Model selection in regression based on pre-smoothing

AERTS, Marc; HENS, Niel & Simonoff, Jeffrey S. (2010) Model selection in regression based on pre-smoothing. In: JOURNAL OF APPLIED STATISTICS, 37 (9). p. 1455-1472.

DOI: 10.1080/02664760903046086
Handle: http://hdl.handle.net/1942/11222

Model Selection in Regression Based on Presmoothing

Marc Aerts [a]**, Niel Hens [a], and Jeffrey S. Simonoff[b]

[a] *Interuniversity Institute for Biostatistics and Statistical Bioinformatics,*

Hasselt University, Campus Diepenbeek, Agoralaan 1, B-3590 Diepenbeek, Belgium
;[b] *Leonard N. Stern School of Business, New York University, 44 West 4th Street;New York, NY 10012-0258*

March 9, 2012

Abstract

In this paper we investigate the effect of presmoothing on model selection. Christóbal Christóbal et al. (1987) showed the beneficial effect of presmoothing for estimating the parameters in a linear regression model. Here, in a regression setting, we show that smoothing the response data prior to model selection by Akaike's Information Criterion can lead to an improved selection procedure. The bootstrap is used to control the magnitude of the random error structure in the smoothed data. The effect of presmoothing on model selection is shown in simulations. The method is illustrated in a variety of settings, including the selection of the best fractional polynomial in a generalized linear model.

Akaike Information Criterion; fractional polynomial; latent variable model; model selection; presmoothing

1 Introduction

Based on observations $(\boldsymbol{x_i}, y_i), i = 1, ..., n$, consider the regression model

$$\boldsymbol{y} \sim f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta}), \tag{1}$$

where

$$\boldsymbol{y} = (y_1, \ldots, y_n)^T, \;\; \boldsymbol{\theta} = (\theta(\boldsymbol{x}_1), \ldots, \theta(\boldsymbol{x}_n))^T, \;\; \boldsymbol{\eta} = (\eta(\boldsymbol{x}_1), \ldots, \eta(\boldsymbol{x}_n))^T.$$

Here f denotes the joint density of $\boldsymbol{y}$ (given $\boldsymbol{x}$), $\boldsymbol{\theta}$ the parameter of interest and $\boldsymbol{\eta}$ a nuisance parameter. We also assume that $\boldsymbol{\theta}$ is in some way related to $E(\boldsymbol{y})$, more precisely we assume that there exists a function g such that

$$E(\boldsymbol{y}) = g(\boldsymbol{\theta}; \boldsymbol{x}).$$

The aim is to select an optimal or a few good models amongst a set of candidate models. Several model selection criteria have been developed, in different settings and with different types of complexities in data and models (see e.g. Akaike, 1973; Takeuchi, 1976; Schwarz, 1978; Spiegelhalter et al., 2002; Pan, 2001a,b; Hens et al., 2006), to accomplish this.

Assume we start from a collection of models, in particular we consider models of the form (1) . The well-known AIC criterion (Akaike, 1973)

$$\text{AIC} = -2L(\boldsymbol{y}; \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}}) + 2K, \tag{2}$$

with $L(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta})$ denoting the loglikelihood of the model and $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\eta}})$ the maximum likelihood (ML) estimator of $(\boldsymbol{\theta}, \boldsymbol{\eta})$, originates from information theory. Here K stands for the total number of estimated parameters,

**Corresponding author. Email: marc.aerts@uhasselt.be

nuisance parameters included. The second term in the AIC formula is often interpreted as a penalization for complexity. The AIC was designed to be an approximately unbiased estimator of the expected *Kullback-Leibler* (KL) information. In general, the KL information between model f_0 (denoting the 'true' model) and model f (the approximating model (1)) is defined as (ignoring an 'historical' factor 2)

$$I(f_0, f) = E\left[\log\left\{\frac{f_0(\boldsymbol{y})}{f(\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{\eta})}\right\}\right]$$

(expectation with respect to the true model) and can be interpreted as the information loss using f to approximate f_0, or as the distance from f_0 to f. This KL distance is not a metric, but it has the property that $I(f_0, f) \geq 0$ with equality only if $f \equiv f_0$.

The basic idea of the presmoothed AIC is to replace the observed value $\boldsymbol{y}$ by the estimated value $\boldsymbol{y}^S = \widehat{E(\boldsymbol{y})} = \hat{g}(\boldsymbol{x};\lambda)$ using a nonparametric regression model with smoothing parameter λ (e.g. local polynomials, penalized regression splines, and so on). Next, the 'smoothed' AIC is calculated for all candidate models using the "pseudo-data" $(\boldsymbol{x}, \boldsymbol{y}^S)$

$$\text{AIC}_S = -2L_S(\boldsymbol{y}^S; \hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\eta}}_S) + 2K, \tag{3}$$

with $(\hat{\boldsymbol{\theta}}_S, \hat{\boldsymbol{\eta}}_S)$ the MLE's of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ based on $\boldsymbol{y}^S = (y_1^S, \ldots, y_n^S)^T$. These presmoothed AIC values are then used to select the final model, or to compute an averaged model (Burnham and Anderson, 2002). The likelihood function L_S might differ from the likelihood of the original data. The rationale of this method is as follows: replacing the data by an estimated curve filters out most of the error structure and more clearly exhibits the optimal parametric mean structure, as a function of $\boldsymbol{x}$. It is clear of course that by presmoothing focus is on model selection of the mean structure, so on the main parameter $\boldsymbol{\theta}$. There is some analogy with the beneficial effect of presmoothing for linear regression estimators as shown in Christóbal Christóbal et al. (1987), Faraldo Roca and González Manteiga (1987) and Janssen et al. (2001).

Presmoothing can also be motivated in the following way. It is well-known that model selection is a highly variable process, in the sense that small perturbations in the data can lead to very different models being chosen. Ye (1998) showed that the cost of model selection (in the sense of overestimation of the strength of the fit of the chosen model) is directly related to stability of the model selection procedure when the data are perturbed, which is itself related to the strength of the structure in the data relative to the noise. The goal of presmoothing is to effectively increase the signal in the data relative to the noise by using estimated curve values as the response data, thereby increasing the stability of the model selection when the data are perturbed, and reducing the cost of model selection.

The paper is organized as follows. In Section 2 we consider the basic implementation of the method in the setting of linear regression. Simulations illustrate the performance of the basic method. An interesting application is the selection of the optimal fractional polynomial (Royston and Altman, 1994). Section 3 highlights some shortcomings and complications of the basic method. A bootstrap approach is proposed to solve these difficulties. The more general setting of categorical response data and generalized linear models is studied in Section 4. This smoothed latent variable approach is illustrated on a example of age stratified seroprevalence data on Hepatitis A. A final discussion section indicates some possible extensions and topics for further research.

2 Presmoothing Data Prior to Model Selection

In this section we focus on the smoothed AIC criterion (3) in the setting of linear regression with normal error structure.

2.1 Akaike Information Criterion Based on Presmoothed Data

Consider classical regression and suppose data are generated by a true model

$$\boldsymbol{y} \overset{f_0}{\sim} N_n(\boldsymbol{\mu}_0, \sigma_0^2 I_n),$$

where $\boldsymbol{\mu}_0 = (\mu_0(1), \ldots, \mu_0(n))^T$, N_n denotes an n-variate normal distribution and I_n the $n \times n$ identity matrix . Consider the approximating, or candidate, family of models

$$\boldsymbol{y} \overset{f}{\sim} N_n(\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2 I_n),$$

where $\boldsymbol{\mu}(\boldsymbol{\theta}) = (\mu(\boldsymbol{x}_1;\boldsymbol{\theta}),,\ldots,\mu(\boldsymbol{x}_n;\boldsymbol{\theta}))^T$.

