K_1/2)\left[g''_{\beta_0}(z) + 2g'_{\beta_0}(z)f'_{\beta_0}(z)/f_{\beta_0}(z)\right] \\ V(z) &= K_2 \mathrm{Var}(Y \mid \beta_0 X = z)/f_{\beta_0}(z) \\ K_1 &= \int u^2 K(u)du \\ K_2 &= \int K^2(u)du. \end{aligned}$$

with f_{β_0} denoting the density of the true index $\beta_0 X$. For a Poisson regression, $\mathrm{Var}(Y \mid \beta_0 X = z) = g_{\beta_0}(z)$.

For the two-steps estimator of the regression, a similar property can be provided using the approach of Härdle and Stoker (1989) developed for the case when an Average Derivative Estimator (ADE) of β is available. They show, indeed, that if $\tilde{\beta}_n$ is a root-n consistent estimator of β and if h'_n is a bandwidth of order $n^{-1/5}$ used to construct the Nadaraya-Watson estimator $\hat{g}_{\tilde{\beta}_n,h'_n}$, as in (1.10), then the resulting regression shares the property:

$$n^{2/5}[\hat{g}_{\tilde{\beta}_n,h'_n}(\tilde{\beta}_n x) - g_{\beta_0}(\beta_0 x)] \xrightarrow{\mathcal{L}} \mathcal{N}(\omega(\beta_0 x), V(\beta_0 x)), \tag{1.16}$$

where the bias $\omega(\beta_0 x)$ and the variance $V(\beta_0 x)$ have the same expression as in (1.15). The only difference with our framework is that they use, for $\tilde{\beta}_n$, an ADE, which is root-n consistent but not asymptotically efficient and relies on the continuity assumption of each of the regressors X. In our case, the M-estimator $\hat{\beta}_n$ does not rely on such condition and is asymptotically efficient.

Using the above results for the one and two steps regression estimators, asymptotic confidence intervals can be constructed for $\hat{R}(x)$, but this involves the nonparametric estimation of the density and regression function and their derivatives appearing in the bias and variance formulae, which is not easy in practice.

The aim of our paper is threefold. First, we investigate, by Monte-Carlo experiments the finite sample properties of the above M-estimators in the Poisson regression case. In particular, it appears that the one-step estimator performs better than the two-steps and that the asymptotic results for the variance of $\hat{\beta}_n$, provided by (1.13), should not be used unless a huge number of observations is available. In addition, in practical situations, the derivation of this variance is often intractable and depends on many unknown quantities. This suggests that a bootstrap approach could provide an easier approximation for the variance of $\hat{\beta}_n$. Secondly, for computing the one-step estimator with real data, we propose a practical and robust method (with respect to the numerical instability). The solution of (1.11) involves rather intricate nonlinear numerical optimization procedures which are numerically unstable when trying to optimize simultaneously in β and in h. In particular, direct search optimization procedures, like the simplex algorithm, provide local optima thus that starting values are influential. The idea is to use the two-steps method for providing the initial guess for the starting values of (β, h) in the algorithm and then to use a variant of Breiman's (1996) bagging method to stabilize the estimation of β_0. The variance of our estimator is estimated through a bootstrap method. Then, even for moderate sample sizes, confidence intervals for β_0 can easily be derived. Finally, we illustrate our method by applying it to real data.

The paper is organized as follows. Section 2 explains how the one-step and two-steps M-estimation methods can be implemented. Section 3 investigates the finite sample properties of our estimators in a particular Monte-Carlo scenario. Section 4 indicates how to perform a bootstrap algorithm in this framework, in order to provide a numerically stable estimator of β_0 and reasonable estimates of its variance. Then Section 5 illustrates with a real data example and Section 6 concludes.

2 Implementation of the M-estimation Method

2.1 The one-step method

As already pointed out, for the practical implementation of estimators in semiparametric or non-parametric models, an important issue is the choice of the bandwidth parameter. For the one-step method, the bandwidth is obtained, from a theoretical point of view, by solving (1.11), which, for the Poisson regression, turns out to be:

$$(\hat{\beta}_n, \hat{h}_n) = \arg\max_{h,\beta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \log(\hat{g}_{\beta,h}^{(-i)}(\beta X_i)) - \hat{g}_{\beta,h}^{(-i)}(\beta X_i) \right\}. \tag{2.1}$$

The optimization can be performed on grids of values for β and for h but the procedure would be very expensive in terms of computation time if these grids are too large. We prefered to use a "direct search" algorithm based on the Nelder-Mead simplex idea which, in fact, acts as a "dynamic" grid search. This algorithm needs reasonable starting values for $(\hat{\beta}_n, \hat{h}_n)$ because it provides local minima.

For $\hat{\beta}_n$, an initial guess could be provided by the consistent ADE estimator when all the explanatory variables X are continuous (see *e.g.* Powell, Stock and Stoker, 1989), then the initial value for $\hat{h}_n$ could be obtained by cross-validation for this fixed value of $\hat{\beta}_n$. In the case of discrete and continuous regressors, the initial guess values for β can be provided either by the direct (noniterative) semiparametric estimation method proposed by Horowitz and Härdle (1996) or by the two-steps method described in the next section. The last approach, which is easier to implement, will be followed below.

2.2 The two-steps approach

As presented above, the maximization problem (1.5) has to be solved using a fixed appropriate bandwidth h_n. We propose to define an estimator $\hat{\beta}_n$ by solving, at a first step,

$$\hat{\beta}_n = \arg\max_{\beta} \frac{1}{n} \sum_{i=1}^{n} \left\{ Y_i \log(\hat{g}^{(-i)}_{\beta,h_n}(\beta X_i)) - \hat{g}^{(-i)}_{\beta,h_n}(\beta X_i) \right\}. \tag{2.2}$$

where h_n is a pilot bandwidth determined by an automatic adaptive rule. The rule is based on the very simple idea that, for a given β, the bandwidth used to estimate g_β should be optimal for estimating the marginal density of βX . The advantage of this approach is that in many cases, a simple and fast to compute rule of thumb is available. When a Gaussian kernel is used, Silverman's (1986) normal reference rule can be used:

$$h_n = h_{\beta,n} = 1.06\, s_{\beta X}\, n^{-1/5}, \tag{2.3}$$

where $s_{\beta X}$ is the standard deviation of the values $\beta X_1, \ldots, \beta X_n$.

Plugging the value of h_n in (2.2) yields the estimator $\hat{\beta}_n$. Here again, the simplex direct-search method could be used. From a numerical point a view, the optimization problem is only in β here, therefore, it is much more stable and faster than the one-step problem (2.1), where we optimize in (β, h). So here, many starting values for $\hat{\beta}_n$ can be tried to fix the optimum of the objective function.

Formally, the rule of thumb (2.3) is optimal for estimating the density of βX, only if the random variable βX is normally distributed, but in practice, this rule is a reasonable choice for many distributions as far as they are unimodal. Note also that a robust version of (2.3) is available (see Silverman, 1986). In addition, this rule provides the optimal order of $n^{-1/5}$ corresponding to the order for h_n in the one-step problem (see Delecroix, Hristache, Patilea, 1999). Finally, in the Monte-Carlo experiment, we will see that, in the chosen scenario, the estimator defined by (2.2) performs pretty well even in moderate sample size. In any case, it provides a reasonable starting value for computing our one-step estimator. Note that in order to solve (2.2), an initial guess for β could be provided by a parametric estimator (for example, GLM).

The value of $\hat{h}_n$ needed to estimate the regression function is now obtained by solving, at a second step:

$$\hat{h}_n = \hat{h}_{\hat{\beta}_n,n} \quad = \quad \arg\max_h \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i \log(\hat{g}^{(-i)}_{\hat{\beta}_n,h}(\hat{\beta}_n X_i)) - \hat{g}^{(-i)}_{\hat{\beta}_n,h}(\hat{\beta}_n X_i)\right\}, \tag{2.4}$$

which is a rather simple one-dimensional optimization problem. The two-steps method ends up with the value of $(\hat{\beta}_n, \hat{h}_n)$ from which the regression function can be estimated by using (1.10).

The two-steps method is much faster and easier to implement than the one-step method described above since it tackles the optimization problem in (β, h) separately for β and for h using the "trick" (2.3). This is of course true, when a simple rule is available to fix a reasonable bandwidth for estimating the marginal density of βX. If this is not the case, the simple rule (2.3) could be replaced by either some cross-validation criterion for estimating the density of βX or by the Härdle and Marron's (1995) rule for selection of the bandwidth of a kernel regression. It is not clear then, if we would still gain any computing time, but we would certainly gain in the numerical stability of the procedure. Indeed, the optimization procedure to get $(\hat{\beta}_n, \hat{h}_n)$ in (2.1) involves the computation of many Nadaraya-Watson type estimators where the denominators, in (1.6), could be near or even equal to zero for many observed βX_i (in particular for some combinations of large β and small h), unless good starting values for the variables are provided. This kind of numerical problem is avoided in a cross-validation done for the density estimation of βX. The properties of the two methods are investigated and compared in the Monte-Carlo experiment described in the next section. The one-step method behaves better, but the two-steps method provides also sensible results for estimating both β_0 and h.

