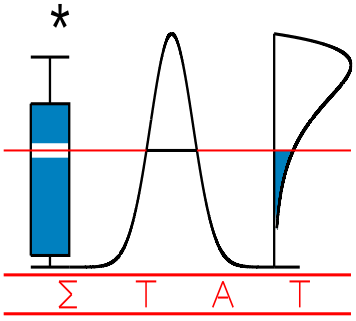
T E C H N I C A L

R E P O R T

0466

MODEL SELECTION FOR INCOMPLETE AND DESIGN-BASED SAMPLES

HENS, N., AERTS, M. and G. MOLENBERGHS



I A P S T A T I S T I C S

N E T W O R K

INTERUNIVERSITY ATTRACTION POLE

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

Model Selection for Incomplete and Design-Based Samples

Niel Hens, Marc Aerts, Geert Molenberghs

Center for Statistics, Limburgs Universitair Centrum, Diepenbeek, Belgium.

Summary. The Akaike Information Criterion, AIC, is one of the most frequently used methods to select one or a few good, *optimal* regression models from a set of candidate models. In case the sample is incomplete, the naive use of this criterion on the so-called complete cases can lead to the selection of poor or inappropriate models. A similar problem occurs when a sample based on a design with unequal selection probabilities, is treated as a simple random sample. In this paper we consider a modification of AIC, based on reweighing the sample in analogy with the weighted Horvitz-Thompson estimates. It is shown that this weighted AIC criterion provides better model choices for both incomplete and design-based samples. The use of the weighted AIC criterion is illustrated on data from the Belgian Health Interview Survey, which motivated this research. Simulations shows its performance in a variety of settings.

Keywords: Akaike Information Criterion, Complex Designs, Missing Data, Model Selection, Weighted Likelihood

1. Introduction

In a regression analysis, starting from a rich enough family of models and based on the data at hand, one or a few good models can be selected, e.g. using the Akaike Information Criterion (AIC). In case of missing data, simple deletion of the subsample of incomplete observations and treating the resulting subsample of so-called *complete cases* as a simple random sample has been shown to possibly lead to biased estimates, even when using a correct model (see e.g. Little 1992, Zhao *et al.* 1996). A similar problem occurs when the observations come from a complex survey design, i.e. when sampling from a finite population with unequal selection probabilities. Indeed, the probability that an observation is incomplete can also be considered as a selection probability for that observation to be included in the sample or not. Analyzing such design-based data as a simple random sample can also introduce bias (Horvitz and Thompson 1952).

There is a vast literature on parametric and nonparametric models in case of incomplete or design-based samples, but most of it concerns estimation (assuming a correct model) rather than model selection. The naive use of model selection criteria however turn outs to be unreliable in case of the aforementioned complications in the data. Indeed, treating the complete cases or the design-based sample as just a simple random sample can invoke some effects to appear or disappear and thus suggest another (incorrect) model to be more adequate for the data at hand.

In the context of incomplete data, selection methods like the predictive divergence for incomplete observations (PDIO, Shimodaira 1994) and the complete data AIC (AICcd, Cavanaugh and Shumway 1998) have been proposed. These methods rely on modelling the complete data likelihood, which introduces an additional model selection problem, namely

the selection of an appropriate model for the missingness mechanism (if not missing completely at random). In this paper we focus on selecting appropriate models for the measurement part, while treating the missingness mechanism as a nuisance. We propose a modification of the AIC-criterion for regression models, based on reweighing the complete cases by their inverse selection probabilities. The latter selection probabilities, if unknown, are preferably estimated non-parametrically (using e.g. splines), in this way avoiding the selection of a parametric model with its assumptions for the missingness process. This weighing of completely observed cases can be seen as an implicit imputation of missing observations and is valid when the probability to be missing depends upon the observed values but not on the unobserved values (MAR in the terminology of Little and Rubin 1987).

For the closely related situation of design-based samples, model selection has not been really investigated. In the next section, the motivating study illustrates both complications of missingness and design-based sampling. In Section 3, the weighted AIC-criterion is introduced and discussed, mainly for parametric models, but its applicability is also extended to nonparametric models. Indeed, analogous to the selection of an optimal model from a set of parametric candidate models, one can choose the optimal smoothing parameter in nonparametric regression based on an AIC criterion, as shown by Hurvich *et al.* (1998). We will modify this criterion to handle incomplete and design-based samples. An application to the cervix cancer screening data is shown in Section 4 while, in Section 5, a simulation study shows the improved performance of the modified AIC-criterion. Finally, Section 6 discusses some other weighted model selection criteria and possible avenues of further research.

2. HIS Example: Cervix Cancer Screening

To outline an evidence-based health policy, one is often interested in the profiles of persons who are at risk to obtain certain diseases and do not respond to prevention programs, e.g. cervix cancer screening. In the Belgian Health Interview Survey (HIS) of 1997, one of the questions investigated is in what respect the group of women, aged 25-64, not having a smear is different from the group of women that did have a smear taken in the past three years. For this purpose discrimination based on civil status, drug consumption, age, educational level and financial status was of interest. In this particular dataset, two complications arise. Firstly, sampling in the HIS was based on a combination of stratification, multistage sampling and clustering (Kish 1995). Secondly, about 30% of the 2893 women had one or more missing covariates for the variables of interest. These design issues, together with the likely occurrence of data to be missing, are inherent to surveys and should be taken into account when selecting an optimal model from a candidate set of models.

In Table 3 an overview of twelve different models, based on the variables given in Table 1, is given together with the original AIC-criterion and three weighted versions. The first modification, 'AIC_{W_1}', corrects for the survey design, the second version, 'AIC_{W_2}', corrects for incomplete data and the combination of both can be found in version, 'AIC_{W_1,W_2}'. Table 3 shows that different models are chosen by the different versions of the AIC-criterion; so it indicates that ignoring missingness or ignoring the sampling design can possibly lead to inappropriate model choices. We refer to Section 4 for a more thorough discussion.

Based on a theoretical justification, the weighted AIC's are defined in the next section.

Table 1. HIS Example: Variables used in the candidate models.

Variable	Abbreviation	Coding
Screening Status	SC	binary
Civil Status	CS	nominal
Drug Consumption	DR	ordinal
Age	Age	continuous
Educational Level	EL	nominal
Financial Status	FS	nominal

3. Weighted Akaike Information Criterion

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. 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).

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

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

with $L(\boldsymbol{\theta}, \boldsymbol{\eta})$ 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, 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 Information* (KL). 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)

$$I(f_0, f) = E\{ \log(\frac{f_0(\boldsymbol{y})}{f(\boldsymbol{y}; \boldsymbol{\theta}, \boldsymbol{\eta})})\} \tag{3}$$

(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 $f \equiv f_0$.

3.1. Missing Data

In case of missing data, the naive use of only complete cases in the definition of $I(f_0, f)$ can lead to serious deficiencies in its applicability to measure the distance between models (and

consequently also in the use of its empirical version, the AIC criterion). For simplicity, let us consider classical regression and suppose data are generated by a true model

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{\mu}_0, \sigma_0^2 I_n), \tag{4}$$

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

$$\boldsymbol{y} \sim \mathcal{N}_n(\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma^2 I_n), \tag{5}$$

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

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

$$E\{\sum_{i=1}^{n} \log f(y_i; \mu(\boldsymbol{x_i}), \sigma^2)\} = -\frac{n}{2}\log(2\pi\sigma^2) - E\left[\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}\right]/(2\sigma^2). \tag{6}$$

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

$$I(f_0, f) = \frac{n}{2}\log(\sigma^2/\sigma_0^2) + n\{\frac{\sigma_0^2}{\sigma^2} - 1\} + \{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T\{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}/(2\sigma^2). \tag{7}$$

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 the missingness process. For $i = 1, \ldots, n$, define the indicator $\delta_i = 1$ if $(\boldsymbol{x_i}, y_i)$ is fully observed and 0 otherwise. In general it is possible that $\pi_i = P(\delta_i = 1) = \pi(\boldsymbol{x_i}, y_i, z_i)$, so the probability that the ith observation is not fully observed is allowed to depend on $\boldsymbol{x_i}$, y_i or even on the value z_i of another, completely ignored variable. In this paper we restrict attention to the MAR setting, implying that π_i does not depend on z_i, that it additionally does not depend on $\boldsymbol{x_i}$ (resp. y_i) in case $\boldsymbol{x_i}$ (resp. y_i) might be missing.

