Statistica Sinica **7**(1997), 327-338

ROBUSTNESS ASPECTS OF MODEL CHOICE

Elvezio Ronchetti

University of Geneva

Abstract: Model selection is a key component in any statistical analysis. In this paper we discuss this issue from the point of view of robustness and we point out the extreme sensitivity of many classical model selection procedures to outliers and other departures from the distributional assumptions of the model. First, we focus on regression and review a robust version of Mallows's C_P as well as some related approaches. We then go beyond the regression model and discuss a robust version of the Akaike Information Criterion for general parametric models.

Key words and phrases: Akaike criterion, autoregressive models, competing models, crossvalidation, diagnostics, information theory, Mallows's C_P, M-estimators, nonnested hypotheses, outliers, robust Akaike criterion, robust C_P, robust regression, robust tests, Schwartz criterion, time series, variable selection, weighted prediction error.

1. Introduction

Model selection is a key component in any statistical analysis. Typically the choice of the final model(s) is an iterative procedure based on subject matter knowledge and on formal selection criteria. In this paper we focus on the robustness issue of the latter. Classical model selection procedures are based on classical estimators and tests. Consider for instance Mallows's C_P (Mallows (1973)), a powerful selection procedure in regression. The C_P statistic is an estimate of a measure of adequacy for prediction and one tries to find submodels of the full model with C_P values close to p or smaller than p (the number of parameters in submodel P). Since the C_P statistic is based on least squares estimation, it is very sensitive to outliers and other departures from the normality assumption on the error distribution.

Whereas in the past 30 years many robust alternatives to the classical estimation procedures have been devised (see for instance Huber (1981); Hampel, Ronchetti, Rousseeuw and Stahel (1986)), the associated model selection problem has been somewhat neglected. On the other hand, the need for robust selection procedures is obvious because one cannot estimate the parameters robustly and apply unmodified classical selection procedures. This state of affairs is perhaps one of the reasons for the widespread false prejudice that "a 'robustnik' ('robustnitsa') never changes his (her) model" (see Hampel (1991)).

The paper is organized as follows. In Section 2, we discuss robust model selection for regression. We review a robust criterion for prediction and its estimated version which leads to a robust C_P statistic. The robust model selection procedure based on this statistic can be used with a large variety of robust estimators for the parameters, including M-estimators, GM-estimators (e.g. bounded influence estimators), and one-step M-estimators with a high breakdown starting point. The robust RC_P allows us to choose the best models which fit the *majority of the data* by taking into account the presence of outliers and possible departures from the normality assumption on the error distribution. Other proposals available in the literature will also be reviewed. In Section 3 we consider general parametric models and discuss a robust version of the Akaike Information Criterion with an application to autoregressive models. In Section 4 we focus on model choice via testing of non-nested hypotheses and in Section 5 we mention two main research directions.

2. Regression Models

2.1. Mallows's C_P

Mallows's C_P (Mallows (1973)) is a powerful technique for model selection in regression. The C_P statistic is defined by $C_P = RSS_P/\hat{\sigma}^2 - n + 2p$, where RSS_P is the residual sum of squares for submodel P, p is the dimension (the number of parameters) of submodel P, n is the number of observations, and $\hat{\sigma}^2$ is an estimate of the error variance σ^2 which is usually computed in the full model. C_P is an estimate of a measure of adequacy for prediction given by the scaled sum of squared errors. If submodel P is correct then C_P will tend to be close to p or smaller than p. Therefore, a simple plot of C_P versus p will point out immediately the better submodels.

Since the C_P statistic is based on least squares estimation (via RSS_P and $\hat{\sigma}^2$), it is very sensitive to outliers and other departures from the normality assumption on the error distribution. The following simple example shows the drastic effect of contamination on the model selection procedure.

We consider bivariate data and in this case there is no need for a model selection procedure. Our purpose is only to highlight the problem which appears in more complex situations. Two observations were generated at each x_i according to $y_{ij} = -x_i + e_{ij}, i = 1, \ldots, 19,\ j = 1, 2$, where the e_{ij} are i.i.d. normally distributed with expectation 0 and variance $(.1)^2$. The two observations corresponding to $x_{10} = 0$ are then set equal and moved between –1.5 and 3.0. Figure 1 shows the situation.