3 Simulations

In this section, we investigate, by Monte-Carlo experiments, the finite sample properties of the one and two-steps estimators from the perspective of estimating both the direction vector β_0 and the regression function R itself.

As far as the estimation of β_0 is concerned, we evaluate the performances of our estimators by computing, as usual, their Monte-Carlo bias, variance and MSE:

$$\text{bias}(\hat{\beta}_n) = \frac{1}{M}\sum_{m=1}^{M}(\hat{\beta}^{(m)}_n - \beta_0) \tag{3.1}$$

$$\mathrm{MSE}(\hat\beta_n) = \frac{1}{M}\sum_{m=1}^{M}(\hat\beta_n^{(m)} - \beta_0)^2 \tag{3.2}$$

$$\mathrm{Var}(\hat\beta_n) = \frac{1}{M}\sum_{m=1}^{M}(\hat\beta_n^{(m)} - \bar{\hat\beta}_n)^2, \tag{3.3}$$

where M is the number of Monte-Carlo replications and $\hat\beta_n^{(m)}$ is the value of $\hat\beta_n$ obtained for the mth sample of size n. These statistics will be reported in the tables below. To appreciate and compare the values of bias$(\hat\beta_n)$, it is useful to compute the Monte-Carlo standard errors of the bias:

$$\mathrm{std}(\mathrm{bias}) = \frac{1}{\sqrt{M}}\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat\beta_n^{(m)} - \bar{\hat\beta}_n)^2} = \sqrt{\frac{\mathrm{Var}(\hat\beta_n)}{M-1}}. \tag{3.4}$$

Finally, we will also provide the asymptotic distribution of $\hat\beta_n$ for our Monte-Carlo scenario. This allows to compare these asymptotic results with the obtained Monte-Carlo sampling distributions. We will compare the Monte-Carlo variances $\mathrm{Var}(\hat\beta_n)$ with the asymptotic efficiency bound given in (1.14) and perform some tests for the normality of the sampling distributions of $\hat\beta_n$, for various finite sample sizes.

When the estimation of the regression itself is concerned, we have many choices to appreciate the global quality of the fit. In what follows, we will concentrate on two measures of the goodness of fit: the average squared error (ASE) and the average Kullback-Leibler distance (AKL). The same kind of global measures of goodness of fit for Poisson regression is used in Climov, Hart, Simar (2000).

Let f_X denote the density of the covariate X, μ be an arbitrary function from $I\!R$ to $I\!R$ and let β be any k-vector. The L_2 distance between the true model (g_{β_0}, β_0) and a candidate (μ, β) is the integrated square error (ISE):

$$\begin{aligned}\mathrm{ISE}(\mu,\beta) &= \mathrm{E}\left\{[\mu(\beta X) - g_{\beta_0}(\beta_0 X)]^2\right\}\\ &= \int \left\{[\mu(\beta x) - g_{\beta_0}(\beta_0 x)]^2\right\} f_X(x)\,dx,\end{aligned}$$

and an empirical version of it is given by the average squared error:

$$\mathrm{ASE}(\mu,\beta) = \frac{1}{n}\sum_{i=1}^{n}[\mu(\beta X_i) - g_{\beta_0}(\beta_0 X_i)]^2. \tag{3.5}$$

This quantity can be evaluated for (μ,β) equal to the estimates $(\hat g_{\hat\beta_n,\hat h_n}, \hat\beta_n)$ obtained for a given data file.

In the case of a Poisson regression, the Kullback-Leibler discrepancy between the true model (g_{β_0}, β_0) and a candidate (μ,β) can be computed as:

$$\mathrm{KL}(\mu,\beta) = \mathrm{E}\left\{\log\frac{f_0(Y; g_{\beta_0}(\beta_0 X))}{f_0(Y;\mu(\beta X))}\right\}$$

$$= \int \left\{ \mu(\beta x) - g_{\beta_0}(\beta_0 x) + g_{\beta_0}(\beta_0 x) \log \frac{g_{\beta_0}(\beta_0 x)}{\mu(\beta x)} \right\} f_X(x)\, dx,$$

where $f_0(y; \lambda)$ denotes the Poisson discrete density with mean λ. An empirical version of it, which does not involve knowledge of $f_X(x)$, is the average KL distance:

$$\mathrm{AKL}(\mu, \beta) = \frac{1}{n} \sum_{i=1}^{n} \left[\mu(\beta X_i) - g_{\beta_0}(\beta_0 X_i) + g_{\beta_0}(\beta_0 X_i) \log \frac{g_{\beta_0}(\beta_0 X_i)}{\mu(\beta X_i)} \right]. \qquad (3.6)$$

Again, this quantity will be be evaluated for (μ, β) equal to the estimates $(\hat{g}_{\hat{\beta}_n, \hat{h}_n}, \hat{\beta}_n)$ obtained for a given data file.

In the Monte-Carlo experiments, we will report the mean values of the ASE and AKL over all the simulated Monte-Carlo samples.

3.1 Monte-Carlo set-up

The simulation scenario can be described as follows. The simulated data (X_i, Y_i), $i = 1, \ldots, n$, are independent and identically distributed. X_i is a bivariate vector having independent components $X_{ij} \sim \mathcal{N}(0, 1)$, $j = 1, 2$ and Y_i is Poisson distributed with conditional mean depending on the index $Z_i = (\beta_0 X_i)$. The direction vector used to generate the data is $\beta_0 = (1, 9)^T$, where the first component of β_0 is fixed at the value 1 for identifiability reasons. Note that this scenario is not particularly favorable for the estimation of β_0, since most of the values of Z_i are concentrated near zero. A uniform distribution for X_i would be more favorable. The main reason for choosing this particular scenario is that the required mathematics for the evaluation of the asymptotical variance of $\hat{\beta}_n$ is not very complicated in this case, so that we can come up with a final value of the asymptotical variance which will be compared to the empirical value.

To describe the dependence of the Poisson conditional mean on the index Z, we choose a quadratic link function:

$$g_{\beta_0}(\beta_0 x) = (\beta_0 x)^2.$$

Increasing sample sizes were used, $n = 50, 100, 200, 400, 800, 1000$, in order to investigate the asymptotics. Due to computing time limitations, we restrict the Monte-Carlo (MC) experiment to $M = 500$ replications. The Nadaraya-Watson estimates are computed with a standard normal kernel.

The Monte-Carlo variances of $\hat{\beta}_n$ can be compared to their theoretical asymptotic counterparts. In the Appendix A we evaluate the expression of the functions $g_\beta(\beta x)$ in this set-up which are used in Appendix B to compute the asymptotic variance of $\sqrt{n}(\hat{\beta}_n - \beta_0)$, as given by (1.14). It turns out that, in our Monte-Carlo scenario the vector β_0 has only one unknown component and we have:

$$\Sigma = \frac{1+\beta_0^2}{4}. \tag{3.7}$$

3.2 Practical implementation and results

We carried out the estimation of h and β exactly as described in Section 2. Through all the Monte-Carlo experiments, even with carefully chosen initial values, the highly non-linear optimization procedures used here can produce local minima which are numerical outliers. This numerical instability can be explained as follows. In the evaluation of the objective function, many Nadaraya-Watson leave-one-out estimators (1.6) have to be computed, whose denominator (and numerator) could be almost equal or equal to zero. This is particularly true in our MC-setup since vey few points Z_i are generated in the tails.

We propose to eliminate the MC-samples providing numerical outliers by an “automatic” adaptive rule. At the end of the MC-loop, we eliminate the samples which provided outlying values for $\hat{\beta}_n$ and /or for $\hat{h}_n$. We define an outlier, as usual, as being a value outside the whiskers of a boxplot: any value larger (smaller) than the 3rd quartile (1st quartile) plus (minus) 1.5 times the interquartile range. This procedure is followed in all the simulations done below and also in the boostrap algorithm of Section 4. The percentage of samples eliminated by this method ranged from 2 to 8 % in our Monte-Carlo experiments, depending on the sample size. The number of the remaining MC-samples in each case is reported in the tables below.

