Pacific Symposium on Biocomputing 4:214-325 (1999)

COMPUTING MINIMUM DESCRIPTION LENGTH FOR ROBUST LINEAR REGRESSION MODEL SELECTION

GUOQI QIAN
Department of Statistical Science, La Trobe University,
Melbourne, VIC 3083, Australia

A minimum description length (MDL) and stochastic complexity approach for model selection in robust linear regression is studied in this paper. Computational aspects and implementation of this approach to practical problems are the focuses of the study. Particularly, we provide both algorithms and a package of S language programs for computing the stochastic complexity and proceeding with the associated model selection. A simulation study is then presented for illustration and comparing the MDL approach with the commonly used AIC and BIC methods. Finally, an application is given to a physiological study of triathlon athletes.

1 Introduction

A powerful statistical tool for quantitative investigations in health and biological sciences is linear regression, where the simultaneous effects of a set of variables on a response variable can be analyzed. An important task in linear regression analysis is to screen a large number of potential explanatory variables to select that subset of them which fit the information contained in the response variable both efficiently and concisely. This is important because it may cause not only serious computation round errors, but also key statistical evidence undetectable if a regression model contains many irrelevant or superfluous explanatory variables. An equally important task is to see how the selected model is affected by outliers in the data. Namely, the model should be robust to radical change of a small portion of the data or a small change in all of the data. A natural solution for robust model selection can be obtained using the information-theoretic approaches such as Algorithmic Probability (ALP)(Solomonoff 1964), Minimum Message Length (MML)(Wallace and Freeman 1987), and Minimum Description Length (MDL) and Stochastic Complexity (Rissanen 1986, 1987 and 1996).

Using the MDL and stochastic complexity approach, Qian and Künsch (1998 a&b) derived a new variable selection criterion for robust linear regression. This criterion chooses such a subset of the explanatory variables relative to which the stochastic complexity of the data attains the minimum. The stochastic complexity of the data relative to the underlying regression model was shown to be approximated by the robust fitting error of the model plus the model complexity — a term depending on the robustness and the signal-

to-noise ratio of the model, and the weighted magnitude of the explanatory variables. Thus the new criterion substantially generalizes those classic model selection criteria such as AIC and BIC where the model complexity depends only on the number of parameters. Asymptotic study reveals that the new criterion selects with probability one the true model if it exists and can be finitely parameterized; and it has the ability of avoiding the two pitfalls of either over-fitting or under-fitting that plague many model selection criteria like AIC and BIC.

The current paper focuses on the computational aspects and the real applications of the stochastic complexity criterion for multiple robust regression model selection. Specifically, we will address the methods and their properties of computing the robust parameter estimates, the weight function and the criterion function that are involved in the model selection procedure. We will also introduce a package of S language programs called `msrob` we have written for the computations. We will then present a simulation study to compare the new criterion with the commonly used AIC and BIC. Finally, we will give an application for determining an athlete's total time in a triathlon from those candidate variables measuring the athlete's gross physical characteristics, the training load and the physiological makeup.

Some other closely related works are Baxter and Dowe(1996) and Dom (1996). In Baxter and Dowe(1996) the problem of order selection for the polynomial regression models is studied in a non-robust context and using the MML principle which is developed by Wallace and co-workers since 1968. Dom (1996) also studied mostly the non-robust polynomial regression order selection but using the MDL principle. A polynomial regression model concerns the relationship between a response variable and a polynomial function of certain explanatory variable. So the statistical problems studied in these two papers are very different from ours which concerns the significant relationship between a response and certain subset of many explanatory variables in a robust framework. MML and MDL both use the code length as a criterion function for model selection. But there are also many significant differences between the two principles. This relationship will not be expounded further here.