For each of the 10 values of y we recompute the C_P statistic for the straight line $(\beta_0 + \beta_1 x)$ and for the parabola $(\beta_0 + \beta_1 x + \beta_2 x^2)$. The full model is $\beta_0 + \beta_1 x + \gamma z$, where $z = 1$ if $x = 0$ and $z = 0$ otherwise. Table 1 gives the values

of the classical C_P statistic and for comparison those of the robust version RC_P which is discussed in subsection 2.2.

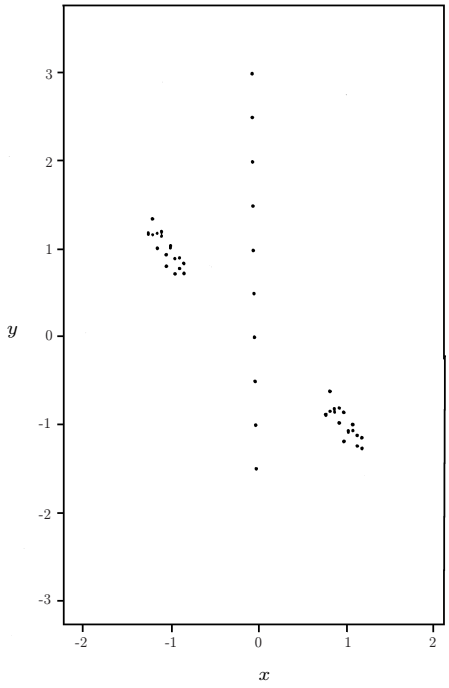

Figure 1. Sensitivity analysis on model selection. Data and positions for bogus points.

Table 1. Values of the C_P and RC_P statistics for the straight line $(\beta_0 + \beta_1 x)$ and the parabola $(\beta_0 + \beta_1 x + \beta_2 x^2)$ for different values of $y_{10,1} = y_{10,2}$.

	C_P		RC_P
$y_{10,j}$	st. line $(p = 2)$	parabola $(p = 3)$	st. line $(p = 2)$
−1.5	413.4	62.2	9.3
−1.0	186.4	35.8	9.3
−0.5	48.9	13.2	9.3
0.0	1.1	2.9	0.2
0.5	42.8	14.1	6.9
1.0	174.0	37.0	6.9
1.5	394.9	63.5	6.9
2.0	705.4	91.3	6.9
2.5	1105.4	119.8	6.9
3.0	1595.0	148.5	6.9

Let us first look at the fourth row of Table 1. When $y_{10,j} = 0$, we see from Figure 1 that the 38 points form an almost straight line. Both the classical C_P

and the robust RC_P consider the straight line ($p = 2$) as an appropriate model. This is what we expect in this situation. When we move $y_{10,j}$ away from 0, the classical C_P *rejects the straight line* and indicates the need for a model with an additional parameter (the full model with the dummy variable z). Of course, among the models with three parameters, the model with a dummy variable is better than the quadratic model. This shows the high sensitivity of the classical C_P selection procedure. Indeed, a very small change in the observations at $x = 0$ has already the effect of changing the selected model. On the other hand, the robust RC_P still considers the straight line as an appropriate model. This provides very useful information. It says that a *robust straight line* is a good model *for the majority of the data* and that the additional parameter suggested by the classical procedure is due to a (few) outlier(s). The weights associated with the robust fit will highlight these points. In this way, the statistician has the choice between a number of competing models (in this case just two: the full model with a dummy variable ($p = 3$) fitted by least squares and a straight line ($p = 2$) fitted by a robust procedure). In general this choice will depend on subject matter and on the region where he or she wants to make the prediction.

2.2. A robust version of mallows's C_P

Consider the linear model $y_i = x_i^T\beta+\epsilon_i$, where x_i^T is the ith row of the design matrix X. An M-estimator $\hat{\beta}_P$ for model P with p parameters is the solution of the equation

$$\sum_i \eta(x_i, (y_i - x_i^T\beta)/\sigma)x_i = 0, \tag{2.1}$$

for some function $\eta(x, r)$. Define the weights $\hat{w}_i = w(x_i, (y_i - x_i^T\hat{\beta}_P)/\sigma) = \eta(x_i, (y_i - x_i^T\hat{\beta}_P)/\sigma)/((y_i - x_i^T\hat{\beta}_P)/\sigma)$ and the rescaled mean squared weighted prediction error

$$\Gamma_P = \frac{1}{\sigma^2}\mathrm{E}\Big[\sum_i \hat{w}_i^2(\hat{y}_i - \mathrm{E}y_i)^2\Big], \tag{2.2}$$

where $\hat{y}_i = x_i^T\hat{\beta}_P$ is the fitted value for submodel P, and $\mathrm{E}[y_i]$ is the expected value under the full model, which is assumed correct. The weights in (2.2) carry valuable diagnostic information. They are different for each model since an observation can be outlying with respect to one model and have full weight in another. The weighting scheme has the effect of downweighting the outlying observations with respect to model P and limiting their influence on Γ_P and therefore on the model selection procedure.