For this setting, $E\{\log f(\boldsymbol{y};\boldsymbol{\theta},\boldsymbol{\eta})\}$ can be written as (ϕ denoting the univariate normal density)

$$E\{\sum_{i=1}^{n}\log\phi(y_i;\mu(\boldsymbol{x_i}),\sigma^2)\} = -\frac{n}{2}\log(2\pi\sigma^2) - E\left[\{\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\theta})\}^T\{\boldsymbol{y}-\boldsymbol{\mu}(\boldsymbol{\theta})\}\right]/(2\sigma^2).$$

Using an analogous expression for $E\{\log f_0(\boldsymbol{y})\}$, it is easy to verify that

$$I(f_0,f) = \frac{n}{2}\log(\sigma^2/\sigma_0^2) + \frac{n}{2}\left\{\frac{\sigma_0^2}{\sigma^2}-1\right\} + \{\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta})\}^T\{\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta})\}/(2\sigma^2). \tag{4}$$

It follows that this measure is minimized as a function of σ^2 and $\boldsymbol{\mu}(\boldsymbol{\theta})$ (and equals 0) by taking $\sigma^2=\sigma_0^2$ and $\boldsymbol{\mu}(\boldsymbol{\theta})=\boldsymbol{\mu}_0$.

Now, let us introduce presmoothing based on a linear smoother. Define

$$\boldsymbol{y}^S = S_\lambda \boldsymbol{y},$$

with S_λ the smoother matrix. In the case that $\text{rank}(S_\lambda)=n$, we have that (see e.g. Chapter 6 in Ruppert et al., 2003)

$$\boldsymbol{y}^S \overset{f_0}{\sim} N_n(S_\lambda\boldsymbol{\mu}_0, \sigma_0^2 S_\lambda S_\lambda^T).$$

Approximating this smoothed 'true' model f_0^S by the smoothed approximate model f^S

$$\boldsymbol{y}^S \overset{f^S}{\sim} N_n(S_\lambda\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2 S_\lambda S_\lambda^T),$$

would lead to exactly the same KL distance $I(f_0,f) = I(f_0^S,f^S)$. This is not unexpected since a linear transformation of the type $\boldsymbol{Ay}$ with $\boldsymbol{y}$ multivariate normal and with $\boldsymbol{A}$ of rank n results again in multivariate normal data with accordingly transformed mean and covariance structure. In general, however, S_λ is singular with rank close to sample size n for so-called full-rank smoothers and with rank considerably less than n for low-rank smoothers (see e.g. Ruppert et al., 2003). Consequently the multivariate distribution of $\boldsymbol{y}^S$ is degenerate (at least one component of $\boldsymbol{y}^S$ can be written as a linear combination of the others).

Therefore we consider, as a simplification, a first basic implementation in which we ignore the dependence structure. We work with an 'independence true model' $\tilde{f}_0$

$$\boldsymbol{y}^S \overset{\tilde{f}_0}{\simeq} N_n(S_\lambda\boldsymbol{\mu}_0, \sigma_0^2 D_\lambda),$$

and its corresponding approximate model $\tilde{f}^S$

$$\boldsymbol{y}^S \overset{\tilde{f}^S}{\simeq} N_n(S_\lambda\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2 D_\lambda),$$

with $D_\lambda = \text{diag}(s_1^2,\ldots,s_n^2)$ where s_i^2 is the ith diagonal element of $S_\lambda S_\lambda^T$.

Some straightforward calculations show that $I(\tilde{f}_0,\tilde{f}^S)$ equals

$$\frac{n}{2}\log(\sigma^2/\sigma_0^2) + \frac{n}{2}\left\{\frac{\sigma_0^2}{\sigma^2}-1\right\} + \{\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta})\}^T S_\lambda^T D_\lambda^{-1} S_\lambda\{\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta})\}/(2\sigma^2).$$

First of all note that $I(\tilde{f}_0,\tilde{f}^S)\geq 0$ with equality only if $\boldsymbol{\mu}_0=\boldsymbol{\mu}(\boldsymbol{\theta})$ and $\sigma_0^2=\sigma^2$, so only if $f\equiv f_0$. Assuming that both bias terms $\parallel (S_\lambda - I_n)\boldsymbol{\mu}_0 \parallel^2$ and $\parallel (S_\lambda - I_n)\boldsymbol{\mu}(\theta) \parallel^2$ are negligible, the only difference with $I(f_0,f)$ as shown in (4), is the diagonal matrix D_λ^{-1}. Expressing distance as the square root of a positive definite quadratic form allows for a geometrical interpretation based on the eigenvalues s_i^{-2} and eigenvectors of D_λ^{-1}. The half-length from the origin $\boldsymbol{\mu}_0=\boldsymbol{\mu}(\boldsymbol{\theta})$ on the hyperellipsoid, defined by $\{\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta})\}^T D_\lambda^{-1}\{\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta})\}$, in the direction of the ith observation is equal to $s_i\sum_{j=1}^n((\mu_0(j)-\mu(\boldsymbol{x}_j;\boldsymbol{\theta}))/s_j)^2$. For instance, in the case of a single covariate and using a local linear smoother with bandwidth h_n (the λ in our notation), it asymptotically holds that $s_i^2 \sim C_i/nh$ for some constants C_i, leading to a half-length of the order $\sqrt{nh_n}$. An optimal bandwidth $h_n \sim n^{-1/5}$ (see e.g. formula (5.13) on page 152 in Simonoff, 1996) shows that the

half-lengths grow with n, showing the way in which this distance measure magnifies the difference between the true model f_0 and the approximating model f.

Ignoring the bias of the smoother, as in the discussion above, we further simplify the smoothed AIC criterion (3) by taking L_S the likelihood associated with the model

$$\boldsymbol{y}^S \overset{\tilde{f}^S}{\simeq} N_n(\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2 I_n), \tag{5}$$

resulting in

$$\mathrm{AIC}_S = n \log(\hat{\sigma}_S^2) + 2K, \tag{6}$$

where

$$\hat{\sigma}_S^2 = \sum_{i=1}^{n} (y_i^S - \mu(\boldsymbol{x}_i; \hat{\boldsymbol{\theta}}_S))^2 / n$$

with $\hat{\boldsymbol{\theta}}_S$ the ML estimator for $\boldsymbol{\theta}$ using the likelihood (5). That is, model selection proceeds by first fitting a nonparametric smoothed model to the data, and then the usual AIC measure based on the resultant fitted values is used to compare models.

In the next section we study this basic smoothed AIC criterion. It is based on a substantially simplified normal likelihood. Note however that even if the response data are not normally distributed, y_i^S, for each $i = 1, \ldots, n$, will be approximately normally distributed for most smoothers.

In Section 3 we reconsider the simplifications made above, and propose a bootstrap-based approach to overcome shortcomings related to them.

2.2 Simulations

We consider two scenarios: in scenario A the family of candidate models consists of models with linear, quadratic and interaction terms in two explanatory variables; in scenario B the family of candidate models is the set of fractional polynomials (Royston and Altman, 1994) in a regression setting.

2.3 Scenario A

In a first scenario, uniform $[0, 10]$ x-values were generated, together with (independently) Bernoulli(0.5) z-values. Given x and z, response y-values were generated from a normal distribution with mean $\mu_0(x, z)$ and variance σ_0^2, which will be specified later on. Samples $\{(x_i, z_i, y_i), i = 1, \ldots, n\}$ were generated with fixed design $\{x_i, z_i, i = 1 \ldots, n\}$. The candidate set of models consists of all submodels (hierarchical in x-effect) of

$$\mu(x, z) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 xz.$$

Four presmoothing strategies were considered based on:

(1) a penalized regression spline $y \sim s(x)$, with the level of smoothing chosen using generalized cross-validation (Eilers and Marx, 1996);

(2) a Generalized Additive Model (GAM, see Hastie and Tibshirani, 1990) with penalized splines built from $y \sim s(x) + z + zs(x)$ according to Wood and Augustin (2002): a term is dropped if i) the estimated degrees of freedom (edf) for the term is close to their lower limit (e.g., 1 for a univariate smooth and 'close to' specified as edf less than 1.6), and ii) the 95% confidence region for the smooth includes zero everywhere, and iii) the GCV score for the model goes down if the term is removed from the model; this approach is a "middle way" between the GAM framework based on backfitting (being flexible and efficient, but facing difficulties when it comes to model selection and inference) and the computationally very expensive approach of the generalized spline smoothers of Gu and Wahba (1991);

(3) a GAM model with penalized splines built from $y \sim s(x) + z + zs(x)$ using the corrected AIC-criterion as proposed by Hurvich et al. (1998), and

(4) using a GAM model $y \sim s(x) + z + zs(x)$ with penalized splines without model building.