2 The Stochastic Complexity Criterion

When studying the dependence of a response variable y on a p-dimensional explanatory variable x, a linear model is usually assumed between y and x. Namely, for a sample of independent observations $(x_1^t, y_1), \cdots, (x_n^t, y_n)$ from

(x^t, y), we assume

$$y_i = x_i^t \beta + r_i \tag{1}$$

where β is a p-dimensional unknown parameter and r_i is the error with mean 0 conditional on x_i. Provided that the model (1) is valid, information about the indicated dependence can be obtained from a statistical inference of β based on the data. For validity of the model (1), we usually include in (1) all the explanatory variables available in the first consideration in practice, which results in a so-called full model. The validation of the full model usually can be carried out based on the proper subject knowledge. However, if the full model retains many explanatory variables, its statistical inference is typically inefficient and non-informative. Therefore, a variable selection procedure is indispensable for proceeding with a good regression analysis. With such a procedure, any important explanatory variables should not be missed out, while at the same time no superfluous variables should be included in the model.

Of many attractive information-theoretic approaches, we choose to use the MDL and the associated stochastic complexity. It is formalized by identifying a model with the length of an instantaneously decipherable code which is obtained from an optimal two-step coding scheme determined by this model. For a parametric model, the two-step scheme first encodes the parameter space, then encodes the data for each fixed parameter value. The shortest code length obtained in such a way is called the stochastic complexity of the data relative to the employed model. According to the MDL principle, the smaller the stochastic complexity the better is the corresponding model.

From Rissanen (1996) and Qian and Künsch (1998a) it follows that the stochastic complexity relative to a class of parametric probability densities can be expressed as the minus maximum log-likelihood for the data plus a model complexity term determined by the Fisher information and the maximum likelihood estimator (MLE) of the parameter. This result can be directly applied to a regression model (1) if the ordinary least squares method is used, i.e, the error r_i is given a normal distribution. But the parameter estimation and model selection based on least squares can be seriously affected by one or few outliers in the data. Thus, in robust regression, one only assumes r_i to follow some distribution in an infinite dimensional neighbourhood of the normal. An optimal representation of this neighbourhood is known to be the so-called least favorable distribution (cf. Hampel *et al.*(1986,section 7.4d) and Huber (1964)). When using the least favorable distribution to describe the data, the length of the code constructed will be robust against a radical change of a small portion of the data or a small change in all of the data. Thus, the model selection procedure will also be robust based on the robust code for the data. With

this argument and other ideas underlying the two-step coding scheme, it has been shown that the stochastic complexity of $Y_n = (y_1, \cdots, y_n)^t$ relative to the regression model (1) can be well approximated by

$$SC(Y_n|X_n) = \sum_{i=1}^{n} \rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\ln E\rho_c^{''} + \frac{1}{2}\ln|X_n^t W_n^2 X_n| + \ln\prod_{j=1}^{p}\frac{|\hat{\beta}_j| + n^{-1/4}}{\sigma} \quad (2)$$

plus terms irrelevant to model selection. The technical detail for the derivation of equation (2) can be found in Qian and Künsch (1998b). In equation (2), $\rho_c(t) = \frac{1}{2}t^2$ for $|t| < c$ and $c|t| - \frac{1}{2}c^2$ for $|t| \geq c$ is the Huber function used to prevent the model selection from being heavily affected by outliers in the data, and $\rho_c^{''}(t) = 1$ for $|t| < c$ and 0 for $|t| \geq c$. The constant c in the Huber function is used to adjust the degree of efficiency of the associated robust estimation procedure. The expectation $E\rho_c^{''} = (2\Phi(c) - 1)/(2\Phi(c) - 1 + 2c^{-1}\phi(c))$, where Φ and ϕ are respectively the cumulative distribution and the density function of standard normal, is obtained by taking the expectation with respect to the least favorable distribution for the error term in equation (1). In addition, $X_n = (x_1, \cdots, x_n)^t$ is an $n \times p$ design matrix, $W_n = diag(w_1, \cdots, w_n)$ with $w_i = w(x_i) \in (0, 1]$ a weight function measuring the outlyingness of x_i, and σ measures the scale of $w(x_i)r_i$. The M-estimator $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_p)$ is defined by

$$\hat{\beta} = \arg\min_{\gamma}\sum_{i=1}^{n}\rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\gamma)\}. \quad (3)$$

It can be shown that $\hat{\beta}$ is also the MLE relative to the least favorable distribution. Since the objective is to select an optimal model, those irrelevant terms in the stochastic complexity can be removed.