3.2.1 Two-steps method

In Table 1, we present the performances for the estimation of β_0 using the two-steps approach. Under the heading AVar and r we report the corresponding values of the theoretical asymptotic variances computed with (3.7) and, respectively, the ratio of the empirical variance to AVar. We display also some information on the bandwidth values chosen at this first step: the averages of all the MC-values of $h_{\hat{\beta}_n,n}$, computed by the rule of thumb (2.3) at $\hat{\beta}_n$, and their MC-standard deviations. We report also, for comparison, $h_{0,n}$, the theoretical value of the corresponding bandwith provided by the same rule of thumb at the true value β_0 and for the true variance of $\beta_0 X$ in our chosen scenario:

$$h_{0,n} = 1.06\sqrt{\mathrm{Var}(\beta_0 X)}\; n^{-1/5} = 1.06\sqrt{1+\beta_0^2}\; n^{-1/5}. \tag{3.8}$$

From this table, it appears that the bias in estimating β_0 is quite negligible when n increases but the variance of our estimator is larger than the bounds given by the column AVar, in particular for small n. We will see below that the one-step procedure behaves better.

	$\hat{\beta}_n$						$h_{\hat{\beta}_n,n}$		$h_{0,n}$	MC
n	bias	std(bias)	Var	MSE	AVar	r	mean	std		
50	-0.0606	0.0785	2.8463	2.8500	0.4100	6.94	4.3421	0.9345	4.390	463
100	-0.0579	0.0371	0.6498	0.6532	0.2050	3.17	3.7968	0.4382	3.821	472
200	-0.0018	0.0231	0.2593	0.2593	0.1025	2.53	3.3153	0.2458	3.326	488
400	0.0216	0.0145	0.1024	0.1029	0.0512	2.00	2.9036	0.1498	2.896	487
800	-0.0098	0.0090	0.0391	0.0392	0.0256	1.53	2.5156	0.0862	2.521	483
1000	0.0043	0.0087	0.0368	0.0368	0.0205	1.80	2.4107	0.0769	2.411	491

Table 1: *Monte-Carlo simulations: first-step estimation of $\beta_0 = 9$ in a two-steps procedure. MC is the number of the remaining MC-samples in the analysis from* 500 *replications.*

In any case, this very simple estimator of β_0 provides sensible results and can certainly serve as a first guess for computing our one-step estimator.

The results of the second step are summarized in Table 2. The values of $\hat{h}_n$ were found by a simple one-dimensional optimization procedure by solving (2.4), with h constrained to be positive. We report the MC-averages of the optimal bandwidths, along with their MC-standard deviations. The quality of the regression fit may be appreciated through the MC-averages and standard deviations of the ASE and AKL criteria. These values will allow a comparison with the performances of the one-step estimator below. It will be seen that the performances of our two-steps regression estimator are less good but, still, they are quite reasonable. Thus, the value of $\hat{h}_n$ obtained through this procedure can again serve as initial value for the one-step algorithm.

n	$\hat{h}_n$		AKL$(\hat{g}_{\hat{\beta}_n,\hat{h}_n}, \hat{\beta}_n)$		ASE$(\hat{g}_{\hat{\beta}_n,\hat{h}_n}, \hat{\beta}_n)$	
50	0.8924	(0.3210)	0.3322	(0.1884)	62.9579	(32.8283)
100	0.6856	(0.1839)	0.1728	(0.0490)	43.1134	(19.2534)
200	0.5780	(0.1308)	0.1034	(0.0243)	29.9574	(10.5182)
400	0.4838	(0.0876)	0.0603	(0.0117)	20.7421	(6.7120)
800	0.4099	(0.0543)	0.0352	(0.0062)	13.6227	(3.9465)
1000	0.3988	(0.0543)	0.0304	(0.0054)	12.2049	(3.4723)

Table 2: *Results for the second step: bandwiths and goodness of fit measures: averages over the MC-replications, standard deviations are between parenthesis.*

Figure 1 displays the global performances of our two-steps procedure in terms of the sample size n. We can see how the values $\hat{\beta}_n$ tend to be more and more concentrated around the true value $\beta_0 = 9$ as the sample size increases. Also, the box-plots of the values $\hat{h}_n$, AKL and ASE are more concentrated towards 0 as n increases.

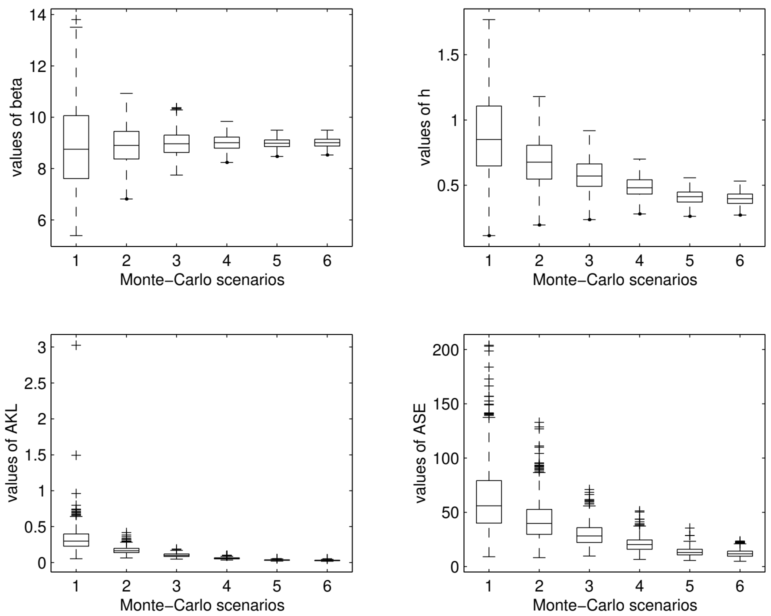

Figure 1: *Two-steps procedure: box-plots of the values of* $\hat{\beta}_n, \hat{h}_n, AKL$ *and* ASE*, over the MC-replications. The sample sizes are* $n = 50, 100, 200, 400, 800$ *and* 1000 *respectively.*

3.2.2 One-step method

The implementation of the one-step method follows the procedure explained in Section 2. The initial values for β and h are given by the two-steps procedure described above.

Table 3 presents the results of the MC-experiment: the performances of $\hat{\beta}_n$ and the average and standard deviation values of $\hat{h}_n$. For comparison, the average values of the optimal bandwidth $\hat{h}^0_{\beta_0,n}$ are also reported, where $\hat{h}^0_{\beta_0,n}$ is defined as:

$$\hat{h}^0_{\beta_0,n} = \arg\max_h \frac{1}{n}\sum_{i=1}^n \left\{ Y_i \log(\hat{g}^{(-i)}_{\beta_0,h}(\beta_0 X_i)) - \hat{g}^{(-i)}_{\beta_0,h}(\beta_0 X_i) \right\}. \tag{3.9}$$

This is a simple one-dimensional optimization procedure, where h is constrained to be positive.

The results can be compared with those of Table 1. The two procedures (one-step and two-stepss) seem to provide the same quality for the estimation of β_0 as far as the bias is concerned. But for our chosen scenario, the variance (and MSE), is better for the one-step estimator, though the empirical MC-variances are still larger than the theoretical asymptotic values AVar. The normality of the sampling distribution of $\hat{\beta}_n$ is investigated below.

Table 4 presents the performance of the fit of the regression by reporting the averages of ASE and AKL criteria, which can be compared with those obtained in Table 2 for the two-steps procedure. In view of the AKL criterion, which is the most sensible performance

	$\hat{\beta}_n$						$\hat{h}_n$		$\hat{h}^0_{\beta_0,n}$	MC
n	bias	std(bias)	Var	MSE	AVar	r	mean	std		
50	-0.0063	0.0511	1.2407	1.2408	0.4100	3.03	0.8065	0.2819	0.8928	476
100	0.0137	0.0337	0.5422	0.5424	0.2050	2.64	0.6575	0.1891	0.6977	479
200	0.0115	0.0204	0.1966	0.1968	0.1025	1.92	0.5520	0.1285	0.5891	475
400	-0.0107	0.0124	0.0746	0.0747	0.0512	1.45	0.4751	0.0851	0.4904	483
800	-0.0016	0.0087	0.0373	0.0373	0.0256	1.45	0.4118	0.0564	0.4114	489
1000	-0.0120	0.0068	0.0218	0.0219	0.0205	1.06	0.3896	0.0514	0.3969	475

Table 3: *Monte-Carlo simulations: estimation of $\beta_0 = 9$ using the one-step procedure. MC is the number of the remaining MC-samples from* 500 *replications.*

measure in the Poisson setting, the one-step procedure provides, in our MC-scenario, slightly better global fit of the regression than the two-steps approach but the order of magnitude of the measures of the quality of the fit are comparable.

