On Model Selection in Robust Linear Regression

by

Guoqi Qian [1] [2]
and
Hans R. Künsch

Research Report No. 80
November 1996

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

---

[1] Research supported by Swiss National Science Fund 2100-043210.95.

[2] Seminar für Statistik, ETH, and School of Statistics, La Trobe University, Melbourne 3083, Australia

On Model Selection in Robust Linear Regression

Guoqi Qian [‡§]
and
Hans R. Künsch
Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

November 1996

Abstract

Several model selection criteria which generally can be classified as the penalized robust method are studied in this paper. Particularly we derive a criterion based on Rissanen's stochastic complexity. Some asymptotic properties concerning strong consistency of selecting the optimal model by these criteria are given under general conditions. Other features like robustness against outliers and effect of signal-to-noise ratio are also discussed. Finally, examples and simulations are presented to evaluate their finite sample performance. The robust procedure used in this paper considers the gross error in both the response and the independent variables through a generalized Huberization.

Key words and phrases: model selection, robust regression, stochastic complexity, consistency, AIC, BIC, cross validation, signal-to-noise ratio.

1 Introduction

Choosing a model is often an important goal of a statistical analysis. Clearly, exploratory investigations and subject knowledge are important parts for the formulation of model classes. But once the possible model classes have been set up, it is most useful to have an efficient and objective criterion to derive model selection from the data alone. In this paper we study the model selection problem in robust linear regression. Compared with the rich literature on robust estimation and testing, only few papers have been devoted to this so far. See Ronchetti (1985), Hampel (1983), Machado (1993), Ronchetti and Staudte (1994), and Ronchetti, Field and Blanchard (1996) where robust versions of AIC, BIC, C_p and cross-validation are proposed.

It is widely known that in linear regression model selection based on least squares, AIC and the cross-validation method with a fixed size of the associated validation sample are not consistent in selecting the true model if it can be finitely parameterized: they tend to choose a model with too many independent variables. On the other hand, BIC and the cross-validation method with an increasing validation sample size (cf. Shao (1993)) are consistent if a true model with a finite number of independent variables exists. However, the penalty term in both BIC and AIC depends only on the number of independent variables. They ignore for instance how large the contribution of an independent variable to the predictor is and how closely related the independent variables are.

We think that when proposing a model selection procedure in robust linear regression, we should consider at least three issues. First, the criterion proposed should take into account the possibility that both response and predictors of some observations may contain gross errors. Therefore the criterion should not choose a complicated model in order to fit also a small number of outliers. The second issue is that the criterion proposed should be consistent if a finite dimensional true model exists, and it should possess some asymptotic optimality properties if the true model is infinite dimensional. The third one concerns the effect of the signal-to-noise ratio on the empirical performance of the criterion. Here, the effect of the signal-to-noise ratio means that a model selection criterion is not likely to pick out a term in a regression model whose coefficient is relatively small compared to the dispersion parameter of the model.

In this paper, we consider model selection procedures based on penalizing robust fitting errors. We will call them penalized robust deviance procedures. They include many of the procedures considered previously as well as a new one based on the ideas of stochastic complexity of Rissanen (1983, 1989, 1996). In order to describe them, we introduce the following framework.

Suppose the observations $(x_1^t, y_1), \cdots, (x_n^t, y_n)$ with $x_i \in \mathcal{R}^p$ and $y_i \in \mathcal{R}$ are i.i.d. following an unknown distribution $F_\beta(x, y)$, such that the usual regression model is satisfied:

$$y_i = x_i^t\beta + r_i \text{ with } E(r_i|x_i) = 0. \tag{1.1}$$

Thus $\mu_i = E(y_i|x_i) = x_i^t\beta$, and the r_i's are i.i.d. with $E(r_i) = 0$ and $cov(x_i, r_i) = 0$. Note that we do not assume that r_i is independent of x_i. In particular, the variance of r_i may depend on x_i. For simplicity of presentation, we assume that the p components of x contains all independent variables available in the data so that many components of β may be zero. Then (1.1) also gives the full model. Clearly the set of all possible models can be identified with $\mathcal{A} = \{\alpha : \text{any non-empty subset of } \{1, \cdots, p\}\}$. Each α, of size p_α, in $\mathcal{A}$ corresponds to a predictor $x_\alpha^t\beta_\alpha$ and vice versa. Here x_α is a sub-vector of x indexed by α and β_α is similarly defined. Given a vector β, $\mathcal{A}$ can be divided into two subsets:

1. $\mathcal{A}_c = \{\alpha : \beta_i = 0 \quad \text{for any} \quad i \notin \alpha\}$;
2. $\mathcal{A}_w = \{\alpha : \beta_i \neq 0 \quad \text{for some} \quad i \notin \alpha\}$.

Clearly $\mathcal{A}_w$ represents all the wrong predictors and $\mathcal{A}_c$ represents all the correct predictors. But many models in $\mathcal{A}_c$ include irrelevant variables and thus are too complicated. The optimal model is defined as a correct model in $\mathcal{A}_c$ with the smallest dimension. For simplicity we assume that such an optimal model is unique. It is easy to see that this is the case if the components of x are linearly independent.

For the above setup, we study procedures to select a predictor of the following form

$$\hat{\alpha} = \arg\min_{\alpha} \left\{ \sum_{i=1}^{n} \rho\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} + C(n, \alpha) \right\}. \tag{1.2}$$

Here $C(n, \alpha)$ is a penalty term measuring the complexity of a model α, $w(x) \in (0, 1]$ is a weight function measuring the outlyingness of the independent variable x, σ measures the scale of $w(x_i)r_i$ and $\hat{\beta}_\alpha$ is the M-estimator corresponding to the robust deviation $\rho(\cdot)$. It is defined by

$$\hat{\beta}_\alpha = \arg \min_{\gamma \in \mathcal{R}^{p_\alpha}} \sum_{i=1}^{n} \rho\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\gamma)\}. \tag{1.3}$$

Choosing this estimator implies that the first term in (1.2) is the robust fitting error of a model α. In particular, it guarantees that this robust fitting error decreases when additional independent variables are included in the model. The problem is then to choose a penalty term $C(n, \alpha)$ as the model complexity so that the resulting model selection criterion (1.2) performs satisfactorily with respect to the three issues we just discussed.

In practice, σ will be estimated with a robust estimate obtained by using essentially Huber's proposal 2 (Huber 1981, p.137) or Hampel's median absolute deviation (Hampel, 1974, p.388) in the full model. The weights $w(x_i)$ will also be based on the full model, see section 3 below. The robust function $\rho(\cdot)$ used in this paper will be Huber's function defined as

$$\rho_c(t) = \begin{cases} \frac{1}{2}t^2, & |t| < c \\ c|t| - \frac{1}{2}c^2, & |t| \geq c \end{cases} .$$

This function was chosen not only because of the minimax property of the associated least favorable distribution for the gross error model (cf. Theorem 1 and its corollary of Huber(1964)), but also because it allows an information theoretic interpretation for one of the criteria derived later. From (1.3) it follows that the M-estimator $\hat{\beta}_\alpha$ is also the MLE in a contaminated normal model with heteroscedastic errors whose scale depends on the independent variables, compare (2.8) below.

In section 2 we will discuss different choices of the penalty term and their relationship with robust versions of AIC, Mallows C_p, cross-validation and BIC methods. One particular choice will be derived using the stochastic complexity theory of Rissanen (1989, 1996). This will give a more general and precise formulation comprising such features of the model as robustness, design matrix and signal-to-noise ratio. In section 3 we study the robustness property for the M-estimator $\hat{\beta}_\alpha$ and selection of the weight function $w(\cdot)$. In section 4 we provide some asymptotic properties of the penalized robust deviance criterion in terms of selecting correct models and the optimal model. Only very weak conditions are required for these results to be true. In section 5 other aspects of our model selection procedure like the effect of letting one response observation go to $\pm\infty$ and the breakdown property are discussed. It is found that the penalized robust deviance criterion behaves well in the presence of outliers. We further present a simulation study in section 6 to show the effectiveness of our robust procedure. Finally, all the proofs are given in the appendix.

2 The Penalty Term and Model Complexity

2.1 Robust AIC, BIC, Mallows C_p and Cross-Validation

A basic difficulty in model selection comes from the fact that by using more and more complex models, the fit for the available data improves, but the predictive power for future data gets worse after some point. Clearly, the robust fitting error $\sum_{i=1}^n \rho_c\{w(x_i)(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)/\sigma\}$ will decrease if additional independent variables, including those with corresponding β components equal to zero, are included in the model. Thus the robust fitting error alone cannot serve as a model selection criterion. When a model is estimated from the available data, whether or not to accept this model should depend on its ability to predict future data. Akaike (1973, 1974) used the relative entropy between the true probability density and the estimated one as a measure of predictability. He showed that an asymptotically unbiased estimator of an essential part of this relative entropy can be obtained as the negative log-likelihood plus a penalty term equal to the dimension of the parameter in the estimated model. This is the familiar AIC criterion, which is also equivalent to Mallows (1973) C_p statistic. For the ordinary linear regression case, this relative entropy is equivalent to the true error rate of the predictor which is defined and studied by Efron (1983,1986). The above estimator becomes unbiased for the true error rate in this case.

In robust regression the linear model (1.1) is considered with the error r_i following Huber's least favorable distribution, conditional on the independent variables x_i. Applying the idea underlying AIC to this distribution and using an asymptotic equivalence derived by Stone (1977), but taking expectations under the standard normal model, Ronchetti (1985) obtains the following asymptotic unbiased estimator which is a robust version of AIC (Actually $w(x) = 1$ is used in his paper, but the general derivation is basically the same.):

$$RAIC(\alpha) = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} + \frac{E_\Phi \psi_c^2}{E_\Phi \psi_c'} p_\alpha \tag{2.1}$$

where

$$\psi_c(t) = \rho_c'(t) = \begin{cases} -c, & t \le -c \\ t, & |t| < c \\ c, & t \ge c \end{cases}, \quad \psi_c'(t) = \begin{cases} 1, & |t| < c \\ 0, & |t| \ge c \end{cases}$$

and Φ is the standard normal distribution function. Hampel (1983) suggests a different penalty term based on heuristic arguments. His robust version of AIC is as follows:

$$HAIC(\alpha) = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} + \frac{1}{2}\left\{\frac{E_\Phi \psi_c^2}{E_\Phi \psi_c'} + \frac{E_\Phi \psi_c^2}{(E_\Phi \psi_c')^2}\right\} p_\alpha. \tag{2.2}$$

The criteria (2.1) and (2.2) can also be regarded as robust versions of Mallows C_p statistic due to their equivalence for $c = \infty$. In section 4 we will prove that under quite general conditions

$$\sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma} r_i\} - |O(\log\log n)| \quad \text{a.s.}$$

if α is a correct model. This implies that the robust AIC or C_p as defined by (2.1) or (2.2) is likely to select a model with superfluous variables, since the penalty term $C(n, \alpha)$ here is of the form $const \cdot p_\alpha$ which is smaller than $|O(\log\log n)|$ in magnitude.

In order that the selected model does not overfit the data, the penalty term $C(n, \alpha)$ should be an increasing function with respect to p_α and must be greater than $O(\log\log n)$ in magnitude. Schwarz (1978) uses a Bayesian approach to derive a penalty term of the form $\frac{1}{2} p_\alpha \log n$ for general parametric models. The resulting model selection criterion is usually called BIC or SIC. Machado (1993) derives a robust version of BIC based on objective functions defining M-estimators for a parametric model. Its special case called Huber SIC for robust regression model is defined by

$$RBIC(\alpha) = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} + \frac{1}{2} p_\alpha \log n \tag{2.3}$$

with $w(x) = 1$. It is an obvious extension if we also consider robustness against outliers in the independent variables by allowing $w(x) \in (0, 1]$. While we will see later that the robust BIC criterion defined by (2.3) is consistent in selecting the optimal model, it is a bit unsatisfactory that the penalty term is simply determined by the dimension of the predictor and the sample size. This form cannot provide detailed information about the effect of the selected model on the predictability of future data. Below we will present a penalty term with more comprehensive information about the model, based on a newly developed theory of stochastic complexity (Rissanen (1989, 1996), Qian and Künsch (1996)).

Cross-validation gives a different estimator for the predictability of the selected model on future data. A robust version of cross-validation statistic can be defined as

$$RCV(\alpha) = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha^{(i)})\} \tag{2.4}$$

where $\hat{\beta}_\alpha^{(i)}$ is the M-estimator based on all the observations except (x_i^t, y_i). We will not study this criterion rigorously in this paper. However, by the heuristics of the influence function IF (Hampel, 1974)

$$\hat{\beta}_\alpha - \hat{\beta}_\alpha^{(i)} \approx \frac{1}{n}(IF(x_i, y_i) - \frac{1}{n-1}\sum_{j \neq i} IF(x_j, y_j)),$$

we expect that asymptotically

$$\sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha^{(i)})\} = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)\} + O(1) \quad \text{a.s.} \tag{2.5}$$

if α is a correct model and the M-estimator has a bounded influence. Thus cross-validation is expected to behave similarly as the robust AIC. For recent work on robust regression model selection by cross-validation, we refer to Ronchetti, Field and Blanchard (1996).

2.2 A Stochastic Complexity Criterion

Stochastic complexity assesses the fit of statistical models by their ability to compress the data. This is measured by the length needed to encode the data by the instantaneously decipherable code which is optimal for a model. The associated principle of minimum description length states that the shorter the code length the better is the selected model. The optimal codes are obtained in two steps: In the first step one encodes the parameters of the model and in the second step one encodes the data conditional on the employed parameter. The shortest code length obtained in a model class is called the stochastic complexity of the data relative to this model class.