Note that each term in (2) has a clear interpretation. The first term in (2) is the sum of the robustified fitting errors which shows the goodness of robust fit to the observations. It will decrease if additional explanatory variables are included in the model. This implies that the more explanatory variables are included in (1) the shorter is the code length for encoding the data. But the stochastic complexity also depends on other terms in (2) representing the model complexity. The second term gives the cost of using a robust method, which is 0 if $c = +\infty$ and negative otherwise. Note that $c = +\infty$ corresponds to the least squares method which is non-robust. Thus a robust method is preferred. The third term gives the weighted magnitude of the explanatory

variables and the last one the generalized signal-to-noise ratio. Therefore, the model complexity in (2) is much more comprehensive than that in many other criteria, e.g. AIC, BIC and Mallows' C_p, where it depends only on the dimension of the parameter. One can also see that the model complexity in (2) depends on the Fisher information $I_n(\beta) = \sigma^{-2}(E\rho_c^{''})X_n^t W_n^2 X_n$.

The expression (2) has to be modified to be invariant. Qian and Künsch (1998b) proposed the following modification

$$SC^{'}(Y_n|X_n) = \sum_{i=1}^{n} \rho_c\{\frac{w_i}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\ln E\rho_c^{''} + \frac{1}{2}\ln|X_n^t W_n^2 X_n| + \ln\prod_{j=2}^{p}\left(\frac{|\hat{\beta}_j|}{\sigma} + s_{x(j)}^{-1}n^{-1/4}\right), \quad (4)$$

where $s_{x(j)}^2 = (\sum_{i=1}^n w_i^2)^{-1}\sum_{i=1}^n w_i^2(x_{ij} - \bar{x}_{\cdot j})^2$ and $\bar{x}_{\cdot j} = (\sum_{i=1}^n w_i^2)^{-1}\sum_{i=1}^n w_i^2 x_{ij}$. The quantity $s_{x(j)}^2$ can be regarded as an estimate for the variance of the j-th component of x. Assuming that $x_{i1} \equiv 1$, i.e. the regression contains an intercept, and that the p components of x are linearly independent and $w(x)$ is invariant, it can be shown that $SC^{'}(\cdot)$ is invariant under both scale and shift transformations of y and x.

Suppose that the regression model (1) is the full model under consideration, the set of all candidate models can be identified with $\mathcal{A} = \{\alpha :$ any non-empty subset of$\{1, \cdots, p\}\}$ or a subset of $\mathcal{A}$. Each α in $\mathcal{A}$ corresponds to a sub-model of (1) which contains those components of x indexed by α, and vice versa. Based on (4), we propose the following model selection procedure:

1. For each candidate model $\alpha \in \mathcal{A}$, compute $SC^{'}(Y_n|X_{\alpha n})$, where $X_{\alpha n}$ consists of those columns of X_n indexed by α.

2. Select the model α^* which minimizes $SC^{'}(Y_n|X_{\alpha n})$ among all candidate models in $\mathcal{A}$.

By an asymptotic expansion for $SC^{'}(Y_n|X_{\alpha n})$, it can be found, under some very general regularity conditions, that the stochastic complexity (4) for a model that incorrectly describes the dependence between y and x exceeds that for a correct model by a term of order $O(n)$ with probability 1; and the stochastic complexity for a correct model exceeds that for the simplest correct model by a term of order $O(\log n)$ with probability 1. Therefore, the proposed procedure above selects with probability 1 the simplest model of those in $\mathcal{A}$ which correctly describes the dependence between y and x. We refer to Qian and Künsch (1998b) for a rigorous proof of this result. In addition, it can be

shown using section 6.3 of Hampel et al. (1986) that the above procedure is robust with bounded influence against outliers of both y and x provided that the weight function $w(x)$ is properly chosen.

3 Computing the Stochastic Complexity

To compute the stochastic complexity (4), we must be able to compute $\hat{\beta}$ and σ. In addition, we should have a procedure for choosing the weight function $w(x)$ and the tuning parameter c.