Ronchetti and Staudte (1994) define a robust version of C_P as follows:

$$RC_P = \frac{W_P}{\hat{\sigma}^2} - (U_P - V_P), \tag{2.3}$$

where $W_P = \sum_i \hat{w}_i^2 r_i^2 = \sum_i \hat{w}_i^2 (y_i - \hat{y}_i)^2$ is the weighted residual sum of squares, and $\hat{\sigma}^2$ is a robust and consistent estimator of σ^2 in the full model given by $\hat{\sigma}^2 = W_{full}/U_{full}$. V_P and U_P are constants given by $V_P = \mathrm{tr}(\, RM^{-1}QM^{-1}\,)$ and $U_P - V_P = \mathrm{E}\|\eta\|^2 - 2\,\mathrm{tr}(\, NM^{-1}\,) + \mathrm{tr}(\, LM^{-1}QM^{-1}\,)$, where $M = \mathrm{E}[\eta'(x,\epsilon)xx^T]$ with η' denoting the derivative of η with respect to its second argument, $Q = \mathrm{E}[\eta^2(x,\epsilon)xx^T]$, $\|\eta\|^2 = \sum_{1\le i\le n} \eta^2(x_i,\epsilon_i)$, $N = \mathrm{E}[\eta^2\eta' xx^T]$, $L = \mathrm{E}[w'\epsilon(w'\epsilon + 4w)xx^T] = \mathrm{E}[((\eta')^2 + 2\eta' w - 3w^2)xx^T]$, and $R = \mathrm{E}[w^2 xx^T]$.

If model P holds, $\hat{\sigma}^2 \approx W_P/U_P$ and $RC_P \approx V_P$. Therefore, models with values of RC_P which are close to V_P or smaller than V_P will be preferred to others, and a plot of RC_P versus V_P will aid in this selection. Mallows (1973), p. 665, pointed out that in the classical C_P plot the variance of the slope of the line joining the points (p,C_p) and (d,d), where d is the number of parameters in the full model, is $2/(d-p)$. In our experience the variability of the RC_P line is a little greater, so $2/(d-p)$ can be used as a rough lower bound for the variability of RC_P. The plot of RC_P versus V_P can be easily integrated in existing computer packages. It is recommended to plot the classical C_P versus p next to the robust plot, so that it is clear which models are suggested by both procedures. For Huber type estimators V_P is a fixed multiple (≈ 1) of the dimension p of the subset P, and $V_P \approx p$. However, for estimators whose weight function depends on the explanatory variables, such as the Mallows' type estimators, the value of V_P could vary for different models of the same dimension. Examples can be found in Ronchetti and Staudte (1994) and Sommer and Staudte (1995).

Note that when the weights are identically 1, W_P becomes the residual sum of squares of a least squares fit, $V_P = p$, $U_P = (n-p)$, and RC_P reduces to Mallows's C_P.

2.3. Other results

In subsection 2.2 we discussed a direct robustification of Mallows's C_P. A different approach is to look at Mallows's C_P as a special case of Akaike's Information Criterion (Akaike (1973)) applied to regression models. Following this idea, Ronchetti (1982, 1985) proposed to apply to regression a robust version of Akaike's Criterion. This will be discussed in subsection 3.1 in the framework of general parametric models. We will then come back to the special case of regression.

A paper related to the results of subsection 2.2 is Léger and Altman (1993). The authors approach the problem from a diagnostic point of view. They argue that, whereas the influence of individual cases on the parameters of the selected model is often assessed as part of the model building process, such conditional

measures fail to evaluate the influence of the cases on the variable selection process. Hence they extend influence measures based on distances between predicted values to model selection problems in a manner that accounts for the selection process. Such influence measures are useful diagnostic tools which complement Cook's and other similar distances. They help to identify influential observations in the model selection but are not used to downweigh such observations in the process in order to achieve a robust model selection and a robust estimation of the parameters.