n	AKL($\hat{g}_{\hat{\beta}_n,\hat{h}_n}, \hat{\beta}_n$)		ASE($\hat{g}_{\hat{\beta}_n,\hat{h}_n}, \hat{\beta}_n$)	
50	0.2790	(0.0989)	59.3787	(28.1859)
100	0.1722	(0.0469)	43.3698	(17.9833)
200	0.1012	(0.0216)	30.0675	(10.3864)
400	0.0600	(0.0116)	20.9637	(6.9062)
800	0.0355	(0.0061)	14.0052	(4.0600)
1000	0.0300	(0.0049)	12.5413	(3.5662)

Table 4: *One-step procedure: goodness of fit measures for the regression, averages over the MC-replications, standard deviations are between parenthesis.*

It is certainly worth to have an idea of the sampling distribution of $\hat{\beta}_n$. Figure 2 displays the box-plots of the MC-values of $\hat{\beta}_n$ for increasing values of n. The picture does not show real departure from symmetry for the distribution of $\hat{\beta}_n$. Figure 3 shows the sampling densities of the standardized $\hat{\beta}_n$, *i.e.* $\sqrt{\mathrm{AVar}^{-1}}\,(\hat{\beta}_n - \beta_0)$ compared to the theoretical limit $\mathcal{N}(0,1)$. Here, a substantial difference remains even for large n. The following Kolmogorov-Smirnov test for the normality was performed:

$$H_0 \;:\quad \sqrt{\mathrm{AVar}^{-1}}\,(\hat{\beta}_n - \beta_0) \sim \mathcal{N}(0,1). \tag{3.10}$$

The p-values for the tests are equal to 0, 0, 10^{-05}, 0.0189, 0.0115 and 0.0478, for the sample sizes n =50, 100, 200, 400, 800 and 1000 respectively: the asymptotic distribution is rejected at any sensible level for n smaller than 400. This is mainly due to the underestimation of the variance of $\hat{\beta}_n$ by AVar. For larger n, the p-values remain very small, for the same reason

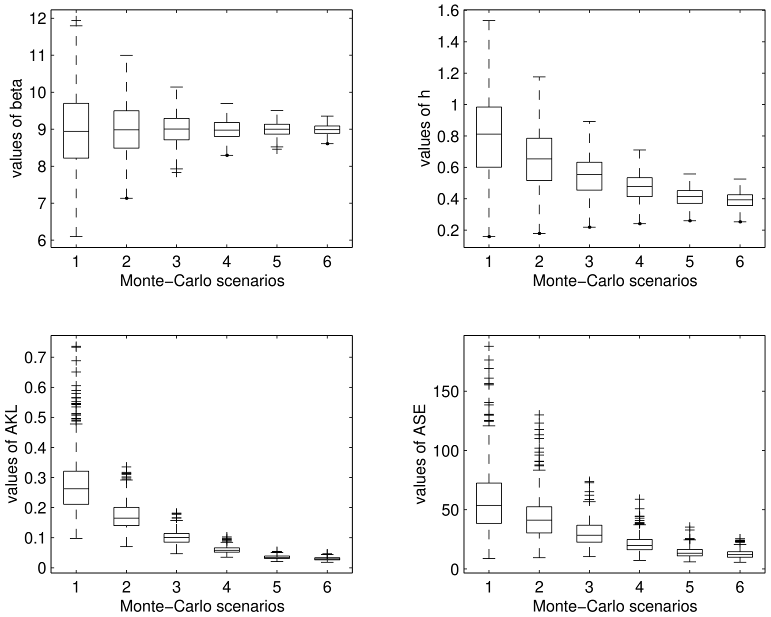

Figure 2: *One step procedure: box-plots of the values of* $\hat{\beta}_n, \hat{h}_n, AKL$ *and* ASE*, over the MC-replications. The sample sizes are* $n = 50, 100, 200, 400, 800$ *and* 1000 *respectively*

although for $n = 1000$, we are not far from the 5% level. This is important, since it shows, at least in our MC-scenario, that confidence intervals or testing using the asymptotic result could be misleading.

The same comparison is considered for the differences $(\hat{\beta}_n - \beta_0)$ standardized by the MC-estimation of the variance. Figure 4 shows that here, the $\mathcal{N}(0,1)$ approximation is better. The Kolmogorov-Smirnov test of normality confirms this impression. We test:

$$H_0 \;:\quad \sqrt{\mathrm{Var}^{-1}}(\hat{\beta}_n - \beta_0) \sim \mathcal{N}(0,1). \tag{3.11}$$

providing the p-values, 0.7555, 0.8844, 0.8546, 0.4718, 0.8620 and 0.0857, for n =50, 100, 200, 400, 800 and 1000 respectively. Here, the normality assumption of the standardized error term cannot be rejected even for $n = 50$, at the level 5%.

The message of this experiment is the following, the asymptotic result could not be used as such, even for large values of n. In addition, it must be pointed out that the computations of AVar, in Appendix A and B are usually untractable. So, in any case, the asymptotic result is rarely usefull in practice. But the Monte-Carlo experiment suggests an alternative, available for moderate to large values of n. Indeed, we can use a normal approximation for the sampling distribution of $\hat{\beta}_n$ for doing inference on β, but using a better estimator of its variance. In practice, the most natural analog of the Monte-Carlo estimator used above is the bootstrap estimator of the variance of $\hat{\beta}_n$. Such a bootstrap procedure will be described

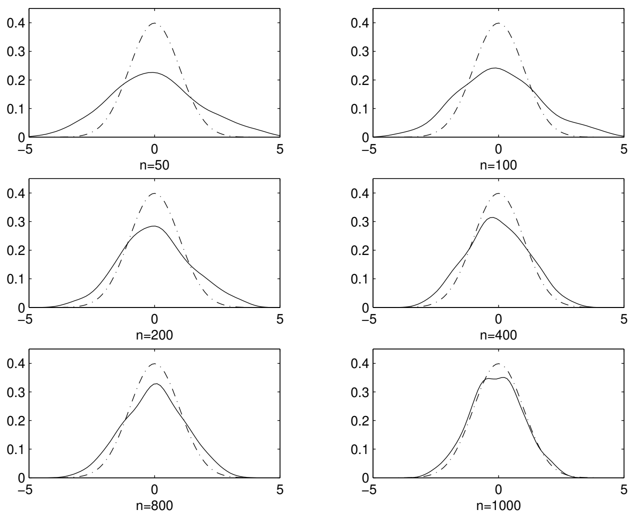

Figure 3: *Estimated densities (solid lines) of $\hat{\beta}_n - \beta_0$ standardized by the asymptotic AVar, in a one-step procedure, over the MC-replications, compared with the $\mathcal{N}(0,1)$ (dash-doted lines).*

in the next section.

4 A Bootstrap Algorithm

We will use the bootstrap for two reasons: first, we want to estimate the variance of our estimator $\hat{\beta}_n$ and secondly, we would like to numerically stabilize its value. The latter point can be viewed as a variant of the bagging ("bootstrapping and averaging") procedure proposed in Breiman (1996): we compute B bootstrap estimators $\{\hat{\beta}_n^{*b}, b = 1, \ldots, B\}$ by using the algorithm proposed below, then we eliminate the numerical outliers, by the same procedure as above for our Mone-Carlo experiment. The final corrected estimator of β_0 is $\hat{\beta}_{c,n}$, the mean of the remaining bootstrap values. The variance will be estimated as usual by $\mathrm{Var}(\hat{\beta}_n^*)$ the empirical variance of the remaining bootstrap values.

Then, due to the normality of the sampling distribution around the true value β_0, even for moderate n, an approximate $(1-\alpha) * 100\%$ confidence interval for β_0 is provided by

$$\beta_0 \in \left[\hat{\beta}_{c,n} \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\beta}_n^*)}\right], \tag{4.12}$$

where z_q is the q-percentile of a standard normal distribution.