Computing the M-estimator $\hat{\beta}$. From (3) it follows that $\hat{\beta}$ is the solution of

$$\sum_{i=1}^{n} \frac{w_i}{\sigma} \psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}x_i = 0, \tag{5}$$

where $\psi_c(t) = \rho_c^{'}(t) = t$ for $|t| < c$ and $c \cdot sign(t)$ for $|t| \geq c$. Define $u_i = w_i^2 v_i$ with $v_i = \psi_c\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}/\{\frac{w_i}{\sigma}(y_i - x_i^t\beta)\}$. The equation (5) is equivalent to

$$\frac{1}{\sigma^2}\sum_{i=1}^{n} u_i(y_i - x_i^t\beta)x_i = 0. \tag{6}$$

It follows from (6) that

$$\hat{\beta} = (\sum_{i=1}^{n} u_i x_i x_i^t)^{-1}(\sum_{i=1}^{n} u_i y_i x_i). \tag{7}$$

So $\hat{\beta}$ can be computed with a recursive procedure provided that σ, w_i's and c are given. Namely, starting from an initial value of β, we compute the weights u_i's, then compute a new value of β from (7). Continue this process until the difference between two successive computations is negligible. The above procedure is referred to be the iteratively reweighted least squares(IRLS) method. By Huber (1981, section 7.8) it can be shown that the IRLS method used here is convergent provided that the design matrix X_n has full rank.

Computing an estimator of σ. The scale parameter σ is treated as a nuisance parameter in our selection procedure. It could be estimated differently for each candidate model considered. But this way entails encoding the parameter σ and including its code length in formulating the stochastic complexity (4), which is not the case for our approach. Thus we are apt to a simpler way

to estimate σ from the full model and to use the same estimate for all the candidate models. This will also ensure a desirable property that the accumulated robust fitting error, i.e. the first term of (4), decreases as additional explanatory variables are included in the model. Usually, a robust estimate of σ can be obtained by using essentially Huber's proposal 2 (Huber 1981, p.137) or Hampel's median absolute deviation (Hampel 1974, p.388). Using the former method, $\hat{\sigma}$ is the solution of the equation

$$\sum_{i=1}^{n} \psi_c^2\{\frac{w_i}{\sigma}(y_i - x_i\hat{\beta})\} = (n-p)\gamma(c). \tag{8}$$

where $\gamma(c) = 2\Phi(c) - 1 - 2c\phi(c) + 2c^2(1-\Phi(c))$ is chosen in such a way that it is the expectation of the left hand side of (8) if $w_i(y_i - x_i\beta) = w_i r_i$ has a $\mathcal{N}(0,\sigma)$ distribution. Using the v_i's defined above, the equation (8) can again be solved by a convergent recursive method. When using Hampel's method, σ is estimated by

$$1.4826 \times \text{median}_i\{w_i(y_i - x_i\hat{\beta})\}.$$

Choosing the weight function $w(\cdot)$. Ideally $w(x)$ should be determined by a model which correctly describes the dependence between y and x. But whether a model is correct or not is unknown before proceeding with the model selection. In addition, the penalty of using a wrong model for determining $w(x)$ is not given in the criterion (4). Due to these facts, we suggest that $w(x)$ be determined based on the full model. Based on the full model, we proposed that

$$w(x) = w_b(x^tBx) \quad \text{where } w_b(t) = \min(1, \tfrac{b}{\sqrt{t}}) \tag{9}$$

with b chosen a priori (e.g. $b = p$) and B a positive definite matrix determined by

$$\frac{2\Phi(c) - 1}{2\Phi(c) - 1 + 2c^{-1}\phi(c)} \frac{1}{n} \sum_{i=1}^{n} w_b(x_i^tBx_i)^2 x_i x_i^t = B^{-1}. \tag{10}$$

By using (9) and (10), the M-estimator $\hat{\beta}$ possesses a robustness property called the bounded self-standardized sensitivity. The expression (9), such a form is often used for the weight function in robust statistics, implies that the influence of x will be weighted down if x^tBx is larger than a given value b. Clearly, the matrix B can be computed with a recursive procedure once b and c are fixed. But this procedure may not converge since the solution B of (10) may not exist or may be multiple. Empirical study shows that the procedure is convergent if b is large enough, but all w_i's equal 1 if b is too large. A further investigation is needed for this problem.