Table 1: Scenario A, $\mu_0^1(x,z)$: the number of times each model has been selected, using the original data and presmoothed data from the four smoothing strategies.

	1	x	z	x, x^2	x, z	$x, z,$ xz	$x, x^2,$ z	$x, x^2,$ z, xz
Original Data	0	259	0	511	44	35	93	58
(1)	0	14	0	888	3	0	63	32
(2)	0	10	0	752	5	7	110	116
(3)	0	15	0	663	8	16	155	143
(4)	0	5	0	47	7	26	99	816

Note that since all of these methods include only a smooth curve for x, $s(x)$, none of the model building techniques associated with them address the question of interest here; that is, whether a linear or quadratic term in x is necessary.

When evaluating the properties of model selection criteria, an important distinction is between the situation where the 'true' model is among the candidate models, and the situation where it is not (McQuarrie and Tsai, 1998). AIC is known to be an inconsistent model selection criterion (a consistent model selection criterion is one that, with probability approaching 1 as the sample size increases, chooses the 'true' model, when it is among the candidate models). It is, however, asymptotically efficient (an efficient model selection criterion is one that chooses the model with prediction error asymptotically indistinguishable from that of the best model among all candidate models, when the 'true' model is not among the candidate models), which is in most cases to be preferred (since in practice the true model is unknown and typically too complex to be part of the candidate set of models, and a predictive criterion is natural in the regression context). Theoretically, however, it is worthwhile to consider both situations where the generating model is part of the set of candidate models and where it is not. Therefore, we consider two different mean structures for the normal distribution from which response y-values are generated, $\mu_0^1(x,z) = -3 + 3x + 5x^2$ and $\mu_0^2(x,z) = -3 - 3\log(x+1) + 5x^2$, while we take $\sigma_0 = \exp(5)$.

For $\mu_0^1(x,z)$, Table 1 shows the selection results for 1000 simulated samples of size $n = 50$. Strategy (1) outperforms all others. The true model with x- and x^2-effects is chosen much more often than based on the original data. Strategy (1) assumes no z-effect (as in the generating model), an assumption that makes model building unnecessary, but which has to be checked in real data analyses. But strategies (2) and (3) also lead to an increased selection of the true model. Model selection after smoothing without model building (strategy 4), however, results in the selection of models that are far too complex. This already illustrates a crucial point. Replacing the original responses with presmoothed data, without any careful consideration or model building, might falsely turn very small effects as produced by the smoother into relevant effects (since noise has essentially been removed). Of course, selection of (too) complicated models is not disadvantageous in and of itself. Therefore, we also compared the different strategies by looking at the mean averaged squared error (MASE), again based on 1000 simulated samples of size $n = 50$,

$$\text{MASE} = \frac{1}{1000}\sum_{r=1}^{1000}\left\{\frac{1}{n}\sum_{i=1}^{n}(\hat{\mu}^{(r)}(x_i,z_i) - \mu_0(x_i,z_i))^2\right\}.$$

Here, $\hat{\mu}^{(r)}(x_i,z_i)$ denotes the fitted value within simulation run r and $\mu_0(x_i,z_i)$ the true generating model. In Table 2, MASE-values together with bias and variance decomposition confirm the performance of the different methods. There is a large decrease in bias and applying methods (1), (2) and (3) reduces the variability, while using method (4) results in an increased variability compared to model selection based on the original data. Different simulation settings (different sample sizes n, different values of σ_0) show similar results: strategies (1) to (4) always keep their relative ordering from best (strategy 1) to worst (strategy 4), and the classical AIC on the original data is often worse than strategies (1) to (3) but sometimes close and even better than some of the strategies (1) to (3) (especially for n large). This is further illustrated in the next setting.

Table 2: Scenario A, $\mu_0^1(x,z)$: MASE and bias and variance decomposition, using the original data and presmoothed data from the four smoothing strategies.

	MASE	bias^2	var
Original Data	1942.41	27.63	1914.79
(1)	1295.03	2.41	1292.62
(2)	1573.91	2.69	1571.22
(3)	1646.58	2.69	1643.90
(4)	2108.53	2.68	2105.85
Correct	1230.81	2.38	1228.44

In a second setting, we took $\mu_0^2(x,z)$, i.e. the generating model is not included in the set of candidate models. Figure 1 shows MASE results as a function of the sample size $n \in \{50, 100, 150, 200\}$, and for $\sigma_0 \in \{\exp(5), \exp(6), \exp(7)\}$. This figure shows that strategy (1) (indicated as gam(x) in the figure) is performing very well, especially for σ_0 small or n large. It gets close to the MASE of the (estimated) true model $\mu_0^2(x,z)$. Model selection based on presmoothed data according to strategy (2) and (3) is better than based on the original data, except when the variance is very large. If the sample size gets large, all methods seem to converge. Strategy (4) is no longer included.

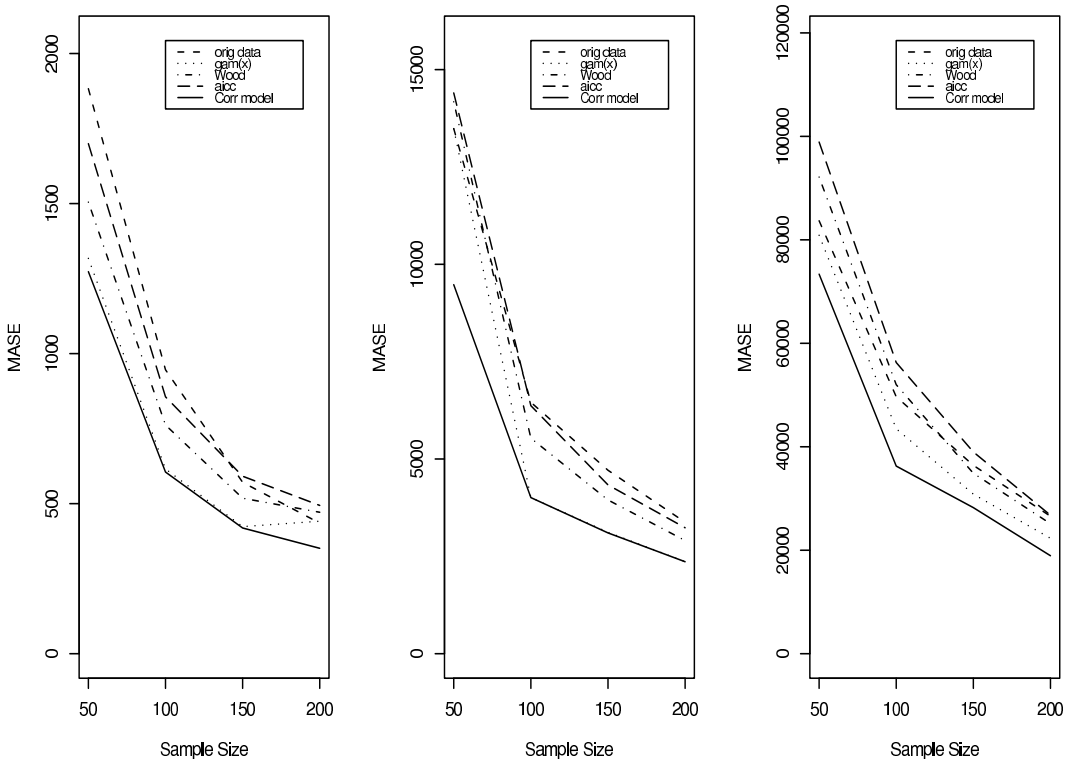

Figure 1: Scenario A, $\mu_0^2(x,z)$: MASE-values based on model selection using AIC for the original data, the GAM(x)-presmoothed data (strategy 1), the Wood-presmoothed data (strategy 2), the AIC$_c$-presmoothed data (strategy 3) and the correct model $\mu_0^2(x,z)$. The horizontal axis of each plot indexes sample size n, while σ_0 increases from exp(5) (left) to exp(6) (middle) to exp(7) (right panel).

