Goodness-of-fit methods for probabilistic index models

Jan De Neve[1,*], Olivier Thas[1,2] and Jean-Pierre Ottoy[1]

[1] Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure Links 653, B-9000 Gent, Belgium

[2] Centre for Statistical and Survey Methodology - School of Mathematics and Applied Statistics, University of Wollongong, NSW 2522, Australia.

[*] corresponding author; e-mail: JanR.DeNeve@UGent.be

Running head. Goodness-of-fit of probabilistic index models

Keywords. goodness-of-fit; lack-of-fit; linear smoothers; probabilistic index models.

Abstract. A class of semiparametric regression models, called probabilistic index models, has been recently proposed. Because these models are semiparametric, inference is only valid when the proposed model is consistent with the underlying data-generating model. However, no formal goodness-of-fit methods for these probabilistic index models exist yet. We propose a test and a graphical tool for assessing the model adequacy. Simulation results indicate that both methods succeed in detecting lack-of-fit. The methods are also illustrated on a case study.

1 Introduction

Recently, Thas et al. (2012) proposed a class of semiparametric regression models, called *probabilistic index models* (PIM). These models focus on the probabilistic index, which is defined as the probability $\mathrm{P}(Y \preccurlyeq Y') \equiv \mathrm{P}(Y < Y') + 0.5\mathrm{P}(Y = Y')$, with Y and Y' two independent random

variables whose distribution may depend on a (fixed or random) covariate vector, say $\boldsymbol{X}$ and $\boldsymbol{X}'$ respectively. In particular, let $(Y, \boldsymbol{X})$ and $(Y', \boldsymbol{X}')$ be independent and identically distributed random variables with density $f_{Y\boldsymbol{X}}$. Then a PIM is defined as

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}'\right) = m(\boldsymbol{X}, \boldsymbol{X}'; \boldsymbol{\beta}) = g^{-1}(\boldsymbol{Z}^T\boldsymbol{\beta}), \quad (\boldsymbol{X}, \boldsymbol{X}') \in \mathcal{X}. \tag{1}$$

Here $\mathcal{X}$ denotes the set of all possible pairs $(\boldsymbol{X}, \boldsymbol{X}')$ for which the model is defined, g is a link function, and $\boldsymbol{Z}$ is a p-vector depending on $(\boldsymbol{X}, \boldsymbol{X}')$; for continuous predictors, often $\boldsymbol{Z} = \boldsymbol{X}' - \boldsymbol{X}$. Because PIMs are semiparametric, inference is only valid when model (1) is consistent with the data-generating model. In this paper we propose a goodness-of-fit (GOF) test, and a graphical diagnostic tool which compares the model predictions with a nonparametric estimator of the probabilistic index.

To illustrate our setting we consider the Childhood Respiratory Disease Study (CRDS), which is also analysed in Thas et al. (2012, Section 6.1). The response variable is the forced expiratory volume (FEV in litres). The age (AGE in years) and smoking indicator (SMOKE = 1 if the child smokes, SMOKE = 0 if the child does not smoke) are recorded for 654 children of ages 3–19 years. When analysing the effect of smoking on the lung capacity, age may be a confounder, and therefore should be taken into account. A part of the data are illustrated in Figure 1, which shows nonparametric density estimates of the FEV distributions for several combinations of smoking status and age. We fit a linear PIM with logit link, i.e.

$$\begin{aligned} &\mathrm{logit}\left[\mathrm{P}\left\{\mathrm{FEV} \preccurlyeq \mathrm{FEV}' \mid (\mathrm{SMOKE}, \mathrm{AGE}), (\mathrm{SMOKE}', \mathrm{AGE}')\right\}\right] \\ &= \beta_1(\mathrm{SMOKE}' - \mathrm{SMOKE}) + \beta_2(\mathrm{AGE}' - \mathrm{AGE}). \end{aligned} \tag{2}$$

It holds that $\hat{\beta}_1 = -0.46$ (standard error: 0.25 and $p = 0.064$) and $\hat{\beta}_2 = 0.56$ (standard error: 0.028 and $p < 0.0001$); we refer to Appendix A for a summary of the estimation theory. The estimated probability that FEV is larger for a smoking child as compared to a non-smoker of the same age is $\hat{\mathrm{P}}\{\mathrm{FEV} \preccurlyeq \mathrm{FEV}' \mid (0, \mathrm{AGE}), (1, \mathrm{AGE})\} = \mathrm{expit}(-0.46) = 39\%$. It is unlikely that a smoker has a better pulmonary function than a non-smoker of the same age. The effect is not significant at the 5% level of significance, which is surprising, as it is expected that smoking affects a child's lungs. So perhaps the data contain no evidence for this hypothesis or the study is underpowered. However, the lack of significance may also arise when the model does not fit

the data properly. Before drawing conclusions about the effect of smoking on the lung function, it is therefore important to first assess the GOF of model (2).

In Section 2 the GOF methods are developed. Section 3 assesses the performance of the GOF test in a simulation study. In Section 4 the CRDS example is revisited while Section 5 contains the discussion.

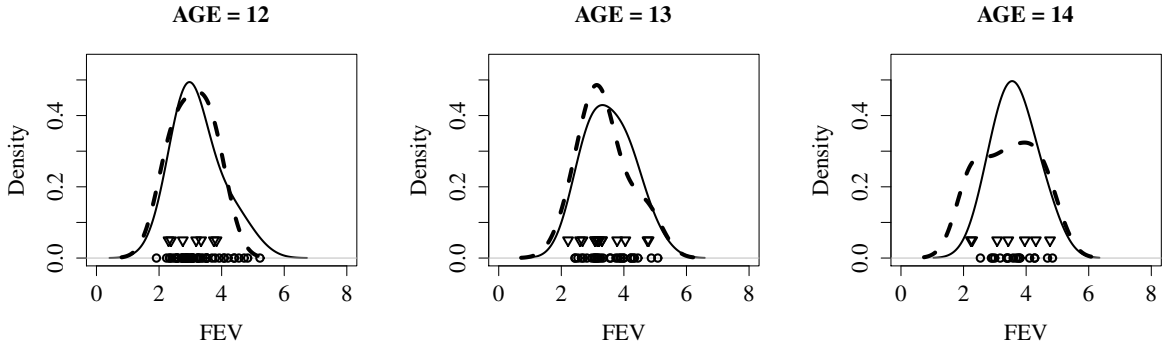

Figure 1: Kernel density estimates of the FEV distributions and individual sample observations for smokers (∇ - - -) and non-smokers ($\circ$ —).