The bootstrap algorithm could be done nonparametrically by drawing sample values with replacement from the pairs $\{(X_i, Y_i), i = 1, \ldots, n\}$, but we will gain in precision if we take into account the semiparametric structure of the Poisson regression model. The algorithm is described as follows:

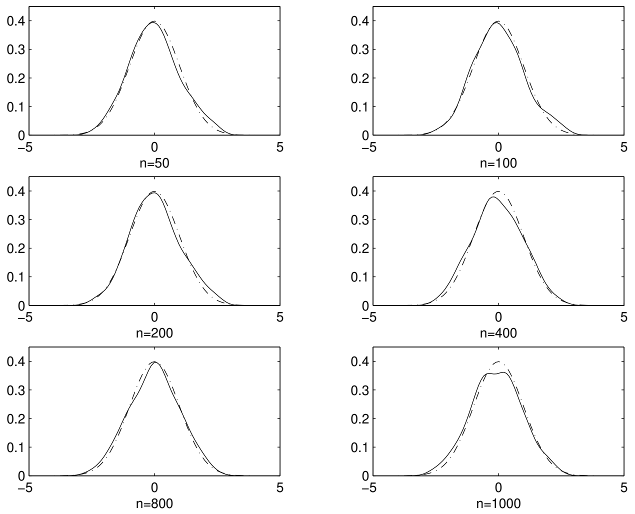

Figure 4: *Estimated densities (solid lines) of $\hat{\beta}_n - \beta_0$ standardized by the empirical variance* $\mathrm{Var}(\hat{\beta}_n)$*, in a one-step procedure, over the MC-replications, compared with the* $\mathcal{N}(0,1)$ *(dash-doted lines).*

[0] With the original sample $\{(X_i, Y_i), i = 1, \ldots, n\}$, compute the initial estimates by using the procedure of Section 2: an initial guess for (β_0, h) given by the two-steps procedure (2.2) and (2.4), then the one-step estimator by solving (2.1). This provides $\hat{\beta}_n$ and $\hat{h}_n$. Set the bootstrap counter $b = 1$.

[1] Generate a sample of vectors $\{X_i^{*b}, i = 1, \ldots, n\}$ by sampling with replacement in the original sample values $\{X_i, i = 1, \ldots, n\}$. Estimate the regression function at the obtained points:

$$\hat{g}_i^{*b} = \hat{g}_{\hat{\beta}_n, \hat{h}_n}(\hat{\beta}_n X_i^{*b}), \ i = 1, \ldots, n,$$

by using the Nadaraya-Watson formula (1.12).

[2] Generate the values of Y_i^{*b} as a Poisson r.v. with mean $\hat{g}_i^{*b}$, for $i = 1, \ldots, n$.

[3] Compute the bootstrap estimates $(\hat{\beta}_n^{*b}, \hat{h}_n^{*b})$ with the bootstrap sample $\{(X_i^{*b}, Y_i^{*b}), i = 1, \ldots, n\}$, by solving the one-step problem (2.1) with $(\hat{\beta}_n, \hat{h}_n)$, computed in step [0], as starting values for the optimization algorithm.

[4] Repeat the loop [1]–[3], for $b = 1 \ldots, B$. This provides the empirical bootstrap values $\{(\hat{\beta}_n^{*b}, \hat{h}_n^{*b}), b = 1 \ldots, B\}$.

After this bootstrap loop, we compute $\hat{\beta}_{c,n}$, the mean of the bootstrap values not considered as numerical outliers due to the outlying value of either $\hat{\beta}_n^{*b}$ or $\hat{h}_n^{*b}$. Then, $\hat{h}_{c,n}$, the value of h

corresponding to $\hat{\beta}_{c,n}$, is provided by cross validation, by solving the simple unidimensional optimization procedure (2.4) at the value $\hat{\beta}_{c,n}$.

The procedure is illustrated with one typical sample of our Monte-Carlo scenario, with $n = 100$ and $B = 500$ bootstrap replications. The estimation of β_0 provided at the initial step [0] is $\hat{\beta}_n = 9.2643$ and $\hat{h}_n = 0.7732$. The bootstrap correction (10% of bootstrap samples considered as outliers) is $\hat{\beta}_{c,n} = 9.2288$ and $\hat{h}_{c,n} = 0.7714$. In this case, the corrected values are similar to the original ones, but the corrected values are much more numerically stable. The bootstrap provides also an estimator of the variance. We obtain here $\mathrm{Var}(\hat{\beta}_n^*) = 0.8063$. The corresponding 95% confidence interval for β_0 is $[7.4689, 10.9887]$.

The estimate of the regression function with the same sample, is displayed at Figure 5: the top plot is not available in practice, with real data, since both β_0 and g are unknown. The estimate in the bottom plot behaves reasonably well, as expected from our Monte-Carlo experiment in the preceding section. Notice the bad behavior of our estimate in the tails: this is mainly due to our chosen scenario for generating the values of X_i, there are very few observations near the borders.

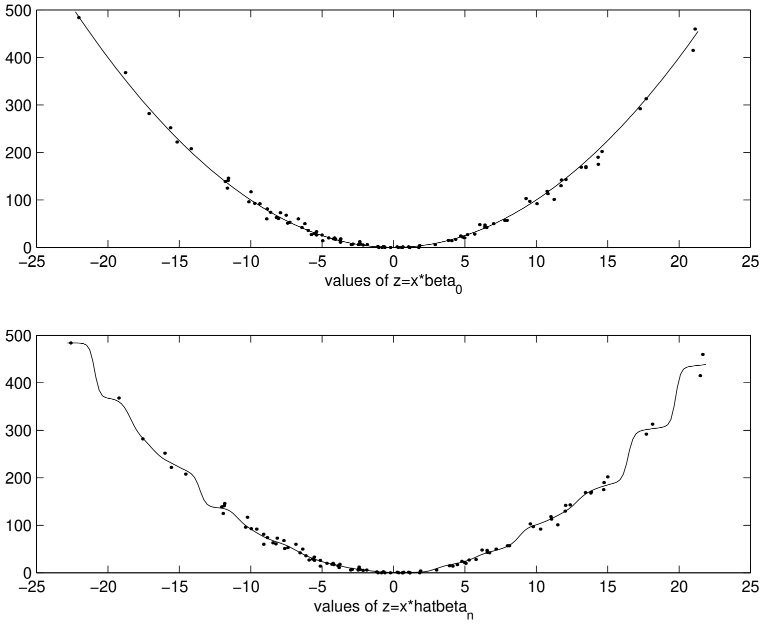

Figure 5: *Estimation of the regression function for one typical sample of size* $n = 100$ *in our Monte-Carlo scenario. Top plot: the true function* z^2*, the dots are the pairs* $(\beta_0 X_i, Y_i)$ *with* $\beta_0 = 9$*. Bottom plot: the estimate* $\hat{g}_{\hat{\beta}_{c,n},\hat{h}_{c,n}}$ *with* $\hat{\beta}_{c,n} = 9.2288, \hat{h}_{c,n} = 0.7714$*, the dots are the pairs* $(\hat{\beta}_{c,n} X_i, Y_i)$*.*

Of course, we cannot draw final conclusions after this illustration with one particular sample, even if we tried many other generated samples obtaining essentially the same results.

The bootstrap algorithm for $n = 100$, with $B = 500$, took 2h40min on a Pentium III, 450 Mghz machine, using Matlab.

5 An Illustration with Real Data

We applied the method described in Section 4 to a real dataset[1] containing the number of suicides in $n = 121$ Austrian municipalities (Carinthia region) from 1980 to 1995, which represents the response variable, denoted by Y. The dataset also contains demographic and geographic information about every municipality for the year 1991, which is a census year. The idea is to explain the variation of the number of suicides by some socio-demographic explanatory factors. In this study we considered three explicative variables: the difference between the migration to and from the municipality (X_1), the mean altitude measured from the sea level (X_2) and the density of the population (X_3). Two municipalities (Klagenfurt and Villach) with outlying values for the population density and the migration variables were excluded from the dataset. The plot of the 119 remaining observations from the Figure 6 gives a quick representation of the two-dimensional associations between the chosen variables. Increasing the density seems to increase the number of suicides. As expected, the municipalities with high population density are mostly located at lower altitude and correspond to cities with positive migration gradient. The apparent negative effect of the altitude X_2 on Y may be due, in part, to an indirect effect of the density X_3.

Assuming that these data were generated by a Poisson SIM, defined in Section 1, the statistical problem is to estimate the link function and the parameters of the linear index $\beta_0 X = \beta_{01}X_1 + \beta_{02}X_2 + \beta_{03}X_3$. As explained above, for identifiability reasons, we choose to fix $\beta_{01} = 1$, so that only the last two components of X have to be estimated. In order to facilitate the relative comparison of the effects of the explanatory variables, the X-values have been standardized to have mean 0 and variance 1. The magnitude of β_{0j}, $j = 2, 3$ measures the change in X_j, in standardized units, required to match the effect of a standardized unit change in X_1. Let $\hat{\beta}_n$ denote the estimator of the vector $(\beta_{02} \;\; \beta_{03})$.

We first evaluated the index parameter using the parametric GLM model with exponential link function (this a parametric model often used in this situation):

$$Y|X = x \sim Po(\exp(\beta_0 X)),$$

where $\beta_0 X = \beta_{00} + \beta_{01}X_1 + \beta_{02}X_2 + \beta_{03}X_3$. The GLM estimator of β_0 is $\hat{\beta}_{GLM} = (2.5728 \; -0.2751 \; -0.2971 \; 0.3620)$ with corresponding estimated variance $(0.0007 \; 0.0004 \; 0.0009 \; 0.0003)$.