Table 3: Scenario B, $\mu_0^1(x,z)$: selected powers within the family of fractional polynomials, using the AIC-criterion (left) and AIC$_S$-criterion (right).

AIC						AIC$_S$				
$p_1 \backslash p_2$	0.0	0.5	1.0	2.0	3.0	$p_1 \backslash p_2$	0.5	1.0	2.0	3.0
-2.0	1	4	28	134	87	-2.0			7	7
-1.0	2	3	10	29	36	-1.0		4	9	8
-0.5	2	13	14	11	39	-0.5	2	1	16	19
0.0		11	14	10	36	0.0	7	11	28	24
0.5		29	42	15	31	0.5	11	26	68	43
1.0			49	*16*	31	1.0		48	*297*	159
2.0				12	8	2.0			40	45
3.0					118	3.0				120

2.4 Scenario B

In a second scenario we generate data according to the same setting as Scenario A with $\mu_0^1(x,z)$, but now the family of candidate models is the family of fractional polynomials of degree 1 and 2 within the recommended grid $\{-2,-1,-0.5,0,0.5,1,2,3\}$ (Royston and Altman, 1994) . Note that the true model is a fractional polynomial of degree 2 with $p_1 = 1$ and $p_2 = 2$. Since the candidate models here include no z-effect, only presmoothing strategy (1) was applied.

Table 3 shows an overview of the powers chosen by AIC and AIC$_S$. Powers not chosen by any of the selection criteria were omitted from the table. All fractional polynomials selected using the AIC-and AIC$_S$-criterion are of degree 2. It can be seen that the generating model, which is contained in the set of the candidate models $(p_1,p_2) = (1,2)$, was chosen 16 times by the AIC-criterion while it was chosen 297 times using AIC$_S$. Moreover there is a clear concentration around the true combination $(p_1,p_2) = (1,2)$ for AIC$_S$, whereas AIC leads to a larger spread over different and typically lower powers. Many other simulations with different true functions were performed, but they essentially showed the same results.

We also compared the MASE of the best models selected by AIC$_S$ with the ones selected by AIC, by looking at the ratio MASE(AIC)/MASE(AIC$_S$). In about 85%, these ratios are larger than 1, indicating an improved model choice when using the AIC$_S$-criterion.

Using $\mu_0^2(x,z)$ as generating model, again a large majority (84%) of the ratios is larger than 1. The smallest value was 0.58, the largest 204.33, again indicating that the models selected by presmoothing outperform the ones selected by original data.

The main conclusions of these simulations on the basic presmoothing method, under Scenario A and B, can be summarized as follows. The method clearly shows some potential to improve model selection in a regression setting. This improvement is most apparent in the case where there are only one or two explanatory variables and the family of candidate models is a very 'rich family' (like the family of fractional polynomials). In the case where there are many explanatory variables and the family of candidate models does not contain enough models to allow flexible curvatures, the preliminary smoothing has to be guided carefully, and some model building is necessary. We do not recommend a blind use of the basic method proposed in this section, but suggest to use it always in comparison with classical AIC (applied on the original data). The new approach can reassure the selection based on classical AIC. In case they point at different models, one can examine in more detail what causes this disagreement. Whether they agree or disagree, in both case we recommend the use of the modified approach presented in the next section.

Finally, it is obvious that by focusing on the mean structure, the method cannot be used to compare models with e.g. the same mean structure but different variance structures.

3 Presmoothing and Bootstrapping

The basic smoothed AIC approach (6) treats the smoothed data as independent, homoscedastic normal data, hereby ignoring that i) S_λ is singular, ii) $y_i^S, i = 1,\ldots,n$, are not independent, iii) $y_i^S, i = 1,\ldots,n$, do not have the same variance and iv) a smoother is biased. In the following section, we propose to generate new

bootstrap samples, conditional on the original sample, and to apply AIC on these bootstrap samples to overcome these limitations. We first restrict attention to the case of continuous response. Section 4 treats the case of categorical outcomes.

3.1 A Bootstrap Approach

Consider the additive location model

$$\boldsymbol{y} = g(\boldsymbol{\theta}; \boldsymbol{x}) + \boldsymbol{\varepsilon}, \tag{7}$$

and the residuals resulting from an estimated smooth fit (with smoother matrix S_λ)

$$\boldsymbol{e} = \boldsymbol{y} - S_\lambda \boldsymbol{y}.$$

Define new observations, centered at the smooth fit, together with "controlled" error, using a constant $0 \leq c \leq 1$,

$$\tilde{\boldsymbol{y}}_S(c) = S_\lambda \boldsymbol{y} + c\, \boldsymbol{e}. \tag{8}$$

Equation (8) mimics the location model (7) and allows one to control the signal-to-noise ratio by the constant c. Both terms, the systematic component $S_\lambda \boldsymbol{y}$ and the error component $c\, \boldsymbol{e}$ have distributional properties driven by that of the original sample. The bootstrap allows us to disconnect this relation, thus avoiding most of the drawbacks of the basic smoothed AIC approach. If the original error $\boldsymbol{\varepsilon}$ has an i.i.d. structure, a nonparametric bootstrap approach can be defined as

$$\boldsymbol{y}^*_S(c) = S_\lambda \boldsymbol{y} + c\, \boldsymbol{e}^*, \tag{9}$$

using resampled residuals $\boldsymbol{e}^* = (e^*_1, \ldots, e^*_n)^T$, taking randomly with replacement from the set $\{e_1, \ldots, e_n\}$.

Once bootstrap data $\boldsymbol{y}^*_S(c)$ are generated, model selection can be based on the AIC-criterion (mimicking (2))

$$\text{AIC}^* = -2L(\boldsymbol{y}^*_S(c); \hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\eta}}^*) + 2K,$$

with $(\hat{\boldsymbol{\theta}}^*, \hat{\boldsymbol{\eta}}^*)$ the MLE's of $(\boldsymbol{\theta}, \boldsymbol{\eta})$ based on $\boldsymbol{y}^*_S(c)$ and using the log-likelihood L of the original data $\boldsymbol{y}$.

Conditional on the original sample these new smoothed and bootstrapped data $\boldsymbol{y}^*_S(c)$ reflect approximately i) the right location, by consistency of $S_\lambda \boldsymbol{y}$ (see e.g. Ruppert et al., 2003), and ii) the right i.i.d. error structure, by consistency of the nonparametric bootstrap (see e.g. Efron and Tibshirani, 1998). By taking $0 < c < 1$, one can control the level of error and consequently, to some extent, the quality of the data in order to select the most appropriate model.

Figure 2 illustrates this intuitively appealing idea. The black solid line is the true generating function (a fitted spline model). The dashed line is the corresponding estimated spline model based on the original data, shown in the left upper panel. The right upper panel and the lower left and right panels show bootstrap data generated according to (9) with c taking decreasing values 1, 0.5 and 0.25. The idea is that the data shown in the lower panels, according to smaller values of c, more clearly show the true underlying mean function.

Alternatively other bootstrap based error structures can be used, e.g. based on the parametric bootstrap, e.g. for normal i.i.d. errors

$$\boldsymbol{e}^* \sim N_n(0, \hat{\sigma}^2_S I_n),$$

where

$$\hat{\sigma}^2_S = \frac{1}{n} \boldsymbol{y}^T (I - S_\lambda)^T (I - S_\lambda) \boldsymbol{y},$$

the variance estimator based on the smoother. The parametric bootstrap also easily allows more complicated (e.g. correlated) error structures.

An interesting question is how to determine an optimal value for the parameter c in (9). In the next section we try to get more insight in the role of this parameter c.

