Model Selection for Regression Analyses with Missing Data

M. Aerts[1], N. Hens[1] and G. Molenberghs[1]

[1] Center for Statistics, Limburgs Universitair Centrum, Universitaire Campus, B-3590 Diepenbeek, Belgium

Abstract: The Akaiki Information Criterion, AIC, is one of the leading selection methods for regression models. In case of partially missing covariates with missingness probability depending on the response, regression estimates based on the so-called complete cases are known to be biased. In this contribution it is shown that model selection using AIC-values based on the complete cases can lead to the choice of wrong or less optimal models. In analogy with the weighted Horvitz-Thompson estimator, we propose a weighted version of AIC. It is shown that this weighted AIC criterion improves model choices.

Keywords: Akaiki Information Criterion; Missing Data; Model Selection; Weighted Likelihood

1 Introduction

Let $(x_1, z_1, y_1), ..., (x_n, z_n, y_n)$ be a sample where y denotes a response variable and x and z covariate variables. Here we focus on the case that, for a fixed value of x and z, the response y is normally distributed with variance σ^2. Suppose we want to select an optimal model from a set of K candidate models for the mean function $\mu(x, z) = E(y|x, z)$. A well-established method is selecting the model k which minimizes the AIC criterion (Akaike 1973, Linhart and Zucchini 1986, Burnham and Anderson 1998, Hurvich and Tsai 1989):

$$AIC = -2\log(\text{likelihood of model } k) + 2 \times (\#\text{ parameters of model } k), \tag{1}$$

where the likelihood is evaluated at the corresponding ML-estimator. For a normal error structure, this simplifies to (ignoring some constant terms, not depending on k):

$$AIC = n \log \hat{\sigma}_k^2 + 2p_k, \tag{2}$$

where $\hat{\sigma}_k^2$ is the ML variance estimator based on model k and p_k is the number of regression coefficients in model k.

In a missing data context, covariate x or response y may be missing. We assume z is always observed. Let $\delta_i = 1$ if the ith observation is completely observed and $\delta_i = 0$ otherwise. Furthermore, let the selection probabilities

$\pi_i = P(\delta_i = 1|y_i, x_i, z_i)$ reflect the missing at random (MAR) missingness mechanism (Rubin 1976). So, $\pi_i = P(\delta_i = 1|y_i, z_i)$ in the missing covariate case and $\pi_i = P(\delta_i = 1|x_i, z_i)$ in case the response y is subject to missingness. For missing covariate data, Flanders and Greenland (1991) and Zhao and Lipsitz (1992) suggested a weighted estimator in the spirit of Horvitz and Thompson (1952), based on the weighted likelihood or weighted least squares criterion for the complete cases (CC) with weights equal to $1/\hat{\pi}_i$, where $\hat{\pi}_i$ is an appropriate estimator for the selection probabilities π_i. Wang et al. (1997) proposed to use a nonparametric kernel smoother to estimate the selection probabilities while fitting the regression curve with a parametric model and Wang et al. (1998) proposed a weighted local linear estimator for $\mu(x)$ while using local linear estimates for $\pi(y_i)$.
Model selection for incomplete data has not received much attention in the literature. Cavanaugh and Shumway (1998) derived and investigated a variant of AIC motivated by the same principle as the 'predictive divergence of incomplete observations'. Hens, Aerts and Molenberghs (2004) proposed modifications of several model selection criteria using weighting likelihood ideas and compared it to "model selection after imputation" methods. A similar weighted Akaiki information criterion in the context of robust model selection and robust regression models has been proposed by Agostinelli (2002).

2 Modified AIC criterion

We focus on the weighted AIC criterion applied to normal response data as described in the previous section. Weighting in (2) each complete case contribution to the loglikelihood with weight $1/\hat{\pi}_i$ leads to the criterion

$$AIC_W = (\sum_{i=1}^{n} \delta_i/\hat{\pi}_i) \log \hat{\sigma}^2_{W,k} + 2p_k \tag{3}$$

where $\hat{\sigma}^2_{W,k}$ is the ML variance estimator based on the weighted (normal) likelihood.