The use of complete cases (CC) only (those for which $\delta_i = 1$) (and hence ignoring the missing data mechanism) is translated in a replacement of (6) by

$$E\{\sum_{i=1}^{n} \delta_i \log f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)\} = -\frac{E\{\text{trace}(D)\}}{2}\log(2\pi\sigma^2) - E\left[\{\boldsymbol{y} - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T D\{y - \boldsymbol{\mu}(\boldsymbol{\theta})\}\right]/(2\sigma^2) \tag{8}$$

where $D = \text{diag}(\delta_1, \ldots, \delta_n)$. As a function of σ^2 and $\boldsymbol{\mu}(\boldsymbol{\theta})$, and using a saturated model $\boldsymbol{\mu}(\boldsymbol{\theta}) = \boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ for the mean function, this expression (8) is maximized and the corresponding CC version of the KL distance

$$\begin{aligned} I_{CC}(f_0, f) &= E\{\sum_{i=1}^{n} \delta_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\} \\ &= \frac{E\{\text{trace}(D)\}}{2}\log(\frac{\sigma^2}{\sigma_0^2}) + E\left[\{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T D\{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}\right]/(2\sigma^2) \\ &\quad + E\{\boldsymbol{z}^T D\boldsymbol{z}\}\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2} - 1\right) + E\{\boldsymbol{z}^T D\}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta}))\left(\frac{\sigma_0}{\sigma^2}\right), \end{aligned} \tag{9}$$

(with $\boldsymbol{z} = (\boldsymbol{y} - \boldsymbol{\mu}_0)/\sigma_0$) is minimized at

$$\tilde{\theta}_i = \frac{E\{y_i \pi_i\}}{E\{\pi_i\}} = \mu_0(i) + \frac{\mathrm{Cov}(y_i, \pi_i)}{E\{\pi_i\}} \tag{10}$$

and

$$\tilde{\sigma}^2 = \frac{\sum_{i=1}^{n} E[\pi_i \{y_i - \tilde{\theta}_i\}^2]}{\sum_{i=1}^{n} E\{\pi_i\}}. \tag{11}$$

In the above expressions and in the what follows, moment related operators like the expectation E or the covariance (Cov) act on the random variables y_i and δ_i and treat $\boldsymbol{x_i}$ as nonrandom.

First of all, under a MCAR (missing completely at random) mechanism, $\pi_i = \pi$ and the above solutions simplify and are equal to the 'true' values, $\mu_0(i)$ and σ_0^2 respectively. The same holds in the MAR case that y_i is missing with probability $\pi_i = \pi(\boldsymbol{x_i})$, only depending on $\boldsymbol{x_i}$. If however π_i does depend on y_i in a way that $\mathrm{Cov}(y_i, \pi_i) \neq 0$, $I_{CC}(f_0, f)$ reaches a different minimum at (10) and (11). In fact, since by definition $I_{CC}(f_0, f_0) = 0$, this minimal value is negative (which is undesirable for a distance measure). If e.g. y_i and π_i are positively correlated, then $\tilde{\mu}_i > \mu_0(i)$. This is to be expected since observations with smaller values of y_i are discarded with higher probability. Also for nonsaturated models for $\boldsymbol{\mu}(\boldsymbol{\theta})$, such kind of anomalies can be shown.

The AIC criterion (2) based on the complete cases is given by

$$\mathrm{AIC}_{CC} = -2 \sum_{i=1}^{n} \delta_i \log[f(y_i; \mu(\boldsymbol{x_i}; \hat{\boldsymbol{\theta}}_{CC}), \hat{\sigma}^2_{CC})] + 2K, \tag{12}$$

where $\hat{\boldsymbol{\theta}}_{CC}$ and $\hat{\sigma}^2_{CC}$ are the ML estimators, maximizing the CC-loglikelihood (as described by the first term in (12)). For classical regression and ignoring constants, this can be simplified to

$$\mathrm{AIC}_{CC} = \left(\sum_{i=1}^{n} \delta_i \right) \log(\hat{\sigma}^2_{CC}) + 2K. \tag{13}$$

In case of MCAR, criterion (12) (or 13) is an approximately unbiased estimate of $I_{CC}(f_0, f)$ and is expected to behave appropriately (the missingness just results in an implicit sample size reduction). But for the MAR setting with missingness probabilities depending on the response, nothing guarantees that the above AIC criteria will serve any longer as useful model selection criteria.

The shortcomings of a CC approach, as described above, can be circumvented by a simple modification of the KL distance $I_{CC}(f_0, f)$ and corresponding AIC_{CC} criterion. This modification is inspired by the technique of weighted estimation. Assuming a correct model is used, Flanders and Greenland (1991) and Zhao and Lipsitz (1992) showed that the use of weighted estimators, solving the weighted estimating equations (WEE)

$$\sum_{i=1}^{n} w_i \Psi(y_i; \boldsymbol{\theta}, \boldsymbol{\eta}) = 0, \tag{14}$$

with Ψ the derivative of the log(quasi)likelihood and with weights w_i inversely proportional to the missingness probabilities, are consistent and asymptotically unbiased. The idea of WEE was inspired by the Horvitz-Thompson estimator in the closely related setting of design-based samples with unequal selection probabilities (see Horvitz and Thompson 1952). In Section 3.2, we further exploit this setting and its similarity with missing data for model selection.

Analogous to (14), a weighted KL distance can be defined as

$$I(f_0, f; w) = E\{\sum_{i=1}^{n} w_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\}. \tag{15}$$

Taking the weights

$$w_i = \delta_i/\pi_i, \tag{16}$$

the deficient distance $I_{CC}(f_0, f)$ is rectified and turned into the original data KL distance ('original' referring to the 'full' data, before introducing missingness). Indeed,

$$E\{\sum_{i=1}^{n} \frac{\delta_i}{\pi_i} \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\} \quad = \quad \sum_{i=1}^{n} E\{\log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\}.$$

In a similar way, the weighted AIC criterion

$$\text{AIC}_W = -2\sum_{i=1}^{n} w_i \log[f(y_i; \mu(\boldsymbol{x_i}; \hat{\boldsymbol{\theta}}_W), \hat{\sigma}^2_W)] + 2K, \tag{17}$$

with w_i as in (16) and with $\hat{\boldsymbol{\theta}}_W$ and $\hat{\sigma}^2_W$ the weighted ML estimators (maximizing the weighted maximum likelihood), is expected to behave appropriately, i.e. to correct for the missing data. Indeed, denote $\hat{\boldsymbol{\theta}}_{\text{O}}$ and $\hat{\sigma}^2_{\text{O}}$ the ML estimators based on the original data, and consider the Taylor expansion (linear terms canceling out)

$$\begin{aligned} & -2\sum_{i=1}^{n} w_i \log[f(y_i; \mu(\boldsymbol{x_i}; \hat{\boldsymbol{\theta}}_{\text{O}}), \hat{\sigma}^2_{\text{O}})] \qquad (18) \\ \approx \quad & \text{AIC}_W - 2\left((\hat{\boldsymbol{\theta}}_{\text{O}} - \hat{\boldsymbol{\theta}}_W)\ (\hat{\sigma}^2_{\text{O}} - \hat{\sigma}^2_W)\right) \mathcal{I}_n(\hat{\boldsymbol{\theta}}_W, \hat{\sigma}^2_W) \left((\hat{\boldsymbol{\theta}}_{\text{O}} - \hat{\boldsymbol{\theta}}_W)\ (\hat{\sigma}^2_{\text{O}} - \hat{\sigma}^2_W)\right)^T, \end{aligned}$$

where the matrix $\mathcal{I}_n$ is the matrix of second derivatives of the weighted log-likelihood, evaluated at $(\hat{\boldsymbol{\theta}}_W, \hat{\sigma}^2_W)$. The expected value of the left-hand side equals the expected value of the AIC criterion based on the original data. Since both estimates, the 'original' $(\hat{\boldsymbol{\theta}}_{\text{O}}, \hat{\sigma}^2_{\text{O}})$ and the 'weighted' $(\hat{\boldsymbol{\theta}}_W, \hat{\sigma}^2_W)$, are estimating the same parameter (being the true value $(\boldsymbol{\theta}_0, \sigma^2_0)$ in case the model under consideration is a correct model), the second term in the right hand side is negligible, at least in a first order approximation.