Choosing the tuning parameter c. The smaller the parameter c is, the more robust is the model selection procedure, but at the same time the procedure is also less efficient. We will choose the well-known value 1.345 for c so that $\hat{\beta}$ has efficiency 0.95 when r_i follows a normal distribution. See Huber (1981,p.91) and Hampel et al. (1986, p.399) for detail.

4 Software for implementing the stochastic complexity criterion

The `S` language (Becker, Chambers and Wilks, 1988) provides a very flexible environment for analyzing data. We have written a package of `S` functions, called **`msrob`**, for the robust regression model selection using the stochastic complexity criterion and some other related criteria. There are two key functions in this package: **`xrlm.select`** and **`xrlm`**. The function **`xrlm.select`** is used to select the optimal regression model by one of the following four criteria: stochastic complexity, Ronchetti's robust AIC (Ronchetti, 1985), Hampel's robust AIC (Hampel, 1983) and robust BIC (Machado, 1993). The function **`xrlm`** is used to fit a robust regression model according to (3). The package **`msrob`** can be obtained free of charge via the World Wide Web address **`http://lib.stat.cmu.edu/S/msrob`** or by sending an e-mail message containing the text "send **`msrob`** from `S`" to **`statlib@stat.cmu.edu`**.

5 Simulation and Example

Simulation results. We carried out a simulation study to evaluate the robustness performance of our stochastic complexity criterion. For purpose of comparison, results for three other criteria were also obtained. The three criteria are the two versions of the robust AIC given by Ronchetti(1985) and Hampel(1983) and the robust BIC by Machado (1993). In the study we considered $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + r$ as the full model. So there were in total $2^6 = 64$ possible sub-models with an intercept term. The sample size n was chosen to be 30. The six explanatory variables X_1 to X_6 were generated independently and uniformly on $[0, 1]$ except that the first observation of each X_i was 3 and the second was 5. Thus, the first two sample points were leverage points and they had large influence on the regression procedure. Six distributions for the error r were chosen to represent various deviation from normality. They are standard normal $\mathcal{N}(0, 1)$, student's t with 3 degrees of freedom, Cauchy ($t_{(1)}$), log-normal with mean 0 and scale 1 which is asymmetric, slash which is a standard normal divided by a uniform on $[0, 1]$, and contaminated ε-normal $0.9\mathcal{N}(0, 1) + 0.1\mathcal{N}(0, 3)$. The observations

Table 1: Frequencies of Different Models Being Selected in 200 Simulations

Model Category	Error Distribution $\mathcal{N}(0,1)$	$t_{(3)}$	Cauchy	Log-N(0,1)	Slash	ε-N
Stochastic Complexity Criterion						
True	143	117	30	129	7	135
Other correct	55	49	9	44	5	50
Incorrect	2	34	161	27	188	15
Ronchettis's Robust AIC						
True	118	111	42	122	20	122
Other correct	81	72	17	64	8	70
Incorrect	1	17	141	14	172	8
Hampel's Robust AIC						
True	126	116	40	127	17	129
Other correct	72	64	16	55	7	63
Incorrect	2	20	144	18	176	8
Machado's Robust BIC						
True	168	134	28	151	6	156
Other correct	27	19	7	15	2	27
Incorrect	5	47	165	34	192	17

of Y were obtained from

$$Y = 1 + 2.5X_1 + 3X_2 - 3X_3 + r \tag{11}$$

with r generated from one of the six error distributions. The coefficient values were so selected that they would give t-values of about 4 if r were normally distributed. It is clear that the model (11) is the true model. But other models containing X_1, X_2 and X_3 are also correct models. We carried out 200 simulation runs. Table 1 gives the frequencies of selecting the three types—true, other correct and incorrect—of models by each of the four criteria.