For a class of parametric densities $\mathcal{M} = \{f(z^n|\theta)\}$ where $\theta \in \Theta \subset \mathcal{R}^k$, Qian and Künsch (1996) obtained the following approximation of the stochastic complexity of z^n relative to $\mathcal{M}$

$$SC(z^n|\mathcal{M}) = -\log f(z^n|\hat{\theta}) + \frac{k}{2}\log \rho_{n1} + \frac{1}{2}\log |I_n(\hat{\theta})| + \sum_{i=1}^{k} \log(|\hat{\theta}_i| + n^{-1/4}). \tag{2.6}$$

Here $\hat{\theta}$ is the maximum likelihood estimate of θ;

$$I_n(\theta) = E_\theta \left[\frac{-\partial^2 \log f(z^n|\theta)}{\partial\theta\partial\theta^t} \right]$$

is the expected Fisher information matrix and ρ_{n1} is the maximal eigenvalue of $I_n(\hat{\theta})^{-\frac{1}{2}} J_n(z^n)$ $I_n(\hat{\theta})^{-\frac{1}{2}}$ where

$$J_n(z^n) = -\frac{\partial^2 \log f(z^n|\theta)}{\partial\theta\partial\theta^t}|_{\theta=\hat{\theta}}$$

is the observed Fisher information. The derivation of (2.6) is motivated by Rissanen (1983,1996) which showed the important role played by the Fisher information in determining the optimal quantization of the parameter space. The first term of (2.6) is the optimal code length of the data for a chosen member in $\mathcal{M}$. The remaining terms are called model complexity. They are obtained by encoding the parameter θ to a certain precision by the optimal quantization.

The result (2.6) can be readily extended to the regression model. For the regression model (1.1), the stochastic complexity of $Y_n = (y_1, \cdots, y_n)^t$ relative to $X_n = (x_1, \cdots, x_n)^t$ can be obtained from (2.6) by using the conditional distribution of Y_n given X_n and $\theta = \beta$. However, the stochastic complexity obtained in this way is not useful in practice for selecting an optimal predictor, since it depends on the unknown (conditional) density function of the errors r_i. We could try to estimate this density either parametrically or nonparametrically. But then we would have to include additional penalty terms representing the complexity of the model estimation for the distribution of the errors. We prefer a simpler procedure where we employ Huber's least favorable distribution instead of the unknown true distribution. This means that we do not attempt to find the shortest possible encoding of the data. But this is presumably not necessary for the purpose of finding the best predictor. Note that the expression (2.6) is an approximate code length for arbitrary data z^n, not only for those which are typical under a model distribution. The possibility that the data are not generated by a model distribution is taken into account by computing ρ_{n1} in (2.6).

The essential point when choosing a (conditional) error distribution is that the corresponding MLE should be robust because otherwise there is no chance to obtain a robust model selection procedure. In particular, assuming a (conditional) normal distribution is not possible. We choose Huber's least favorable distribution because of its minimax property (Huber (1964)) although other choices might be possible. It is likely that this choice leads to a code which is not much longer than the optimal code for a wide range of error distributions which might underly the data. In order to obtain protection against outliers in the independent variable x, we have to let the scale of the least favorable distribution depend on x. We consider the following density of r_i given x_i:

$$f(r_i) = (1-\lambda)(\sqrt{2\pi}\sigma)^{-1} w(x_i) \exp\{-\rho_c(\frac{w(x_i) r_i}{\sigma})\} = \frac{w(x_i)}{\sigma} f_0(\frac{w(x_i) r_i}{\sigma}), \tag{2.7}$$

where $f_0(r) = (1-\lambda)(\sqrt{2\pi})^{-1}\exp\{-\rho_c(r)\}$, $0 < \lambda < 1$, $\rho_c(\cdot)$ is the Huber function and $\psi_c(\cdot)$ its derivative. The constants λ and c are connected by $(1-\lambda)^{-1} = 2\Phi(c) - 1 + 2\phi(c)/c$ where ϕ is the density of the standard normal distribution function Φ. It easily follows that $E(r_i|x_i) = 0$ and

$$var(r_i|x_i) = \frac{\sigma^2}{w^2(x_i)}(1-\lambda)(2\Phi(c) - 1 + 4(\frac{1}{c} + \frac{1}{c^3})\phi(c))\}.$$

Thus, r_i has a (conditional) variance inflated by a factor $1/w^2(x_i)$. The choice of $w(x)$ depends on how the corresponding x acts on the regression estimator. Apparently we want to properly weigh down those x points with large influence. At the time being we assume that $w(x)$ is a function determined by the distribution of x in the full model. We will discuss the choice of $w(x)$ and its estimation from the data in more detail in the next section.

To find the stochastic complexity of the data, we need the log-likelihood for the response observations conditional on X_n, which is by (1.1) and (2.7)

$$\ell(Y_n|X_n, \beta, \sigma) = n\log(1-\lambda) - \frac{n}{2}\log 2\pi - n\log\sigma + \sum_{i=1}^n \log w(x_i) - \sum_{i=1}^n \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_i^t\beta)\}. \tag{2.8}$$

We also need the expected Fisher information matrix $I_n(\beta)$ relative to the least favorable distribution (2.7). To find $I_n(\beta)$, note that

$$\frac{\partial \ell}{\partial \beta} = \frac{1}{\sigma}\sum_{i=1}^{n} \psi_c\{\frac{w(x_i)}{\sigma}(y_i - x_i^t\beta)\}w(x_i)x_i$$

and

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta^t} = -\frac{1}{\sigma^2}\sum_{i=1}^{n} \psi_c'\{\frac{w(x_i)}{\sigma}(y_i - x_i^t\beta)\}w^2(x_i)x_i x_i^t.$$

Thus one can readily verify that

$$I_n(\beta) = -E(\frac{\partial^2 \ell}{\partial \beta \partial \beta^t}) = E(\frac{\partial \ell}{\partial \beta}\frac{\partial \ell}{\partial \beta^t}) = \frac{1}{\sigma^2}(E_{f_0}\psi_c')X_n^t W_n^2 X_n, \tag{2.9}$$

where $W_n = diag(w(x_1), \cdots, w(x_n))$. When there is no risk of confusion, we will write for simplicity $E\psi_c{}'$ instead of $E_{f_0}\psi_c'$. With a little calculation we see that

$$E\psi_c' = E\psi_c^2 = \frac{2\Phi(c) - 1}{2\Phi(c) - 1 + 2c^{-1}\phi(c)}. \tag{2.10}$$

Note that the expectation in (2.9) is taken with respect to the assumed model and not with respect to the unknown true distribution, in accordance with the criterion (2.6). Because ψ_c' is bounded by 1, ρ_{n1} in (2.6) is always bounded by a finite number if $E(||w(x_i)x_i||^2) < \infty$. We thus omit this term in the following.

Denoting the maximum likelihood estimator relative to the (conditional) least favorable distribution (2.7), we obtain from (2.6) to (2.10) the following approximation to the stochastic complexity of Y_n relative to the predictor $X_n\beta$ and the (conditional) least favorable distributions for R_n

$$\begin{aligned} SC(Y_n|X_n) &= -\ell(Y_n|X_n, \hat{\beta}, \sigma) + \frac{1}{2}\log|I_n(\hat{\beta})| + \sum_{i=1}^{p}\log(|\hat{\beta}_i| + n^{-1/4}) \\ &= \sum_{i=1}^{n}\rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_i^t\hat{\beta})\} + \frac{p}{2}\log E_{f_0}\psi_c' \\ &+ \frac{1}{2}\log|X_n^t W_n^2 X_n| + \log\prod_{j=1}^{p}\frac{|\hat{\beta}_j| + n^{-1/4}}{\sigma} \\ &+ \text{terms negligible for model selection.} \end{aligned} \tag{2.11}$$

Now we give some interpretations for (2.11). The first term in (2.11) is the robust fitting error which shows the goodness of the robust regression for fitting the observations. The second to the fourth terms give a cost of using the robust procedure and the employed model, which we call the model complexity. It is interesting to note that because $E\psi_c' < 1$ if c is finite, the second term is negative and thus the robustification of the procedure reduces the whole stochastic complexity. This might seem controversial, but it is virtually in accord with the philosophy of robust statistics: It avoids a precise modeling of the error distribution which would have to be highly complex and impractical. Rather the simple least favorable gross error model is used which gives a good description of the data for many possible error distributions. In addition to being affected by the robustness of the procedure, the model complexity is also dependent on the weighted magnitude of X_n, defined as $|X_n W_n^2 X_n|$, and the (generalized) signal-to-noise ratio $(|\hat{\beta}_j| + n^{-1/4})/\sigma$. It thus allows a more detailed quantification of model complexity than those criteria which just consider the number of parameters. Note that for small values of $|\hat{\beta}_j|/\sigma$ the last term acts as a small bonus for the more complex models which include the j-th component. But including this component increases the penalty in $|X_n W_n^2 X_n|$, and the latter is dominant asymptotically. We know that BIC and cross validation with validation sample size comparable to the total sample size tend to underfit because of their insensitivity to variables with small signal-to-noise ratios (Rissanen (1986), Wei (1992), Qian (1994, Section 3.5)) while AIC, cross validation with small size of the validation sample and Mallows C_p tend to overfit. The use of the last term in (2.12) for model selection, which is derived by a stochastic complexity idea, offers a compromise over the two extremes.

Note that we do not set aside any code length for describing the scale parameter σ. This is because we treat it as nuisance parameter for our model selection purpose. In practice σ can be replaced by a robust estimate using, for instance, Huber's proposal 2 or Hampel's median absolute deviation with a little modification for the full model. Namely, estimate σ by finding a stationary point of

$$\sum_{i=1}^{n} \psi_c^2\{\frac{w(x_i)}{\sigma}(y_i - x_i^t\beta)\} = (n-p)\gamma(c)$$

with respect to σ, where the constant $\gamma(c)$ is chosen for Fisher-consistency of σ at the normal distribution, so $\gamma(c) = E_\Phi \psi_c^2(Z) = 2\Phi(c) - 1 - 2c\phi(c) + 2c^2(1-\Phi(c))$; or estimate σ by $1.4826 \times \text{median}_i\{|w(x_i)(y_i - x_i^t\beta)|\}$ where the constant is obtained again by considering the Fisher-consistency at the normal distribution.

From the above derivation it follows that (2.11) holds for any model α in $\mathcal{A}$. By the principle of minimum description length, a criterion for predictor selection in robust linear regression can be proposed based on the stochastic complexity (2.11): Select the model α so that

$$\begin{aligned} SC(Y_n|X_{\alpha n}, \sigma^2) &= \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} + \frac{p_\alpha}{2}\log E_{f_0}\psi_c' \\ &+ \frac{1}{2}\log|X_{\alpha n}^t W_n^2 X_{\alpha n}| + \log\prod_{j=1}^{p_\alpha}\frac{|\hat{\beta}_{\alpha j}| + n^{-1/4}}{\sigma} \end{aligned} \tag{2.12}$$

is minimized. Here $X_{\alpha n}$ consists of the columns of X_n indexed by α.

2.3 Invariance of the Stochastic Complexity Criterion

A model selection criterion should be invariant under linear transformations of the response and/or some independent variables. We assume that the weights $w(x_i)$ are invariant under linear transformations. This holds for our choices of $w(x)$ we will discuss in the next section.

First we consider the effect of shift transformations of y and/or x. If all models contain an intercept, i.e. $x_{i,1} = 1$, then $\hat{\beta}_2, \cdots, \hat{\beta}_{p_\alpha}$ as well as the quantities $\sum_{i=1}^n \rho_c\{w(x_i)(y_i - x_{\alpha i}^t\hat{\beta})/\sigma\}$ and $|X_{\alpha n}^t W_n^2 X_{\alpha n}|$ are invariant. Hence we obtain an invariant criterion if we drop the factor with $j=1$ in the last term of (2.12).

We assume that under a scale transformation of y, $\hat{\sigma}$ changes accordingly. Then we obtain an invariant criterion if we replace the last term in (2.12) by $\log\prod_{j=2}^{p_\alpha}(|\hat{\beta}_{\alpha(j)}|/\sigma + n^{-1/4})$. Finally, in order to obtain a criterion which is invariant also under scale transformations of x, we replace this term by

$$\log \prod_{j=2}^{p_\alpha} \{ \frac{|\hat{\beta}_{\alpha(j)}|}{\sigma} + const\ s(x_{\alpha(j)})^{-1} n^{-1/4} \},$$

where $s(x_{\alpha(j)})^2$ is a weighted estimate of the variance of the j-th component of x_α:

$$s(x_{\alpha(j)})^2 = \frac{\sum_{i=1}^n w^2(x_i)(x_{\alpha(j)i} - \bar{x}_{\alpha(j)})^2}{\sum_{i=1}^n w^2(x_i)}, \quad \bar{x}_{\alpha(j)} = \frac{\sum_{i=1}^n w^2(x_i) x_{\alpha(j)i}}{\sum_{i=1}^n w^2(x_i)}.$$

Concerning the choice of *const* there is some freedom. One possibility we use here is to take $const = 2(\chi^2_{n-1}(1-\delta)/n)^{1/2}$. The factor 2 is motivated by the fact that $2\sigma/s(x_{\alpha,j})$ is the critical value for testing $\beta_j = 0$ when one uses least squares and only the j-th component of x is in the model. The other factor is motivated by the fact that $s(x_{\alpha,j})^{-1}(\chi^2_{n-1}(1-\delta)/n)^{1/2}$ is the upper bound of a $(1-2\delta)100\%$ confidence interval for the inverse standard deviation of $x_{\alpha,j}$ for normal variables. In practice we usually take δ small, say 0.001,0.005 or 0.05.

From the above considerations we propose a modified criterion

$$SC'(Y_n|X_{\alpha n}\beta_\alpha, \sigma^2) = \sum_{i=1}^n \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)\} + \frac{p_\alpha}{2} \log E\psi_c'$$
$$+ \frac{1}{2}\log|X_{\alpha n}^t W_n^2 X_{\alpha n}| + \log \prod_{j=2}^{p_\alpha} \left(\frac{|\hat{\beta}_{\alpha(j)}|}{\sigma} + 2(\chi^2_{n-1}(1-\delta))^{1/2} s(x_{\alpha(j)})^{-1} n^{-3/4} \right). \qquad (2.13)$$

Under the assumption that in all our models $x_{\alpha(1)i} = 1$ for all $i = 1, \cdots, n$, $S'(\cdot)$ is invariant under both scale and shift transformations of x and y.