For a normal regression model with $\mu(\boldsymbol{x_i}, \boldsymbol{\theta}) = \boldsymbol{x_i}\boldsymbol{\theta}$, $i = 1, \ldots, n$, where $\boldsymbol{x_i} = (1\ x_{i1} \ldots n_{ip})$ and $\boldsymbol{\theta} = (\theta_0\ \theta_1 \ldots \theta_p)^T$, the weighted AIC criterion can be rewritten in terms of squared residuals

$$\text{AIC}_W = (\sum_{i=1}^{n} w_i) \log\left(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i}\right) + 2(p+2), \tag{19}$$

where e_i are the residuals from the fitted model, using weighted ML. In the context of robust model selection procedures, Agostinelli (2002) introduced a robust modification of

the Akaike Information Criterion (AIC), based on the weighted likelihood methodology. He proposed a similar weighted AIC_W criterion, but with weights downplaying the contribution of highly influential outliers.

Of course, typically the missing probabilities are unknown and have to be estimated, introducing essentially two further complications: i) finding appropriate estimates $\hat{\pi}_i$ which is again a model selection problem and ii) the effect on the characteristics of AIC_W when using weights

$$\hat{w}_i = \delta_i/\hat{\pi}_i. \tag{20}$$

Regarding the first complication, we suggest the use of a nonparametric or flexible semiparametric estimator (generalized additive models (gam) or e.g. regression trees for more complicated data structures, as illustrated in Section 4 and Section 5). This avoids the need for another model selection step. It is also important to note that, since the estimation of the missingness probabilities is a step *prior to* the envisaged model selection exercise, and hence is common to all candidate models under consideration, it has no effect on the penalization term in the expression of AIC_W. Concerning the second complication: rather than focusing on a theoretical study of the effect of estimating π_i on the expected value of AIC_W (a Taylor expansion immediately shows highly 'untractable' bias expressions), we opted for examining the finite sample performance of AIC_W with estimated weights by a simulation study (see Section 5).

In analogy to its expression based on the original data (Hurvich and Tsai 1989), we define a bias-corrected weighted AIC as

$$\text{AIC}_W^{cor} = \text{AIC}_W + \frac{2K(K+1)}{\sum_{i=1}^n w_i - K - 1}. \tag{21}$$

This small-sample correction (second-order bias adjustment) has been especially recommended in a setting where there are many parameters in relation to the size of the sample n (for more details see Burnham and Anderson 2002). Its performance in some simulations is briefly discussed in Section 5.1.3.

3.2. Design-Based Samples

Assume a finite population consisting of N units with measurements $\mathcal{M} = \{y_1, \ldots, , y_N\}$. A particular sampling plan leads to the random variable $\delta_i = 1$ if the ith unit is included in the sample (and 0 otherwise) with $n = \sum_{i=1}^N \delta_i$ the total sample size. The selection probabilities are defined as $\pi_i = P(\delta_i = 1)$, for $i = 1, \ldots, N$. The choice $\pi_i = n/N$ corresponds to a simple random sample. In this finite population setting, only the δ_i are to be considered as random; the set $\mathcal{M}$ is to be considered as unknown but fixed.

Supposing that the population $\boldsymbol{y} = (y_1, \ldots, y_N)^T$ is a single realization of a true 'super-population' model $f_0(\cdot)$, using the approximating model $f(\cdot; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)$ and treating the sample indicated by the δ_i as a random sample, a KL distance similar to the $I_{CC}(f_0, f)$ measure in (9) can be defined as (with now the expectation E with respect to the δ_i's, conditional on the 'realized' population)

$$\begin{aligned} I_{DB}(f_0, f) &= E\{\sum_{i=1}^N \delta_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]\} &(22)\\ &= \sum_{i=1}^N \pi_i \log[(f_0(y_i)/f(y_i; \mu(\boldsymbol{x_i}; \boldsymbol{\theta}), \sigma^2)]. &(23)\end{aligned}$$

For true and approximating models as in (4) and (5), with now $\boldsymbol{\mu} = (\mu(\boldsymbol{x}_1;\boldsymbol{\theta}),,\ldots,\mu(\boldsymbol{x}_N;\boldsymbol{\theta}))^T$ and $\boldsymbol{\mu}_0 = (\mu_0(1),\ldots,\mu_0(N))^T$ and with $\boldsymbol{z} = (\boldsymbol{y}-\boldsymbol{\mu}_0)/\sigma_0$ as before, we get

$$\begin{aligned} I_{DB}(f_0,f) &= \frac{\text{trace}(\Pi)}{2}\log(\frac{\sigma^2}{\sigma_0^2}) + \{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}^T\Pi\{\boldsymbol{\mu}_0 - \boldsymbol{\mu}(\boldsymbol{\theta})\}/(2\sigma^2) \\ &\quad + z^T\Pi z\frac{1}{2}\left(\frac{\sigma_0^2}{\sigma^2}-1\right) + z^T\Pi(\boldsymbol{\mu}_0-\boldsymbol{\mu}(\boldsymbol{\theta}))\left(\frac{\sigma_0}{\sigma^2}\right). \end{aligned} \tag{24}$$

As an example, consider a simple two-valued true superpopulation model

$$\boldsymbol{\mu}_0 = (\mu_0(1),\ldots,\mu_0(N_1),\mu_0(N_1+1),\ldots,\mu_0(N))^T = (\mu_1,...,\mu_1,\mu_2,...,\mu_2)^T$$

with $\mu_1 \neq \mu_2$, and the incorrect constant model $\boldsymbol{\mu}(\theta) = (\theta,...,\theta)^T$. For this incorrect model, the minimal distance $I_{DB}(f_0,f)$ is at least as small as its value at $\tilde{\sigma}^2 = \sigma_0^2$ and

$$\tilde{\theta} = \frac{\sum_{i=1}^N \pi_i y_i}{n}. \tag{25}$$

Using the correct two-parameter mean model with $\sigma^2 = \sigma_0^2$, $I_{DB}(f_0,f)$ is minimized at

$$\tilde{\mu}_1 = \frac{\sum_{i=1}^{N_1}\pi_i y_i}{n_1},\ \ \tilde{\mu}_2 = \frac{\sum_{i=1}^{N_2}\pi_i y_i}{n_2}, \tag{26}$$

where $n_1 = \sum_{i=1}^{N_1}\delta_i$ and $n_2 = \sum_{i=N_1+1}^{N}\delta_i$. Now, in the particular case that the selection probabilities induce a bias resulting in $\tilde{\mu}_1 = \tilde{\mu}_2$, the KL distance $I_{DB}(f_0,f)$ is exactly the same for both models and hence the incorrect model is indistinguishable from the correct model.

Identical to the case of missing data, the weighing of the KL distance and corresponding AIC criterion, with weights as in (16), can be used to correct both measures. Note that in general the selection probabilities can depend on both $\boldsymbol{x_i}$ and y_i. In most applications the selection probabilities π_i are determined by the design of the sample and hence are known.

3.3. Design-Based Samples with Missing Observations

In typical surveys, as in the cervix cancer screening example introduced in Section 2, both complications occur together. In this case δ_i, indicating whether or not the ith unit is in the sample and is fully observed, can be written as

$$\delta_i = \delta_i^D\delta_i^M, \tag{27}$$

where $\delta_i^D = 1$ if the ith unit is included in the sample (as in Section 3.2) and $\delta_i^M = 1$ if the ith unit is fully observed (as in Section 3.1). The weighted AIC (17) can now be based on weights $w_i = \delta_i/\pi_i$ where

$$\pi_i = P(\delta_i = 1) = P(\delta_i^M = 1|\delta_i^D = 1)P(\delta_i^D = 1). \tag{28}$$

These latter probabilities can be estimated by the product of the (known) probabilities $P(\delta_i^D = 1)$ and the (nonparametrically) estimated probabilities $P(\delta_i^M = 1|\delta_i^D = 1)$.