2 Goodness-of-fit methods

2.1 Rationale

We start by considering a single continuous predictor; the extension to multiple predictors is addressed at the end of the section. Let $m_0(X, X')$ be the PIM which is consistent with the data-generating model, to be denoted as the true model, and let $m(X, X'; \boldsymbol{\beta})$ be the PIM that will be fitted to the data, referred to as the working model. The GOF null hypothesis is

$$H_0 : m_0(X, X') = m(X, X'; \boldsymbol{\beta}), \quad (X, X') \in \mathcal{X}, \tag{3}$$

for some $\boldsymbol{\beta} \in \mathbb{R}^p$. We consider a specific setting where the quadratic probit PIM is the true model and the linear probit PIM is the working model

$$m_0(X, X') = \Phi\left\{\beta_1(X' - X) + \beta_2\left(X'^2 - X^2\right)\right\}, \quad m(X, X'; \beta) = \Phi\left\{\beta(X' - X)\right\},$$

with Φ the standard normal distribution function. Consider the following settings: $\beta_1 = 0.3$, β_2 takes the values 0, -0.05 and -0.20 and the predictor X takes n equidistant values in $[-5, 5]$. When $\beta_2 = 0$ there is no quadratic effect and the null hypothesis (3) holds, while when $\beta_2 = -0.05$ ($\beta_2 = -0.20$) there is a weak (strong) quadratic effect and the null hypothesis does not hold.

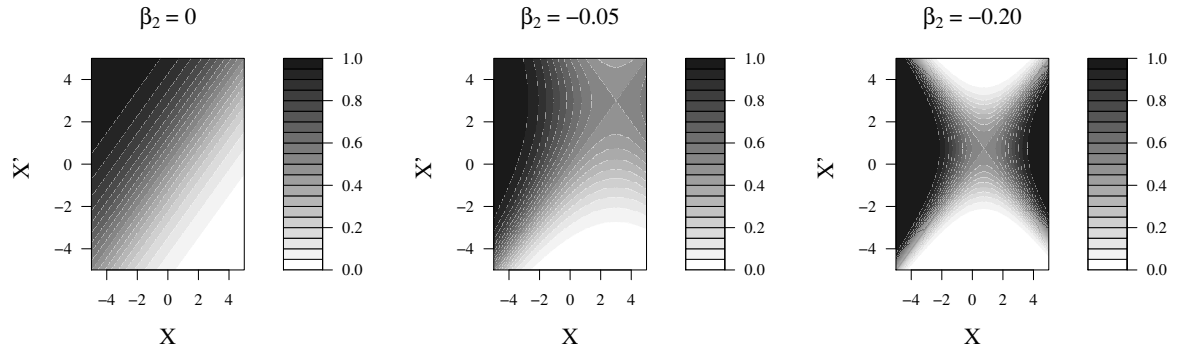

Figure 2: Quadratic probit PIM $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi\{\beta_1 \left(X' - X\right) + \beta_2 \left(X'^2 - X^2\right)\}$, with $\beta_1 = 0.3$ as a function of X. A grey coding is used to indicate the value of $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right)$.

Since a PIM depends on (X, X'), a 3-dimensional plot is needed for visualization; see Figure 2. Although this plot provides all information, it is difficult to interpret, so that we restrict (X, X') to a number of values which are relevant for the interpretation. Let Δ be a fixed value, then we restrict the plot to $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + \Delta\right)$, i.e. the probability that the response increases when the predictor is increased by Δ units. For the example setting, we can write

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + \Delta\right) = \Phi(\tilde{\beta}_1 + \tilde{\beta}_2 X), \quad \tilde{\beta}_1 = \beta_1 \Delta + \beta_2 \Delta^2, \quad \tilde{\beta}_2 = 2\beta_2 \Delta. \tag{4}$$

Equation (4) indicates that the choice of Δ is important. As Δ increases, the difference between $m_0(X, X' = X + \Delta)$ and $m(X, X' = X + \Delta; \beta)$ becomes more pronounced; see Figure 3. Consider the left panel where $\Delta = 1$. When the linear PIM holds, i.e. $\beta_2 = 0$, $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right)$ is fixed at $\Phi(\tilde{\beta}_1) = \Phi(0.3) \approx 62\%$ and independent of X. However, with increasing magnitude of β_2, this probability depends more strongly on the predictor X. When $\beta_2 = -0.20$, for example, it holds that $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right) > 95\%$ for $X < -4$, while for $X > 4$ this becomes $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right) < 7\%$. The restricted probability pro-

vides information on the difference between a quadratic and linear PIM, while retaining a simple interpretation.

If m_0 and $\boldsymbol{\beta}$ are known the plot suggests that comparing $m_0(X, X' = X + \Delta)$ with $m(X, X' = X + \Delta; \boldsymbol{\beta})$ captures information on the adequacy of the model fit. For a point x, consider the difference $\mathcal{R} = m_0(x, x' = x + \Delta) - m(x, x' = x + \Delta; \boldsymbol{\beta})$. If the working model provides a good approximation $\mathcal{R}$ will be close to zero; if the models differ substantially, $\mathcal{R}$ provides information on how to improve the working model. For practical use m_0 can be replaced with a non-parametric kernel estimator, say $\hat{m}_0$, and $\boldsymbol{\beta}$ by a consistent estimator $\hat{\boldsymbol{\beta}}$, but a drawback of this approach is that $\hat{m}_0$ may be biased. We consider a kernel estimator of the *residuals*

$$R(X, X') = \mathrm{I}\left(Y \preccurlyeq Y'\right) - m(X, X'; \hat{\boldsymbol{\beta}}),$$

with $\mathrm{I}\left(Y \preccurlyeq Y'\right)$ denoting the *pseudo-observations*, defined as $\mathrm{I}\left(Y \preccurlyeq Y'\right) = 1$ if $Y < Y'$, $\mathrm{I}\left(Y \preccurlyeq Y'\right) = 0.5$ if $Y = Y'$ and $\mathrm{I}\left(Y \preccurlyeq Y'\right) = 0$ otherwise. Since the conditional expectation under H_0 is zero, there is no bias (le Cessie and van Houwelingen, 1991; Hardle and Mammen, 1993). We obtain a graphical tool by plotting the smoothed residuals as a function of the predictor and we construct a statistical test by considering a quadratic form of these residuals.