Antoch (1986, 1987) developed an algorithm to perform variable selection by using a robust estimator for the parameters. The basic idea is to compute the $\alpha-$trimmed least squares estimators suggested by Koenker and Bassett (1978) for all possible submodels and to compare them with the same estimator obtained in the full model. Then, the submodels which lead to estimates which are "undistinguishable" from that of the full model are considered acceptable.

Other related papers are Sommer and Huggins (1996), Shi and Tsai (1996), and Qian and Künsch (1996).

3. A Robust Akaike Criterion

3.1. General parametric models

In this subsection we discuss a robust version of the Akaike Information Criterion for a general parametric model $\{P_\theta \mid \theta \in \Theta\}$ (cf. Ronchetti (1982, 1985)). Consider n iid observations $z_1, \ldots, z_n$ and denote by L_p the log-likelihood of the model with p parameters. Akaike's Criterion amounts to choosing the model that minimizes $-2L_p+2p$. This procedure may be viewed as an extension of the likelihood principle and is based on a general information theoretic criterion. In fact $2L_p - 2p$ is a suitable estimate of the expected entropy of the model and by the Akaike Criterion the entropy will be, at least approximately, maximized (cf. Akaike (1973)). The criterion can be generalized by replacing $2p$ by αp for a given fixed α (cf. Bhansali and Downham (1977)). The Akaike Criterion is based on the computation of the log-likelihood function at the maximum likelihood estimator for θ. Since it is well known that maximum likelihood estimators are nonrobust for many important parametric models, we prefer to use general M-estimators (Huber (1981)).

A general M-estimator is defined as the minimum with respect to θ of the objective function $\sum_i \tau(z_i, \theta)$, for a given function τ, and satisfies the first order condition

$$\sum_i \psi(z_i, \theta) = 0, \tag{3.1}$$

where $\psi(z,\theta) = \partial\tau(z,\theta)/\partial\theta$. If we choose $\tau(z,\theta) = -\log p_\theta(z)$, where p_θ is the density of P_θ, the objective function equals minus the log-likelihood function, ψ is the score function, and the corresponding M-estimator is the maximum likelihood estimator. Also (2.1) defines an M-estimator for regression based on a special function ψ which takes into account the structure of the regression model.

In order to derive the Akaike Criterion based on a general M-estimator, it is helpful to look at such an estimator as a maximum likelihood estimator with respect to an underlying density $p_\theta(z)$ proportional to $\exp(-\tau(z,\theta))$. (Of course, this is true only when the function τ satisfies certain conditions but this does not affect the result below.) Then, we can write the usual Akaike Criterion based on this density and we obtain the following robust version

$$AICR(p;\alpha_p,\tau) = 2\sum_i \tau(z_i,\hat{\theta}) + \alpha_p, \tag{3.2}$$

where $\hat{\theta}$ is the general M-estimator defined by (3.1), $\alpha_p = 2\,\mathrm{tr}(M^{-1}Q), M = -\mathrm{E}[\partial\psi/\partial\theta], Q = \mathrm{E}[\psi\psi^T]$. This choice of α_p follows from the asymptotic equivalence of the Akaike Criterion given in Stone (1977).

If we apply (3.2) to the linear model of subsection 2.2 by using Huber's estimator (Huber (1981)), that is $\tau(z,\beta) = \rho_c((y - x^T\beta)/\sigma)$, where $\rho_c(r) = r^2/2$ if $|r| \le c$ and $\rho_c(r) = c\,|r| - c^2/2$ otherwise , (3.2) gives the robust criterion for regression proposed by Ronchetti (1982, 1985) (cf. also Hampel, Ronchetti, Rousseeuw and Stahel (1986), p. 366, formula (7.3.14)). In this case $\alpha_p = 2p\mathrm{E}[\psi_c^2]/\mathrm{E}[\psi_c']$, where $\psi_c(r) = \partial\rho_c(r)/\partial r = \max(-c,\min(r,c))$. By a different way, Hampel (1983) obtained in this case the slightly different value $\alpha_p = p(\mathrm{E}[\psi_c^2]/\mathrm{E}[\psi_c'] + \mathrm{E}[\psi_c^2]/(\mathrm{E}[\psi_c'])^2)$ which differs little from $2p$ for the usual values of c between 1.3 and 1.6.