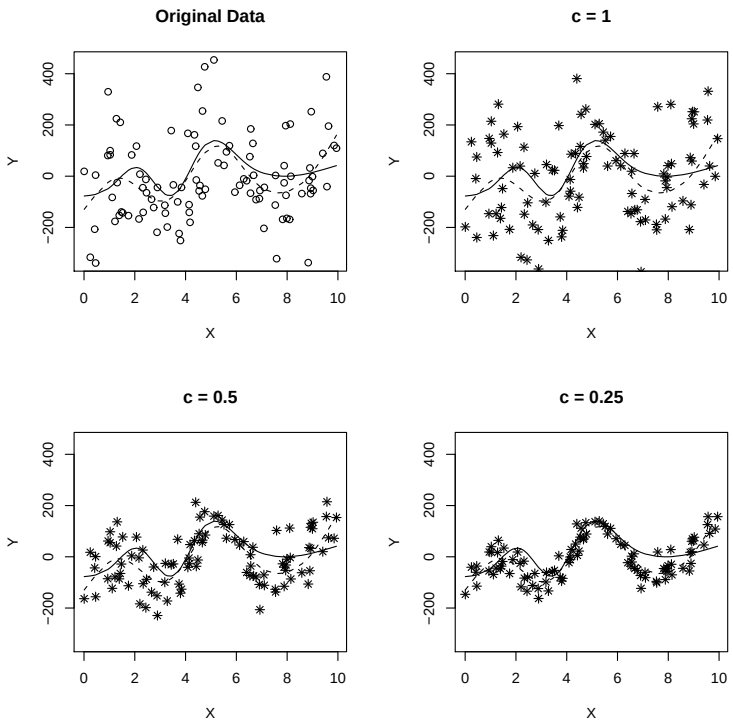

Figure 2: Simulated data according to bootstrap model (9): original data (open circles, left upper panel) and smoothed data with $c \in \{1, 0.5, 0.25\}$ (stars, right upper and left and right lower panels). The solid curve is the true regression function; the dashed curve is the spline fit to the original data.

3.2 Choice of error-control parameter c

For simplicity, we again focus on classical regression (with normal error structure). Reconsider identity (4), and assume a 'true' model $\boldsymbol{\mu}_0 = X_0\boldsymbol{\theta}_0$ for some design matrix X_0 and an approximating model $\boldsymbol{\mu}(\boldsymbol{\theta}) = X\boldsymbol{\theta}$ for some typically different design matrix X. Consider the true values $\boldsymbol{\theta}_0$ and σ_0^2 as fixed. Minimizing the right-hand side of (4) as a function of $\boldsymbol{\theta}$ and σ^2 leads to the minimum

$$\min_f I(f, f_0) = \frac{n}{2}\log\left(1 + \frac{(X_0\boldsymbol{\theta}_0)^T(I-H)X_0\boldsymbol{\theta}_0}{n\sigma_0^2}\right),$$

attained at

$$\boldsymbol{\theta} = (X^TX)^{-1}X^TX_0\boldsymbol{\theta}_0,$$

and

$$\sigma^2 = \sigma_0^2 + \frac{(X_0\boldsymbol{\theta}_0)^T(I-H)X_0\boldsymbol{\theta}_0}{n},$$

where $H = X(X^TX)^{-1}X^T$ is the well-known hat matrix associated with matrix X. Note that if $X = X_0$ (so the approximating model equals the true model), then $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ and $(I-H)X_0$ reduces to the zero-matrix and $\sigma^2 = \sigma_0^2$ and $\min_f I(f, f_0) = 0$, as expected. Note also that if the approximating model is not equal (or does not contain) the true model, the scale on which the distance $\min_f I(f, f_0)$ is measured depends on the value of σ_0^2. If $\sigma_0^2 \to \infty$, then $\min_f I(f, f_0) \to 0$, even if f is a wrong model. In other words, using the AIC estimate for the KL distance, a large value of σ_0^2 will make it more difficult to detect differences between models. Small values of σ_0^2 however magnify the distance scale and will allow AIC, for a fixed sample size n, to more easily select a good model. This confirms the intuition that for data with little noise, better suited models can be selected, as compared to high noise data. With respect to the choice of c in (9), this seems to suggest taking c small.

Conditional on $\boldsymbol{y}$, we can mimic these considerations, but now based on a 'true' generating (bootstrap) model

$$\boldsymbol{y}_S^*(c)|\boldsymbol{y} \overset{f_0^*}{\sim} N_n(S_\lambda\boldsymbol{y}, c^2\hat{\sigma}_S^2 I_n),$$

and an approximating model (the same as on the original data)

$$\boldsymbol{y}_S^*(c)|\boldsymbol{y} \overset{f^*}{\sim} N_n(X\boldsymbol{\theta}^*, \sigma^{*2} I_n).$$

Similar calculations show that the KL distance (conditional on $\boldsymbol{y}$) attains a minimum

$$\min_{f^*} I(f^*, f_0^*) = \frac{n}{2}\log\left(1 + \frac{(S_\lambda\boldsymbol{y})^T(I-H)S_\lambda\boldsymbol{y}}{nc^2\hat{\sigma}_S^2}\right),$$

attained at

$$\boldsymbol{\theta}^* = (X^TX)^{-1}X^TS_\lambda\boldsymbol{y},$$

and

$$\sigma^{*2} = c^2\hat{\sigma}_S^2 + \frac{(S_\lambda\boldsymbol{y})^T(I-H)S_\lambda\boldsymbol{y}}{n},$$

with H the same hat matrix as before.

Of course, here $\min_{f^*} I(f^*, f_0^*)$ depends on $\boldsymbol{y}$ and hence is a random variable. A crucial point is that now $\min_{f^*} I(f^*, f_0^*) > 0$, even if one uses the true model for f^* (i.e. $X = X_0$). Indeed since a smoother has finite sample bias, it holds that $S_\lambda\boldsymbol{y} \neq X_0\boldsymbol{\theta}_0$ with probability one, and consequently $(S_\lambda\boldsymbol{y})^T(I-H)S_\lambda\boldsymbol{y} \neq 0$ with probability one.

Comparing $\min_{f^*} I(f^*, f_0^*)$ and $\min_f I(f, f_0)$, we see that this bias, the difference between $S_\lambda\boldsymbol{y}$ and $X_0\boldsymbol{\theta}_0$, plays a major role. Compared to a (relatively) simple true model (such as $X_0\boldsymbol{\theta}_0$ with $\boldsymbol{\theta}_0$ of limited dimension), the smooth fit $S_\lambda\boldsymbol{y}$ may expose more local and small curvatures and complexities. Since the bootstrap approach (9) treats model f_0^* with $S_\lambda\boldsymbol{y}$ as true model, small values of c would, as discussed above, magnify the distance-scale to an extent that these little and local complexities are getting relevant, when using the estimate AIC^* of $\min_{f^*} I(f^*, f_0^*)$. This latter consideration indicates one should not take c too small, in order to downplay the bias of the smoother. Moreover it indicates that ideally, c depends on the

Table 4: Spline model $\mu_0(x,z)$ as true model: the number of times each model has been selected, using the original data and presmoothed data with controlled level of bootstrap error.

c	1	x	z	x,z	x,x^2	x,z xz	x,x^2 z	x,x^2 x^3	x,x^2 z,xz	x,x^2 x^3,z	x,x^2 z,xz x^2z	x,x^2 x^3,z xz	x,x^2 x^3,z xz,x^2z
Original Sample													
	23	14	13	3	28	1	1	8	1	1	5	0	2
Smoothed Sample + $c\times$ Parametric Bootstrap Errors													
0	1	1	0	0	11	0	1	49	0	4	4	5	24
10^{-6}	1	5	0	1	13	1	1	45	1	4	3	4	21
0.001	1	5	0	1	14	1	1	46	1	4	3	4	19
0.01	1	5	0	1	16	1	1	50	2	7	3	2	11
0.1	1	6	0	1	23	1	3	45	3	8	2	1	6
0.2	1	6	0	1	27	1	5	39	3	8	2	1	6
0.5	5	9	2	3	30	3	3	29	6	2	3	1	4
0.8	13	11	2	2	31	6	3	18	6	2	3	0	3
0.9	17	11	1	2	30	6	2	16	6	2	4	0	3
1	20	13	2	1	27	7	2	13	6	2	4	0	3

original sample $\boldsymbol{y}$, and it should reflect the bias $S_\lambda \boldsymbol{y} - X_0\boldsymbol{\theta}_0$. The smaller the bias, the smaller c should be taken. Since the bias is unknown and hard to estimate, this indicates that finding the optimal c is not straightforward. One option is to conduct a kind of sensitivity analysis, by showing how the selection of models change with the value of c ranging on a grid from 0 to 1. In case the sample size allows, one could split the sample in two subsamples: a learning sample to select and build the models and a test sample to select the optimal value of c, minimizing the prediction error. This method will be illustrated in Section 4.2.