[1]The data set was kindly provided by Martin Weichbold, Institut für interdisziplinäre Tourismusforschung - Universität Salzburg. A more substantial study on the evolution of the number of suicides in Austria can be found in Ziegler, Bachleitner and Arminger (1995) or Haller and Lingg (1985).

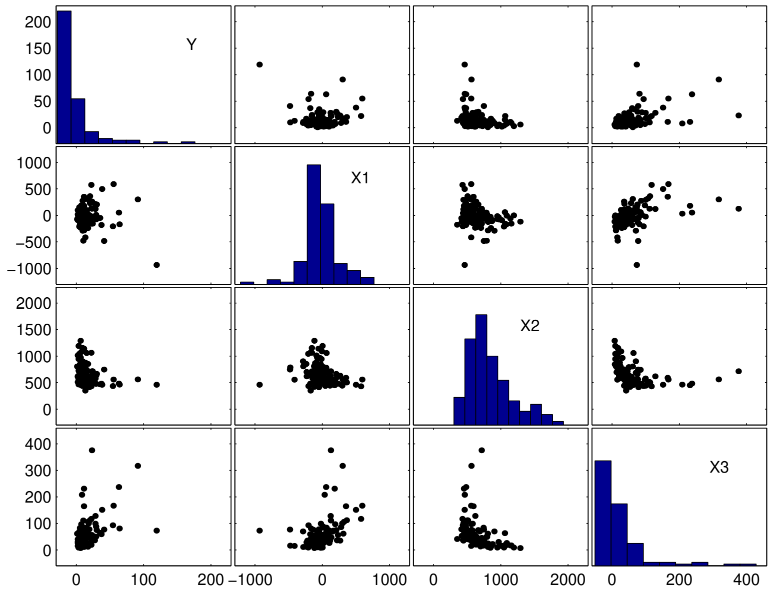

Figure 6: *Draftman plot for the 4 variables considered.*

The estimator of the asymptotic variance of $\hat{\beta}_{GLM}$ was computed as the inverse of the hessian matrix: $\left[\frac{-\partial^2 \ell_n(Y;\hat{\beta}_{GLM})}{\partial\beta\partial\beta'}\right]^{-1}$, where ℓ_n is the sample log-likelihood under GLM. The corresponding 95% confidence intervals are $[2.5207,\ \ 2.6249]$, $[-0.3132,\ \ -0.2369]$, $[-0.3553,\ \ -0.2390]$ and $[0.3298,\ \ 0.3942]$ for β_{00}, β_{01}, β_{02} and respectively β_{03}. Here, we can write the GLM index estimator as follows:

$$2.5728 - 0.2751(X_1 + 1.080X_2 - 1.316X_3),$$

such that the expression between the parenthesis can be directly compared to the SIM index estimator obtained below.

For the semiparametric approach, we obtain the following results. At the step [0] of the algorithm described in Section 4, we used the parametric GLM estimate $(1.080\ \ -1.316)$ as starting value. The two-steps method provides the estimators $\hat{\beta}_n = (0.3078\ \ -1.5598)$ and $\hat{h}_n = 0.3061$. Then, the one-step method provides the initial estimates $\hat{\beta}_n = (0.6862\ \ -1.5648)$ and $\hat{h}_n = 0.0657$. Using these last values for the semiparametric bootstrap procedure described at the steps [1]-[3], with $B = 500$, we obtain the corrected final values $\hat{\beta}_{c,n} = (0.6853\ \ -1.5614)$ and $\hat{h}_{c,n} = 0.0847$. Here, 21.8% of the bootstrap samples were considered as outliers. This number of outliers is higher than what we obtained in our Monte-Carlo sample from Section 4, in part because of the higher dimension of the parameter vector. The bootstrap estimator of the variance is $\text{Var}(\hat{\beta}^*_{n2}) = 0.0006$, $\text{Var}(\hat{\beta}^*_{n3}) = 0.0004$. The corresponding 95% confidence intervals are $[0.6360, 0.7347]$ and $[-1.6027, -1.5201]$ for β_{02}

and respectively for β_{03}.

In the Monte-Carlo setup of Section 4, the semiparametric bootstrap is well suited because we know the real data generating process and the real value of β_0. With real data, a nonparametric bootstrap is certainly more robust concerning the Poisson hypothesis and less dependent on the one-step estimated value of β coming from the original sample. In this case, at step [1] we generate a sample of vectors $(X_i^{*b}, Y_i^{*b}), i = 1, \ldots, n$ by sampling with replacement in the original sample values $(X_i, Y_i), i = 1, \ldots, n$. Here, the one-step initial parameter estimate serves only as a starting value for the optimization process in the bootstrap loop. The nonparametric bootstrap procedure, with $B = 500$, gives corrected final values comparable to those obtained using the semiparametric bootstrap: $\hat{\beta}_{c,n} = (0.6915 \ -1.5542)$ and $\hat{h}_{c,n} = 0.1203$. Here, only 10% of the bootstrap values were considered as outliers. The bootstrap estimator of the variance is $\mathrm{Var}(\hat{\beta}_{n2}^*) = 0.0022$, $\mathrm{Var}(\hat{\beta}_{n3}^*) = 0.0007$. The corresponding 95% confidence intervals are $[0.6006, 0.7824]$ and $[-1.6058, -1.5026]$ for β_{02} and respectively for β_{03}.

As far as the estimation of β is concerned, we can conclude that the order of magnitude of $\hat{\beta}$ is the same for the parametric and the semiparametric approach. In the light of the confidence intervals obtained by the GLM approach, the migration X_1 and the altitude X_2 have a negative significant effect (at 5%) on the number of suicides Y, whereas the density X_3 has a positive significant effect on Y. The variable X_3 has the most important effect on Y among all. These results agree with what we have already observed from the draftman plot of Figure 6. They also agree with those given by the SIM approach, with the only difference that for SIM, the estimated coefficients must be interpreted, as we already pointed out above, relatively to the effect of X_1. Thus, the SIM index estimator says that X_2 has a contribution 0.6853 times larger than the contribution of X_1 on Y, in the same sense as X_1. The effect of X_3 on Y is 1.5614 times larger than the effect of X_1, in the opposite sense as X_1.

The estimate of the regression function is displayed in the Figure 7: the top and middle plots show the Nadaraya-Watson estimator $\hat{g}_{\hat{\beta}_{c,n}, \hat{h}_{c,n}}$ based on the SIM index estimator using the semiparametric and respectively the nonparametric bootstrap procedures. The bottom plot gives the estimated parametric exponential link $\exp(\hat{\beta}_{GLM} X)$, based on the GLM index estimator. Note the fluctuating Nadaraya-Watson estimator, especially at the left extremity, where only a few points are available. These points correspond to the municipalities with extreme high values of density X_3, which can also be observed in the right column of Figure 6.

Climov, Hart and Simar (1999) have pointed out that the optimal value of h minimizing our criterion (2.1) might undersmooth the estimate of the regression function: this is certainly the case here. A more smooth nonparametric estimator could be obtained by minimizing, at the final step of the algorithm, a so-called "Double Smoothing" criterion, which prevents undersmoothing in small or moderate samples (see Climov, Hart and Simar, 1999, for details). We used it at the final step and obtained a higher value for the final bandwidth: $\hat{h}_{c,n} = 0.3475$. The Nadaraya-Watson estimator, using this last value of h, is represented in the Figure 8 in the top plot, together with the exponential link of the GLM, in the bottom plot. The semiparametric estimator and the parametric estimators have a comparable shape in the right region where enough data points are available. The difference between the two link function estimators at the left extremity is due to the sparseness of the data and to the presence of three influential points (which represent cities having low values of Y for the corresponding high values of density X_3). For this particular dataset, a method using a local smoothing parameter would be indicated, in order to account for the regions with different number of points. For example, the method of Härdle and Marron (1995) could be used with two blocks (one for the left region with very few data and a second block for the right region) for computing the bandwidth.