In the next section, we show how the idea of a weighted AIC can be extended to select a smoothing parameter for nonparametric regression.

3.4. Smoothing Parameter Selection using AIC_W

Assume

$$y_i = \mu_0(\boldsymbol{x_i}) + \epsilon_i, \quad i, \ldots, n \tag{29}$$

where $\mu_0(\cdot)$ is an unknown smooth function and $\epsilon_i, i = 1, \ldots, n$, are independent error terms with mean 0 and variance σ_0^2. Different linear smoothers for μ are available: orthogonal series, kernel estimators, splines, ... (see e.g. Simonoff, 1996). The most crucial choice for any smoother is the choice of the smoothing parameter. Hurvich, Simonoff and Tsai (1998) proposed to select this parameter α by minimizing the AIC-criterion

$$\mathrm{AIC}_\alpha = n \log(\hat{\sigma}^2) + \frac{n + \mathrm{trace}(S_\alpha)}{1 - \{\mathrm{trace}(S_\alpha) + 2\}/n}, \tag{30}$$

where S_α is the smoother matrix for which $\hat{\boldsymbol{y}} = S_\alpha \boldsymbol{y}$.

In case of an incomplete or design-based sample, this criterion can be turned into a weighted version

$$\mathrm{AIC}_{\alpha,W} = (\sum_{i=1}^{n} w_i) \log\left(\frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i}\right) + \frac{\sum_{i=1}^{n} w_i + \mathrm{trace}(S_{W,\alpha})}{1 - \{\mathrm{trace}(S_{W,\alpha}) + 2\}/(\sum_{i=1}^{n} w_i)}. \tag{31}$$

where $S_{W,\alpha}$ is the smoother matrix from the weighted fit. Taking $S_{W,\alpha}$ the classical regression 'hat matrix', (31) reduces (up to a constant) to (21).

4. The HIS 1997 Revisited

Since the design of the Health Interview Survey follows a complex multistage probability sampling scheme, it is necessary to incorporate this in the model selection procedure. A second complication is the substantial amount of missing covariate data (about 30%) spread over several covariates. Let us consider the candidate models given in Table 2. In Table 3, the models are ranked according to their AIC-criterion based on the complete cases (second column). For all other columns, the three models with lowest AIC-values are indicated by their ranks.

In the third column, a first weighted version, AIC_{W_1}, takes into account the complex design. Individual weights, W_1, reflecting the stratification at provincial level and the differential selection probabilities within households were available. This results in a somewhat different ordering of the models. The best model now is the model with original rank 8.

Similarly, the fourth column shows the modified AIC-value, AIC_{W_2}, incorporating missing covariate data (assuming MAR). Because of the high dimensional covariate space, a classification tree with surrogate splitting was used to obtain estimates of the missingness probabilities and thus the weights W_2. This leads to only minor changes, as compared to the second column. The best model now is model 2.

In the fifth column both complications have been taken into account by multiplying both weights in AIC_{W_1,W_2}. Again the same models appear to be the best ones; model 8 showing up again, now as the third best model, while model 3 is having the lowest value.

Although the same set of models reappears as the set of best models, this example illustrates that differently weighted AIC criteria can select different models as best ones. Since the choice of the final model or the set of final models used for e.g. model averaging is affected by missing data and by the design, we recommend in general the use of the weighted criteria (at least as a sensitivity tool).

Table 2. His Example: Overview of the candidate models.

Model	Structure
(1)	SC$\sim$ Age+Age^2+log(DR)+CS
(2)	SC$\sim$ Age+Age^2+log(DR)+EL+DR*EL
(3)	SC$\sim$ Age+Age^2+DR+EL+EL*DR
(4)	SC$\sim$ Age+Age^2+log(DR)
(5)	SC$\sim$ Age+Age^2+log(DR)+log(Age)
(6)	SC$\sim$ Age+Age^2+DR
(7)	SC$\sim$ Age+Age^2+CS+CS*Age
(8)	SC$\sim$ CS+Age+EL+DR+Age*EL
(9)	SC$\sim$ Age+Age^2
(10)	SC$\sim$ CS+Age+EL+DR+Age*EL+DR*EL
(11)	SC$\sim$ FS+CS+DR+Age+EL
(12)	SC$\sim$ FS+CS+DR+AGe+Age*FS

Table 3. His Example: The different (weighted) AIC-values and, between brackets, the rank of the three best models.

Model	AIC	AIC$_{W_1}$	AIC$_{W_2}$	AIC$_{W_1,W_2}$
(1)	1489.02*(1)*	975.31	2614.04*(2)*	1451.19
(2)	1489.81*(2)*	969.04	2606.71*(1)*	1441.53*(2)*
(3)	1490.70*(3)*	963.26*(2)*	2617.82*(3)*	1440.44*(1)*
(4)	1492.39	965.66*(3)*	2625.36	1445.89
(5)	1494.10	967.60	2625.73	1447.96
(6)	1495.86	967.64	2632.11	1449.03
(7)	1496.19	984.37	2631.01	1461.50
(8)	1496.84	961.57*(1)*	2628.85	1441.77*(3)*
(9)	1496.97	969.54	2636.47	1451.42
(10)	1502.31	967.35	2632.49	1447.34
(11)	1504.01	970.94	2648.48	1460.69
(12)	1516.75	980.92	2676.15	1477.45

To study the effects of weighing more closely, a simulation study in a variety of settings was conducted. The next section summarizes our main findings. All computations were conducted in R 1.9 (R Development Core Team 2003).

5. Simulations

In the first two scenarios, we consider a setting with missing covariate data. The third scenario focuses on design-based samples and the last scenario on the selection of the smoothing parameter in nonparametric regression.

5.1. Scenario 1: Parametric Model Selection for Incomplete Data

In the initial setting, the set of candidate models contains the true model.

5.1.1. Initial Setting

In this 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 $\mu_0(x, z) = -3 + 3x + 5x^2$ and variance $\sigma_0^2 = \exp(5)$. x-observations were then turned missing with conditional probability (see left bottom panel in Figure 1),

$$\pi(y, z) = 1 - [1 + \exp\{1 - 0.009(y - 300)\}]^{-1}. \tag{32}$$

Not depending on unobserved x-values, the missingness process is MAR. Let n denote the total sample size and n_c the number of complete observations. We generated 1000 different samples $\{(x_i, z_i, y_i), i = 1, \ldots, n\}$, with fixed design $\{x_i, z_i, i = 1 \ldots, n\}$. For each sample, 8 different regression models were fit, all submodels of $\mu(x, z) = \beta_0 + \beta_1 x_1 + \beta_2 x^2 + \beta_3 z + \beta_4 xz$.

Four different 'strategies' are compared: i) AIC on the original data, before introducing missingness (what we would get if no values were missing), ii) (unweighted) AIC on the complete cases only (ignoring missingness), iii) weighted AIC using the true weights (16) and iv) weighted AIC, using the estimated weights (20). The probabilities (32) are estimated by gam estimates $\hat{\pi}(y, z)$ (using the R package `mgcv`). On average 35% of the x-values were missing. In Figure 1, a typical dataset for scenario 1 is shown together with the missingness probabilities and the estimated weights. This latter figure shows a double curve, as a consequence of the additive model in x and z (being binary). The upper part of Table 4 displays the results for $n = 100$. Each column (from 2 to 9) corresponds to a particular model and the numbers show how often the respective model has been selected by AIC under the four strategies mentioned above. Models more complex than the true quadratic model $\{x, x^2\}$ can be considered as correct models, the others as incorrect models. The last rightmost column shows the total number of times a correct model was chosen. The table shows that for the initial setting, the unweighted AIC applied on the complete cases, very often selects the incorrect simpler model $\{x\}$. This is to be expected since the missingness is mainly located at the larger y-values (which of all response values mostly represent the quadratic effect). The weighted versions correct for that, especially the one with true weights which selects about 9% more often a correct model (though it less often selects the true model).