The simulations in the next section illustrate how the smoothed data with bootstrap error lead to an improved selection of models and give some further insights in the optimal choice of c.

3.3 Simulations

We generated 100 samples of size $n = 100$. As in Section 2.3, uniform[0,10] x-values were generated, together with (independently) Bernoulli(0.5) z-values. Given x and z, response y-values were generated from the normal distribution with standard deviation $\sigma_0 = \exp(5)$ and with true mean function $\mu_0(x,z)$ equal to a particular nonlinear function, namely the solid line in Figure 2 corresponding to a fitted spline model. Uniform[0,10] x-values were generated, together with (independently) Bernoulli(0.5) z-values. So, the left upper panel of Figure 2 shows a typical data set, and the other panels show smoothed data for three different values of c, as they are used in the simulations.

As candidate models we consider a family of 13 models, all submodels of the model with terms x, x^2, x^3, z, xz, x^2z. Given the true model is highly nonlinear without z-effect, the cubic model with x, x^2 and x^3 terms can be considered as the best model. Table 4 shows how often each of the different models have been chosen, based on the original data (top line) and based on smoothed data according to model (9) using presmoothing strategy (3) and with the parametric bootstrap error (3.1), for different values of c. Model selection based on the original data seems to lead often to too simple models (the constant model, the model with only a linear x effect or a spurious z effect). The same happens for the smoothed AIC_S with c close to 1. This is expected because the same level of error as in the original data is added to the mean structure. Choices of c close to 0, coinciding with the basic implementation discussed in Section 2, leads to the selection of overly complicated models with spurious z effects (as already noticed in Section 2). When c increases, there is a clear shift to simpler models, with overall dominance of the best cubic model. So, based on Table 4, model selection is optimal and stable for c-values within the range $[0.1, 0.2]$. Similar results were found for the nonparametric approach (9).

Figure 3 shows results of squared bias, variance and MASE using the smoothed AIC_S, relative to the

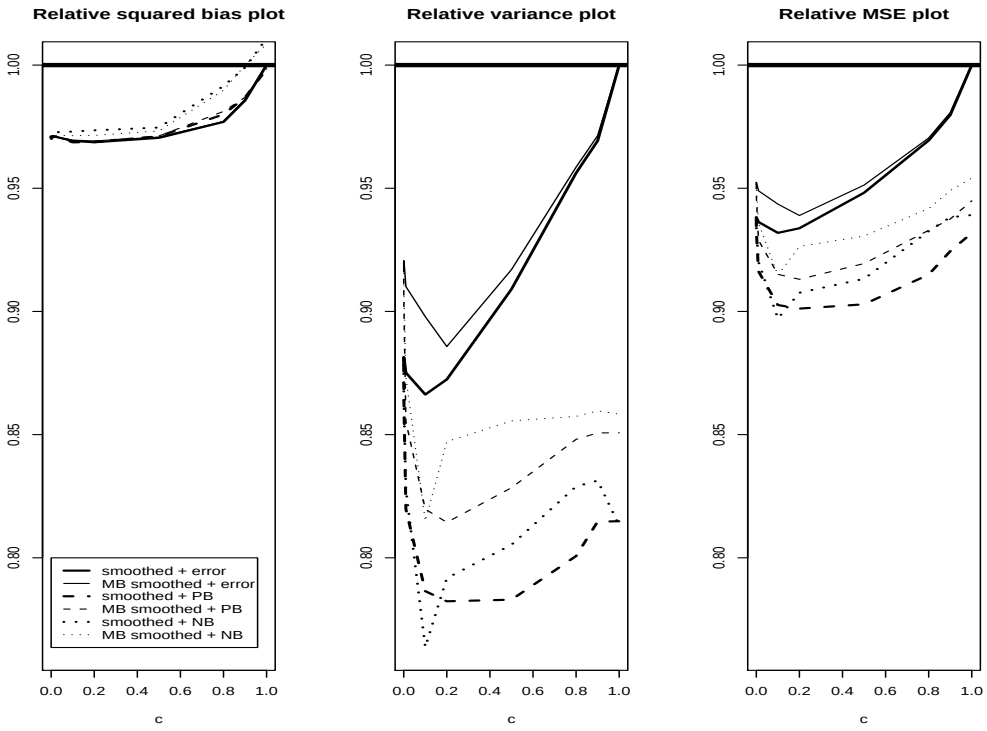

Figure 3: Spline model $\mu_0(x, z)$ as true model: relative squared bias, variance and MASE-values based on model selection using presmoothed data according to strategy 1 (bold lines) and strategy 3 (normal lines) and using fixed error (solid lines) and bootstrap error, nonparametrically (dotted lines) and parametrically (dashed lines). The horizontal axes indicate the level c of error structure.

corresponding values when using AIC based on the original data. Whereas Table 4 only shows results for strategy (3) as discussed in Section 2 and the parametric bootstrap, Figure 3 also includes results for strategy (1) (smoothed), strategy (3) (MB smoothed), for the implementation with fixed residuals (8) (indicated by +error) and the nonparametric and parametric bootstrap implementation (indicated by +NB and +PB respectively). The figures show that especially the variance is reduced by using AIC_S. Since the normal distribution is the right one, there is not much difference between the nonparametric and parametric bootstrap approach. Using the correct univariate smoother leads to better results than when strategy (3) is used. According to the short dashed curve (the setting of Table 4 in the right panel: relative MASE), the choice $c = 0.2$ leads to the best results.

4 Latent Variable Approach for Categorical Data

In this section we discuss the implementation of presmoothing and bootstrapping in the case of categorical outcomes. Of course, in this situation the additive structure of mean plus error structure no longer holds. The approach of categorical outcomes generated by a latent continuous outcome, however, allows us to implement a similar idea as in the previous section.

4.1 Smoothed Latent Variable

Consider a categorical variable y with ordered categories $1, \ldots, J$, and a latent continuous variable

$$y_{L} - \mu(x) \sim G,$$

such that

$$y = j \quad \text{if} \quad \alpha_{j-1} < y_{L} \leq \alpha_j,$$

for cutpoints

$$-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_J = \infty.$$

Then the cumulative distribution of y can be written in terms of the latent distribution G as

$$P(y \leq j|x) = P(y_{L} \leq \alpha_j|x) = G(\alpha_j - \mu(x)).$$

Different choices of the latent distribution G correspond to different generalized logit models. Taking $G = \Phi$ (the standard normal distribution function) leads to the cumulative probit model, taking G the logistic distribution to the proportional odds model, and taking $G(z) = 1 - \exp(-\exp(z))$ coincides with the proportional hazards model. For more details, see Section 7.2 in Agresti (2002) or Section 10.2 in Simonoff (2003).

The analogue of the presmoothing and bootstrap approach (9) using the latent distribution is as follows. Fit the model $G^{-1}(P(y \leq j|x)) = \alpha_j - \mu(x)$ using a smoother (e.g. spline), leading to $\hat{\mu}_S(x)$. Next, generate bootstrap values of the latent variable, using the parametric bootstrap

$$y^*_{L,i}(c) = \hat{\mu}_S(x_i) + ce^*_i, \tag{10}$$

with

$$e^*_i \sim G, \tag{11}$$

and, as before, some value of $0 < c < 1$. Finally, select the parametric model $\mu(x, \theta)$ using AIC on these smoothed and bootstrapped data $y^*_{L,i}(c)$ and taking density g (corresponding to G) in the construction of the log likelihood. Use as a final model

$$G^{-1}(P(y \leq j|x)) = \alpha_j - \mu(x, \theta). \tag{12}$$

Once again a sensitivity analysis can show the dependence on c, and in case a test sample is available, this can be used to select an optimal value of c (minimizing the prediction error). In the next section this approach is illustrated to select the best fractional polynomial to fit Hepatitis A seroprevalence data.