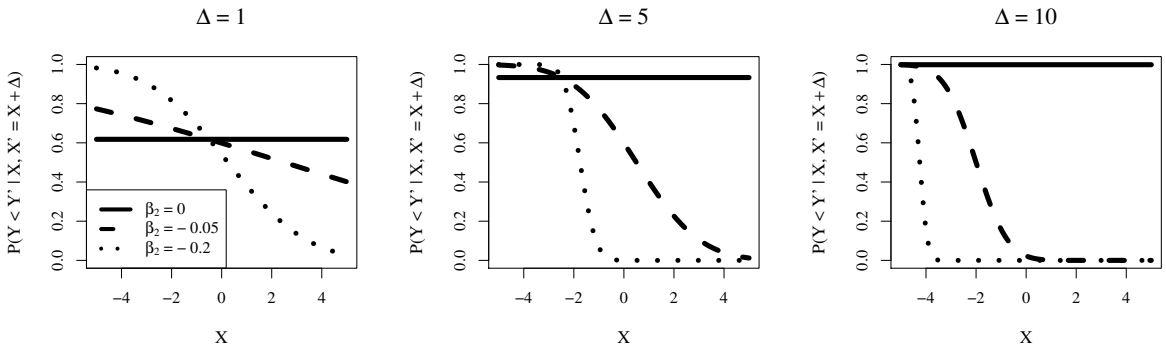

Figure 3: Quadratic probit PIM $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi\{\beta_1(X' - X) + \beta_2\left(X'^2 - X^2\right)\}$ with predictors restricted to $X' = X + \Delta$ and $\beta_1 = 0.3$.

2.2 The goodness-of-fit test

Since a PIM depends on (X, X'), we need to define appropriate *kernels* for our setting. Consider the multiplicative kernel

$$K_{h_1,h_2}(x, x'; X, X') = D\left(\frac{x - X}{h_1(x)}\right) D\left(\frac{x' - X'}{h_2(x')}\right), \tag{5}$$

where h_1 and h_2 are bandwidths and D is a kernel function such as a Gaussian, uniform, or triangular function. Our kernel provides double smoothing, i.e. for each (X, X'), we consider the distance between X and x, and between X' and x'. More weight is given to couples for which X is close to x and X' to x'. If no smoothing is desired, which, for example, may happen when a categorical predictor has sufficient replicates, we write $h_1 = h_2 = 0$ and denote by D the Dirac-delta function. For notional convenience we drop the dependence on h_1 and h_2 and write $K(x, x'; X, X')$ instead of $K_{h_1,h_2}(x, x'; X, X')$.

A Nadaraya–Watson kernel estimator (Nadaraya, 1964; Watson, 1964) of the residuals is defined by

$$\hat{R}(x, x') = \frac{\sum_{(k,l)\in\mathcal{I}_n} R(X_k, X_l) K(x, x'; X_k, X_l)}{\sum_{(k,l)\in\mathcal{I}_n} K(x, x'; X_k, X_l)}, \tag{6}$$

where $\mathcal{I}_n = \{(k, l) \in \mathbb{N}^2 \mid (X_k, X_l) \in \mathcal{X}\}$. The asymptotic null distribution is obtained by a first order Taylor expansion; we refer to Appendix B for details. It holds that, under regularity conditions and H_0, as $n \to \infty$,

$$\frac{\hat{R}(x, x')}{\sqrt{\mathrm{Var}\{\hat{R}(x, x')\}}} \xrightarrow{d} \mathrm{N}(0, 1),$$

where $\xrightarrow{d}$ is used to denote convergence in distribution. These results hold for linear smoothers (i.e. linear transformations of the residuals) in general. Therefore, instead of a local constant smoother (6), which suffers from design and boundary bias, local linear regression may be preferred (Fan and Gijbels, 1996; Wasserman, 2007). As mentioned in Section 2.1, we focus on the probability $\mathrm{P}(Y \preccurlyeq Y' \mid X, X' = X + \Delta)$ for assessing model adequacy, and plot the smoothed residuals $\hat{R}(x, x' = x + \Delta)$ as a function of x. These residuals provide information on the bias of the working model and are bounded in $[-1, 1]$. Figure 4 shows such a plot, based on random samples of size $n = 150$ for the 3 settings described in the left panel of Figure 3 with $\Delta = 1$. The left panel of Figure 4 corresponds to the setting under H_0 and the residuals are close to 0. For a weak

quadratic effect, the middle panel indicates that the fitted model gives biased probabilistic index estimators. For $X < -1$ the probability is underestimated, while for $X > 1$ it is overestimated. The right panel shows a strong quadratic effect for which similar conclusions hold. There is a multiplicity problem, as n confidence intervals are calculated simultaneously. Therefore, these intervals are only indicative, but they may be helpful in interpreting the graphical GOF tool.

For formal hypothesis testing we construct a single quadratic form of the smoothed residuals. Consider a fixed finite number of points, say $x_1, \ldots, x_m$, within the range of X, with $m \leq n$. Let $\boldsymbol{I}$ denote the $|\mathcal{I}_n|$-vector of pseudo-observations $\mathrm{I}(Y_i \preccurlyeq Y_j)$; further let $m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})$ denote the $|\mathcal{I}_n|$-vector with elements $m(X_i, X_j; \boldsymbol{\beta})$ and

$$\boldsymbol{V} = \mathrm{diag}\left(m(X_i, X_j; \boldsymbol{\beta}) - m(X_i, X_j; \boldsymbol{\beta})^2\right), \quad \boldsymbol{H} = -\frac{\partial m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \left(\frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}\right)^{-1} \frac{\partial m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}^{-1},$$

with $\boldsymbol{U}$ as defined in (21) and $\hat{\boldsymbol{R}}$ the m-vector of residuals $\hat{R}(x_i, x_i + \Delta)$. We define the quadratic form

$$S = \hat{\boldsymbol{R}}^T \mathrm{Var}(\hat{\boldsymbol{R}})^{-1} \hat{\boldsymbol{R}}, \tag{7}$$

with $\mathrm{Var}(\hat{\boldsymbol{R}}) = \boldsymbol{K}(\mathrm{diag}(\boldsymbol{1}) - \boldsymbol{H})\boldsymbol{\Sigma_I}(\mathrm{diag}(\boldsymbol{1}) - \boldsymbol{H})^T \boldsymbol{K}^T$, where $\boldsymbol{K}$ denotes the $(m \times |\mathcal{I}_n|)$-matrix of weights $K(x_i, x_i + \Delta; X_k, X_l) / \sum_{(k,l) \in \mathcal{I}_n} K(x_i, x_i + \Delta; X_k, X_l)$. If $n \to \infty$ and m remains fixed and bounded, under H_0

$$S \xrightarrow{d} \chi^2_m, \tag{8}$$

and a consistent estimator of $\mathrm{Var}(\hat{\boldsymbol{R}})$ can be obtained by replacing $\boldsymbol{\Sigma_I}$ with $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{I}}$; see Appendix B. The quadratic form S takes the estimated correlations between the residuals $\hat{R}(x_i, x_i + \Delta)$ and $\hat{R}(x_j, x_j + \Delta)$ into account. In total $m(m-1)/2$ correlations need to be estimated. When m is large relative to the sample size n, the estimated covariance matrix $\mathrm{Var}(\hat{\boldsymbol{R}})$ is not guaranteed to be positive definite. Therefore m should be chosen small relatively to the sample size n and the design points $x_1, \ldots, x_m$ should cover the whole range of X so as to increase the likelihood of detecting departures from the underlying model.