Härdle (1987) investigated the properties of a selection criterion for regression which is asymptotically equivalent to AICR and showed that it is asymptotically optimal in the sense of Shibata (1981).

Hurvich and Tsai (1990) developed a small sample criterion for the selection of least absolute deviations regression models. Their criterion provides an exactly unbiased estimator for the expected Kullback-Leibler information when the underlying distribution of the errors is double exponential. This selection procedures performs better than the usual normal-based Akaike criterion and the robust criterion AICR based on $\tau(z,\beta) = \left|y - x^T\beta\right|$.

Machado (1993) investigated the robustness properties of the Schwarz (1978) criterion defined by the minimization of $-2L_p + p\log(n)$, where L_p is the log-likelihood of the model with p parameters and n is the sample size. He proposed a

robustification of this criterion which corresponds to (3.2) with the same penalty as in the classical case, that is $\alpha_p = p\log(n)$.

3.2. Autoregressive models

Martin (1980) and Behrens (1991) adapted (3.2) to autoregressive models. Their robust criterion is based on M-estimators $\hat{\phi} = (\hat{\phi}_1, \ldots, \hat{\phi}_p)^T$ for autoregressive models defined by $\sum_{t=p+1}^{n} \rho(Z_{t-1,p}, \hat{u}_{t,p}) = \min$, or by the first order conditions $\sum_{t=p+1}^{n} \psi(Z_{t-1,p}, \hat{u}_{t,p}) Z_{t-1,p} = 0$, where $Z_{t-1,p} = (z_{t-1}, \ldots, z_{t-p})^T$, $\hat{u}_{t,p} = (z_t - \hat{\phi}^T Z_{t-1,p})/s$, $\psi(z,u) = \partial\rho(z,u)/\partial u$, and s is a scale estimate. Then, the robust model selection criterion is defined by

$$RAIC(p) = \sum_{t=p+1}^{n} \rho(Z_{t-1,p}, \hat{u}_{t,p}) + \alpha_p. \tag{3.3}$$

Martin (1980) chooses $\alpha_p = 2(p+2)/(n-1)$ and Behrens (1991) chooses essentially $\alpha(p) = \mathrm{tr}(\tilde{M}^{-1}\tilde{Q})$, where $\tilde{M} = \mathrm{E}[\psi'_t Z_{t-1,p} Z^T_{t-1,p}]$, $\tilde{Q} = \mathrm{E}[\psi^2_t Z_{t-1,p} Z^T_{t-1,p}]$, $\psi_t(z,u) = \psi(Z_{t-1,p}, u_{t,p})$, and $\psi'(z,u) = \partial\psi(z,u)/\partial u$. A comparison of (3,2) and (3.3) shows that Behrens's choice of α_p is the natural extension of AICR to autoregressive models. Actually, she shows that this criterion is asymptotically optimal in the sense of Shibata (1981). Martin (1980) presents a few examples which show the performance of the robust criterion in the presence of a small amount of outliers in the data.

4. Model Choice VIA Testing of Non-Nested Hypotheses

Assume under the hypothesis a model F^0_α (with density $f^0(z;\alpha)$) and under the alternative a model F^1_β (with density $f^1(z;\beta)$), where α and β are p x 1 and q x 1 parameter vectors respectively. Cox (1961, 1962) proposed the following statistic

$$U_{\mathrm{Cox}} = n^{-1} \sum_i \log\left[\frac{f^0(z_i;\hat{\alpha})}{f^1(z_i;\hat{\beta})}\right] - \int \log\left[\frac{f^0(z;\hat{\alpha})}{f^1(z;\beta_{\hat{\alpha}})}\right] f^0(z;\hat{\alpha})dz, \tag{4.1}$$

where $\beta_{\hat{\alpha}}$ is the pseudo maximum likelihood estimator defined as the solution in β of $\int \frac{\partial}{\partial\beta} \log f^1(z;\beta) f^0(z;\hat{\alpha})dz = 0$.

Two modifications of U_{Cox} have been proposed by Atkinson (1970) ($\hat{\beta}$ is replaced by $\beta_{\hat{\alpha}}$) and by White (1982) ($\beta_{\hat{\alpha}}$ is replaced by $\hat{\beta}$). In these three cases $\sqrt{n}U_{\mathrm{Cox}}$ is asymptotically normal.