4.2 Data Example: Hepatitis A Seroprevalence

Hepatitis A virus (HAV) is mainly (> 95 %) transmitted by the feco-oral route (e.g. through food and water polluted by faeces containing the virus). Transmission is facilitated by poor hygienic living and housing conditions, and is particularly common in developing countries (see e.g. Hadler, 1991; Beutels et al., 1997). In these countries HAV is mainly a childhood infection, whereas in industrial countries HAV infection occurs during adulthood as well as childhood. In the poorest developing countries, the pattern of high endemicity is characterized by rapid infection at a very young age; over 90% of the children become infected by the age of 5. In 1993 and early 1994 , a study of the prevalence of HAV antibodies was conducted in the Flemish community in Belgium. The purpose of this study was to obtain data on the prevalence of hepatitis A in Flanders and to analyze the epidemiological pattern of HAV. During the study period serum samples were collected from hospitals (non-infectious disease wards) in the Flemish community. The dataset contains the serological results of 3161 Belgian individuals, i.e. a binary response $y = 1$ if infected (and 0 otherwise); together with their age a in years, ranging from 0.5 to 92.5 years. The study group was similar in composition to the Flemish population in terms of age. Here we focus on Belgian males, resulting in 1646 observations, which we consider, for illustrative purposes, as our population, and a random subset of size 200 as our random sample (RS). This sample of size 200 is used to select the model by both methods, with and without presmoothing, and the remaining 1446 observations form the test sample (TS) that will be used to compute the prediction errors (number of misclassified cases) for both methods. When using AIC directly on the original data, all 200 observations of the RS will be used as such. When using AIC on the smoothed data, the sample of size 200 is randomly split in two subsamples of size 100, one of which we call the learning sample (LS), and the other the validation sample (VS). The LS is used to identify the best model by the smoothed AIC using presmoothing strategy (1), for a grid of values for the control parameter c. The VS is subsequently used to identify the optimal value of c.

Shkedy et al. (2006) propose to model the prevalence and force of infection as a function of age, within the framework of fractional polynomials. They discuss several parametric examples from the infectious diseases literature and show that all of these examples can be expressed as special cases of fractional polynomial models. Note that the choice of a parametric model facilitates an easy derivation of secondary epidemiological parameters as the age of maximal force of infection and the basic reproduction number. Here we consider as candidate models the family of fractional polynomials of order two with powers $p_1 \leq p_2$, where

$$p_i = \{\text{from -2 to 3 in steps of 0.1}\}, \quad i = 1, 2,$$

together with the probit link function.

Using AIC on the original data from the RS leads to the optimal powers $p_1 = -1.3$, $p_2 = -1.3$, which identifies the probit model

$$\Phi^{-1}\{P(y = 1)\} = \theta_0 + \theta_1 a^{-1.3} + \theta_2 a^{-1.3} \log(a). \tag{13}$$

This model leads to 26 misclassifications on the LS, 16 on the VS, resulting in a total of 42 on the RS, and leads to 351 misclassifications on the TS.

The upper panels of Figure 4 show the powers selected by the approach based on (10)-(12) with $G = \Phi$, as a function of c (using the learning sample). As c increases, the both powers gradually increase from 1.5 to their maximal value of 3. The left middle panel shows the number of misclassifications using the validation sample, as a function of c. It suggests to take a value c in the neighborhood of 0.5. We took $c = 0.5$, which corresponds to powers 2.4 and 2.4 and probit model

$$\Phi^{-1}\{P(y = 1)\} = \theta_0 + \theta_1 a^{2.4} + \theta_2 a^{2.4} \log(a). \tag{14}$$

The number of misclassifications for this model is equal to 26 on the LS and to 14 on the VS and to 322 on the TS, which is almost 10% less than the model based on the classical AIC. The right middle panel shows the smoothed and bootstrapped latent observations $y^*_{L,i} = \hat{\mu}_S(x_i) + ce^*_i$ (equation (10)) for $c = 0.5$. The fits of both final models (13) and (14) on the LS, together with the (jittered) data, are shown in the left lower panel of Figure 4. The fits are quite different. The model selected by using AIC on the classical data of the RS (dashed line) has an unexpected rise for very small ages. The right lower panel shows the fits of both models on the TS, together with the (jittered) data. Again there is a substantial difference between both

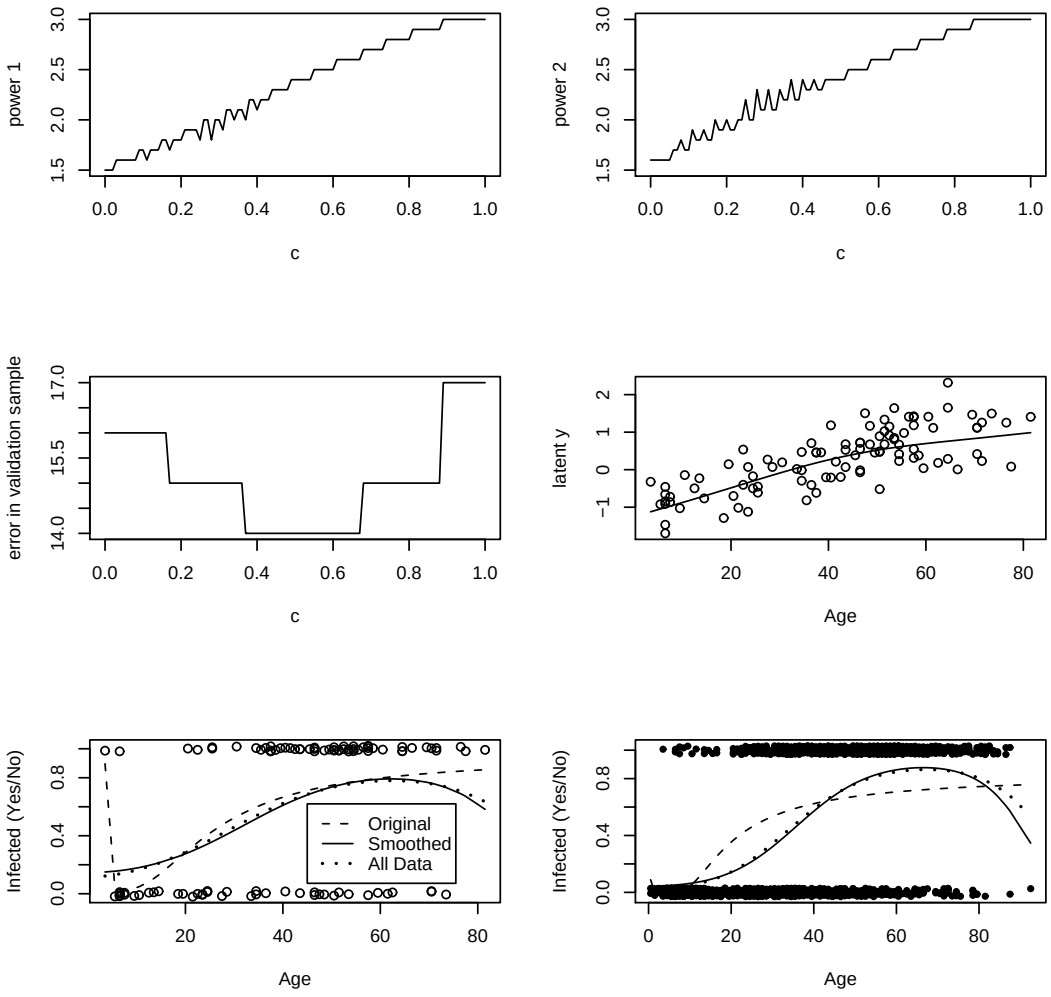

Figure 4: The Hepatitis A Example. Upper panels: the first (left) and second (right) powers of the optimal fractional polynomial, based on the smoothed and bootstrapped data, as a function of c. Left middle panel: the prediction error using the smoothed AIC on the VS. Right middle panel: the latent smoothed and bootstrapped observations which were used to select the optimal model. Lower left panel: fitted optimal fractional polynomials together with the (jittered) data of the LS. Lower right panel: fitted optimal fractional polynomials together with the (jittered) data of the TS.