Our methods can be extended to multiple predictors, say $\boldsymbol{X}^T = (X_1, \ldots, X_d)$, by considering multiplicative kernels

$$K_{\boldsymbol{h}_1, \boldsymbol{h}_2}(\boldsymbol{x}, \boldsymbol{x'}, \boldsymbol{X}, \boldsymbol{X'}) = \prod_{i=1}^{d} K_{h_{1i}, h_{2i}}(x_i, x'_i, X_i, X'_i), \tag{9}$$

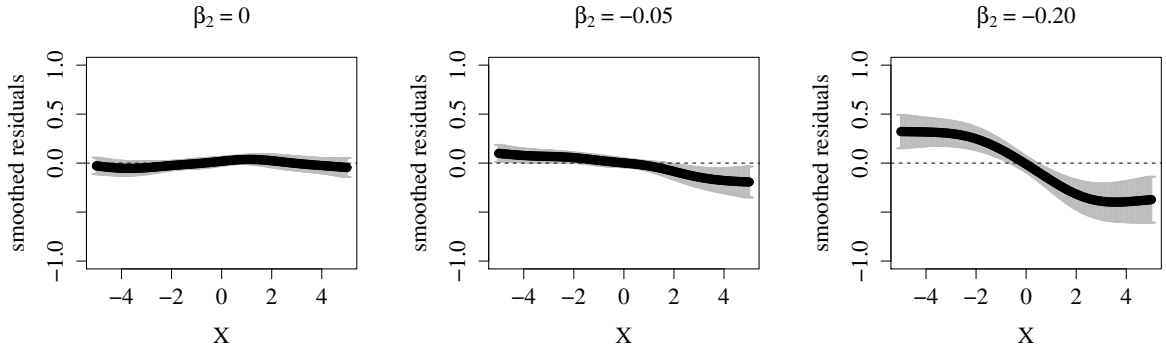

Figure 4: Smoothed residuals $\hat{R}(x, x+\Delta)$ as a function of x according to the different settings of the left panel of Figure 3 with $\Delta = 1$, for a random sample of size $n = 150$, and Gaussian kernel with $h_1 = h_2 = 1.5$. The black dots are the smoothed residuals, and the grey bars indicate pointwise 95% confidence intervals.

where $\boldsymbol{h}_i^T = (h_{i1}, \ldots, h_{id})$. The extension for two predictors is illustrated in Section 3. For high-dimensional data, however, smoothers based on a multiplicative kernel are not always useful in practice due to the curse of dimensionality and the computational burden. Therefore, nonparametric smoothers can be restricted to, for example, additive models.

2.3 Automatic bandwidth selection

It is known that the choice of bandwidth is often more important than the choice of kernel. Bandwidths may be selected in a data-driven fashion by using, for example, cross-validation (CV). The properties of the leave-one-out CV for independent responses has been examined by many authors; see for example Wong (1983). This CV can result in poor bandwidths if responses are dependent and several modifications have been proposed; see for example Chu and Marron (1991). We propose a modification of the leave-one-out CV score, accounting for the sparse correlation of the pseudo-observations. The bandwidth is chosen as the minimizer of

$$\mathrm{CV}(h_1, h_2) = |\mathcal{I}_n|^{-1} \sum_{(i,j)\in\mathcal{I}_n} \left\{ R(X_i, X_j) - \hat{R}_{-(i,j)}(X_i, X_j) \right\}^2, \tag{10}$$

with $\hat{R}_{-(i,j)}(X_i, X_j)$ the smoothed residual obtained by omitting all residuals containing (Y_i, X_i) or (Y_j, X_j).

3 Simulation study

Theoretical properties of S are empirically evaluated by means of simulations. The effect of the choice of bandwidth and Δ on the size and power of the test is examined for single and multiple predictors. The properties of the test with automatic bandwidth selection are also briefly examined.

3.1 A single predictor

3.1.1 Empirical sizes

To examine the null distribution of S more closely we generate data with the simple linear model

$$Y = \alpha X + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2), \tag{11}$$

which embeds the PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi\left\{\beta(X' - X)\right\}, \quad \beta = \alpha/\sqrt{2\sigma^2}, \tag{12}$$

see Thas et al. (2012, Section 4.1). The following parameters are fixed: $\alpha = 0.9\sqrt{2}$ and $\sigma^2 = 9$. The null distribution is examined for different values of the bandwidth h_1 and h_2, where we restrict to $h_1 = h_2$ and denote this by h, different values of Δ, and different sample sizes n. The statistic is based on three design points: $X = -3$, $X = 0$, and $X = 3$ with Gaussian kernel. Based on 1000 Monte Carlo simulation runs, the empirical rejection rates are compared to the nominal significance levels of 1%, 5%, and 10%. The asymptotic chi-squared distribution is used for p-value calculation. Table 1 shows all results.

For a sample size $n = 100$ and a small bandwidth $h = 0.5$ the test is highly conservative, while for a large bandwidth $h = 2.5$ the test is highly liberal. Best results are obtained for an intermediate bandwidth $h = 1.5$. For $n = 250$ and $h = 0.5$ the test is too conservative for $\Delta = 1$ and slightly less conservative for $\Delta = 2$. With $h = 1.5$ the test has approximately a correct size for all Δ, while for $h = 2.5$ the test remains too liberal. For a sample size $n = 500$ and $h = 0.5$ the test is conservative for $\Delta = 1$ and has approximately a correct size for $\Delta = 2$. For $h = 1.5$ the test has approximately a correct size, while for $h = 2.5$ the test remains liberal.
In conclusion, best results are obtained for an intermediate bandwidth of $h = 1.5$, while the choice of Δ is less important.

h	Δ	$n = 100$			$n = 250$			$n = 500$		
		1%	5%	10%	1%	5%	10%	1%	5%	10%
0.5	1	0.0	0.7	3.8	0.2	3.1	7.3	0.3	3.5	7.7
0.5	2	0.0	1.7	5.1	0.3	3.2	9.0	0.5	4.9	9.9
1.5	1	0.5	4.4	8.8	0.4	5.1	9.5	1.2	4.4	11.1
1.5	2	0.3	3.6	9.3	0.6	4.7	11.2	1.2	5.8	11.7
2.5	1	3.4	9.6	15.4	2.6	8.0	14.1	2.3	7.7	13.4
2.5	2	2.3	7.4	14.4	1.9	7.8	13.8	1.8	7.5	13.0

Table 1: Empirical rejection rates (%) at the 1%, 5%, and 10% level of significance and based on 1000 Monte-Carlo simulations for model (12).