However, neither $SC(\cdot)$ nor $SC'(\cdot)$ are invariant under orthogonal transformations of x_α, since $\prod_{j=2}^{p_\alpha} \frac{\hat{\beta}_{\alpha(j)}}{\sigma}$ is not invariant under such a transformation even though $\sum_{i=1}^n \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)\}$ and $|X_{\alpha n}^t W_n^2 X_{\alpha n}|$ are. One may argue whether this is an unpleasant feature of the stochastic complexity criterion or not. In most model selection problems, each independent variable has a specific meaning and there is no physical reason to consider linear combinations of these variables.

3 Robustness of $\hat{\beta}_\alpha$ and choice of $w(x)$

From (1.3) it follows that the M-estimator $\hat{\beta}_\alpha$ can also be obtained by solving

$$\sum_{i=1}^n \frac{w(x_i)}{\sigma} \psi_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \beta_\alpha)\} x_{\alpha i} = 0. \qquad (3.1)$$

For the special case with $w(x) = 1$, this is just the Huber estimator. With general weights it belongs to the class of M-estimators considered in Hampel et al. (1986, Section 6.3), namely it is of the type proposed by Hill and Ryan (see Hill, 1977). In contrast to the more popular proposals by Mallows and Schweppe, it is a maximum likelihood estimator in a regression model (1.1) which is essential for the stochastic complexity criterion.

Corresponding to $\hat{\beta}_\alpha$, let $T(\alpha, F_\beta)$ be the functional defined as the solution of the equation

$$\int \frac{w(x)}{\sigma} \psi_c\{\frac{w(x)}{\sigma}(y - x_\alpha^t T(\alpha, F_\beta))\} x_\alpha dF_\beta(x, y) = 0. \qquad (3.2)$$

We assume that $T(\alpha, F_\beta) = \beta_\alpha$ if α is a correct model, which is called Fisher consistency. Note that $T(\alpha, F_\beta) = \beta_\alpha$ is a solution of (3.2) if α is a correct model and $E_{F_\beta}(\psi_c\{w(x)r/\sigma\}|x) = 0$. Strong consistency and asymptotic normality of $\hat{\beta}_\alpha$ follows from the general results in Maronna and Yohai (1981) under mild regularity conditions on F_β.

Now let us see how outliers in response and independent variables affect $\hat{\beta}_\alpha$. From formula (6.3.3) of Hampel et al (1986), the influence function of $T(\alpha, F_\beta)$ is given by

$$IF(x,y;T,\alpha,F_\beta) = \frac{w(x)}{\sigma}\psi_c\{\frac{w(x)}{\sigma}(y - x_\alpha^t T(\alpha, F_\beta))\}M^{-1}(\alpha, w, F_\beta)x_\alpha, \tag{3.3}$$

where

$$M(\alpha, w, F_\beta) = \int \frac{w^2(x)}{\sigma^2}\psi_c'\{\frac{w(x)}{\sigma}(y - x_\alpha^t T(\alpha, F_\beta))\}x_\alpha x_\alpha^t dF_\beta(x,y).$$

When α is a correct model, it follows that $M(\alpha, w, F_\beta) = E_{F_\beta}(\frac{w^2(x)}{\sigma^2}\psi_c'\{\frac{w(x)}{\sigma}r\}x_\alpha x_\alpha^t)$ does not depend on β.

It follows from (3.3) that if $\|w(x)x\|$ is bounded, then $IF(x,y;T,\alpha,F_\beta)$ is bounded. Here $\|\cdot\|$ indicates the Euclidean norm. Thus the M-estimator $\hat{\beta}_\alpha$ obtained by (3.1) is robust against outliers of both response and independent variables if $w(x)$ is selected so that $\|w(x)x\|$ is bounded.

We now turn to the choice of the weights $w(x)$. The first decision to be made is whether one wants to determine weights individually for each model, or whether one wants to determine them only once from the full model and then use the same weights for each model. We have opted here for the second choice because it is simpler and it guarantees that the robust fitting error always decreases when the model contains more independent variables.

Ideally we would choose the weight function $w(\cdot)$ based on the full model so that $\hat{\beta}$ has some optimality property like minimizing the asymptotic variance. Such results for the Schweppe- and Mallows- type M-estimators can be found in Hampel et al. (1986, section 6.3b). However, it seems rather difficult to do so for our case. But the results for the Schweppe- and Mallows- type M-estimators suggest that we can reasonably restrict $w(x)$ to be of the form

$$w(x) = w_b(\|Bx\|) \quad \text{where} \quad w_b(t) = \min(1, \frac{b}{|t|})$$

with b chosen a priori and B a non-singular matrix to be determined. We will chose B such that $\hat{\beta}$ obtained from (3.1) in the full model has a bounded self-standardized sensitivity $\gamma_s^* = bc$. The self-standardized sensitivity is defined in section 4.2b of Hampel et al. (1986) as

$$\gamma_s^* = \sup_{x,y}(\psi^t E_{F_\beta}[\psi\psi^t]^{-1}\psi)^{1/2}$$

where $\psi(x,y;\beta) = w_b(\|Bx\|)\psi_c\{w_b(\|Bx\|)(y - x^t\beta)/\sigma\}x/\sigma$. Because

$$\sup_{x,y}\|B\psi\|^2 = \frac{c^2}{\sigma^2}\sup_x w_b(\|Bx\|)^2\|Bx\|^2 = \frac{c^2b^2}{\sigma^2},$$

it follows that $\gamma_s^* = bc$ will hold if $E_{F_\beta}[\psi\psi^t] = \sigma^{-2}(B^tB)^{-1}$. This means B is the solution of

$$E_{F_\beta}[w_b(\|Bx\|)^2\psi_c\{\frac{1}{\sigma}w_b(\|Bx\|)r\}^2xx^t] = (B^tB)^{-1}. \tag{3.4}$$

Since the distribution F_β is unknown, we can again use the least favorable distribution (2.7) to represent the distribution of r conditional on x. Then using (2.10) and the empirical distribution for x, (3.4) can be written as

$$\frac{2\Phi(c)-1}{2\Phi(c)-1+2c^{-1}\phi(c)}\frac{1}{n}\sum_{i=1}^{n} w_b(\|Bx_i\|)^2 x_i x_i^t = (B^tB)^{-1}. \tag{3.5}$$

This can be solved for B with a recursive procedure taking, e.g. $b = p$ and $c = 1.345$. In case we assume that r has a normal distribution $N(0, \sigma^2)$ conditional on x, we obtain by an easy calculation from (3.4) the following equation

$$\begin{aligned}(B^tB)^{-1} \quad = \quad & \frac{1}{n}\sum_{i=1}^{n}\{2c^2 w_b(\|Bx_i\|)^2 - w_b(\|Bx_i\|)^4 + 2w_b(\|Bx_i\|)^2(w_b(\|Bx_i\|)^2 - c^2)\cdot \\ & \Phi(\frac{c}{w_b(\|Bx_i\|)}) - 2cw_b(\|Bx_i\|)^3\phi(\frac{c}{w_b(\|Bx_i\|)})\}x_i x_i^t. \end{aligned} \tag{3.6}$$

Equations (3.5) or (3.6) can be solved for B with a recursive procedure once we fix b and c. Note that b and c are not uniquely determined through the bound on the sensitivity. Usually we take $c = 1.345$ and $b = p$ or $b = 1.2p$, For fixed b and c, (3.5) and (3.6) may not have a solution or have more than one solution. Of course when b is large enough, (3.5) and (3.6) have solutions, but all the associated weights equal 1 if b is too large.

Using the weight $w(x)$ determined by (3.4), it can be shown by some algebraic calculation that $\hat{\beta}_\alpha$ also has a bounded self-standardized sensitivity if α is a correct model.

Rather than using an implicit method to find a proper $w(x)$, we could use the following weight function which is easy to calculate and intuitively a natural way to guard against aberrant observations without losing too much efficiency. This weight function is defined as

$$w(x) = \frac{\psi_b(\|x-a\|_g)}{\|x-a\|_g} = \min(1, \frac{b}{\|x-a\|_g}), \tag{3.7}$$

where b is some positive constant and a is a centering vector, and $\|\cdot\|_g$ denotes a norm. In practice a and b may be determined by some diagnostic techniques (e.g. leverages, standardized difference of fitted values DFITS, etc., refer to Staudte and Sheather (1990, section 7.2)) based on the least square fit. For example, suppose we find the leverage values for $x_1^*, \cdots, x_q^*$ are greater than $\frac{1.5p}{n}$ among the n observations. So we suspect they may seriously affect the robustness of $\hat{\beta}$. Then we could choose

$$a = \frac{1}{n-q}\sum_{x_i \notin \{x_1^*,\cdots,x_q^*\}} x_i$$

as the central point of those points not singled out. Similarly $b = \max_{x_i \notin \{x_1^*,\cdots,x_q^*\}} \|x_i - a\|_g$ (or with a slight difference, $b = \min_{1\le i\le q}\|x_i^* - a\|_g$) where $\|x\|_g^2 \overset{def}{=} x^t(X_n^{*t}X_n^*)^{-1}x$ and X_n^* is the sub-matrix of X_n obtained by removing $x_1^*, \cdots, x_q^*$.

We propose the above three ways of defining a weight function but still do not have a universal criterion to say which is the best. Instead, we use them as rules of thumb. It is found, through the examples later in this paper, that weight functions defined by (3.5), (3.6) and (3.7) all weigh down those influential points and do not show a big difference in the situation where only a small number of points are away from the main body of x observations. They all work quite well in this situation. It should be kept in mind, however, that it is a difficult task to detect a large proportion of outliers in high dimensions.

4 Asymptotic Properties

In this section we study asymptotic properties of the Robust AIC and BIC, and the stochastic complexity criterion derived in Section 2. It has been seen that all these criteria are based on minimizing a statistic of the common form

$$\sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)\} + C(n, \alpha), \tag{4.1}$$

where the first term is the robust fitting error of model α and the second term is the complexity of model α. In the following we give some results about the asymptotic expansion on the robust fitting error, the proof of which is deferred to the Appendix. Based on these asymptotic expansions we can obtain a strong consistency property of the model selection criteria using (4.1).

Theorem 4.1 *Assume (x_i^t, y_i) $(i = 1, 2, \cdots)$ are i.i.d. and follow the regression model (1.1). Let $w(x) \in (0, 1]$ be a fixed weight function and σ a fixed scale parameter. Suppose the following conditions are satisfied:*

(A.1). $E(w(x_1)||x_1||) < \infty$.

(A.2). $||T(\alpha, F_\beta) - \hat{\beta}_\alpha|| = o(1)$ *a.s. for any model* $\alpha \in \mathcal{A}$.

(A.3). $E[\psi_c\{w(x_1)r_1/\sigma\}|x_1] = 0$ *a.s.*.

(A.4). $P(|r_1| \le \sigma\nu_0|x_1) > 0$ *a.s. for some* $\nu_0 \in (0, c)$.

(A.5). $P(x_1^t u = 0) < 1$ *for any* $u \neq 0$ *in* $\mathcal{R}^p$.

Then for any model $\alpha \in \mathcal{A}$

$$\sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)\} = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t T(\alpha, F_\beta))\} - |o(n)| \quad a.s.. \tag{4.2}$$

Moreover, for any incorrect model $\alpha \in \mathcal{A}_w$

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \left[\rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t T(\alpha, F_\beta))\} - \rho_c\{\frac{w(x_i)}{\sigma} r_i\}\right] > 0 \quad a.s.. \tag{4.3}$$

Theorem 4.2 *In addition to conditions (A.2), (A.3) and (A.4), suppose that the following conditions are satisfied:*

(A.6). $\sup_x w(x)||x|| < \infty$.

(A.7). $P(y_1 = a|x_1) = 0$ *a.s. for any* $a \in \mathcal{R}$.

Then for any correct model $\alpha \in \mathcal{A}_c$

$$\sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma}(y_i - x_{\alpha i}^t \hat{\beta}_\alpha)\} = \sum_{i=1}^{n} \rho_c\{\frac{w(x_i)}{\sigma} r_i\} - |O(\log\log n)|. \quad a.s.. \tag{4.4}$$

The proofs are given in the Appendix. Note that the conditions above are rather weak. That (A.2) holds, follows from the general results in Maronna and Yohai (1981). From Theorems 4.1 and 4.2 the following results are obviously true.

Corollary 4.1 *If $C(n, \alpha) = o(n)$ a.s., then under conditions in Theorem 4.1 the model selection criterion defined by (4.1) will select a correct model in $\mathcal{A}_c$ almost surely as $n \to \infty$. Accordingly, the same is true for RAIC, HAIC, RBIC and SC defined in Section 2.*

Corollary 4.2 *If in addition $\frac{C(n,\alpha)}{\log\log n} \to \infty$ a.s. and $C(n, \alpha)$ is an increasing function with respect to p_α, then under conditions in Theorem 4.2, the optimal model in $\mathcal{A}_c$ will be selected using (4.1) almost surely as $n \to \infty$. Accordingly, the same is true for RBIC and SC.*

5 Robustness of the Selection Procedure

Since $\rho_c(t)$ is a strictly increasing function of $|t|$, the change in the value of the criterion (4.1) is not bounded when one y-value changes arbitrarily. Nevertheless, the influence due to an arbitrary change of one observation on the model selection procedure minimizing (4.1) is bounded. This can be seen as follows. Assume that $y_1 > x_{\alpha 1}^t \hat{\beta}_\alpha + c\hat{\sigma}/w(x_1)$ for any $\alpha \in \mathcal{A}$. Then for any $y_1' > y_1$

$$\psi_c\{\frac{w(x_1)}{\hat{\sigma}}(y_1' - x_{\alpha 1}^t \hat{\beta}_\alpha)\} = \psi_c\{\frac{w(x_1)}{\hat{\sigma}}(y_1 - x_{\alpha 1}^t \hat{\beta}_\alpha)\} = c.$$

But the equations for $\hat{\beta}_\alpha$ and $\hat{\sigma}$ in Huber's proposal 2 involve only $\psi(w(x_i)(y_i - x_{\alpha i}^t \beta_\alpha)/\sigma)$. Thus when the estimates are unique, they are unchanged if we change y_1 to y_1'. The same result is also true if $\hat{\sigma}$ is Hampel's median absolute deviation and y_1 is sufficiently large. But if the parameter estimates are unchanged, then the criterion changes by $c(y_1' - y_1)w(x_1)/\hat{\sigma}$ for any $\alpha \in \mathcal{A}$. Therefore the selected model does not change if we change y_1 to $y_1' > y_1$ and y_1 is large enough. A similar result holds also if y_1 is small enough. Hence the robustness of the parameter estimates together with the linear growth of ρ in the tails makes the model selection procedure robust.