The other parts of Table 4 show similar results for variations on scenario 1: a smaller sample, larger error variance, larger quadratic effect and more missingness, Figure 2 up to

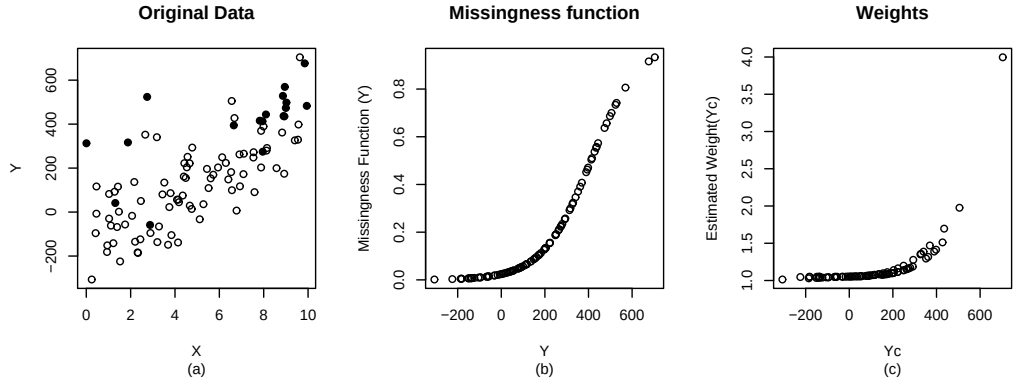


Fig. 1. For an arbitrary chosen sample under scenario 1: (a) original sample, complete cased (white bullets) and unobserved data (black bullets); (b) missingness probabilities; (c) estimated weights.

Table 4. Scenario 1. The numbers indicate how often a model has been selected, for the four strategies. The last column shows how often a correct model has been chosen, out of 1000. This scenario is repeated for different settings.

	1	x	z	x, x^2	x, z	$x, z,$ xz	$x, x^2,$ z	$x, x^2,$ z, xz	correctly classified
Scenario 1: Initial Setting ($n = 100, \sigma_0^2 = \exp(5), \text{slope} = 5, \%(\text{miss}) = 35$)									
Original Data	0	114	0	666	31	18	106	65	837
Complete Cases	0	312	0	452	65	35	91	45	588
True Weighted	0	199	0	371	67	61	129	173	673
Est. Weighted	0	228	0	416	70	56	110	121	647
Scenario 1: Sample Size 50									
Original Data	0	329	0	408	61	48	95	59	562
Complete Cases	1	477	1	267	94	63	61	36	364
True Weighted	0	330	1	247	116	107	92	107	446
Est. Weighted	2	373	0	271	102	90	90	80	441
Scenario 1: Variance exp(5.3)									
Original Data	0	304	0	471	58	35	77	55	603
Complete Cases	0	482	0	279	106	59	48	26	353
True Weighted	0	274	0	239	102	137	89	159	487
Est. Weighted	0	313	0	270	94	115	85	124	479
Scenario 1: Larger Quadratic Effect: slope = 7									
Original Data	0	28	0	728	1	11	144	88	960
Complete Cases	0	258	0	493	66	33	96	54	643
True Weighted	0	129	0	327	77	77	148	242	717
Est. Weighted	0	140	0	386	72	53	154	197	737
Scenario 1: Missingness 50%									
Original Data	0	143	0	649	21	24	105	58	812
Complete Cases	1	457	0	314	74	59	68	27	409
True Weighted	3	151	1	181	113	201	115	235	531
Est. Weighted	2	200	3	224	101	161	114	195	533

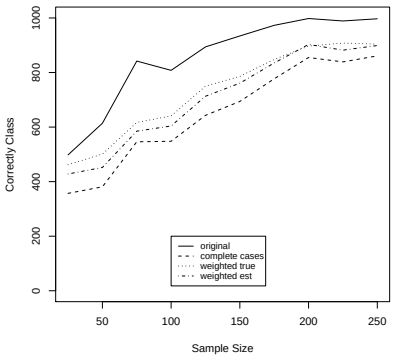


Fig. 2. Correctly selected models for different sample sizes.

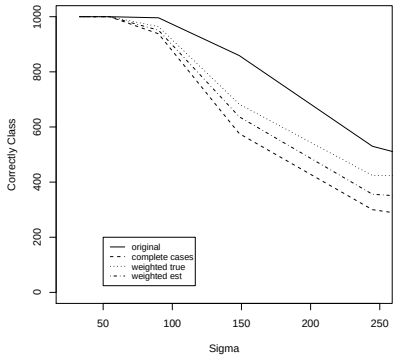


Fig. 3. Correctly selected models for different sigma-values.

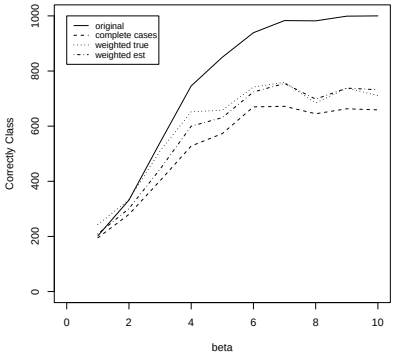


Fig. 4. Correctly selected models for different quadratic effects.

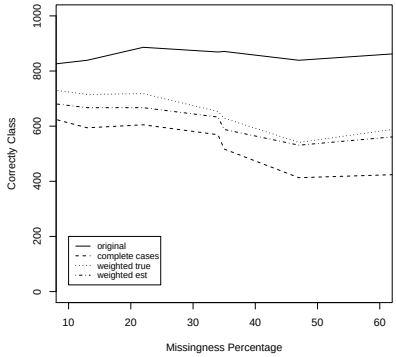


Fig. 5. Correctly selected models for different missingness percentages.

Figure 5 display the number of correct models as a function of sample size n, error variance σ_0^2, quadratic effect of x in $\mu_0(x, z)$ and missingness percentage (by changing the coefficient of y in equation (32)). All curves show the decrease in selecting a correct model when using the unweighted AIC on the complete cases. The difference gets more pronounced for increasing error variance, increasing missingness and increasing quadratic effect of x in $\mu_0(x, z)$. Note that this latter increasing effect implicitly generates more missingness via, on average, increasing response values y (see equation (32)). The use of the weighted version improves the performance of the AIC and the version with known weights is consistently doing better than with estimated weights. One might argue that the gain by using the weighted AIC is not so spectacular but rather moderate. On the other hand, we have to realize that correcting for missing information is often a hard exercise, since information in available data might be very scarce.

Table 5. Scenario 1, initial setting. Model selection using different smoothers to estimate the weights.

	x	x, x^2	x, z	$x, z,$ xz	$x, x^2,$ z	$x, x^2,$ z, xz	correctly classified
Complete Cases	312	452	65	35	91	45	588
NW h=150 (y, z)	255	426	72	45	112	90	628
NW h=150 (y)	253	431	69	47	116	84	631
NW CV (y, z)	221	403	74	59	120	123	646
NW CV (y)	237	417	79	59	112	96	625
gam CV(y, z)	228	416	70	56	110	121	647
gam CV (y)	191	387	74	72	130	146	663
True Weights	199	371	67	61	129	173	673

Table 6. Scenario 1 with sample size 30. Model selection using the corrected AIC-criterion.

	1	x	z	x, x^2	x, z	$x, z,$ xz	$x, x^2,$ z	$x, x^2,$ z, xz	correctly classified
Original Data	0	435	0	392	77	31	40	25	457
Complete Cases	16	616	3	217	80	34	26	8	251
True Weights	6	398	1	260	129	77	61	68	389
Est. Weights	8	442	0	275	122	53	56	63	394

5.1.2. Nonparametric Weighting Methods

Different smoothers can be used to estimate the missingness probabilities $\pi(y, z)$. In scenario 1, equation (32) shows that these probabilities only depend on y. In Section 5.1.1, these probabilities were estimated with a gam model, as a function of both y and z. In this section we illustrate how results differ when using different smoothers: gam using y only, Nadaraya-Watson (NW) kernel estimate using both y and z or y only, with fixed or with data-driven bandwidth (cross-validation).

The results in Table 5 show that the best results are obtained when using the penalized spline as a function of y only. The other numbers are more or less similar. The fixed bandwidth $h = 150$ for the NW-estimator was chosen by visual inspection of some of the generated samples. Main conclusion is that the choice of smoother and smoothing parameter is not unimportant. It is also recommendable to examine the missingness process carefully, so that accurate estimation of the probabilities is possible.