3.1.2 Empirical powers

The results of Section 3.1.1 suggest that good results were obtained for a medium bandwidth. Therefore we restrict the power study to $h_1 = h_2 = 1.5$ in a Gaussian kernel with design points $X = -3$, $X = 0$ and $X = 3$. We generate data according to the model

$$Y = \alpha_1 X + \alpha_2 f(X) + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2). \tag{13}$$

We fix $\alpha_1 = 0.9\sqrt{2}$ and $\sigma^2 = 9$ and consider three cases: a *quadratic model* with $f(X) = X^2$ and $\alpha_2 = -0.05\sqrt{2}$ or $\alpha_2 = -0.125\sqrt{2}$; a *sine model* with $f(X) = \sin(X)$ and $\alpha_2 = -0.6\sqrt{2}$ or $\alpha_2 = -1.2\sqrt{2}$; an *exponential model* with $f(X) = \exp(X)$ and $\alpha_2 = 0.02\sqrt{2}$ or $\alpha_2 = 0.04\sqrt{2}$. The parameter values are chosen so that most empirical powers are bounded away from the trivial powers of 5% and 100%. The PIM corresponding to model (13) is given by

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi\left[\beta_1(X' - X) + \beta_2\{f(X') - f(X)\}\right], \quad \beta_i = \alpha_i/\sqrt{2\sigma^2}, \quad i = 1, 2. \tag{14}$$

We analyse the data with the incorrect working model $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \Phi\{\beta(X' - X)\}$. The three panels starting from the left of Figure 5 show the probability $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X' = X + 1\right)$ as a function of X for the three different models and for different β_2 values.

Table 2 gives the empirical rejection rates based on 1000 Monte Carlo simulations for the different data-generating models at the 5% level of significance. The test succeeds in detecting

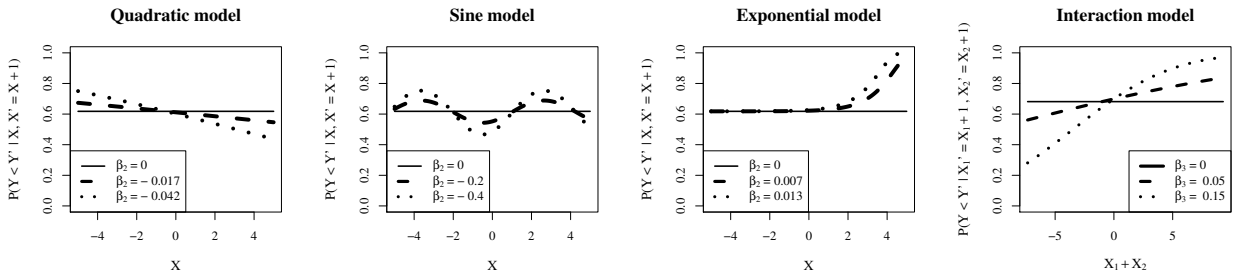

Figure 5: Conditional PIM for $X' = X + 1$ over different values of β_2 for the quadratic, sine, and exponential versions of model (14), and probability $\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1' = X_1 + \Delta_1, X_2' = X_2 + \Delta_2\right)$ as a function of $X_1 + X_2$ for different values of β_3.

lack-of-fit. Under the conditions of the simulation study, for the quadratic and sine model highest powers are obtained with $\Delta = 1$ while for the exponential model this is $\Delta = 2$.

β_2	$n = 100$		$n = 250$		$n = 500$	
	$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$	$\Delta = 1$	$\Delta = 2$
	quadratic model					
-0.017	12.0	11.0	42.1	40.7	78.2	75.5
-0.042	73.2	68.9	99.8	99.5	100.0	100.0
	sine model					
-0.2	14.6	8.7	53.6	36.3	89.2	70.5
-0.4	64.9	39.6	99.7	94.9	100.0	100.0
	exponential model					
0.007	14.0	14.2	49.6	57.2	82.4	89.7
0.013	38.1	42.1	96.9	98.6	100.0	100.0

Table 2: Empirical rejection rates (%) at the 5% level of significance and based on 1000 Monte-Carlo simulations for model (14).

We examined the power of detecting a misspecified link function by simulating data with

model (11) and analysing this data with $\mathrm{P}\left(Y \preccurlyeq Y' \mid X, X'\right) = \mathrm{expit}\{\gamma(X' - X)\}$. The simulation results indicated low to moderate powers (results not shown).

3.2 Multiple predictors

3.2.1 Empirical sizes

Consider the data-generating model

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2),$$

with embedded PIM

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1, X_2, X_1', X_2'\right) = \Phi\left\{\beta_1(X_1' - X_1) + \beta_2(X_2' - X_2)\right\}, \quad \beta_i = \alpha_i/\sqrt{2\sigma^2}, \quad i = 1, 2. \tag{15}$$

The following parameters are fixed: $\alpha_1 = \alpha_2 = 1$ and $\sigma^2 = 9$, corresponding to $\beta_1 = \beta_2 = 0.24$. The predictor X_1 takes n equidistant values in the interval $[-5, 5]$, while $X_2 \sim \mathrm{N}(0, 4)$. The statistic is based on three design points: $(X_1, X_2) = (-3, -2.5)$, $(X_1, X_2) = (0, 0)$, and $(X_1, X_2) = (3, 2.5)$, with Gaussian kernel and with bandwidths $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (1.5, 1.5)$. Different values for $\boldsymbol{\Delta}$ and n are considered. Based on 1000 Monte Carlo simulation runs, the empirical rejection rates are compared to the nominal rejection rates for significance levels of 1%, 5%, and 10%. The results are presented in Table 3. For a sample size $n = 100$ our test is highly conservative, while it becomes less conservative when the sample size increases. For $n = 500$ our test has approximately a correct size for all choices of $\boldsymbol{\Delta}$.

3.2.2 Empirical powers

Consider the data-generating model with interaction

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1 X_2 + \varepsilon, \quad \varepsilon \sim \mathrm{N}(0, \sigma^2). \tag{16}$$

We fix $\alpha_1 = \alpha_2 = 1$ and $\sigma^2 = 9$ and consider different values of α_3. The corresponding PIM is

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1, X_2, X_1', X_2'\right) = \Phi\left\{\beta_1(X_1' - X_1) + \beta_2(X_2' - X_2) + \beta_3(X_1' X_2' - X_1 X_2)\right\}. \tag{17}$$

The data are analyzed with the incorrect working model

$$\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1, X_2, X_1', X_2'\right) = \Phi\left\{\gamma_1(X_1' - X_1) + \gamma_2(X_2' - X_2)\right\}. \tag{18}$$

$\boldsymbol{\Delta}$	$n = 100$			$n = 250$			$n = 500$		
	1%	5%	10%	1%	5%	10%	1%	5%	10%
$(1, 1)$	0.0	0.9	3.4	0.3	2.4	6.1	0.9	4.7	9.0
$(1, 2)$	0.0	1.1	4.5	0.2	2.3	6.9	1.1	4.3	10.2
$(2, 1)$	0.0	1.1	4.3	0.3	3.1	6.6	0.8	3.9	9.3
$(2, 2)$	0.0	0.7	5.3	0.3	3.6	6.8	0.9	3.4	10.8

Table 3: Empirical rejection rates (%) at the 1%, 5%, and 10% level of significance and based on 1000 Monte-Carlo simulations for the model (15).