The same argument works also for more than one outlier as long as the estimators are unique. But typically the outliers must be very large until we reach the point where the selected model does not change any more. This is related to the well known problem of the low breakdown point of M-estimators if the number of independent variables is large. One way out is to use a bounded function ρ in (4.1). The problem then is to compute the estimator (1.3): There will be many local minima, and we will have to find the global one. A different, but equally difficult approach would be to use squared errors like for least squares, but to sum in (4.1) only over a subset containing say 90% of the data and to extend the minimization also over this subset. This would be the most direct implementation of the idea that one wants to find the model which fits best for the majority of the data. But this raises a number of questions like computation and stability which are beyond the scope of this paper.

6 Examples and Simulation Study

We will illustrate in this section the following four aspects: choices of weight functions $w(x)$, behavior of robust parameter estimators, effect of small signal to noise ratios and robustness properties of the model selection both in the case of a single outlier and in the case of heavy-tailed errors. Two types of regression models are employed in our examples. In the first, the full model is a quadratic: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + r$. The second one contains two groups with a dummy variable x_2 for the second group , and the full model contains all interactions $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + r$. There are 31 observations for every sample used in this section. The independent variable x_1 is generated from a uniform distribution on $[1, 4]$ except that the first observation is an outlier equal to -2.0. The variable x_2 is generated from a Bernoulli distribution. The data for x_1 and x_2 are shown in Figure 1 and Figure 2.

Choosing the weight function $w(x)$. For the quadratic full model, the weight of the observation at $x_1 = -2.0$ is 0.099, 0.1 and 0.088 respectively for each of the three weight functions (3.5), (3.6) and (3.7). The weights of other observations are all equal to 1 no matter which of the three weight functions is used. It is seen that the results for the three different methods are very close. For the interaction full model, the weight of the first observation at $x_1 = -2.0$ is 0.648 and 0.708 respectively for each of the methods (3.5) and (3.6), while the other weights all equal to 1. So these two methods also yield very similar results. But when we use the method (3.7), the weights for the interaction full model are 0.482, 0.537, 0.660, 0.885 and 0.848 respectively for observations 1, 5, 8, 30 and 31, and 1 for all others. This significant difference is caused by the fact that five points are labeled high leverage by the detection rule used in method (3.7). When applying (3.5), we set $b = p$ and $c = 1.345$, and when applying (3.6) we set $b = 1.2p$ and $c = 1.345$. The iterative procedure with $b = p$ did not converge in the case of equation (3.6).

Parameter estimation and model selection. Suppose the true regression model is

$$y = 2 + 3x_1 + r \tag{6.1}$$

where r is a standard normal random variable. We generate a random sample of size 31 for y and replace the first observation y_1 by -30 so y_1 may be regarded as an outlier. Now we fit a straight line model to the values of y and x_1 according to (3.1). The weights for independent variables are computed from the quadratic full model and by (3.5), where $c = 1.345$ and $b = p$. The scale parameter σ is estimated by the modified Huber's proposal 2 given in Section 2.2, which gives $\hat{\sigma} = 1.213$. We also fit a quadratic model and a quadratic model without the first order term, using both the robust method (3.1) and least squares method. The results are shown in Figure 1, in which LS denotes for least squares fit and RLS for the robust fit. It is seen that the robust parameter estimates for the straight line model are quite close to their true values.

Based on the above robust estimates we proceed with the model selection from four candidate models: constant, straight line, full quadratic and quadratic without the first order term. Values of the four criteria RAIC (2.1), HAIC (2.2), RBIC (2.3) and RSC (2.13) are computed for the four candidate models. Note that here $\hat{\sigma}$ used is 1.201, which is obtained based on the quadratic full model. The model selection results are listed in Table 1.

From Table 1 we see that all the four criteria select the same and also the true model. The second best model selected by the stochastic complexity criterion is the full model so also a correct one. But the other three criteria select a wrong model as the second best model.

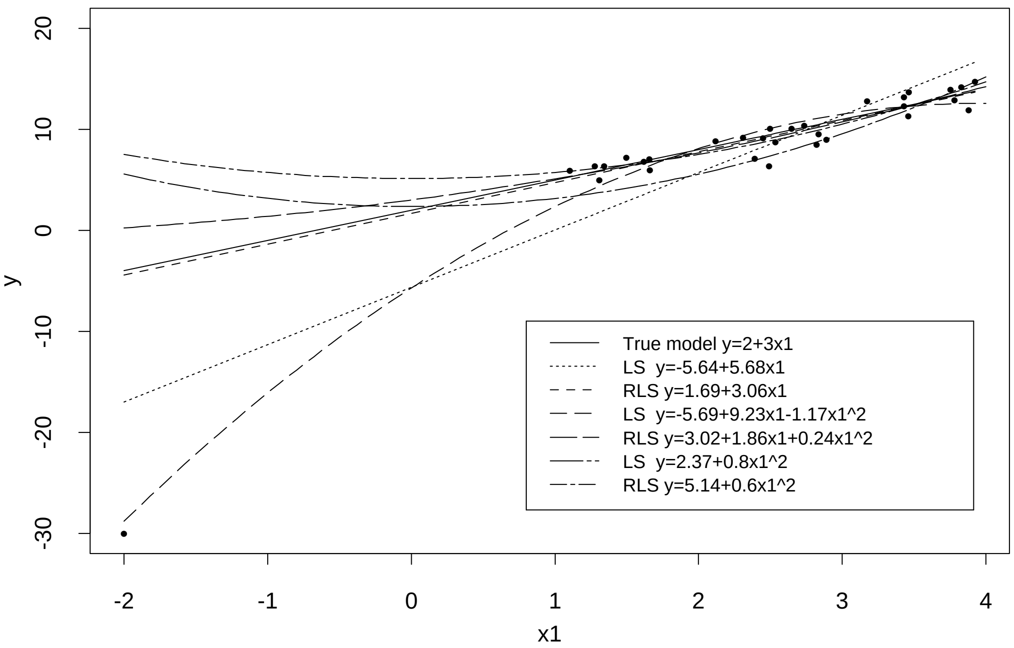

Figure 1. Comparison between Robust and Least Squares Fit

If the data are generated from the following regression model

$$y = 2 + 3x_1 + 4x_2 + r \tag{6.2}$$

with r being a standard normal random variable and $y_1 = -30$ being regarded as an outlier, the corresponding results for robust parameter estimation and model selection are given in Figure 2 and Table 2. Note that in getting these results we assume the interaction full model but the other settings are the same as those for (6.1). We see that similar conclusions can be drawn.

Table 1: *Criteria Values of Model Selection for Model (6.1)*

	Criteria Values			
Candidate Model	RSC	RAIC	HAIC	RBIC
β_0	63.48	62.77	62.87	63.62
$\beta_0 + \beta_1 x_1$ (optimal)	17.89	15.25	15.44	16.95
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$(correct)	17.91	15.89	16.17	18.44
$\beta_0 + \beta_2 x_1^2$	18.21	15.57	15.76	17.27

Table 2: *Criteria Values of Model Selection for Model (6.2)*

	Criteria Values			
Candidate Model	RSC	RAIC	HAIC	RBIC
β_0	88.67	87.95	88.05	88.81
$\beta_0 + \beta_1 x_1$	71.07	68.35	68.53	70.05
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	37.81	33.23	33.51	35.79
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	38.42	34.09	34.46	37.50
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2$	41.12	36.75	37.03	39.31
$\beta_0 + \beta_2 x_2$	86.38	84.19	84.38	85.89
$\beta_0 + \beta_2 x_2 + \beta_3 x_1 x_2$	56.94	52.37	52.65	54.93
$\beta_0 + \beta_3 x_1 x_2$	67.94	65.34	65.52	67.04

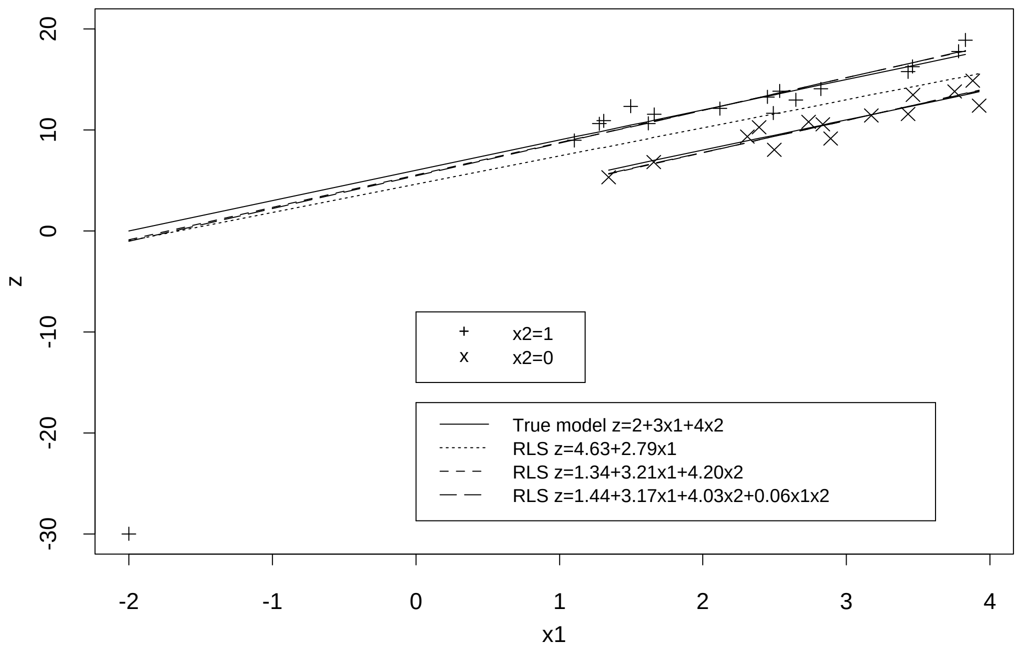

Figure 2. Robust Fit Using x1 and x2

Consistency, robustness and effect of small signal to noise ratio. To see the performance of the different model selection procedures, we run a series of simulations to compute the empirical probability distributions of selecting different candidate models by the four criteria. To carry out such a simulation we use an original regression model to generate a set of data, then contaminate the data by changing one y-value at a certain x-value. Table 3 gives an overview about the true models used in our simulation study, the x-values at which the data are contaminated, the corresponding full models which are used in the selection and the identification number of each situation. Note that $x_{1,1} = -2.0$ so far. But we will change it to 6.0 in some cases for a more comprehensive evaluation.

Table 3: *Overview of Original Models Used in Simulation*

Full model	True Model	Outlier Position		
		$x_{1,1} =$ -2.0	$x_{1,16} =$ 2.54	$x_{1,1,} =$ 6.0
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	$E(y) = 2 + 3x_1$	(i)		
	$E(y) = 2 + 3x_1 + 0.5x_1^2$	(ii)		
	$E(y) = 2 + 3x_1$		(iii)	
	$E(y) = 2 + 3x_1$			(iv)
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	$E(y) = 2 + 3x_1 + 4x_2$	(v)		
	$E(y) = 2 + 1.5x_1 + 1.5x_2$	(vi)		
	$E(y) = 2 + 3x_1 + 4x_2$		(vii)	
	$E(y) = 2 + 3x_1 + 4x_2$			(viii)

There are in total 8 original models in Table 3. For each of these 8 models, 31 observations of y are generated from a normal distribution with variance 1 for a total of 50 times, to get 50 samples of y. Then we replace the observation y_i with $i = 1$ or $i = 16$ depending on the corresponding outlier position value in Table 3 in each of the 50 samples with one value from the set $\{E(y_i) + 3k, k = -50, -49, \cdots, 50\}$ which has 101 points. We conduct a model selection by both robust and non-robust procedures for each of the 8 true models and for each of the 50 times 101 data sets generated in this way.

Note that models (ii) and (vi) are very different from the other models in the sense that there is a term with small signal-to-noise ratio for these two models. As a rule of thumb we know the ratio statistic $\hat{\beta}/s_{\hat{\beta}}$, a quantity closely related to the signal-to-noise, roughly has a normal distribution and can be used to test $H_0 : \beta = 0$ against its alternative. Without loss of too much precision we can use $s_{\hat{\beta}_i} = \hat{\sigma}(X^t W^2 X)^{-1}_{ii}$. With this setting we can get $(\hat{\beta}_1/s_{\hat{\beta}_1}, \hat{\beta}_2/s_{\hat{\beta}_2}) = (2.059, 1.767)$ and (3.192,0.944) respectively for models (ii) and (vi) relative to the full model. So model selection for these two models are much more difficult than for the other six models. When data are generated from model (ii) one should choose model (ii) itself as the optimal model. But when data are generated from model (vi) it seems also reasonable in practice to choose $\beta_0 + \beta_1 x_1$ as the optimal model.

In our model selection the candidate models are all submodels of the full model which contain the intercept term. The model selection criteria used are non-robust stochastic complexity criterion (NSC), AIC, BIC, stochastic complexity criterion (RSC), Ronchetti's robust AIC (RAIC), Hampel's robust AIC (HAIC) and robust BIC (RBIC). Note that in computation all non-robust criteria here can be regarded as special situations of the robust criteria by setting the tuning parameter c equal to ∞ and the weight function $w(x) = 1$. The simulation results for the model selection are shown in Table 4 and Figures 3 to 10.