5.1.3. Corrected AIC

For small sample sizes, the use of the corrected AIC-criterion (21) is recommended. The results in Table 6 are based on the corrected AIC-criterion for the initial setting of Scenario 1 but with $n = 30$. The improvement is considerable. The true model is chosen more often using the weighted AIC, especially when the weights are estimated (this latter phenomenon was also noticeable in Table 4).

5.2. Scenario 2: Generating Model Not Included

We now consider the (more realistic) setting that the set of candidate models does not contain the true model. The response y is generated as in scenario 1, but now with mean

function $\mu_0(x) = -3 - 3\log(x+1) + 5x^2$. The same set of candidate models is considered. Since now direct comparison with the true model, nor a categorization in correct or incorrect models is possible anymore, we computed the average of the fitted values based on the selected model, together with 95% pointwise confidence intervals, using AIC on the original data, (unweighted) AIC on the complete cases, and weighted AIC on the complete cases. The resulting curves are shown in Figure 6 together with the true underlying function $\mu_0(x)$ (solid curve). Again, as before, gam was used to estimate the weights. The middle figure clearly shows the bias when using the unweighted AIC on the complete cases. The use of the weighted AIC nicely corrects the average best model in the direction of the true underlying curve. In Figure 7, boxplots of the simulated average squared errors, $(1/n)\sum_{i=1}^{n}(\hat{\mu}(x_i) - \mu_0(x_i))^2$ for the three methods, confirm the correction provided using the weighted AIC-criterion.

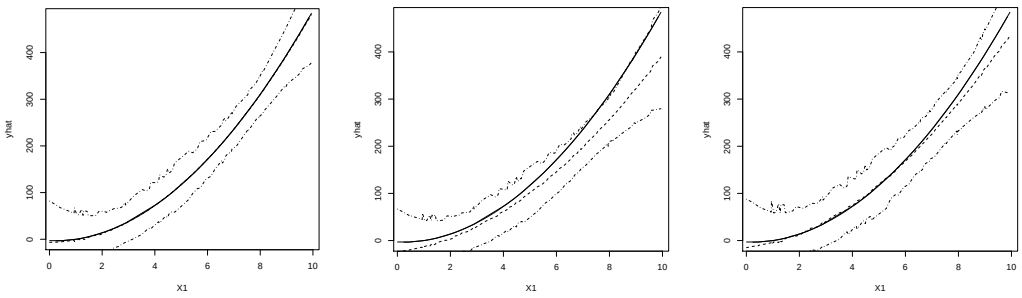


Fig. 6. Average best model with 95% pointwise confidence intervals for the original data (left), the complete cases with unweighted AIC (middle) and with weighted AIC (right). The solid curve is the true function $\mu_0(x)$

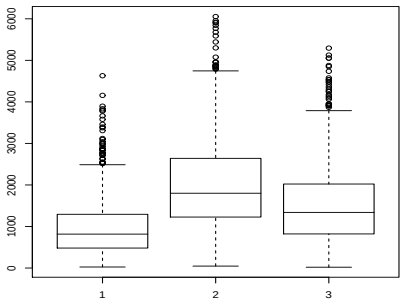


Fig. 7. Boxplots of the simulated ASE-values for the original data (left), the complete cases with unweighted AIC (middle) and with weighted AIC (right).

5.3. Scenario 3: Model Selection for Design-Based Samples

To illustrate the use of the weighted AIC for design-based samples, a population $\{y_1, \ldots, y_N\}$ of size $N = 1500$ was generated, as a single realization from the superpopulation model f_0, being a normal distribution with variance σ_0^2 and mean $\mu_0(i) = \mu_1$ for $i = 1, \ldots, 500$ (group 1), $\mu_0(i) = \mu_2$ for $i = 501, \ldots, 1000$ (group 2), $\mu_0(i) = \mu_3$ for $i = 1001, \ldots, 1500$ (group 3).

In a first setting 1000 samples were taken by dividing this population into three strata based on the ordered population y values: the 200 smallest y-values, the middle 900 y-values and the 400 largest y-values. The sample was then taken as follows: a population unit i (y_i) is selected for the sample with probability $p_1 f$ when it belongs to the first or third stratum and with probability $p_2 f$ when it belongs to the second stratum. When $p_1 < p_2$, this results in an oversampling of the second stratum.

The (single) population was generated with $\mu_2 = \mu_3 = \kappa = -\mu_1$ with $\kappa > 0$. The simulation parameters κ, σ_0, f, p_1 and p_2 were set to different values as shown in Table 7. For each of the samples, 5 different models were fit: (1) $\mu_i = \mu, i = 1, \ldots, 3$, (2) $\mu_1 = \mu_2 \neq \mu_3$, (3) $\mu_1 \neq \mu_2 = \mu_3$, (4) $\mu_1 = \mu_3 \neq \mu_2$, and (5) $\mu_i \neq \mu_j$ for $i \neq j$. Model (3) is the true model, model (5) is another correct model. The other models assume $\mu_1 = \mu_2$ or $\mu_1 = \mu_3$ and are incorrect (for $\kappa \neq 0$).

In a first setting, where $\{\kappa, \sigma_0, f\} = \{0.5, 3, 0.5\}$, sampling was done according to different choices of (p_1, p_2), ranging from simple random sampling $p_2/p_1 = 1$ to highly unequal stratified sampling $p_2/p_1 = 11$. The results in Table 7 show an improved selection for the AIC_W-criterion compared to the AIC-criterion. Models (3) and (5) are chosen more frequently by the AIC_W-criterion.

Increasing σ_0 (more noise) results in model (1) to be chosen more frequently. Also to be expected, a larger choice of κ (group 1 more different) leads more often to correct model choices. The fraction parameter f was initially chosen 0.5, resulting in a sample of size 225. To reflect the behavior for a smaller sample, f was set to 0.2, resulting in a larger variability due to the smaller sample size (= 90). For all variations of the basic setting, AIC_W improves the selection from slightly to substantially (according to the ratio p_2/p_1), except for $\kappa = 1$.

In a second setting, the same population was taken but now design-based sampling was based on two strata, the 300 largest y-values of the third group and the remaining 1200 y-values. Sampling was done as follows: a population unit i is selected with probability $p_1 f$ when it belongs to the first stratum and with probability $p_2 f$ when it belongs to the second stratum. If $p_1 < p_2$ this results in an undersampling of units in the third group with the larger y values. The results for 1000 such samples are shown in Table 8, again for the same basic setting and variations thereof. One can see that the AIC-criterion very often chooses the incorrect model (4) $\mu_1 = \mu_3 \neq \mu_2$ and the AIC_W-criterion corrects this choice to model (3) $\mu_1 \neq \mu_2 = \mu_3$, which is the true model. For all variations of this setting, the AIC_W outperforms AIC in all cases. The differences are much more pronounced as in previous setting. One can also observe that the number of times a correct model is selected by the AIC_W-criterion is more or less the same for all different choices of (p_1, p_2). When sampling probabilities are equal and thus a simple random sample is taken, the choices made using AIC and AIC_W are essentially the same.

Table 7. Scenario 3, first setting: The number of models chosen by AIC and AIC_W, for different variations of the basic setting and different choices of p_1 and p_2.