From Table 1 we see all the four criteria perform quite well even when the error distribution is considerably deviated from normal (i.e. $t_{(3)}$, log-normal and ε-normal). The relative frequencies of selecting the true model is between 55.5% and 78% for these three error distributions. (Compare with 59% and 84% for the normal error.) They are between 4% and 23.5% in selecting the incorrect models. But when the error distribution is Cauchy or slash, neither of the criteria works well in selecting the correct models. This is probably because Cauchy and slash deviate so much from normal that their population

expectations do not exist. Thus a more robust and efficient procedure would be required for this situation. When comparing these four criteria with each other, we see the AIC methods usually have lower frequencies of selecting the true model but higher frequencies of selecting other superfluous correct models than the other two criteria. The stochastic complexity criterion may have little lower frequencies of selecting the true model than the BIC method, but it also has lower frequencies of selecting the incorrect models so has a more stable performance. Since in practice one generally does not know which candidate model is exactly the true model, to reduce the chance of selecting an incorrect model is as important as to enhance the chance of selecting the true one. From this point of view we would prefer the stochastic complexity criterion to the robust BIC. Actually a further simulation study by us reveals that the stochastic complexity method performs more stable than the robust BIC especially when the β values in the true model have more moderate t-values mentioned above, namely, when the signal-to-noise ratio becomes weaker.

An actual example. To illustrate the application to practical problems for our proposed method, we present a real data example arising in a physiology study of triathlon athletes. The data used in this example were taken from Kohrt et al. (1987) who studied the performance of a group of 65 male athletes in half-triathlon event over a 6-week period. The data can also be found in Glantz and Slinker (1990, pp.647-648). There are 10 variables in the data: half-triathlon performance time (t min.), age (A years), weight (W kg.), years triathlon experience (E years), amount of training running (T_R km/week), biking (T_B km/week), and swimming (T_S,km/week), and maximum oxygen consumption while running (V_R mL/min/kg), biking (V_B mL/min/kg), and swimming (V_S mL/min/kg). These 10 variables represent the athletes' half-triathlon performance, gross physical characteristics, training, and exercise capacity.

The objective of the study is to see which variables determine best the athletes' final time when they compete in the triathlon. This was addressed by conducting a variable selection on the full regression model

$$t = \beta_0+\beta_1 A+\beta_2 W+\beta_3 E+\beta_4 T_R+\beta_5 T_B+\beta_6 T_S+\beta_7 V_R+\beta_8 V_B+\beta_9 V_S+r. \quad (12)$$

We applied to the variable selection the stochastic complexity criterion as well as Ronchetti's and Hampel's robust AIC and Machado's robust BIC. There were in total $2^9 = 512$ sub-models for selection if only considering those including an intercept term. Table 2 lists the 8 best sub-models selected from these 512 models by each of the four criteria. In the table, each set of the 8

Table 2: Eight Best Models Selected by Each Criterion in the Example

Stochastic Complexity	Ronchetti's Robust AIC
$A + E + T_R + T_B + V_R$	$A + E + T_R + T_B + V_R$
$A + E + T_R + T_B + V_R + V_B$	$A + E + T_R + T_B + V_R + V_B$
$A + E + T_B + V_R$	$A + E + T_R + T_B + T_S + V_R$
$A + E + T_R + T_B + T_S + V_R$	$A + E + T_R + T_B + V_R + V_S$
$A + E + T_R + T_B + V_R + V_S$	$A + E + T_R + T_B + T_S + V_R + V_B$
$A + E + T_S + V_R + V_B$	$A + W + E + T_R + T_B + V_R$
$A + W + E + T_R + T_B + V_R$	$A + E + T_R + T_B + V_R + V_B + V_S$
$A + E + T_B + V_R + V_B$	$A + W + E + T_R + T_B + V_R + V_B$
Hampel's Robust AIC	**Machado's Robust BIC**
$A + E + T_R + T_B + V_R$	$A + E + T_R + T_B + V_R$
$A + E + T_R + T_B + V_R + V_B$	$A + E + T_B + V_R$
$A + E + T_R + T_B + T_S + V_R$	$A + E + T_R + T_B + V_R + V_B$
$A + E + T_R + T_B + V_R + V_S$	$A + E + T_R + T_B + T_S + V_R$
$A + W + E + T_R + T_B + V_R$	$A + E + T_R + T_B + V_R + V_S$
$A + E + T_R + T_B + T_S + V_R + V_B$	$A + W + E + T_R + T_B + V_R$
$A + E + T_R + T_B + V_R + V_B + V_S$	$E + T_S + V_R + V_B$
$A + W + E + T_R + T_B + V_R + V_B$	$A + E + T_S + V_R + V_B$