In each bar-plot in Figure 3 to Figure 10, there are 101 bars. The x-coordinate of each of these bars is used to indicate the outlier value of y_1 (or y_{16}). Each bar gives the distribution in terms of frequency of selecting each candidate model (or a group of them) based on the data having the indicated outlier and by the criterion indicated by the title. Note that cases (i), (ii), (iii) and (iv) share the same class of candidate models indicated by the legend in Figure 3; and cases (v), (vi), (vii) and (viii) share another set of candidate models indicated by the legend in Figure 7. Because the results for non-robust model selection are generally bad, we plotted them only for cases (i) and (v). The marginal frequency refers to the summation over the x-axis of the frequencies of selecting a candidate model. So Table 4 gives a measure of overall performance against an outlier of y by different methods. Figures 3 to 10 and Table 4 support the following arguments:

(1). The robust methods give a significant improvement over non-robust methods in terms of the number of times of selecting correct models.

(2). For all cases except (ii) and (vi), i.e. whenever the signal to noise ratio is significantly large, all robust methods have quite high frequencies (marginal frequencies between 60% and 94%) of selecting the optimal models, and quite low frequencies (marginal frequencies less than 22%) of selecting incorrect models.

(3). For case (ii) where the term x_1^2 has a small signal-to-noise ratio but the test statistic mentioned above is still quite large, the stochastic complexity method is much better than the other robust methods in terms of the number of times of selecting the optimal model and of selecting an incorrect model.

(4). For case (vi) where the term x_2 has both a small signal-to-noise ratio and a small test statistic value, it is quite difficult to compare the four methods with each other. On the one hand, they select the optimal model for approximately the same number of times. The difference is less than 5%. On the other hand, the stochastic complexity method tends to not select the term x_2 in the model, which is consistent with the analysis of signal-to-noise ratio and hypothesis testing and seems to be more reasonable in practice.

(5). In general, the robust AIC methods are likely to overfit the model and the robust BIC is likely to underfit the model. The stochastic complexity method seems to take account of the information given by the signal-to-noise ratio and hypothesis testing, It thus falls between the two extremes.

(6). All the four robust methods are competitive with each other. None of the methods is universally better than the others.

Regarding a model selection procedure as a way of finding a model with the best prediction ability, one can compare the four robust model selection methods in terms of their final prediction error (FPE) on response observations excluding the outliers. Namely use $\sum_{i=2}^{31}(E(y_i) - \hat{y}_i(\hat{\alpha}))^2$ for all the cases except (iii) and (vii) and $\sum_{i\neq 16}(E(y_i) - \hat{y}_i(\hat{\alpha}))^2$ for cases (iii) and (vii), where $\hat{\alpha}$ represents the model selected by one of these four criteria. The comparison is possible since here we know the values of $E(y_i)$'s. The comparison results for cases (i), (ii), (v) and (vi) are plotted in Figure 11, where the x-axis represents the outlier position, and the y-axis gives the difference of the average FPE's between one method and the stochastic complexity method. The average FPE is obtained from the 50 simulations running at each outlier position. From Figure 11, we also see that the four model selection methods are very competitive with each other.

Robustness against non-normality. Consider the two models $y = 2 + 3x_1 + 4x_2 + r$ and $y = 2 + 3x_1 + r$ where the signal-to-noise ratio for every term is significantly large under normality assumptions as indicated previously. We study how the four robust model selection methods are affected if r is from a distribution having thicker tails than those of the normal distribution. The y observations are obtained by generating r from standard normal, student's t with 3 d.f., Cauchy ($t_{(1)}$), lognormal with mean 0 and scale 1 which is asymmetric, slash which is a standard normal divided by a uniform on [0,1], and contaminated normal $0.9N(0,1)+0.1N(0,3)$. We conduct model selection for these cases based on 1000 simulations. The results are given in Tables 5 and 6. For cases where r is from standard normal, $t_{(3)}$, lognormal or contaminated normal, all the four methods perform very well and about the same conclusions as from the previous examples can be obtained.

It is interesting to note that model selection on those cases where r is generated from the Cauchy or the slash distribution does not perform as well as expected. A possible explanation may be, as indicated in Ronchetti et. al. (1996), that under slash or Cauchy errors the t-values for the non-zero β parameters are much smaller than under normality. So the evidence for the true parameters in the data may be very weak.

7 Conclusion

We have discussed a class of penalized robust criteria for model selection in linear regression. Particularly, we have derived one such criterion using the newly developed stochastic complexity theory. We have proved that under very general conditions the penalized robust criteria have a strong consistency property. We also show that the influence on these criteria by a small number of outliers is bounded. The simulation results give further support for our model selection procedure.

Appendix: Proof of Theorems 4.1 and 4.2

We start with a Lemma which bounds the error in linearizing ρ.

Lemma A.1 *For any $\nu \in (0,c)$ there exists an $\varepsilon = \varepsilon(\nu) > 0$ such that*

$$\rho_c(\Delta+h) - \rho_c(\Delta) - h\psi_c(\Delta) \geq \varepsilon I(|\Delta| \leq \nu)\min(|h|, h^2)$$

for any Δ and h in $\mathcal{R}$. Here $I(\cdot)$ is the indicator function.

Proof: We classify all the possible values of Δ into three subsets: $|\Delta| > \nu$, $0 \leq \Delta \leq \nu$ and $-\nu \leq \Delta \leq 0$. Then we prove the inequality for each subset.

By the convexity of ρ_c, $\rho_c(\Delta+h) - \rho_c(\Delta) - h\psi_c(\Delta) \geq 0$ which covers the case $|\Delta| > \nu$.

If $0 \leq \Delta \leq \nu$, we classify all the possible values of h into three categories $|\Delta+h| < c$, $\Delta + h \geq c$ and $\Delta + h \leq -c$. Then when $|\Delta+h| < c$, we have

$$\rho_c(\Delta+h) - \rho_c(\Delta) - h\psi_c(\Delta) = \frac{1}{2}(\Delta+h)^2 - \frac{1}{2}\Delta^2 - \Delta h = \frac{1}{2}h^2.$$

When $\Delta + h \geq c$, $h > 0$ and thus we have

$$\begin{aligned}\rho_c(\Delta+h) - \rho_c(\Delta) - h\psi_c(\Delta) &= c(\Delta+h) - \frac{1}{2}c^2 - \frac{1}{2}\Delta^2 - \Delta h \\ &= (c-\Delta)h - \frac{1}{2}(c-\Delta)^2 \geq \frac{1}{2}(c-\Delta)h \geq \frac{1}{2}(c-\nu)h.\end{aligned}$$

When $\Delta + h \leq -c$, $h < 0$ and thus we have

$$\begin{aligned}\rho_c(\Delta+h) - \rho_c(\Delta) - h\psi_c(\Delta) &= -c(\Delta+h) - \frac{1}{2}c^2 - \frac{1}{2}\Delta^2 - \Delta h \\ &= -(c+\Delta)h - \frac{1}{2}(c+\Delta)^2 \geq -\frac{1}{2}(c+\Delta)h \geq \frac{1}{2}c|h|.\end{aligned}$$

For the situation $-\nu \leq \Delta \leq 0$, we can use the same arguments as in the situation $0 \leq \Delta \leq \nu$. Then the lemma follows if we define $\varepsilon = \min\{\frac{1}{2}, \frac{1}{2}(c-\nu)\}$.

$\square$

Next we prove the law of the iterated logarithm for the Huber estimator of location. For this we define

$$\gamma(z,h) = \frac{\psi_c(z) - \psi_c(z-h)}{h} \ (h \neq 0), \quad \gamma(z,0) = 0.$$

Then we have

Lemma A.2 *Let Z_i $(i = 1, 2, \cdots)$ be i.i.d. with $E[\psi_c(Z_i)] = 0$ and $P[Z_i = \pm c] = 0$. If $\{\Delta_n\}$ is a sequence of solutions to $\sum_{i=1}^n \psi_c(Z_i - \Delta_n) = 0$ which converges to 0 a.s., then*

$$\frac{1}{n}\sum_{i=1}^n \gamma(Z_i, \Delta_n) \to P(|Z_i| \leq c) \quad a.s..$$

If in addition $P[|Z_i| \leq c] > 0$, then

$$\Delta_n = O(\sqrt{\frac{\log\log n}{n}}) \quad a.s..$$

Proof: First we note that $|\gamma(z,h)| \leq 1$ for all z and all h and $\gamma(z,h) = I(|z| \leq c)$ if $|h| < |z-c|$ and $|h| < |z+c|$. Hence

$$|\gamma(z,h) - I(|z| \leq c)| \leq 2\{I(|z+c| \leq h) + I(|z-c| \leq h)\}.$$

Let

$$N_{\pm k} = \{\omega | \frac{1}{n}\sum_{i=1}^n I(|Z_i \pm c| \leq \frac{1}{k}) \not\to P(|Z_i \pm c| \leq \frac{1}{k})\}$$

and

$$N_0 = \{\omega | \frac{1}{n}\sum_{i=1}^n I(|Z_i| \leq c) \not\to P(|Z_i| \leq c)\} \cup \{\omega | \Delta_n \not\to 0\}.$$

Because $P(\bigcup_{k=-\infty}^\infty N_k) = 0$, it is sufficient for the first part to show that for $\omega \notin \bigcup_{k=-\infty}^\infty N_k$, $\frac{1}{n}\sum_{i=1}^n \gamma(Z_i, \Delta_n)$ converges to $P(|Z_i| \leq c)$.

Let $\varepsilon > 0$ be given. Then choose k such that $P(|Z_i \pm c| \leq \frac{1}{k}) \leq \varepsilon$ (which is possible because $P(Z_i = \pm c) = 0$). Now choose n_0 such that for all $n \geq n_0$:

$$\begin{aligned} |\Delta_n| &\leq \frac{1}{k}, \\ |\frac{1}{n}\sum_{i=1}^n I(|Z_i \pm c| \leq \frac{1}{k}) - P(|Z_i \pm c| \leq \frac{1}{k})| &\leq \varepsilon, \\ \frac{1}{n}\sum_{i=1}^n I(|Z_i| \leq c) - P(|Z_i| \leq c)| &\leq \varepsilon. \end{aligned}$$

This implies that for $n \geq n_0$

$$\begin{aligned} & |\frac{1}{n}\sum_{i=1}^n \gamma(Z_i, \Delta_n) - P(|Z_i| \leq c)| \\ \leq & |\frac{1}{n}\sum_{i=1}^n \gamma(Z_i, \Delta_n) - \frac{1}{n}\sum_{i=1}^n I(|Z_i| \leq c)| + |\frac{1}{n}\sum_{i=1}^n I(|Z_i| \leq c) - P(|Z_i| \leq c)| \\ \leq & \frac{2}{n}\sum_{i=1}^n I(|Z_i + c| \leq \frac{1}{k}) + \frac{2}{n}\sum_{i=1}^n I(|Z_i - c| \leq \frac{1}{k}) + \varepsilon \\ \leq & 2P(|Z_i + c| \leq \frac{1}{k}) + 2P(|Z_i - c| \leq \frac{1}{k}) + 5\varepsilon \leq 9\epsilon. \end{aligned}$$

For the second part, note that by the definition of $\gamma(Z,h)$ and Δ_n we have

$$\sum_{i=1}^{n}\gamma(Z_i,\Delta_n)=\frac{\sum_{i=1}^{n}\psi_c(Z_i)-\sum_{i=1}^{n}\psi_c(Z_i-\Delta_n)}{\Delta_n}=\frac{\sum_{i=1}^{n}\psi_c(Z_i)}{\Delta_n}.$$

Thus form the first part it follows that

$$\Delta_n=\frac{\sum_{i=1}^{n}\psi_c(Z_i)}{\sum_{i=1}^{n}\gamma(Z_i,\Delta_n)}\to\frac{\frac{1}{n}\sum_{i=1}^{n}\psi_c(Z_i)}{P(|Z_1|\le c)}\quad \text{a.s..}$$

Since $|\psi_c|\le c$, the second part follows by applying the law of the iterated logarithm to $\psi_c(Z_i)$ $(i=1,2\cdots)$.

□

For the proofs of Theorems 4.1 and 4.2, we may and will assume without loss of generality that $\sigma=1$.

Proof of Theorem 4.1: By definition of $\hat{\beta}_\alpha$, we have for any $\alpha\in\mathcal{A}$

$$\sum_{i=1}^{n}\rho_c\{w(x_i)(y_i-x_{\alpha i}^t\hat{\beta}_\alpha)\}-\sum_{i=1}^{n}\rho_c\{w(x_i)(y_i-x_{\alpha i}^tT(\alpha,F_\beta))\}\le 0. \tag{a.1}$$

Now write $\Delta_i=w(x_i)(y_i-x_{\alpha i}^tT(\alpha,F_\beta))$ and $h_i=w(x_i)x_{\alpha i}^t(T(\alpha,F_\beta)-\hat{\beta}_\alpha)$. By Lemma A.1, it follows that

$$\sum_{i=1}^{n}\rho_c\{w(x_i)(y_i-x_{\alpha i}^t\hat{\beta}_\alpha)\}-\sum_{i=1}^{n}\rho_c\{w(x_i)(y_i-x_{\alpha i}^tT(\alpha,F_\beta))\}\ge\sum_{i=1}^{n}\psi_c(\Delta_i)h_i.$$

Since $|\psi|_c\le c$, we obtain by applying the Cauchy-Schwarz inequality

$$|\sum_{i=1}^{n}\psi_c(\Delta_i)h_i|\le c\sum_{i=1}^{n}|h_i|\le c||T(\alpha,F_\beta)-\hat{\beta}_\alpha||\sum_{i=1}^{n}w(x_i)||x_{\alpha i}||.$$

Then by conditions (A.1) and (A.2) and the strong law of large numbers, the last term in the above inequality is of order $o(n)$ a.s.. From this and inequality (a.1), it follows that (4.2) is true.

To prove (4.3), we write $h_i^{'}=w(x_i)(x_i^t\beta-x_{\alpha i}^tT(\alpha,F_\beta))=w(x_i)x_i^tu(\alpha)$ where $u(\alpha)=\beta-T^*(\alpha,F_\beta)$ and $T^*(\alpha,F_\beta)_j=T(\alpha,F_\beta)_j$ for $j\in\alpha$ and $T^*(\alpha,F_\beta)_j=0$ for $j\notin\alpha$. Note that $u(\alpha)\not\equiv 0$ if $\alpha\in\mathcal{A}_w$, i.e. α is an incorrect model.