		AIC						AIC_W					
p_1	p_2	1	2	*3*	4	*5*	*Cor*	1	2	*3*	4	*5*	*Cor*
Basic													
0.05	0.55	321	110	445	107	17	462	128	192	277	133	270	547
0.10	0.50	284	101	498	92	25	523	155	146	424	136	139	563
0.20	0.40	191	116	594	63	36	630	156	132	572	60	80	652
0.30	0.30	133	108	639	64	56	695	125	108	648	63	56	704
$\sigma_0 = 4$													
0.05	0.55	467	108	301	115	9	310	134	205	281	189	191	472
0.10	0.50	428	117	325	118	12	337	209	199	328	161	103	431
0.20	0.40	331	121	450	75	23	473	259	144	471	72	54	525
0.30	0.30	305	136	445	86	28	473	295	137	455	86	27	482
$\kappa = 1$													
0.05	0.55	13	31	817	25	114	931	27	89	397	62	425	822
0.10	0.50	6	8	841	11	134	975	9	23	604	20	344	948
0.20	0.40	2	5	850	2	141	991	2	6	786	2	204	990
0.30	0.30	0	1	842	0	157	999	0	1	840	0	159	999
$f = 0.2$													
0.05	0.55	494	113	249	133	11	260	116	211	240	204	229	469
0.10	0.50	481	142	241	128	8	249	227	193	280	189	111	391
0.20	0.40	440	130	304	112	14	318	351	158	321	129	41	362
0.30	0.30	364	133	360	123	20	380	368	130	364	118	20	384

Table 8. Scenario 3, second setting: The number of models chosen by AIC and AIC_W, for different variations of the basic setting and different choices of p_1 and p_2.

		AIC						AIC_W					
p_1	p_2	1	2	*3*	4	*5*	*Cor*	1	2	*3*	4	*5*	*Cor*
Basic													
0.05	0.55	92	120	56	596	136	192	66	175	510	52	197	707
0.10	0.50	189	19	392	381	19	411	46	171	590	12	181	771
0.20	0.40	126	131	651	31	61	712	60	197	615	7	121	736
0.30	0.30	133	108	639	64	56	695	125	108	648	63	56	704
$\sigma_0 = 4$													
0.05	0.55	162	266	27	389	156	183	156	307	377	56	104	481
0.10	0.50	370	59	215	349	7	222	144	276	475	28	77	552
0.20	0.40	289	168	472	44	27	499	137	283	500	14	66	566
0.30	0.30	305	136	445	86	28	473	295	137	455	86	27	482
$\kappa = 1$													
0.05	0.55	0	0	316	599	85	684	0	0	613	3	384	997
0.10	0.50	0	0	757	64	179	936	0	0	709	0	291	1000
0.20	0.40	0	3	845	1	151	996	0	2	775	0	223	990
0.30	0.30	0	1	842	0	157	999	0	1	840	0	159	999
$f = 0.2$													
0.05	0.55	336	138	108	385	33	141	243	254	356	77	70	426
0.10	0.50	439	64	219	270	8	227	263	236	395	62	44	439
0.20	0.40	359	167	381	76	17	398	250	240	439	46	25	464
0.30	0.30	364	133	360	123	20	380	368	130	364	118	20	384

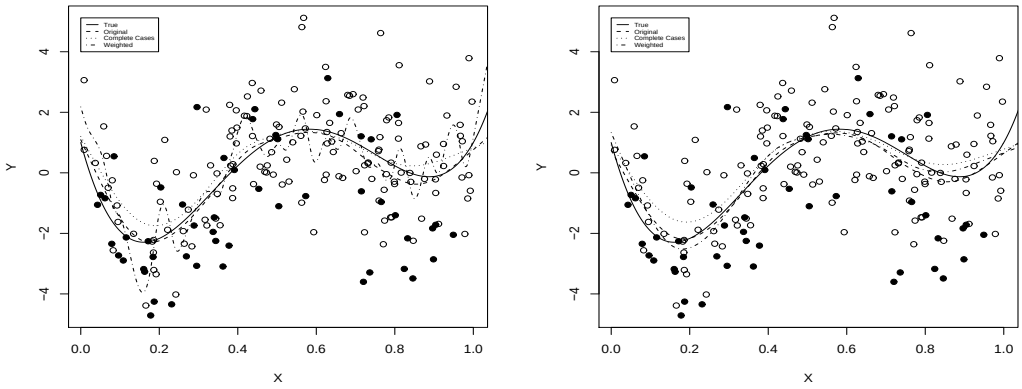


Fig. 8. Simulated dataset with spline fits according to the different methods together with the true function, using the ML variance estimator $\hat{\sigma}^2_{ML}$ (left panel) and the unbiased variance estimator $\hat{\sigma}^2_U$(right panel).

5.4. Scenario 4: Smoothing Parameter Selection in Nonparametric Regression for Incomplete Data

For this scenario, $n = 200$ x-values were generated from uniform$[0, 1]$, and corresponding y-values from a normal distribution with mean $\mu_0(x) = 1 - 48x + 218x^2 - 315x^3 + 145x^4$ and variance $\sigma_0^2 = 0.4$ Range(y). This corresponds to one of the simulation settings used in Hurvich *et al.* (1998). Next, x observations were turned missing with probability

$$\pi(y) = [1 + \exp\{2 - 0.1(y - 2)^2]^{-1}. \tag{33}$$

For each of the 1000 generated samples $\{Y_i, i = 1, \ldots, n\}$ with a fixed design $\{x_i, i = 1 \ldots, n\}$, a smoothing spline was fitted (using `smooth.spline` in R) according to three methods, and with smoothing parameter selected by AIC (as introduced by Hurvich *et al.* 1998). The first method is based on the original data, while the second method is based on the complete cases only and finally the third method weights the complete cases (at the model selection stage and at the final fitting stage) with $\hat{w}_i = 1/\hat{\pi}_i$ where $\hat{\pi}_i$ is the estimated probability for a complete case to be observed. The estimation of π_i is also based on a smoothing spline with smoothing parameter again determined by AIC.

The left panel in Figure 8 displays an arbitrary sample together with the fitted splines. The white dots indicate the observed data, while the black dots show the unobserved or missing data. The spline using the weights tends to severely undersmooth.

In this context, Wahba (1990) uses the unbiased variance estimator

$$\hat{\sigma}^2_U = \frac{y^T(I - S_\alpha)^2 y}{\text{trace}(I - S_\alpha)}, \tag{34}$$

where S_α is the smoother matrix. The use of $\hat{\sigma}^2_U$ instead of $\hat{\sigma}^2_{ML}$ is equivalent to an extra penalization of $-n \log(\text{trace}(I - S_\alpha))$, which corrects for undersmoothing, as can be seen for the fit of a random sample in the right panel of Figure 8. This is also confirmed by

Table 9. The average number of parameters using variance estimator $\hat{\sigma}^2_{ML}$ or $\hat{\sigma}^2_U$.

	$\hat{\sigma}^2_{ML}$	$\hat{\sigma}^2_U$
Original Data	8.33	6.99
Complete Cases	7.55	6.31
Weighted	18.31	9.00

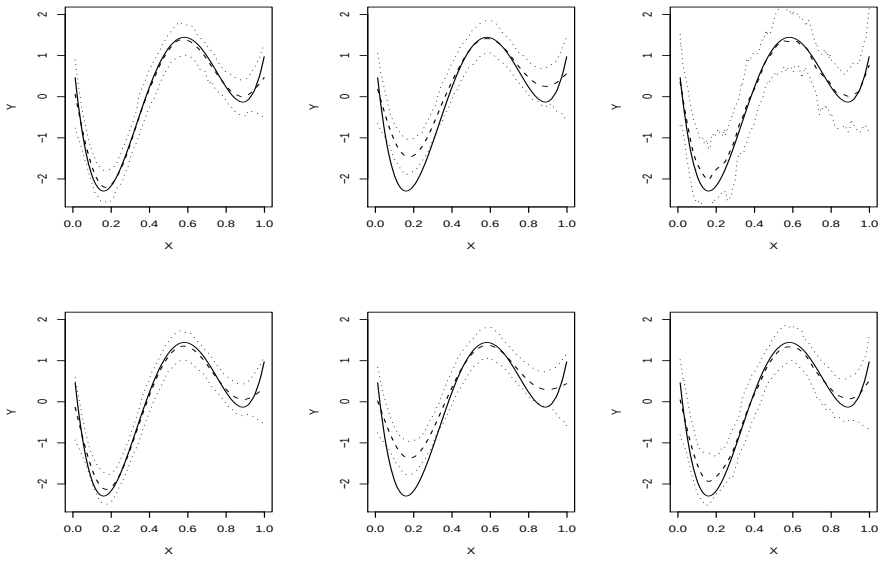


Fig. 9. Average of the fitted values based on the chosen models over simulation runs together with the true function and 95% confidence intervals. From left to right: the original data, the complete cases and the weighted complete cases.