The right panel of Figure 5 plots $\mathrm{P}\left(Y \preccurlyeq Y' \mid X_1' = X_1 + \Delta_1, X_2' = X_2 + \Delta_2\right)$ as a function of the sum $\Delta_2 X_1 + \Delta_1 X_2$ when $\boldsymbol{\Delta} = (1, 1)$ and for different values of β_3. Table 4 gives the empirical rejection rates at the 5% significance level, based on 1000 Monte Carlo simulation runs. The statistic is based on three design points: $(X_1, X_2) = (-3, -2.5)$, $(X_1, X_2) = (0, 0)$, and $(X_1, X_2) = (3, 2.5)$, with Gaussian kernel and bandwidth $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (1.5, 1.5)$.

The test succeeds in detecting an omitted interaction and under the conditions of the simulation study highest powers are obtained for $\boldsymbol{\Delta} = (1, 2)$ or $\boldsymbol{\Delta} = (2, 2)$.

$\boldsymbol{\Delta}$	$(1, 1)$	$(1, 2)$	$(2, 1)$	$(2, 2)$	$(1, 1)$	$(1, 2)$	$(2, 1)$	$(2, 2)$	$(1, 1)$	$(1, 2)$	$(2, 1)$	$(2, 2)$
β_3	$n = 100$				$n = 250$				$n = 500$			
0.05	1.7	2.8	1.6	2.9	28.4	52.4	21.7	44.8	58.3	83.2	54.4	82.6
0.15	42.9	71.6	40.1	75.8	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0

Table 4: Empirical rejection rates (%) at the 5% level of significance and based on 1000 Monte-Carlo simulations for model (17).

3.3 Automatic bandwidth selection

To examine the null distribution of S when the bandwidth is selected based on the modified cross-validation score (10), we reconsider the simulation step-up from Section 3.1.1 with $\Delta = 1$.

Because $|\mathcal{I}_n| = O(n^2)$, which is computationally very demanding for large samples, we restrict the sum in (10) to the subset $\mathcal{I}_{\text{sub}} = \{(i,j) \mid \Delta - 0.05 < X_j - X_i < \Delta + 0.05\}$. For $n = 250$ and $n = 500$ the sum is even restricted to a random sample of size 100 from $\mathcal{I}_{\text{sub}}$. The candidate set of bandwidths is restricted to $\{0.5, 1.5, 2.5\}$ with $h_1 = h_2$. To examine the empirical powers, we reconsider the quadratic model from Section 3.1.2 with $\Delta = 1$. Table 5 gives the empirical rejection rates. For all sample sizes the test is liberal. As compared to Table 1 the results are slightly better with $h_1 = h_2 = 1.5$ and worse with $h_1 = h_2 = 0.5$ or 2.5. For $n = 500$ the empirical rejections rates are close to their nominals for 1% and 5% but too liberal for 10%. The automatic cross-validation results in some power loss as compared to Table 2.

$n = 100$			$n = 250$			$n = 500$			β_2	$n = 100$	$n = 250$	$n = 500$
Empirical type I error									Empirical power quadratic model			
1%	5%	10%	1%	5%	10%	1%	5%	10%	−0.017	14.5	37.5	67.9
2.9	7.4	14.1	2.1	5.5	12.0	0.9	5.8	12.6	−0.042	68.4	87.6	96.1

Table 5: Empirical type I error (%) at the 1%, 5%, and 10% level of significance and empirical powers (%) at the 5% level of significance when the bandwidth is automatically selected with the modified cross-validation score. All results are based on 1000 Monte-Carlo simulations.

3.4 Assessing goodness-of-fit with the graphical tool

In Figure 6 we show the GOF plots for 4 simulated dataset with sample size $n = 150$ for the quadratic, sine, exponential, and interaction model respectively; the Gaussian kernel is used with $h = 1.5$ and $\Delta = 1$. The GOF plots show similar shapes as Figure 5, indicating that GOF plots are informative on how the true model differs from the working model.

4 Case study

We return to the CRDS example. Since most (89%) of the smoking children are between 10 and 16 years old, we restrict the conclusion to that age class. In model (2) the effect of the smoking

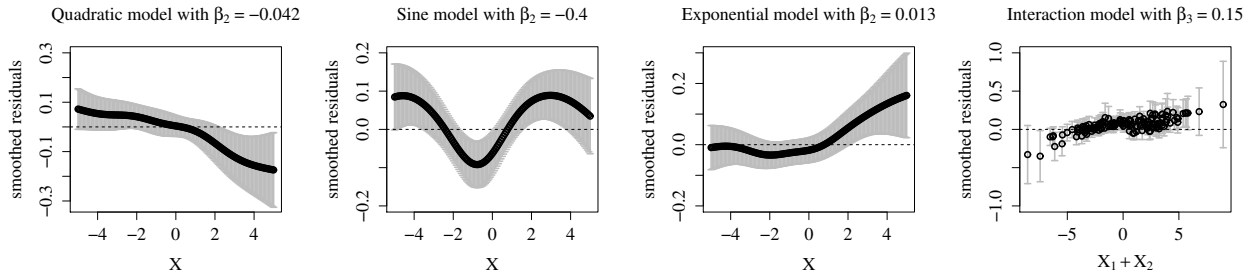

Figure 6: GOF plots for the quadratic, sine, exponential, and interaction models (14) and (17) respectively, for a random sample of size $n = 150$.