By applying Lemma A.1 with $\nu=\nu_0$, it follows that for some $\varepsilon(\nu_0)>0$,

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\Big[\rho_c\{w(x_i)(y_i-x_{\alpha i}^tT(\alpha,F_\beta))\}-\rho_c\{w(x_i)r_i\}\Big]&\ge\frac{1}{n}\sum_{i=1}^{n}\psi_c\{w(x_i)r_i\}h_i^{'}\\ &+\varepsilon(\nu_0)\frac{1}{n}\sum_{i=1}^{n}I(|w(x_i)r_i|\le\nu_0)\min(|h_i^{'}|,h_i^{'2})\stackrel{def}{=}T_1+T_2.\end{aligned} \tag{a.2}$$

For given x_i's, each term in T_1 has conditional mean zero by condition (A.3). Moreover because of $|\psi_c|\le c$ and the Cauchy-Schwarz inequality $E(|\psi_c(w(x_i)r_i)h_i^{'}|)\le cE(w(x_i)||x_i||)||u_\alpha||$ which is finite by condition (A.1). Hence by the strong law of large numbers $T_1=o(1)$ a.s..

For T_2, observe that

$$E(I(|w(x_i)r_i|\le\nu_0)\min(|h_i^{'}|,h_i^{'2}))=E(P(|w(x_i)r_i|\le\nu_0|x_i)\min(|h_i^{'}|,h_i^{'2})).$$

Now by (A.4) and (A.5) $P\{P(|w(x_i)r_i| \leq \nu_0|x_i)\min(|h_i^{'}|, h_i^{'2}) > 0\} > 0$. In addition, $E(P(|w(x_i)r_i| \leq \nu_0|x_i)\min(|h_i^{'}|, h_i^{'2})) < \infty$ by the Cauchy-Schwarz inequality and condition (A.1). Thus by the strong law of large numbers, we have a.s.

$$\lim_n T_2 = E(P(|w(x_i)r_i| \leq \nu_0|x_i)\min(|h_i^{'}|, h_i^{'2})) > 0.$$

Hence we have proved (4.3).

□

Proof of Theorem 4.2: By using Lemma A.1, we get

$$\rho_c\{w(x_i)(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} - \rho_c\{w(x_i)r_i\} \geq \psi_c\{w(x_i)r_i\}w(x_i)x_{\alpha i}^t(\beta_\alpha - \hat{\beta}_\alpha)$$

for any correct model $\alpha \in \mathcal{A}_c$. Note that $x_{\alpha i}^t\beta_\alpha = x_i^t\beta$ if $\alpha \in \mathcal{A}_c$. From this inequality and equation (1.3) and (3.1), it follows that

$$\begin{aligned}
0 &\leq \sum_{i=1}^n \rho_c\{w(x_i)r_i\} - \sum_{i=1}^n \rho_c\{w(x_i)(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} \\
&\leq \sum_{i=1}^n \psi_c\{w(x_i)r_i\}w(x_i)x_{\alpha i}^t(\hat{\beta}_\alpha - \beta_\alpha) \\
&= \sum_{i=1}^n \left[\psi_c\{w(x_i)r_i\} - \psi_c\{w(x_i)(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\}\right] w(x_i)x_{\alpha i}^t(\hat{\beta}_\alpha - \beta_\alpha) \\
&= \sum_{i=1}^n \gamma\{w(x_i)r_i, w(x_i)x_{\alpha i}^t(\hat{\beta}_\alpha - \beta_\alpha)\}(w(x_i)x_{\alpha i}^t(\hat{\beta}_\alpha - \beta_\alpha))^2 \\
&\leq B^2||\hat{\beta}_\alpha - \beta_\alpha||^2 \sum_{i=1}^n \gamma\{w(x_i)r_i, w(x_i)x_{\alpha i}^t(\hat{\beta}_\alpha - \beta_\alpha)\}
\end{aligned}$$

where $B = \sup_i w(x_i)||x_i|| < \infty$ by (A.6). The last inequality is a consequence of the Cauchy-Schwarz inequality. Applying similar arguments as used for Lemma A.2, we see that under conditions (A.2), (A.3), (A.6) and (A.7),

$$\frac{1}{n}\sum_{i=1}^n \gamma\{w(x_i)r_i, w(x_i)x_{\alpha i}^t(\hat{\beta}_\alpha - \beta_\alpha)\} \to E[P(|w(x_1)r_1| \leq c|x_1)] \quad \text{a.s..}$$

In the same way we have also $||\hat{\beta}_\alpha - \beta_\alpha|| = O(\sqrt{\frac{\log\log n}{n}})$ a.s. under conditions (A.2) to (A.4) and (A.6) to (A.7). This implies that

$$0 \leq \sum_{i=1}^n \rho_c\{w(x_i)r_i\} - \sum_{i=1}^n \rho_c\{w(x_i)(y_i - x_{\alpha i}^t\hat{\beta}_\alpha)\} = O(\log\log n) \quad \text{a.s.,}$$

which proves (4.4).

□

Acknowledgment: The first author would like to thank Professors C. Field, G. Gabor, F. Hampel and W. Stahel for their stimulating discussions and useful references before and during the writing of this paper.

References

[1] Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Proc. 2nd Int. Symp. Info. Theory,* 267-281, eds. B.N. Petrov and F. Csaki, Akademia Kiado, Budapest.

[2] Akaike, H. (1974). A new look at statistical model identification. *IEEE Trans. Automatic Control* **19**, 716-723.

[3] Efron, B. (1983). Estimating the error rate of a prediction rule: Improvement on cross-validation. *J. Amer. Statist. Assoc.* **78** , 316-331.

[4] Efron, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81**, 461-470.

[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] Hill, R.W. (1977). *Robust regression when there are outliers in the carriers.* Ph.D. thesis, Harvard University, Cambridge, Mass..

[9] Huber, P.J. (1964). Robust estimation of a location parameter. *Ann. Math. Stat.* **35**, 73-101.

[10] Huber, P.J. (1981). *Robust Statistics.* Wiley, New York.

[11] Machado, J.A.F. (1993). Robust Model Selection and M-estimation. *Econ. Theory* **9**, 478-493.

[12] Mallows, C. L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

[13] Maronna, R.A. and Yohai, V.J. (1981). Asymptotic behavior of general M-estimates for regression and scale with random carriers. *Z. Wahrsch. Verw. Gebiete.* **58** 7-20.

[14] Qian, G. (1994). *Statistical modeling by stochastic complexity.* Ph.D. thesis, Dalhousie University, Halifax, Canada.

[15] Qian, G. and Künsch, H. (1996). *Some Notes On Rissanen's Stochastic Complexity,* preprint.

[16] Rissanen, J. (1983). A universal prior for integers and estimation by minimum description length. *Ann. Statist.* **11**, 416-431.

[17] Rissanen, J. (1986). Order estimation by accumulated prediction errors. In *Essays in Time Series and Allied Processes* (J. Gani and M.B. Priestley, eds.). *J. Appl. Probab.* **23A** 55-61.

[18] Rissanen J. (1989). *Stochastic Complexity in Statistical Inquiry.* World Scientific Publishing Co. Pte. Ltd., Singapore.

[19] Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Trans. Inform. Theory.* **42**, 40-47.

[20] Ronchetti, E. (1985). Robust model selection in regression. *Stat. Prob. Lett.* **3** 21-23.

[21] Ronchetti, E., Field, C. and Blanchard, W. (1996). Robust linear model selection by cross-validation. To appear in *J. Amer. Statist. Assoc.*

[22] Ronchetti, E. and Staudte, R.G. (1994). A robust version of Mallows's C_p. *J. Amer. Statist. Assoc.* **89**, 550-559.

[23] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

[24] Shao, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88**, 486-494.

[25] Staudte, R.G. and Sheather, S.J. (1990). *Robust Estimation and Testing.* Wiley, New York.

[26] Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Statist. Soc. B* **39**, 44-47.

[27] Wei, C.Z. (1992). On the predictive least squares principle. *Ann. Statist.* **20**, 1-42.

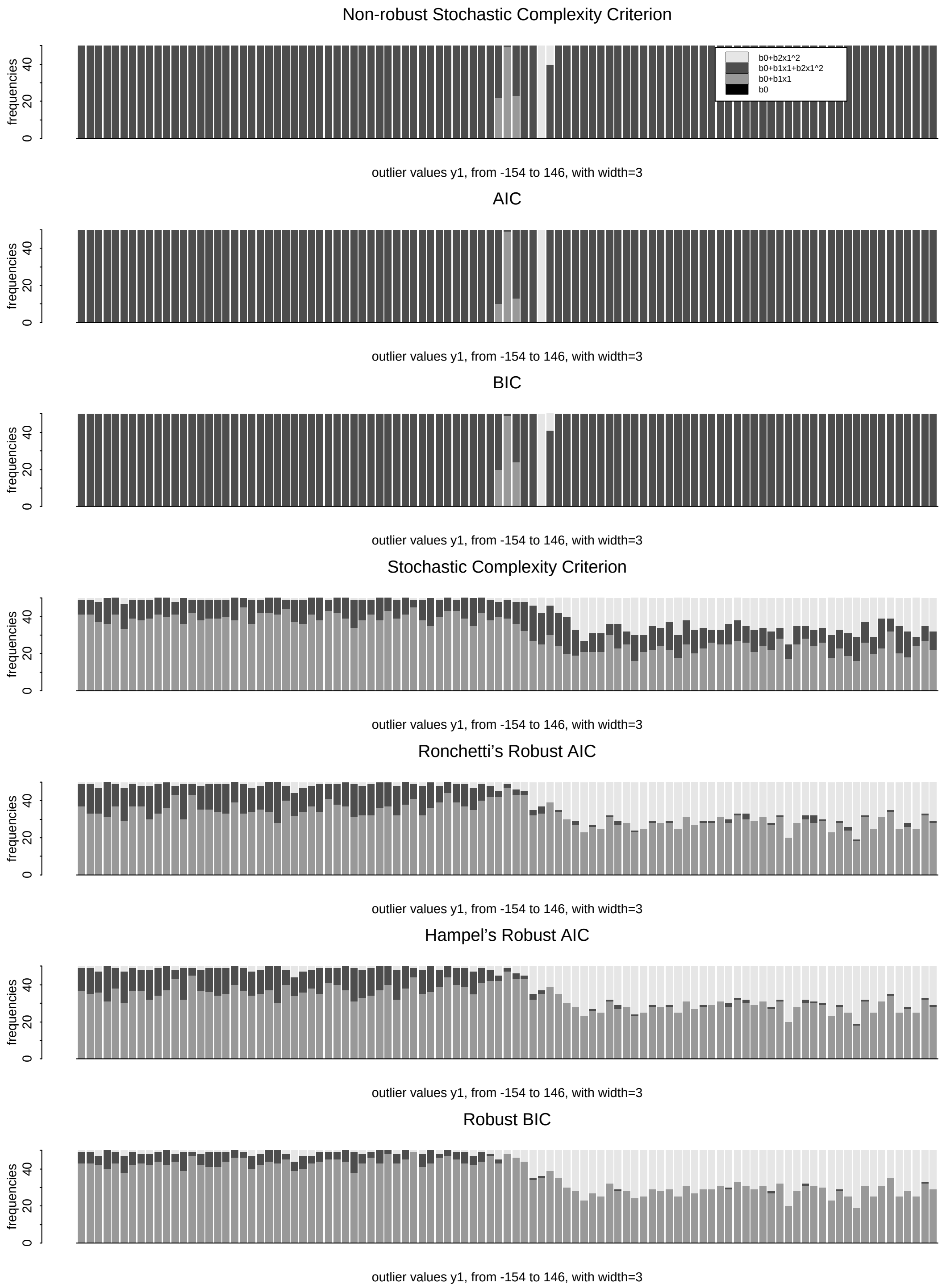

Figure 3. Empirical Distributions of Model Selection for Model (i)

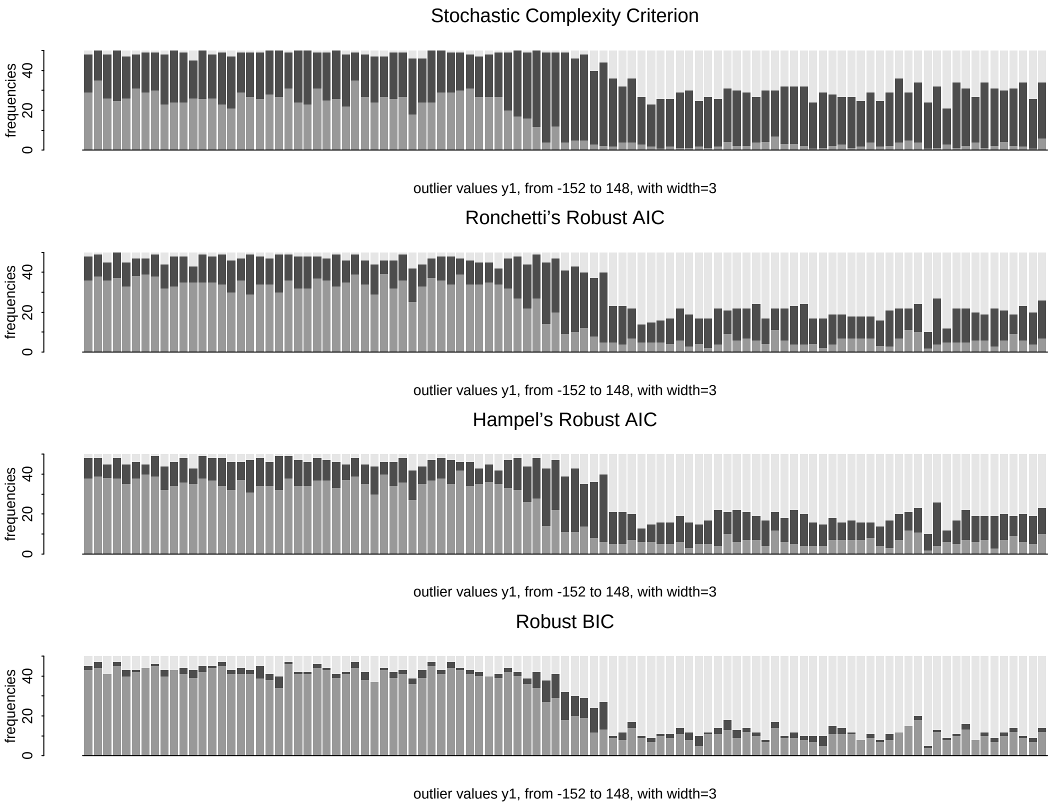

Figure 4. Empirical Distributions of Model Selection for Model (ii)