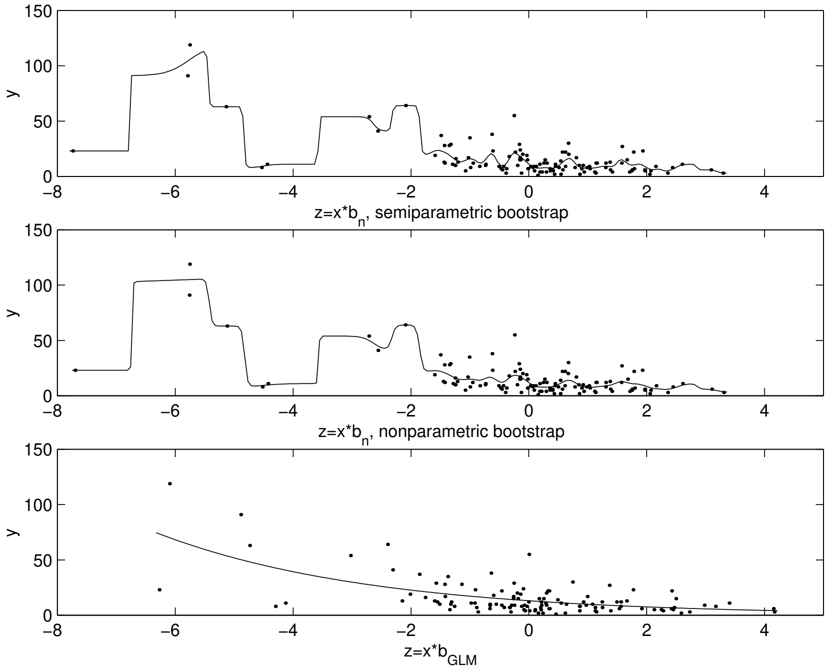

Figure 7: *Regression function estimate: the upper and middle plots present the estimate based on the SIM using the semiparametric and respectively nonparametric bootstrap approach and the lower plot shows the estimate based on GLM.*

To see how well a regression estimate fares on a dataset, we used two performance measures: the Poisson deviance and the Pearson X^2 statistic. The Poisson deviance is defined as twice the difference between the maximum likelihood achievable (in the full model with as many parameters as observations) and that achieved by the model under investigation

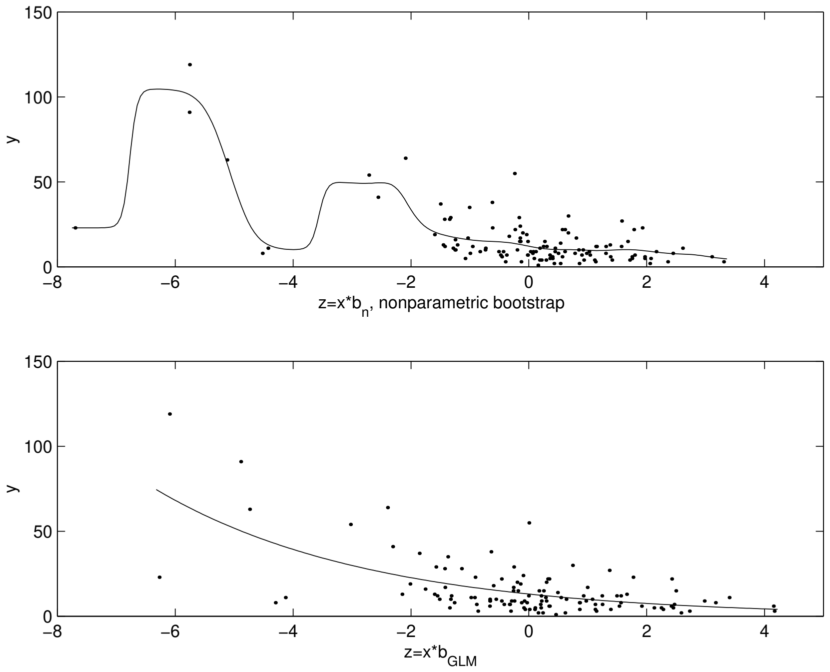

Figure 8: *Regression function estimate with the final h minimizing the Double Smoothing criterion: the upper plot presents the estimate based on the SIM using the nonparametric bootstrap and the lower plot shows the estimate based on GLM.*

(see McCullagh and Nelder, 1989). Here we will only consider the second part (as the full likelihood is the same for both SIM and GLM models), *i.e.*

$$D_P(\mu, \beta) = \frac{1}{n} \sum [\mu(\beta X_i) - Y_i \log \mu(\beta X_i)]. \tag{5.13}$$

The other measure of discrepancy is the generalized Pearson X^2 statistic, which takes the form:

$$D_C(\mu, \beta) = \frac{1}{n} \sum \frac{[Y_i - \mu(\beta X_i)]^2}{V(\mu(\beta X_i))}, \tag{5.14}$$

where $V(\mu(\beta X_i))$ is the estimated variance function for the distribution concerned. In our case, for a Poisson distribution with $V(\mu) = \mu$, we obtain the original Pearson X^2 statistic. These two deviances are evaluated for (μ, β) equal to the SIM and GLM estimates $(\hat{g}_{\hat{\beta}_{c,n},\hat{h}_{c,n}}, \hat{\beta}_{c,n})$ and respectively $(\exp(\hat{\beta}_{GLM} X), \hat{\beta}_{GLM})$ obtained above.

For the SIM, using the semiparametric bootstrap approach, the Poisson deviance is -30.5818 and the Pearson X^2 deviance is 2.8898. For the nonparametric bootstrap approach under SIM, the Poisson deviance is -30.3168 and the X^2 deviance is 3.5498. For GLM we obtained -28.5501 and 7.4653 for the Poisson and respectively the Pearson X^2 deviances.

In the light of these deviance values and of the Figures 7 and 8, these results can be interpreted as follows. The deviance values measure how well a regression estimate fits a dataset but they give no information concerning the smoothness of the estimate. In estimating the link function, a crucial step is the choice of the smoothing parameter, which involves the usual compromise between a model's smoothness and how closely it fits the data. So, in order to compare different link function estimators, two aspects have to be taken into account: the smoothness and the fit to the data. The nonparametric regression estimator based on criterion (2.1) gives the "best" fit (the smallest values for both the Poisson and the Pearson X^2 deviances), but it is very unstable, it tends to interpolate the data. The nonparametric regression estimator with the final h given by the Double Smoothing criterion presents an acceptable degree of smoothness. For this application, in the view of Figure 8 and of the deviance values, it appears that the exponential link function describes well the data, especially for the municipalities with small or moderate values of density (the right part of the plots). The negative values of the index estimator z correspond to large cities, with high values of the density variable X_3 and this part is less well explained by the exponential model. Here, the nonparametric regression estimator fits the observed data better than its parametric counterpart, which we expected given the flexibility of the semiparametric modelisation.

6 Conclusions

We investigate by a simulation experiment, the finite sample properties of $\hat{\beta}_n$, the maximum likelihood index estimator in single index Poisson regression model. We also propose a comparison of a two-steps and a one-step method. The one-step method provides better estimates but the two-steps method may be useful for providing initial values of the parameters in the numerically intricate optimization procedure. The asymptotic normality is achieved even with moderate sample sizes.

The numerical instability of the method, inherent to single-index semiparametric estimation, is corrected by using a variant of the bagging method. This gives a numerically robust estimator of the index vector. The bootstrap loop is also used for providing a reasonable estimator of the variance of our index estimator. Inference is then available, even with moderate sample sizes, using the normal approximation. The procedure is illustrated with a real data example.

Appendix

A Derivation of the $g_\beta(\beta x)$ functions in the Monte-Carlo scenario

The Monte-Carlo set-up was defined in Section 3.1. In this appendix, we derive the analytic form for the functions $g_\beta(x_1 + \beta x_2) = E(Y|X_1 + \beta X_2 = x_1 + \beta x_2)$.

The joint density of the vector (X_1, X_2, Y) evaluated at (x_1, x_2, y) is

$$f_{XY}(x_1, x_2, y) = \frac{\exp(-g_{\beta_0}(x_1 + \beta_0 x_2))g_{\beta_0}(x_1 + \beta_0 x_2)^y}{y!} f_X(x_1, x_2) \tag{A.1}$$

Consider the following linear transformation: $(X_1, Z, Y) = (X_1, X_1 + \beta X_2, Y)$. The joint density of the transformed vector evaluated at (x_1, z, y) is

$$\frac{1}{\beta} f_{XY}(x_1, \frac{z - x_1}{\beta}, y).$$

The conditional density of Y given $Z = z$ is then:

$$\begin{aligned} & P(Y = y|X_1 + \beta X_2 = z) \qquad\qquad (A.2) \\ = & \frac{\int \frac{\exp(-g_{\beta_0}(x_1+\beta_0\frac{z-x_1}{\beta}))\left[g_{\beta_0}(x_1+\beta_0\frac{z-x_1}{\beta})\right]^y}{y!} f_X(x_1, \frac{z-x_1}{\beta})\, dx_1}{\sum_{y\in N} \int \frac{\exp(-g_{\beta_0}(x_1+\beta_0\frac{z-x_1}{\beta}))\left[g_{\beta_0}(x_1+\beta_0\frac{z-x_1}{\beta})\right]^y}{y!} f_X(x_1, \frac{z-x_1}{\beta})\, dx_1} \end{aligned}$$