status on the pulmonary function of a child is not significant. Smoothed residuals were constructed with a Gaussian kernel, $\boldsymbol{\Delta} = (1, 1)$ and bandwidths $\boldsymbol{h}_1 = \boldsymbol{h}_2 = (h, 0)$ with $h \in \{0.5, 1, 1.5\}$, for which the optimal bandwidth was selected based on the cross-validation score (10) with the sum restricted to a random sample of size 100 of $\mathcal{I}_{\text{sub}} = \{(i, j) \mid \text{AGE}_j - \text{AGE}_i = 1\}$. The binary predictor SMOKE has sufficient replicates, and smoothing is unnecessary. Similar to the right panel of Figure 6, the left panel of Figure 7 plots the residuals as a function of the sum SMOKE+AGE for model (2) with $h = 1$. This plot indicates that for the younger children, the probability $\text{P}\{\text{FEV} \preccurlyeq \text{FEV}' \mid (0, \text{AGE}), (1, \text{AGE} + 1)\}$ is underestimated, while for the older it is overestimated. The statistical test with design points $(\text{SMOKE}, \text{AGE}) = (0, 10)$ and $(0, 14)$ confirms this: $S = 10.1$ and $p = 0.006$. The plot suggests that this probability depends on the sum SMOKE+AGE. Therefore we fit an interaction model which takes this into account

$$\begin{aligned} &\text{logit}\left[\text{P}\{\text{FEV} \preccurlyeq \text{FEV}' \mid (\text{SMOKE}, \text{AGE}), (\text{SMOKE}', \text{AGE}')\}\right] \\ &= \beta_1(\text{SMOKE}' - \text{SMOKE}) + \beta_2(\text{AGE}' - \text{AGE}) + \beta_3(\text{SMOKE}' \times \text{AGE}' - \text{SMOKE} \times \text{AGE}), \end{aligned} \tag{19}$$

with estimates $\hat{\beta}_1 = 5.3$ (standard error: 1.04 and $p < 0.0001$), $\hat{\beta}_2 = 0.61$ (standard error: 0.03 and $p < 0.0001$), and $\hat{\beta}_3 = -0.46$ (standard error: 0.08 and $p < 0.0001$). All effects are now highly significant. The middle panel of Figure 7 gives the GOF plot with $h = 1$. Based on the GOF test there is no convincing evidence for lack-of-fit: $S = 0.36$ and $p = 0.84$. It may well be that including additional predictors further improves the model fit. Figure 1 suggested that an interaction effect should be included in the model. The estimated effect of smoking, in terms of

the probabilistic index, is given by

$$\text{logit}\left[\hat{\text{P}}\left\{\text{FEV} \preccurlyeq \text{FEV}' \mid (0, \text{AGE}), (1, \text{AGE})\right\}\right] = 5.3 - 0.46\text{AGE}. \tag{20}$$

The probability for having a better pulmonary function for the smoking child decreases with increasing age. The right panel of Figure 7 shows this probability as a function of AGE. At the age of ten, for example, the estimated probability is 68% with confidence interval $[53\%, 80\%]$. This probability indicates that the lung function is better for smoking children, which seems unreasonable. However, children who smoke at the age of ten likely only just started smoking and the smoking did not affect the lungs yet. By the age of 16 this probability decreased to 12%, indicating it is highly unlikely that a smoking child has a better lung function, demonstrating the adverse effects of smoking; the confidence interval for this probability is $[7\%, 21\%]$.

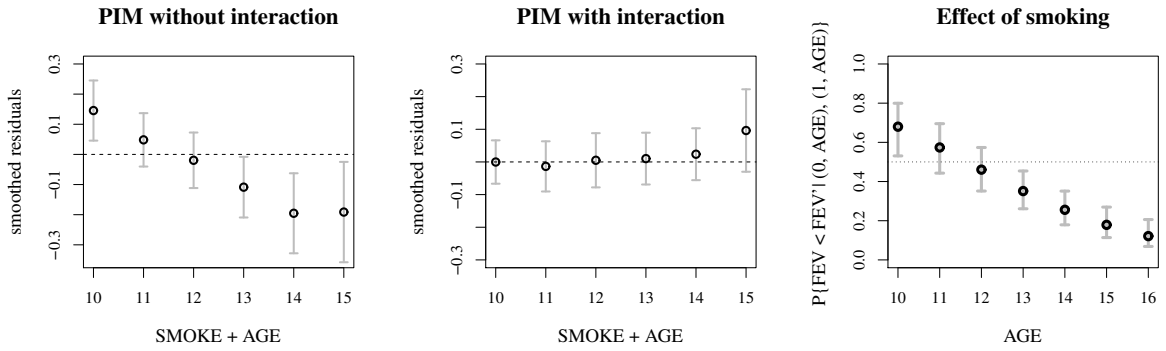

Figure 7: Left: GOF plot for model (2); middle: GOF plot for model (19); right: $\text{P}\{\text{FEV} \preccurlyeq \text{FEV}' \mid (0, \text{AGE}), (1, \text{AGE})\}$ as a function of AGE for model (19). The grey bars indicate the pointwise 95% confidence intervals.

5 Discussion

An informative GOF plot together with a formal GOF test for PIMs is proposed. The GOF test has good power properties and the plot provides information on how the model can be improved. The GOF tools are consistent with the interpretation of a PIM, where the probability $\text{P}\left(Y \preccurlyeq Y' \mid \boldsymbol{X}, \boldsymbol{X}' = \boldsymbol{X} + \boldsymbol{\Delta}\right)$ serves as a basis. The parameter $\boldsymbol{\Delta}$ should be chosen such that this

probability has a meaningful interpretation; for future research it can be interesting to focus on an adaptive selection of $\boldsymbol{\Delta}$. The residuals are based on smoothers and the size of the test particularly depends on the choice of bandwidth. For an intermediate bandwidth, the empirical type I error is close to its nominal value. We proposed a modified cross-validation score to select the bandwidth automatically. The corresponding size remains slightly liberal, even for large sample sizes. It may be of interest to extend the wild bootstrap method of Hardle and Mammen (1993) to our pseudo-observations setting, as this might improve the small-sample behaviour of the test. Our test has good power for detecting an omission of a quadratic, sine, and exponential term as well as an omission of an interaction effect, while having low to moderate power for detecting a misspecified link function. However, for most PIMs the interpretation of the parameters is independent of the link function.

Many GOF statistics use all residuals to form a Cramér–von Mises, Anderson–Darling or Kolmogorov–Smirnov type of test. Because the pseudo-observations are sparsely correlated, the distribution theory of such test statistics is much harder than for many other types of regression models. By constructing our test statistic as a quadratic form which uses only a limited number of design points, some technical difficulties are avoided. Future research may focus on extending our method so as to use all residuals. It is anticipated that this would make the method even more sensitive for detecting a wider range of model departures.

As PIMs are a relatively new class of regression models, non-parametric regression estimators have not been described yet. In this paper an initial step is taken by considering kernel smoothers for the construction of the test statistic. In future work this will be studied in more detail so that PIMs can include genuine nonparametric regression estimators.

6 Acknowledgements

Research supported by IUAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy). The authors would like to thank the referees and associate editor for their constructive comments.