3 Unknown weights

In some settings (e.g. a two-stage design), the selection probabilities are known and do not have to be estimated. In many missing data problems, however, the unknown weights π_i, which can be considered as nuisance parameters, have to be estimated. This estimator has to be consistent, otherwise it will adversely affect the model selection procedure. So if we estimate π_i with a parametric model, we are faced with an additional model selection problem. Hens, Aerts and Molenberghs (2004) suggest the use of a nonparametric estimator, e.g. a kernel smoother as used in Wang et

al. (1998). In the next section we illustrate the applicability of the method in a small simulation study.

4 Simulation Study and Discussion

Observations for a continuous explanatory variable X are generated from a uniform distribution on the interval $[0, 10]$, Z observations are generated from a Bernoulli distribution with probability 0.50. Conditionally upon X, Y observations are generated from a normal distribution with mean $\mu(x) = -3+3x+5x^2$ and variance $\sigma^2 = \exp(5)$. X observations are then turned into 'missing' with conditional probability $\pi(x) = [1+\exp\{1-0.009(y-300)\}]^{-1}$. We generated 1000 different samples $\{Y_i, i = 1, \ldots, n\}$ with a fixed design $\{x_i, z_i, i = 1 \ldots, n\}$ of sample size $n = 100$. For each sample, 8 different regression models were fit, i.e. all submodels of $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z + \beta_4 XZ$.

Model Method	1	X	Z	X, X^2	X, Z	X, X^2, Z	X, Z, XZ	X, X^2, Z, XZ
ALL	0	125	0	647	30	128	13	57
CC	0	340	0	432	71	75	38	44
TW	0	197	0	366	74	116	69	178
EW	0	269	0	422	73	97	52	87
E2	0	220	0	396	78	103	66	137

TABLE 1. Simulation study with 8 candidate models: number of AIC selected models

Method	Correct	Incorrect
ALL	832	168
CC	551	449
TW	660	340
EW	606	394
EW2	636	364

TABLE 2. Simulation study with correctly and incorrectly classified models: number of AIC selected models

Table 1 shows, for each candidate model, the number of times it is has been selected as best model by the AIC criterion (2) or (3), for 5 different methods: ALL stands for an unweighted analysis based on all data (as if no data were missing); CC for an unweighted analysis on the complete cases only (excluding the observations with a missing X-value); TW for a

weighted analysis with true known missingness probabilities $\pi(x)$; EW for a weighted analysis with kernel estimated probabilities $\pi(x)$ using a fixed bandwidth and finally, EW2 for a weighted analysis with kernel estimated probabilities using a cross-validation data-driven choice of the smoothing parameter.
A comparison of the first two rows shows the effect of ignoring the missingness by using an unweighted AIC criterion on the complete cases. The weighted criterion (3) improves the selection of correct models, as shown in the last three rows of Table 1 and Table 2. In Table 2, all more complex models containing the true model as a submodel are collapsed in a category "correct model".
The last two lines illustrates the importance of using a data-driven smoothing parameter, when estimating the missingness probabilities $\pi(x)$.

References

Agostinelli, C. (2002). Robust model selection in regression via weighted likelihood methodology. *Stat. & Prob. Letters*, **56**, 289–300.

Akaiki, H. (1973). Information theory as an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*. Petrov, B.N., and Csaki, F. (eds.), Akademiai Kiado, 267–281.

Burnham, K.P. and Anderson, D.R. (1998). *Model Selection and Inference*. New York: Springer-Verlag.

Cavanaugh, J.E. and Shumway, R.H. (1998). An Akaiki information criterion for model selection in the presence of incomplete data. *J. Statist. Plan. Inf.*, **67**, 45–65.

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

Hens, N., Aerts, M. and Molenberghs, G. (2004). Regression model selection for incomplete and non-random samples. Technical Report.

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

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

Linhart, H. and Zucchini, W. (1986). *Model Selection*. New York: Wiley.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, **63**, 581–592.

Wang, C.Y., Wang, S., Gutierrez, R.G., and Carroll, R.J. (1998). Local linear regression for generalized linear models with missing data. *Ann. Statist.*, **26**, 1028–1050.

Wang, C.Y., Wang, S., Zhao, L-P., Ou, S-T. (1997). Weighted semiparametric estimation in regression analysis with missing covariate data. *J. Amer. Statist. Assoc.*, **92**, 512 –525.

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