By interchanging the order of integration and summation, as all the functions involved are positive and integrable, we have:

$$\begin{aligned} & P(Y = y|X_1 + \beta X_2 = z) \qquad\qquad (A.3) \\ = & \int \frac{\exp(-g_{\beta_0}(x_1 + \beta_0\frac{z-x_1}{\beta}))g_{\beta_0}(x_1 + \beta_0\frac{z-x_1}{\beta})^y}{y!} r(x_1, z, \beta) dx_1 \end{aligned}$$

where

$$r(x_1, z, \beta) = \frac{f_X(x_1, \frac{z-x_1}{\beta})}{\int f_X(x_1, \frac{z-x_1}{\beta}) dx_1}.$$

Now, the link function g_β has the following expression:

$$\begin{aligned} g_\beta(z) &= E[Y|X_1 + \beta X_2 = z] \\ &= \sum_{y\in N} y P(Y = y|X_1 + \beta X_2 = z) \\ &= \sum_{y\in N} y \int \frac{\exp(-g_{\beta_0}(x_1 + \beta_0\frac{z-x_1}{\beta}))\left[g_{\beta_0}(x_1 + \beta_0\frac{z-x_1}{\beta})\right]^y}{y!} r(x_1, z, \beta) dx_1. \end{aligned}$$

By interchanging again the order of integration and summation we obtain:

$$g_\beta(z) = \int g_{\beta_0}\left(x_1 + \beta_0 \frac{z - x_1}{\beta}\right) r(x_1, z, \beta) dx_1. \tag{A.4}$$

When X_1 and X_2 are two independent $\mathcal{N}(0,1)$ variables , $r(x_1, z, \beta)$ has the following expression:

$$r(x_1, z, \beta) = \frac{\exp\left[-\frac{1}{2}\left(x_1^2 + (\frac{z-x_1}{\beta})^2\right)\right]}{\int \exp\left[-\frac{1}{2}\left(x_1^2 + (\frac{z-x_1}{\beta})^2\right)\right] dx_1},$$

which, after some calculations may be written as

$$r(x_1, z, \beta) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\frac{\beta^2}{\beta^2+1}}} \exp\left[-\frac{1}{2\frac{\beta^2}{\beta^2+1}}\left(x_1 - \frac{z}{\beta^2+1}\right)^2\right].$$

The ratio $r(x_1, z, \beta)$ is thus the density function of a normal variable, say U, evaluated at x_1, where

$$U \sim \mathcal{N}\left(\frac{z}{\beta^2+1}, \frac{\beta^2}{\beta^2+1}\right).$$

Substituting into (A.4), we have

$$g_\beta(z) = \mathrm{E}\left[g_{\beta_0}\left(U + \beta_0 \frac{z - U}{\beta}\right)\right]. \tag{A.5}$$

Now, for the "true" link function $g_{\beta_0}(z) = z^2$ we may deduce the form of the g_β functions for any value of β:

$$\begin{aligned} g_\beta(z) &= \mathrm{E}[Y|X_1 + \beta X_2 = z] \\ &= \mathrm{E}\left[(U + \frac{\beta_0}{\beta}(z - U))^2\right] \\ &= \frac{(\beta - \beta_0)^2}{\beta^2+1} + \left[z \frac{\beta\beta_0 + 1}{\beta^2+1}\right]^2 \end{aligned} \tag{A.6}$$

Note that if $\beta = \beta_0$ we obtain $g_\beta(z) = z^2$, hence we recover the "true" link function.

B Derivation of $\mathrm{Var}(\hat{\beta}_n)$ in the Monte-Carlo scenario

¿From equation (1.14), we know that, in order to derive the value of Σ we have to compute (see A.6):

$$\frac{\partial}{\partial\beta} g_\beta(\beta x) = \frac{\partial}{\partial\beta}\left\{\frac{(\beta - \beta_0)^2}{\beta^2+1} + \left[(X_1 + \beta X_2)\frac{\beta\beta_0 + 1}{\beta^2+1}\right]^2\right\},$$

and determine the resulting value for $\beta = \beta_0$. It is clear that

$$\frac{\partial}{\partial \beta}\left\{\frac{(\beta-\beta_0)^2}{\beta^2+1}\right\}$$

vanishes for $\beta = \beta_0$, and thus we have:

$$\begin{aligned}
\frac{\partial}{\partial \beta}\, g_\beta(\beta x)\Big|_{\beta=\beta_0} &= \left\{2(X_1+\beta X_2)\frac{\beta\beta_0+1}{\beta^2+1}\right\}\Big|_{\beta=\beta_0} \cdot \frac{\partial}{\partial \beta}\left\{(X_1+\beta X_2)\frac{\beta\beta_0+1}{\beta^2+1}\right\}\Big|_{\beta=\beta_0} \\
&= 2(X_1+\beta_0 X_2)\left\{X_2+(X_1+\beta X_2)\frac{-\beta_0}{\beta_0^2+1}\right\} \\
&= 2(X_1+\beta_0 X_2)\left\{\frac{X_2}{\beta_0^2+1}-\beta_0\frac{X_1}{\beta_0^2+1}\right\}.
\end{aligned}$$

It follows then from (1.14) that:

$$\begin{aligned}
\Sigma^{-1} &= \mathrm{E}\left\{\frac{1}{(X_1+\beta_0 X_2)^2}\cdot 4(X_1+\beta_0 X_2)^2\cdot\left(\frac{X_2}{\beta_0^2+1}-\beta_0\frac{X_1}{\beta_0^2+1}\right)^2\right\} \\
&= 4\mathrm{E}\left\{\left(\frac{X_2}{\beta_0^2+1}-\beta_0\frac{X_1}{\beta_0^2+1}\right)^2\right\}
\end{aligned}$$

In our simulation setting we considered X_1 and X_2 are two independent normal variables, thus the distribution of the linear combination above is

$$\frac{1}{1+\beta_0^2}X_2-\frac{\beta_0}{1+\beta_0^2}X_1 \sim \mathcal{N}\left(0, \frac{1}{1+\beta_0^2}\right),$$

so that,

$$\Sigma^{-1} = \frac{4}{1+\beta_0^2}.$$

We thus obtain that for the SIM with Poisson conditional distribution, the asymptotic variance of the PML estimator $\hat{\beta}_n$ has the following expression:

$$\sqrt{n}(\hat{\beta}_n-\beta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \frac{1+\beta_0^2}{4}). \tag{B.1}$$

References

[1] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 26, 123–140.

[2] Climov,D., Hart, J. and L. Simar (2000). Automatic smoothing and estimation in single index Poisson regression. Discussion Paper no. 0014, Institut de Statistique, Université Catholique de Louvain.

[3] Delecroix, M. and M. Hristache (1999). M-estimateurs semi-paramétriques dans les modèles à direction rèvélatrice unique. *Bull. Belg. Math. Soc.* **6**, 161-185.

[4] Delecroix, M., Hristache, M. and V. Patilea (1999). Optimal smoothing in semiparametric index approximation of regression functions. Cahiers du CREST no. 9952, Paris.

[5] Haller, R. and A. Lingg (1985). Voralberger Suizidstudie, unpublished manuscript.

[6] Härdle, W., Hall, P. and H. Ichimura (1993). Optimal smoothing in single-index models. *Ann. Statist.* **21**, 157-178.

[7] Härdle, W. and J.S. Marron (1995) Fast and Simple scatterplot smoothingv. *Comp. Statist. Data Anal.***20**, 1-17.

[8] Härdle, W. and T.M. Stoker (1989). Investigating smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Assoc.* **84**, 986-995.

[9] Horowitz, J. L. and W. Härdle (1996). Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates *J. Amer. Statist. Assoc.* **91**, 1632-1640.

[10] Ichimura, H. (1993). Semiparametric leastsquares (SLS) and weighted SLS estimation of single-index models. *Journal of Econometrics*, **58**, 71-120.

[11] McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models.* London: Chapman and Hall.

[12] Newey, W.K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics.* **5**, 99-135.

[13] Powell, J.L., Stock, J.H. and T.M. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrics*, **57**, 1403-1430.

[14] Sherman, R.P. (1994). U-processes in the analysis of a generalized semiparametric regression estimator. *Econometric theory*, **10**, 372-395.

[15] Silverman, B.W. (1986), *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

[16] Ziegler, A., Bachleitner, R. and G. Arminger (1995). Pseudo-Maximum-Likelihood-Schtzung und Regressionsdiagnostik fr Zhldaten: Suizid durch Tourismus. *Allg. Statistisches Archiv*, 79, 170–195.