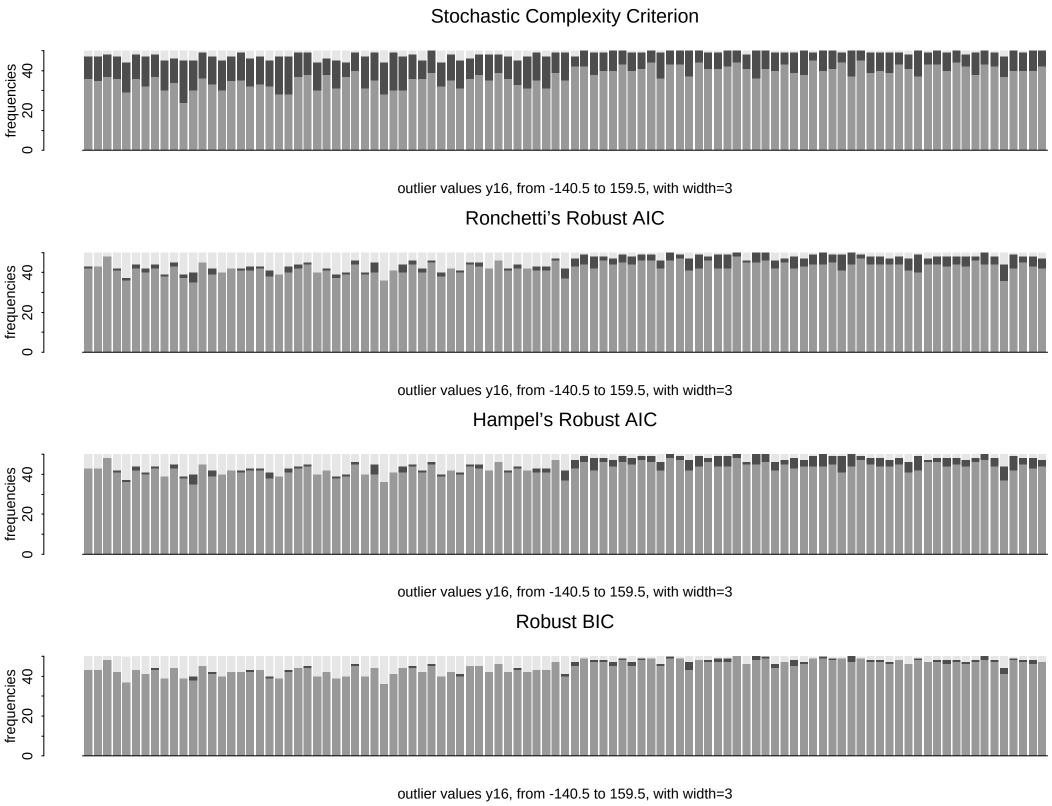

Figure 5. Empirical Distributions of Model Selection for Model (iii)

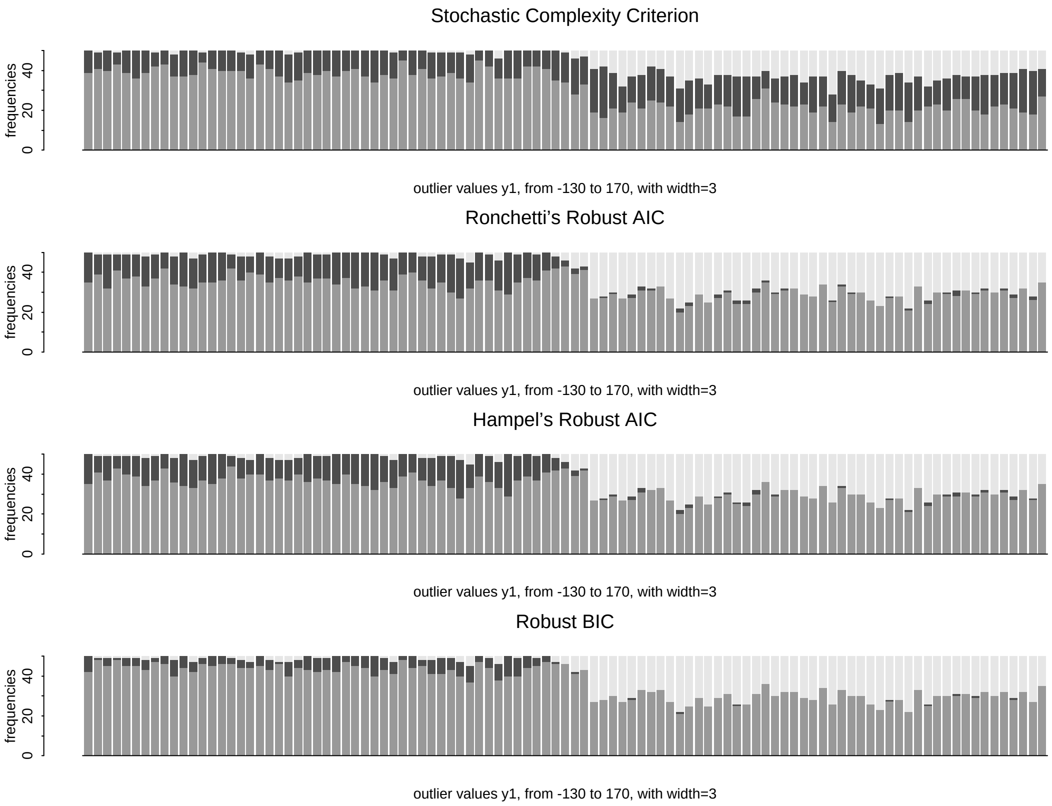

Figure 6. Empirical Distributions of Model Selection for Model (iv)

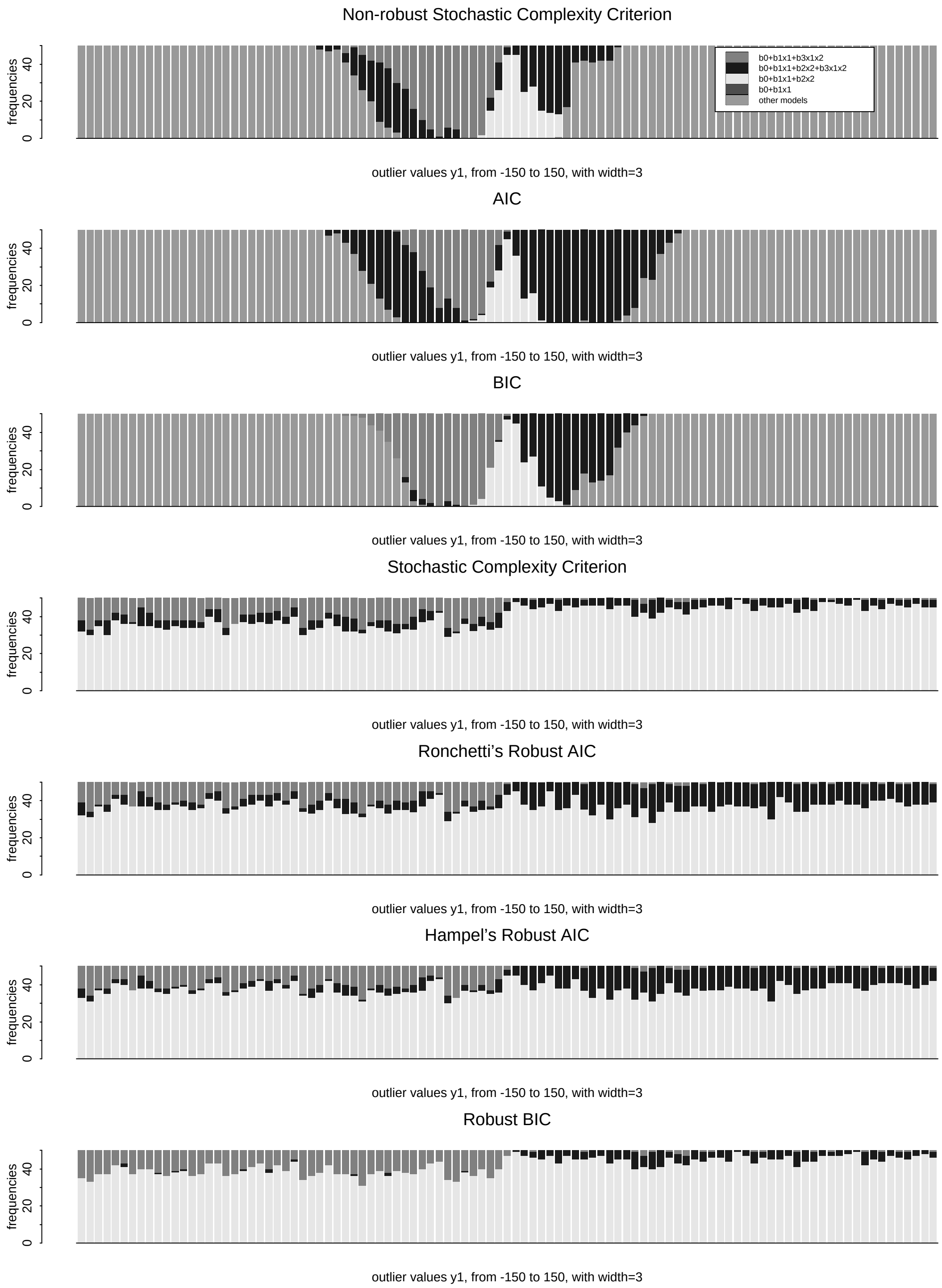

Figure 7. Empirical Distributions of Model Selection for Model (v)

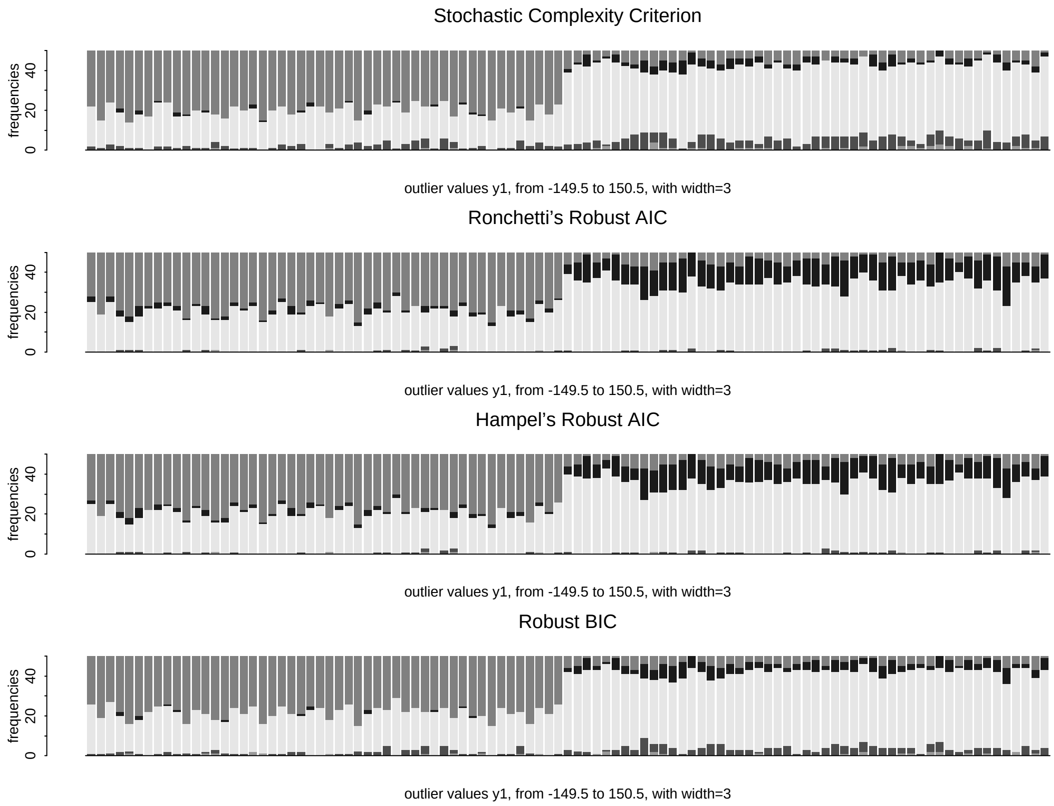

Figure 8. Empirical Distributions of Model Selection for Model (vi)