The tests based on these statistics are widely used as model selection criteria in a variety of situations including, for instance, the statistical analysis of income distributions. It is well known that they suffer from two major problems, namely the lack of accuracy of the asymptotic approximation of the sample distribution

of the test statistic (Atkinson (1970), Godfrey and Pesaran (1983) and the lack of robustness against misspecifications of the underlying distributions Hall (1985)). The following tables taken from Victoria-Feser (1997) illustrate these problems. Table 2 compares the exact finite sample level of the Cox and Atkinson tests (obtained by simulation) with the asymptotic approximation for samples of size 200. It is clear that the accuracy of the latter is poor. Table 3 shows the actual levels (obtained by simulation) of the same tests in the presence of a contaminated underlying distribution. The lack of robustness appears very clearly even for small amounts of contamination. Similar results are found for other situations and are reported in Victoria-Feser (1997).

Table 2. Asymptotic nominal level and finite sample level (%) of the Cox and Atkinson test (Pareto against exponential)

As. Level	Finite Sample (Cox)	Finite Sample (Atk.)
1	1.2	2.0
3	2.0	2.4
5	2.7	2.9
10	4.1	4.5

Table 3. Actual levels (in %) of the Atkinson statistic with contamination (Pareto against exponential)

Amount of	Nominal levels (in %)			
contamination	1%	3%	5%	10%
0%	2.1	3.1	3.5	5.2
1%	2.9	3.7	5.4	10.3
2%	5.2	7.3	8.9	10.9
3%	6.3	8.7	10.3	14.7
4%	9.4	12.5	14.5	18.3
5%	10.3	15.4	17.1	21.9
6%	13.1	18.5	22.5	27.6
7%	15.1	20.7	23.5	29.9

The nonrobustness of these tests can be explained easily by computing the influence function of their level. This function describes the bias on the level of a small amount of contamination in the underlying distribution of the observations (cf. Hampel, Ronchetti, Rousseeuw and Stahel (1986), Ch. 3). Victoria-Feser (1997) shows that the level influence function of these tests is in general unbounded. This explains the large bias which appears in Table 3. It is due to two factors: the nonrobustness of the test statistic and the nonrobustness of the parameter estimation (unboundedness of the influence function of the maximum likelihood estimator $\hat{\alpha}$).

Since Cox's test can be viewed as a parametric scores test for a compound model, Victoria-Feser (1997) derived a robust version by using results on robust tests for general parametric models obtained by Heritier and Ronchetti (1994). Table 4 shows the remarkable stability of the robust test under contamination. As a side effect, the robust version clearly also improve the accuracy of the asymptotic approximation of the sample distribution.

Table 4. Actual levels (in %) of the robust Atkinson statistic with contamination (Pareto against exponential)

Amount of	Nominal levels			
contamination	1%	3%	5%	10%
0%	1.3	3.5	5.5	10.2
1%	1.8	3.1	5.3	10.3
2%	0.6	2.5	4.6	10.1
3%	1.2	3.3	5.1	10.3
4%	1.7	4.2	6.0	12.6
5%	0.5	2.3	4.5	10.3
6%	1.4	3.6	5.4	10.7
7%	0.9	3.2	5.4	9.4

5. Conclusions

In this review we discussed the robustness issue in the selection of a statistical model and summarized the basic concepts which can be applied in order to robustify the classical selection procedures. Much work remains to be done. Three main directions appear to be important. The first one concerns robust model selection in time series. Some results for autoregressive models are discussed in subsection 3.2. It seems important to develop a robust selection procedure for general ARMA models. The second one is related to the use of cross-validation and other resampling techniques as model selection procedures. It appears important to integrate the robustness aspect in this type of procedures. A first step in the framework of linear models is given by Ronchetti, Field and Blanchard (1997) who robustify a least squares selection procedure based on cross-validation proposed by Shao (1993). The third one is an extensive numerical comparison by Monte Carlo of different robust model selection procedures. Results for a few special situations can be found in Härdle (1987), Hurvich and Tsai (1990), Behrens (1991), and Machado (1993).

Acknowledgements

The author would like to thank the Editor, the Associate Editor, and two Referees for their helpful comments.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In 2*nd International Symposium on Information Theory* (Edited by B. N. Petrov and F. Csáki), 267-281. Akadémiai Kiadó, Budapest.

Antoch, J. (1986). Algoritmic development in variable selection procedures. *Proceedings of COMPSTAT* 1986, 83-90, Physica-Verlag, Heidelberg.