Appendices

A Estimation theory

Let $(Y_1, \boldsymbol{X}_1), \ldots, (Y_n, \boldsymbol{X}_n)$ denote a random sample from a distribution with density $f_{Y\boldsymbol{X}}$, then a consistent estimator of $\boldsymbol{\beta}$, say $\hat{\boldsymbol{\beta}}$, can be obtained by solving the estimating equations

$$\boldsymbol{U}_n(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{I}_n} \boldsymbol{U}_{ij}(\boldsymbol{\beta}) = \sum_{(i,j)\in\mathcal{I}_n} \frac{\partial m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \frac{\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})}{m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})\left\{1 - m(\boldsymbol{X}_i, \boldsymbol{X}_j; \boldsymbol{\beta})\right\}} = \mathbf{0}. \quad (21)$$

The estimator $\hat{\boldsymbol{\beta}}$ has an asymptotic multivariate normal distribution and a consistent estimator of the corresponding variance-covariance matrix, say $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$, is provided by the sandwich estimator

$$\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}} = \left(\sum_{(i,j)\in\mathcal{I}_n} \frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} \right)^{-1} \left(\sum_{(i,j)\in\mathcal{I}_n} \sum_{(k,l)\in\mathcal{I}_n} \phi_{ijkl} \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}}) \boldsymbol{U}_{kl}^T(\hat{\boldsymbol{\beta}}) \right) \left(\sum_{(i,j)\in\mathcal{I}_n} \frac{\partial \boldsymbol{U}_{ij}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} \right)^{-1^T},$$

where the indicator ϕ_{ijkl} is defined as $\phi_{ijkl} = 1$ if $\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)$ and $\mathrm{I}\left(Y_k \preccurlyeq Y_l\right)$ are correlated, and $\phi_{ijkl} = 0$ otherwise; we refer to Thas et al. (2012, Section 3) for more details.

B Smoothed residuals

Let $\boldsymbol{I}$ denote the $|\mathcal{I}_n|$-vector of pseudo-observations $\mathrm{I}\left(Y_i \preccurlyeq Y_j\right)$, $m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})$ the $|\mathcal{I}_n|$-vector with elements $m(X_i, X_j; \boldsymbol{\beta})$, and $\boldsymbol{V}$ the diagonal matrix with elements $m(X_i, X_j; \boldsymbol{\beta})\{1 - m(X_i, X_j; \boldsymbol{\beta})\}$. Following le Cessie and van Houwelingen (1991), we consider two first order Taylor approximations; using the notation introduced in Appendix A,

$$\boldsymbol{I} - m(\boldsymbol{X}, \boldsymbol{X'}; \hat{\boldsymbol{\beta}}) \approx \boldsymbol{I} - m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta}) - \frac{\partial m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}), \quad \mathbf{0} = \boldsymbol{U}(\hat{\boldsymbol{\beta}}) \approx \boldsymbol{U}(\boldsymbol{\beta}) + \frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Consequently $\boldsymbol{I} - m(\boldsymbol{X}, \boldsymbol{X'}; \hat{\boldsymbol{\beta}}) \approx (\mathrm{diag}(\mathbf{1}) - \boldsymbol{H})\{\boldsymbol{I} - m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})\}$, where

$$\boldsymbol{H} = -\frac{\partial m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \left(\frac{\partial \boldsymbol{U}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^T} \right)^{-1} \frac{\partial m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})^T}{\partial \boldsymbol{\beta}} \boldsymbol{V}^{-1},$$

which is a generalization of the *hat-matrix*. If $\boldsymbol{K}(X_i, X_j)$ denotes the $|\mathcal{I}_n|$-vector with elements $K(X_i, X_j; X_k, X_l) / \sum_{(k,l)\in\mathcal{I}_n} K(X_i, X_j; X_k, X_l)$, then

$$\hat{R}(X_i, X_j) = \boldsymbol{K}(X_i, X_j)^T \{\boldsymbol{I} - m(\boldsymbol{X}, \boldsymbol{X'}; \hat{\boldsymbol{\beta}})\} \approx \boldsymbol{K}(X_i, X_j)^T (\mathrm{diag}(\mathbf{1}) - \boldsymbol{H})\{\boldsymbol{I} - m(\boldsymbol{X}, \boldsymbol{X'}; \boldsymbol{\beta})\}.$$

It holds that

$$\mathrm{E}\{\hat{R}(X_i, X_j)\} \approx 0, \quad \mathrm{Var}\{\hat{R}(X_i, X_j)\} \approx \boldsymbol{K}(X_i, X_j)^T \left(\mathrm{diag}(\mathbf{1}) - \boldsymbol{H}\right) \mathrm{Var}\left(\boldsymbol{I}\right) \left(\mathrm{diag}(\mathbf{1}) - \boldsymbol{H}\right)^T \boldsymbol{K}(X_i, X_j).$$

The central limit theorem of Lumley and Hamblett (2003, p. 13) guarantees that, under H_0

$$\frac{\hat{R}(X_i, X_j)}{\sqrt{\mathrm{Var}\{\hat{R}(X_i, X_j)\}}} \xrightarrow{d} \mathrm{N}(0, 1).$$

A consistent estimator for $\mathrm{Var}\{\hat{R}(X_i, X_j)\}$ can be obtained by substituting $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and $\mathrm{Var}\left(\boldsymbol{I}\right)$ by $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{I}}$, where

$$\left(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{I}}\right)_{(ij),(kl)} = \begin{cases} \left\{\mathrm{I}\left(Y_i \preccurlyeq Y_j\right) - m(X_i, X_j; \hat{\boldsymbol{\beta}})\right\} \left\{\mathrm{I}\left(Y_k \preccurlyeq Y_l\right) - m(X_k, X_l; \hat{\boldsymbol{\beta}})\right\}, & \text{if } \phi_{ijkl} = 1, \\ 0, & \text{if } \phi_{ijkl} = 0. \end{cases}$$

All results also hold when X_i is a d-dimensional predictor.

References

Chu, C.-K. and Marron, J. S. (1991). Comparison of two bandwidth selectors with dependent errors. *Ann. Stat.*, 19:1906–1918.

Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. New York: Chapman & Hall.

Hardle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Stat.*, 21:1926–1947.

le Cessie, S. and van Houwelingen, J. (1991). A goodness-of-fit test for binary regression models, based on smoothing methods. *Biometrics*, 47:1267–1282.

Lumley, T. and Hamblett, N. M. (2003). Asymptotics for marginal generalized linear models with sparse correlations. *UW Biostatistics Working Paper Series*, 207.

Nadaraya, E. (1964). On estimating regression. *Theory Probab. Appl.*, 9:141–142.

Thas, O., De Neve, J., Clement, L., and Ottoy, J.P. (2012). Probabilistic index models (with Discussion). *J. R. Stat. Soc. Ser. B - Stat. Methodol.*, 74:623–671.

Wasserman, L. (2007). *All of Nonparametric Statistics*. New York: Springer.

Watson, G. (1964). Smooth regression analysis. *Sankhya, Series A*, 26:359–372.

Wong, W. H. (1983). On the consistency of cross-validation in kernel nonparametric regression. *Ann. Stat.*, 11:1136–1141.