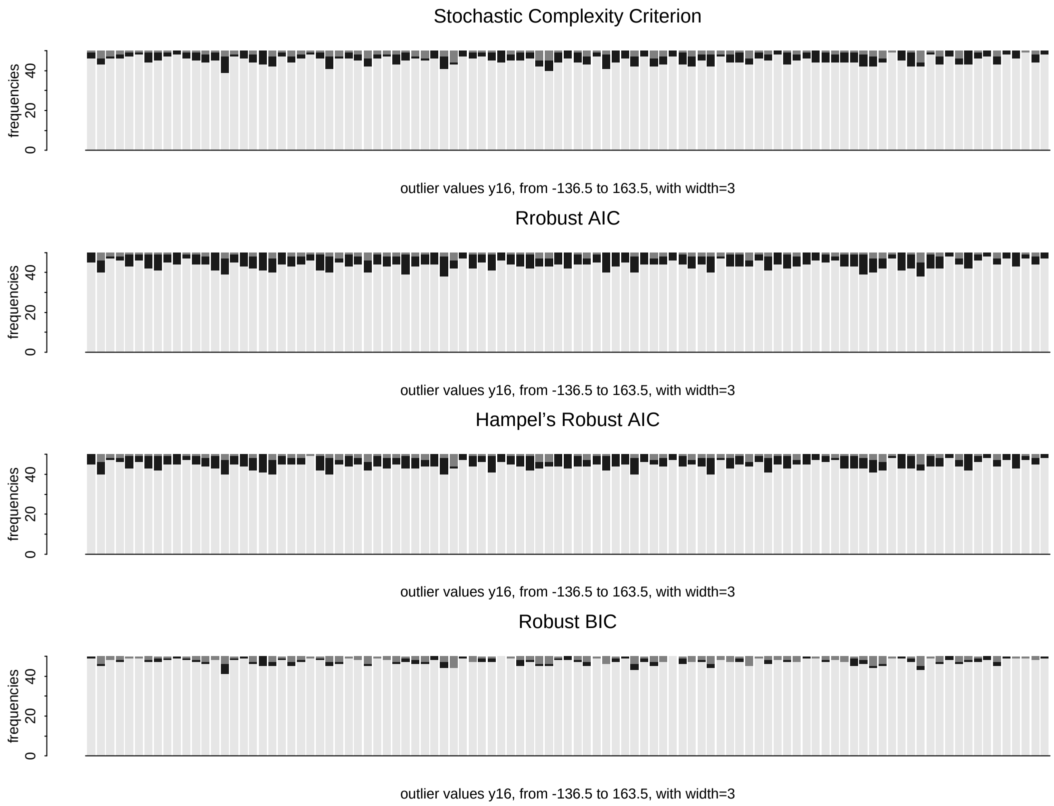

Figure 9. Empirical Distributions of Model Selection for Model (vii)

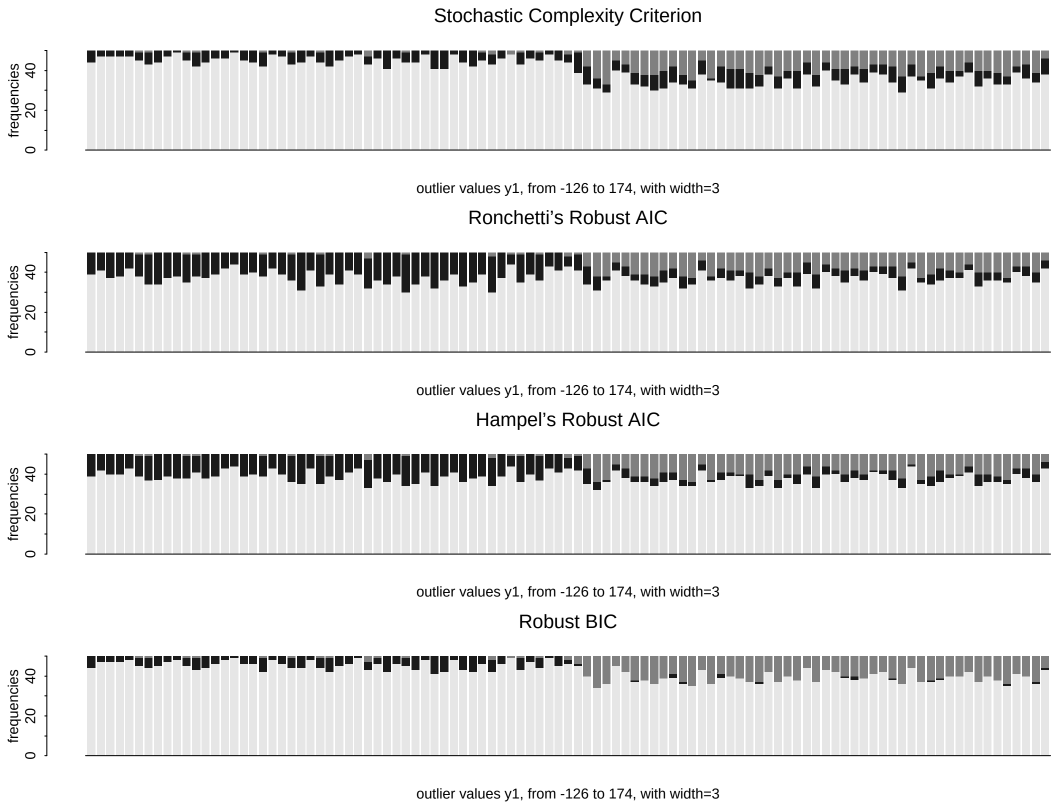

Figure 10. Empirical Distributions of Model Selection for Model (viii)

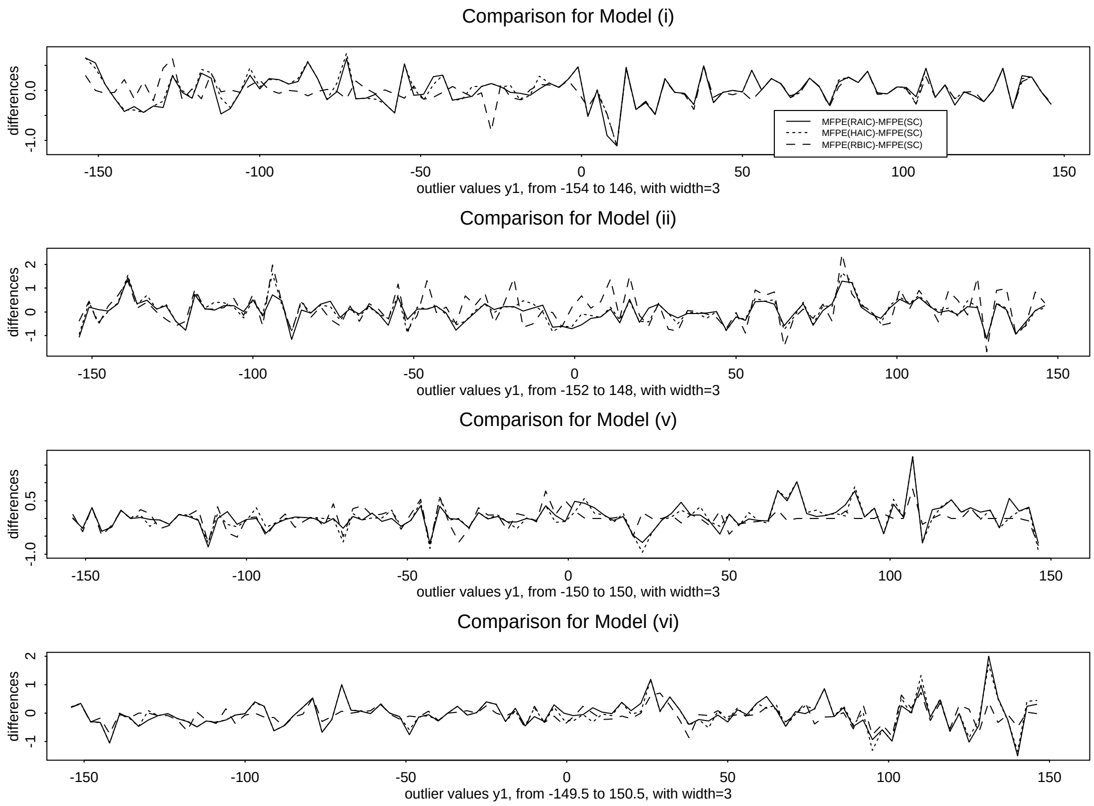

Figure 11. Comparison in Terms of Mean Final Prediction Error

Table 4: *Marginal Frequencies of Model Selection Based on 5050 Simulations*

	non-robust			robust methods			
Selected Model	NSC	AIC	BIC	RSC	RAIC	HAIC	RBIC
Model (i): $E(y) = 2 + 3x_1$							
$\beta_0 + \beta_1 x_1$	94	72	93	3193	3270	3334	3679
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	4896	4928	4898	1061	694	626	275
$\beta_0 + \beta_2 x_1^2$	60	50	59	796	1086	1090	1096
Model (ii): $E(y) = 2 + 3x_1 + 0.5x_1^2$							
$\beta_0 + \beta_1 x_1$	102	76	100	1392	1955	2055	2554
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	4894	4924	4896	2616	1508	1298	288
$\beta_0 + \beta_2 x_1^2$	54	50	54	1042	1587	1697	2208
Model (iii): $E(y) = 2 + 3x_1$, $x_{1,16} = 2.54$							
β_0	3347	2529	3143	0	0	0	0
$\beta_0 + \beta_1 x_1$	1690	2426	1892	3764	4277	4326	4495
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	11	24	11	1094	298	245	69
$\beta_0 + \beta_2 x_1^2$	2	71	4	192	475	479	486
Model (iv): $E(y) = 2 + 3x_1$, $x_{1,1} = 6.0$							
$\beta_0 + \beta_1 x_1$	70	62	79	3046	3269	3347	3711
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	4884	4903	4849	1347	712	633	265
$\beta_0 + \beta_2 x_1^2$	96	85	122	657	1069	1070	1074
Model (v): $E(y) = 2 + 3x_1 + 4x_2$							
other models	3807	3386	3796	0	0	0	0
$\beta_0 + \beta_1 x_1$	0	0	0	0	0	0	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	227	163	223	4021	3694	3792	4214
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	491	1127	469	455	833	717	231
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_3$	525	374	562	574	523	541	605
Model (vi): $E(y) = 2 + 1.5x_1 + 1.5x_2$							
other models	4463	4265	4500	42	7	8	33
$\beta_0 + \beta_1 x_1$	32	2	15	364	47	57	240
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	122	116	128	2757	2710	2787	2943
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	170	476	145	175	692	599	241
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_3$	263	191	262	1712	1594	1599	1593
Model (vii): $E(y) = 2 + 3x_1 + 4x_2, x_{1,16} = 2.54$							
other models	3891	3401	3776	0	0	0	0
$\beta_0 + \beta_1 x_1$	395	586	436	0	0	0	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	676	1002	802	4521	4379	4454	4747
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	13	20	6	359	535	445	112
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_3$	75	41	30	170	136	151	191
Model (viii): $E(y) = 2 + 3x_1 + 4x_2$, $x_{1,1} = 6.0$							
other models	3794	3523	3865	0	0	0	0
$\beta_0 + \beta_1 x_1$	3	0	0	0	0	0	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	253	190	247	4022	3728	3845	4269
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	174	657	232	534	857	720	234
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_3$	826	680	706	494	465	485	547

Table 5: *Effects of Some Thick-tail Error Distributions for model $y = 2 + 3x_1 + 4x_2 + r$ based on 1000 simulations*

Candidate Models	Distributions					
	Normal	$t_{(3)}$	Cauchy	Log N	Slash	ε-Normal
Robust Stochastic Complexity Criterion						
β_0	0	0	10	0	35	0
$\beta_0 + \beta_1 x_1$	0	0	33	0	103	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	887	832	652	843	534	819
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	86	53	33	81	7	88
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2$	27	115	226	76	237	93
$\beta_0 + \beta_2 x_2$	0	0	1	0	2	0
$\beta_0 + \beta_2 x2 + \beta_3 x_1 x_2$	0	0	17	0	20	0
$\beta_0 + \beta_3 x_1 x_2$	0	0	28	0	62	0
RAIC						
β_0	0	0	2	0	9	0
$\beta_0 + \beta_1 x_1$	0	0	13	0	41	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	850	809	635	810	594	767
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	128	96	116	125	81	155
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2$	22	95	197	65	220	78
$\beta_0 + \beta_2 x_2$	0	0	0	0	3	0
$\beta_0 + \beta_2 x2 + \beta_3 x_1 x_2$	0	0	22	0	35	0
$\beta_0 + \beta_3 x_1 x_2$	0	0	15	0	17	0
HAIC						
β_0	0	0	3	0	12	0
$\beta_0 + \beta_1 x_1$	0	0	16	0	46	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	868	821	649	825	600	785
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	108	80	98	108	66	130
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2$	24	99	196	67	218	85
$\beta_0 + \beta_2 x_2$	0	0	0	0	4	0
$\beta_0 + \beta_2 x2 + \beta_3 x_1 x_2$	0	0	22	0	32	0
$\beta_0 + \beta_3 x_1 x_2$	0	0	16	0	22	0
RBIC						
β_0	0	0	8	0	25	0
$\beta_0 + \beta_1 x_1$	0	0	26	0	79	0
$\beta_0 + \beta_1 x_1 + \beta_2 x_2$	941	865	690	892	602	850
$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$	27	27	41	37	18	53
$\beta_0 + \beta_1 x_1 + \beta_3 x_1 x_2$	32	108	190	71	202	97
$\beta_0 + \beta_2 x_2$	0	0	1	0	2	0
$\beta_0 + \beta_2 x2 + \beta_3 x_1 x_2$	0	0	21	0	22	0
$\beta_0 + \beta_3 x_1 x_2$	0	0	23	0	50	0

Table 6: *Effects of Some Thick-tail Error Distributions for model $y = 2 + 3x_1 + r$ based on 1000 simulations*

Candidate Models	Normal	$t_{(3)}$	Cauchy	Log N	Slash	ε-Normal
	Distributions					
Stochastic Complexity Criterion						
β_0	0	0	2	0	24	0
$\beta_0 + \beta_1 x_1$	754	738	660	740	600	666
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	221	208	150	202	142	249
$\beta_0 + \beta_2 x_1^2$	25	54	188	58	234	85
RAIC						
β_0	0	0	2	0	7	0
$\beta_0 + \beta_1 x_1$	855	840	682	842	648	714
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	69	57	84	58	60	117
$\beta_0 + \beta_2 x_1^2$	76	103	232	100	285	169
HAIC						
β_0	0	0	2	0	8	0
$\beta_0 + \beta_1 x_1$	868	847	697	854	658	730
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	55	50	69	46	46	100
$\beta_0 + \beta_2 x_1^2$	77	103	232	100	288	170
RBIC						
β_0	0	0	2	0	21	0
$\beta_0 + \beta_1 x_1$	910	879	744	889	673	793
$\beta_0 + \beta_1 x_1 + \beta_2 x_1^2$	9	18	19	5	18	35
$\beta_0 + \beta_2 x_1^2$	81	103	235	106	288	172