Antoch, J. (1987). Variable selection in linear models based on trimmed least squares estimator. In *Statistical Data Analysis Based on the L_1-norm and Related Methods* (Edited by Y. Dodge), 231-245. North Holland.

Atkinson, A. C. (1970). A method for discriminating between models. *J. Roy. Statist. Soc. Ser. B* **32**, 323-353.

Behrens, J. (1991). *Robuste Ordnungswahl für autoregressive Prozesse.* Ph.D. Thesis, University of Kaiserslautern, Germany.

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika* **64**, 547-551.

Cox, D. R. (1961). Tests of separate families of hypotheses. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 105-123, University of California Press, Berkeley, CA.

Cox, D. R. (1962). Further results on tests of separate families of hypotheses. *J. Roy. Statist. Soc. Ser. B* **24**, 406-424.

Godfrey, L. G. and Pesaran, M. H. (1983). Tests of non-nested regression models: small sample adjustments and Monte Carlo evidence. *J. Econom.* **21**, 133-154.

Hall, A. (1985). A simplified method of calculating the distribution free Cox test. *Econom. Lett.* **18**, 149-151.

Hampel, F. R. (1983). Some aspects of model choice in robust statistics. *Proceedings of the 44th Session of the ISI*, Book 2, 767-771, Madrid.

Hampel, F. R. (1991). Some mixed questions and comments on robustness. In *Directions in Robust Statistics and Diagnostics*, IMA Volumes in Mathematics and Its Applications, **33** (Edited by W. Stahel and S. Weisberg), 101-111. Springer, New York.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley and Sons, New York.

Härdle, W. (1987). An effective selection of regression variables when the error distribution is incorrectly specified. *Ann. Inst. Statist. Math.* **39**, 533-548.

Heritier, S. and Ronchetti, E. (1994). Robust bounded-influence tests in general parametric models. *J. Amer. Statist. Assoc.* **89**, 897-904.

Huber, P. J. (1981). *Robust Statistics.* John Wiley and Sons, New York.

Hurvich, C. M. and Tsai, C. L. (1990). Model selection for least absolute deviations regression in small samples. *Statist. Probab. Lett.* **9**, 259-265.

Koenker, R. and Bassett, G. Jr. (1978). Regression quantiles. *Econometrica* **46**, 33-50.

Léger, C. and Altman, N. (1993). Assessing influence in variable selection problems. *J. Amer. Statist. Assoc.* **88**, 547-556.

Machado, J. A. F. (1993). Robust model selection and M-estimation. *Econometric Theory* **9**, 478-493.

Mallows, C. L. (1973). Some comments on C_P. *Technometrics* **15**, 661-675.

Martin, R. D. (1980). Robust Estimation of Autoregressive Models. In *Directions in Time Series* (Edited by D. R. Brillinger and G. C. Tiao), 228-262. Institute of Mathematical Statistics, Hayward, CA.

Qian, G. and Künsch, H. R. (1996). On model selection in robust linear regression. Research Report 80, ETH Zünch.

Ronchetti, E. (1982). *Robust Testing in Linear Models: The Infinitesimal Approach.* Ph.D. Thesis, ETH Zürich.

Ronchetti, E. (1985). Robust model selection in regression. *Statist. Probab. Lett.* **3**, 21-23.

Ronchetti, E. and Staudte, R. G. (1994). A Robust Version of Mallows's C_P. *J. Amer. Statist. Assoc.* **89**, 550-559.

Ronchetti, E., Field, C. A. and Blanchard, W. (1997). Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **92** (to appear).

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

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

Shi, P. and Tsai, C. L. (1996). A note on the unification of the Akaike information criterion, manuscript.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

Sommer, S. and Staudte, R. G. (1995). Robust variable selection in regression in the presence of outliers and leverage points. *Australian Journal of Statistics* **37**, 323-336.

Sommer, S. and Huggins, R. M. (1996). Variables selection using the wald test and a robust C_p. *Appl. Statist.* **45**, 15-29.

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

Victoria-Feser, M. P. (1997). A robust model choice test for non-nested hypotheses. *J. Roy. Statist. Soc. Ser. B* **59** (to appear).

White, H. (1982). Regularity conditions for Cox's test of non-nested hypotheses. *J. Econometrics* **19**, 301-318.

Department of Econometrics, University of Geneva, CH-1211 Geneva, Switzerland.

(Received February 1995; accepted August 1996)