models. As a gold standard we also selected the best fractional polynomial on the full population (all 1646 data; the unsmoothed and presmoothed choices are virtually indistinguishable on the full data set). The selected powers were 2.0 and 2.1. The fits of this population model on the LS and the TS are shown in bold dotted lines in the left and right lower panel respectively. These fits are remarkably close to the fits based of the fractional polynomial selected by the smoothed and bootstrapped data from the (small) LS and VS. These particular fits show a well-known pattern, reaching a maximum and then slightly decreasing again for the older age group. Under conditions of so-called stationarity, cross-sectional (sero)prevalence data can be interpreted as if they were longitudinal. One would then expect the prevalence to be monotone as a function of age (Anderson and May, 1991). At older ages however, it is quite possible that the seroprevalence declines, due to age-related decline in antibody levels. There is a lot of interest by infectious diseases epidemiologists to see unconstrained and flexible fits for the seroprevalence. The selection of optimal unconstrained models for the seroprevalence as a function of age is therefore important to get more detailed insights in this phenomenon, and presmoothed choice of a fractional polynomial ordered probit model provides a way to do this.

5 Discussion and Further Research

We have shown that a simple application of presmoothing yields a selection criterion AIC_S with improved behavior over the standard AIC criterion. Further, we have shown that using the AIC criterion AIC^* on smoothed and bootstrapped data $\boldsymbol{y}^*_S(c) = S_\lambda \boldsymbol{y} + c\ \boldsymbol{e}^*$ can lead to improved model selection, at least in a setting with a limited number of explanatory variables. The optimal choice of c depends on the original data $\boldsymbol{y}$ and the bias of the smoother. A data-driven method to select $c = c(\boldsymbol{y})$ is a topic of further research. At this point, we recommend using several values for c and to explore in which way different models are selected, as a sensitivity analysis, or to use a test sample if available. Also note that $\min_{f^*} I(f^*, f_0^*) \to \min_f I(f, f_0)$ as $n \to \infty$ and that one might expect c to be taken smaller as n increases. On the other hand, for large n, the classical AIC criterion, used on the original data $\boldsymbol{y}$, may perform so well that it might not be worthwhile to apply the proposed method. A more detailed study on when there is substantial benefit to use the method (in terms of type of outcome data, sample size, number of explanatory variables, etc) is planned as further research.

For higher dimensional problems one could opt to use a multi-dimensional smoother. However this could lead to a loss of effectiveness because of the curse of dimensionality (see e.g. pp. 83-84 in Hastie and Tibshirani, 1990). If the best approximating model does not involve too much in the way of interactions, a GAM could prove to be more effective than a multi-dimensional smoother (and more effective than unsmoothed AIC). The GAM should contain at least those additive terms that correspond to parametric terms included in the most complex parametric model.

In this paper, we focused on uncorrelated data, but investigating its application to data with autocorrelated errors could be worthwhile too. In that case the smoothing parameter method should take autocorrelation into account (Opsomer et al., 2001).

Our method could be worthwhile to compare with simultaneous variable selection and outlier identification in linear regression (Hoeting et al., 1996; Kim et al., 2008). In that case, a robust smoother, for example based on least absolute values rather than least squares, would probably perform better, but this extension is beyond the scope of this paper.

Finally we like to mention that presmoothing can also be applied to other criteria, such as Mallow's C_p, BIC, etc. Some initial analyses and simulations show similar results and conclusions. A deeper study of the performance for other criteria and for different settings, such as multivariate and longitudinal settings, are also topics of further research.

Acknowledgments

We thank Pierre Van Damme and Philippe Beutels, Center for the Evaluation of Vaccination, Faculty of Medicine, University of Antwerp, for making the Hepatitis A seroprevalence data available to us.

We also gratefully acknowledge support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy). The research of Niel Hens has been financially supported by the Fund of Scientific Research (FWO, Research Grant # G039304) of Flanders, Belgium.

References

Agresti, A. (2002). *Categorical Data Analysis*. New York: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In: Csaki F. Petrov, B.N., editor, *In 2nd International Symposium on Information Theory*, Budapest, Akademia Kiado, 267–281.

Anderson, R.M. and May, R.M. (1991) *Infectious Diseases of Humans, Dynamic and Control*. New York: Oxford University Press Inc.

Beutels, M., Van Damme, P., Aelvoet, W., Desmyter, J., Dondeyne, F., Goilav, C., Mak, R., Muylle, L., Pierard, D., Stroobant, A., Van Loock, F., Waumans, P. and Vranckx, R. (1997). Prevalence of Hepatitis A, B and C in the Flemish Population. *Eur. J. Epidem.* , 13, 275–280 .

Burnham, K.P. and Anderson, D.R. (2002). *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.* New York: Springer-Verlag.

Christóbal Christóbal, J.A., Faraldo Roca, P., and González Manteiga, W. (1987). A class of linear regression parameter estimators constructed by nonparametric estimation. *Ann. Statist.*, 15, 603–609.

Efron, B. and Tibshirani, R.J. (1998). *An Introduction to the Bootstrap.* Chapman & Hall, CRC.

Eilers, P.H.C. and Marx, B.D. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statistical Science*, 11, 89–121.

Faraldo Roca, P., and González Manteiga, W. (1987). On efficiency of a new class of linear regression estimates obtained by preliminary nonparametric estimation. *In: New Perspectives in Theoretical and Applied Statistics*, New York: Wiley, 229–242.

Gu, C. and Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. . *SIAM J. Sci. Stat. Comp.*, 12, 383–398.

Hadler, S.C. (1991). Global impact of hepatitis A virus infection: changing patterns. In *Viral hepatitis and Liver Disease* (eds F.B. Hollinger, S.M. Lemon, H.S. Margolis), pp. 14–20. Baltimore: Williams & Wilkins.

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

Hens, N., Aerts, M., and Molenberghs, G. (2006). Model selection for incomplete and design-based samples. *Statistics in Medicine*, 25, 2502-2520.

Hoeting, J., Raftery, A.E. and Madigan, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis*, 22, 252–270.

Hurvich, C.M, Simonoff, J.S., and Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *JRSS-B*, 60, 271–293.

Janssen, P., Swanepoel, J. and Veraverbeke, N. (2001). Efficiency of linear regression estimators based on presmoothing. *Communications in Statistics A*, 30, 2079–2097.

Kim, S., Park, S.H. and Krzanowski, W.J. (2008). Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35, 283–291.

McQuarrie, A.D.R. and Tsai, C.-L. (1998). *Regression and Time Series Model Selection*. Singapore: World Scientific.

Opsomer, J., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science*, 16, 134–153.

Pan, W. (2001). Akaike's information criterion in generalized estimating equations. *Biometrics*, 57, 120–125.

Pan, W. (2001). Model selection in estimating equations. *Biometrics*, 57, 529–534.

Royston, P. and Altman, D. (1994) Regression using fractional polynomials of continuous covariates: Parsimonious parametric modelling. *Applied Statistics*, 43, 429–467.

Ruppert, D., Wand, M.P., and Carroll, R.J. (2003). *Semiparametric Regression*. Cambridge: Cambridge University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461–464.

Shkedy, Z., Aerts, M., Molenberghs, M., Beutels, Ph. and Van Damme, P. (2006). Modeling Age Dependent Force of Infection From Prevalence Data Using Fractional Polynomials. *Statistics in Medicine*, 25, 1577–1591.

Simonoff, J.S. (1996). *Smoothing Methods in Statistics*. New York: Springer-Verlag.

Simonoff, J.S. (2003). *Analyzing Categorical Data*. New York: Springer-Verlag.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *JRSS-B*, 64, 583–639.

Takeuchi, K. (1976). Discussion of informational statistics and a criterion for model fitting. *Suri-Kagaku*, 153, 12–18.

Wood, S.N., and Augustin, N.H. (2002). GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecological Modelling*, 157, 157-177.

Ye, J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93, 120-131.