Table 9. It shows the simulation average of the equivalent number of parameters, selected by the three methods (rows) and for both variance estimators (columns). The models using the unbiased estimator are generally smoother and this reduction in equivalent number of parameters is very substantial for the weighted analysis. Other simulations confirmed this and therefore we certainly recommend the use of the unbiased estimator $\hat{\sigma}^2_U$ for the weighted method.

In Figure 9 the true curve (the solid line) and the simulation average of the fitted curves for all three methods and both variance estimators, together with 95% pointwise confidence intervals, are shown. Again, the beneficial effect on the smoothing when using the unbiased variance estimator is illustrated. The middle panels show that there is substantial bias at both minima, when using the complete cases without weighting. The weighted AIC does correct for bias, as shown in the right panels.

To assess the goodness of fit quantitatively for each of the fits, the average squared error (ASE) was calculated as the simulation average of all values $(1/n)\sum_{i=1}^{n}(\hat{m}_\alpha(x_i) - m_0(x_i))^2$,

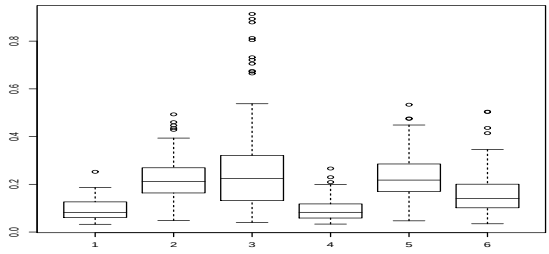


Fig. 10. Boxplots of the simulated ASE-values for the different methods: original data, $\hat{\sigma}^2_{ML}$ (1), complete cases, $\hat{\sigma}^2_{ML}$ (2), weighted complete cases, $\hat{\sigma}^2_{ML}$ (3), original data, $\hat{\sigma}^2_{U}$ (4), complete cases, $\hat{\sigma}^2_{U}$ (5), weighted complete cases, $\hat{\sigma}^2_{U}$ (6).

for each method and each variance estimator. The boxplots in Figure 10 show again that the weighted AIC method is not resulting in an improvement when using $\hat{\sigma}^2_{ML}$, but that it does when using $\hat{\sigma}^2_{U}$.

6. Discussion

The naive use of model selection criteria in case of incomplete and design-based samples can lead to the selection of inappropriate or non-optimal model. In this paper we introduced a weighted Akaike information criterion. The weights are inversely proportional to the selection probabilities and if unknown, can be estimated nonparametrically. Simulations show that the use of this weighted AIC-criterion results in improved model selection. The method can be seen as an implicit nonparametric imputation approach and its application is straightforward. Other options to deal with missingness in the context of model selection are full likelihood methods, that models both measurement and missingness part simultaneously, or first impute missing observations and then select the model based on the augmented dataset. Since both approaches need an additional model to be selected and are not extendable to the analogous setting of design-based samples, these methods were not pursued in this paper.

Next to the AIC, several other model selection criteria have been developed and can be extended to a weighted version to handle incomplete and design-based samples. For a model M with p regression parameters, the Mallows' C_p criterion, developed as an estimator of the relative mean squared error, is very popular for least squares regression. Its definition $\mathrm{Cp} = n\hat{\sigma}^2(M)/\hat{\sigma}^2(F) - (n-2p)$ where $\hat{\sigma}^2(M)$ ($\hat{\sigma}^2(F)$) is the estimated variance based on a reduced model M (respectively full model F), can be modified in the weighted version

$$\mathrm{Cp}_W \quad = \quad \left(\sum_{i=1}^{n} w_i\right) \frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i e_i^{*2}} - \left(\sum_{i=1}^{n} w_i - 2p\right),$$

where e_i and e_i^* are the residuals based on reduced model and full model respectively. Analogously, the Bayesian information criterion $\mathrm{BIC} = n(\log \hat{\sigma}^2_{ML}) + \log(n)K$ (for classical

Table 10. Scenario 1, basic setting. The number of chosen models by the Cp- and BIC-criterion.

	x	x, x^2	x, z	$x, z,$ xz	$x, x^2,$ z	$x, x^2,$ z, xz	correctly classified
Scenario 1: Basic Setting, Cp.							
Original Data	111	662	32	18	108	69	839
Complete Cases	303	455	66	37	91	48	594
True Weights	196	366	68	60	132	178	676
Est.Weights	224	411	70	56	114	126	651
Scenario 1: Basic Setting, BIC.							
Original Data	359	593	21	3	19	5	617
Complete Cases	637	311	32	6	12	2	325
True Weights	420	409	65	30	34	42	485
Est. Weights	499	377	40	22	36	28	441

regression) can be modified in a weighted version

$$\mathrm{BIC}_W \quad = \quad \sum_{i=1}^{n} w_i \left(\log \frac{\sum_{i=1}^{n} w_i e_i^2}{\sum_{i=1}^{n} w_i} \right) + \log \left(\sum_{i=1}^{n} w_i \right) K.$$

We also investigated the performance of these alternative model selectors in a simulation study. As an illustration, Table 10 shows some results for the basic simulation setting of the first scenario. Up to expected differences, like the BIC criterion selecting more simple models, a similar improvement is realized by the weighted selection criteria.

In this paper we focused on the weighted AIC criterion. As BIC, it is a general applicable criterion and has been proven to be very helpful in model selection (see e.g. Burnham and Anderson 2002). Extensions to weighted versions of model selection criteria for generalized estimating equations in the context of clustered data as proposed in Wei (2001a) and Wei (2001b), are topics of current research. Additional further research includes deriving new lack of fit tests when dealing with incomplete and design-based data (e.g. modifications of Aerts *et al.* 1999), and the use of a weighted likelihood ratio test (see e.g. Agostinelli and Markatou 2001) in this context.

Acknowledgements

Financial support from the IAP research network nr P5/24 of the Belgian Government (Belgian Science Policy) is gratefully acknowledged.

References

Aerts, M., Claeskens, G. and Hart, J. (1999) Testing the fit of a parametric function. *Journal of the American Statistical Association*, **94**, 869–879.

Agostinelli, C. (2002) Robust model selection in regression via weighted likelihood methodology. *Statist. Probab. Lett.*, **56**, 289–300.

Agostinelli, C. and Markatou, M. (2001) Test of hypotheses based on the weighted likelihood methodology. *Statistica Sinica*, **44**, 499–514.

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

Burnham, K. and Anderson, D. (2002) *Model selection and multimodel inference: a practical information-theoretic approach*. New York: Springer-Verlag.

Cavanaugh, J. and Shumway, R. (1998) An akaiki information criterion for model selection in the presence of incomplete data. *Journal of Stat. Planning and Inference.*, **67**, 45–65.

Flanders, W. and Greenland, S. (1991) Analytic methods for two-stage case-control studies and other stratified designs. *Stat. in Med.*, **10**, 739–747.

Horvitz, D. and Thompson, D. (1952) A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc.*, **47**, 663–685.

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

Hurvich, C. and Tsai, C.-L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Kish (1995) *Survey Sampling.* New York: Wiley.

Little, R. (1992) Regression with missing x's: A review. *Journal of the American Statistical Association*, **87**, 1227–37.

Little, R. and Rubin, D. (1987) *Statistical Analysis with Missing Data.* Wiley. New York.

R Development Core Team (2003) *R: A language and environment for statistical computing.* Vienna, Austria. ISBN 3-900051-00-3, URL `http://www.R-project.org`.: R Foundation for Statistical Computing.

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

Shimodaira, H. (1994) A new criterion for selecting models from partially observed data. in: Cheeseman, p., oldford, r.w. (eds.), selecting models from data: Artificial intelligence and statistics iv. vol. 89, 21–29.

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

Spiegelhalter, D., Best, N., Carlin, B. 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.

Wahba, G. (1990) *Spline Models for Observational Data.* CBMS-NSF series. SIAM, Philadelphia.

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

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

Zhao, L. and Lipsitz, S. (1992) Design and analysis of two-stage studies. *Stat. in Med.*, **11**, 769–782.

Zhao, L., Lipsitz, S. and Lew, D. (1996) Regression analysis with missing covariate data using estimating equations. *Biometrics*, **52**, 1165–1182.