models is displayed according to the associated criterion values in an ascending order.

From Table 2 we see that all the criteria selected the same best model which includes the five explanatory variables A, E, T_R, T_B and V_R. These five variables are also included in most of the other 28 models. However, each of the other four explanatory variables appears only small number of times in these models. This conclusion is the same as that by Glantz and Slinker (1990, pp. 256-261) who used the Mallows' C_p criterion. From Table 2 we can also see that the robust AIC methods tend to select more complicated models while the robust BIC tends the opposite way. The stochastic complexity method gives an improvement over the robust BIC.

References

1. Baxter, R.A. and Dowe, D.L. (1996). *Model selection in linear regression using the MML criterion.* Technical Report 96/276, Dept. of Computer Science, Monash Univ., Melbourne, Australia.

2. Becker, R., Chambers, J.M. and Wilks, A. (1988). *The New S language.* Wadsworth, Belmont CA.
3. Dom, B.E. (1996). *MDL estimation for small sample sizes and its application to linear regression.* IBM Research Report RJ-10030. June 1996.
4. Glantz, S.A. and Slinker, B.K. (1990).*Primer of Applied Regression and Analysis of Variance.* McGraw-Hill, Inc., New York.
5. Hampel, F.R. (1974). The influence curve and its role in robust estimation. *J. Am. Statist. Assoc.* **69** 383-393.
6. Hampel, F.R. (1983). Some aspects of model choice in robust statistics. *Proceedings of the 44th Session of ISI, Book 2*, Madrid, 767-771.
7. Hampel, F.R., Ronchetti, E. M.,Rousseeuw, P. J. and Stahel, W. A. (1986).*Robust Statistics: The Approach Based on Influence Functions.* Wiley, New York.
8. Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73-101.
9. Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.
10. Kohrt, W.M., Morgan, D.W., Bates, B. and Skinner, J.S. (1987). Physiological responses of triathletes to maximal swimming, cycling, and running. *Med. Sci. Sports Exerc.* **19**, 51-55.
11. Machado, J.A.F. (1993). Robust Model Selection and M-estimation. *EconTher.* **9**, 478-493.
12. Qian, G., and Künsch, H. (1998a). Some notes on Rissanen's stochastic complexity. *IEEE Trans. Inform. Theory.* **44**, 782-786.
13. Qian, G., and Künsch, H. (1998b). On model selection in robust linear regression. *J. Stat. Plan. & Infer.*, in press.
14. Rissanen, J. (1986). Stochastic complexity and modeling. *Annals of Statistics*, **14**, 3, 1080–1100.
15. Rissanen, J. (1987). Stochastic complexity. *J. Roy. Statist. Soc. B* **9** 223-239 and 252-265 (discussions).
16. Rissanen, J. (1996). Fisher information and stochastic complexity, *IEEE Trans. Inform. Theory.* **42**, 40-47.
17. Ronchetti, E. (1985). Robust model selection in regression. *Stat. Prob. Lett.* **3** 21-23.
18. Solomonoff, R.J. (1964). A formal theory of inductive inference I, II. *Information and Control* **7**, 1-22 and 224-254.
19. Wallace, C.S. and Freeman, P.R. (1987). Estimation and inference by compact coding. *J. Roy. Statist. Soc. B* **9** 240-251 and 252-265 (discussions).