ROBUST STEPWISE REGRESSION

C. Agostinelli

2000.10

Dipartimento di Scienze Statistiche
Università degli Studi
Via S. Francesco, 33
35121 Padova

Agosto 2000

Robust Stepwise Regression

Claudio Agostinelli *
Department of Statistics
University of Padova
35121 Padova, Italia

September 8, 2000

Abstract The selection of an appropriate sub-set of explanatory variables to use in a linear regression model is an important aspect of a statistical analysis. Classical stepwise regression could be invalidated by a few outling observations. We introduce a robust F-test in order to perform a stepwise regression that is robust against the presence of outliers. The introduced methodology is asymptotically equivalent to the classical one when no contamination is present. Some examples and simulation are presented.

Keywords: F-test, Robust backward, Robust forward, Robust stepwise, Weighted F-test, Weighted likelihood.

1 Introduction

In order to select from a wide set an appropriate sub-set of explanatory variables to the aim of specified a linear regression model several statistical methods are available. Some of them are the Mallows Cp, (Mallows, 1973), the AIC (Akaike, 1973) and the Cross-validation (Stone, 1974 and Shao, 1993). These methods needs the calculation of all possible linear regression sub-model. Hence when p possible regressors are present the sub-model are

*Address of the author: Claudio Agostinelli, Dipartimento di Statistica, Via San Francesco n. 33, Università di Padova, 35121 Padova, Italia. e-mail: claudio@stat.unipd.it

$2^p - 1$ and the presence of one new variable double the number of sub-model to be considered. Often, the evaluation of all possible sub-model drive the application of the methods to be unfeasible.

From the 60s a method called Stepwise regression (Efroymson, 1960, Goldberger and Jochems, 1961, Goldberger (1961)) that choose a sub-set of explanatory variables exploring only few possible sub-model (Garside, 1965 and Beale, Kendall and Mann, 1967) was developed. Miller (1984) give a comparison between the stepwise method and other model selection procedures.

All these methods could be very sensitive to the presence of a few outling observations. A review of classical robust model choice procedures can be found in Ronchetti (1997). Agostinelli (1999, 2000) use the weighted likelihood approach (Markatou, Basu and Lindsay, 1995, 1998) to define robust version of the Mallows Cp, AIC and Cross-Validation.

In this paper we introduce an F-test function (WF-test) based on weighted likelihood in order to achieve a robust model selection procedure based on stepwise method. The WF-test is asymptotic equivalent to the classical F-test when no contamination is involved. It also agree with the definition of the robust test function proposed in Agostinelli (1998c) and Agostinelli and Markatou (2000) for the weighted likelihood ratio test function, and hence it share the same robustness properties.

Section 2 introduce the weighted likelihood in particular regarding the approach for the linear regression model, in Section 3 the WF-test function is derived and the asymptotic properties are studied. Section 4 discuss the robust stepwise regression and Section 5 presents some examples and a Monte

Carlo simulations.

2 The weighted likelihood approach

Let $x_1, x_2, \ldots, x_n$ be a sample from the random variable X with density $f(\cdot)$ corresponding to the unknown probability measure $F(\cdot)$. We will use the density $m(\cdot;\theta)$ corresponding to the probability measure $M(\cdot;\theta)$ as a model for the random variable X. Note that in the maximum likelihood context $f(\cdot) \equiv m(\cdot;\theta_T)$ (almost surely). Let $u(x;\theta) = \frac{\partial}{\partial\theta}\log m(x;\theta)$ be the score function. Under regularity conditions the maximum likelihood estimator of θ is a solution of the likelihood equation $\sum_{i=1}^{n} u(x_i;\theta) = 0$.

Given any point x in the sample space, Markatou, Basu and Lindsay (1998) construct a weight function $w(x;\theta,\hat{F}_n)$ that depends on the chosen model distribution M and the empirical cumulative distribution $\hat{F}_n(t) = \sum_{i=1}^{n} \mathbf{1}_{x_i<t}/n$; then estimators for the parameter vector θ are obtained as solutions to the set of estimating equations:

$$\frac{1}{n}\sum_{i=1}^{n} w(x_i;\theta,\hat{F}_n)u(x_i;\theta) = 0 \tag{1}$$

The weight function $w(x;\theta,\hat{F}_n)$, by construction, takes values in the interval $[0,1]$ and it is defined as $w(x;\theta,\hat{F}_n) = \min\left\{1, \frac{[A(\delta(x;\theta,\hat{F}_n))+1]^+}{\delta(x;\theta,\hat{F}_n)+1}\right\}$ where $[\cdot]^+$ indicates the positive part.

The quantity $\delta(x;\theta,\hat{F}_n)$ is called Pearson residual and it is defined as $\delta(x;\theta,\hat{F}_n) = \frac{f^*(x)}{m^*(x;\theta)} - 1$, where $f^*(x;\theta) = \int k(x;t,h)\, d\hat{F}_n(t)$ is a kernel density estimator and $m^*(x;\theta) = \int k(x;t,h)\, dM(t;\theta)$ is the smoothed model density. The Pearson residual expresses the agreement between the data and the assumed probability model. The function $A(\cdot)$ is a residual adjustment

function, RAF (Lindsay, 1994) and it operates on Pearson residuals as the Huber ψ-function operates on the structural residuals. When $A(\delta) = \delta$ the weight $w(x; \theta, \hat{F}_n) \equiv 1$, and this corresponds to maximum likelihood. Generally, the weights w use functions $A(\cdot)$ that correspond to a minimum disparity problem. For example, the function $A(\delta) = 2\{(\delta + 1)^{1/2} - 1\}$ corresponds to Hellinger distance while the weight $w(\delta) = 1 - \delta^2/(\delta + 2)^2$, corresponds to the symmetric chi-squared distance. For an extensive discussion of the concept of the RAF see Lindsay (1994).

This weighting scheme provides fully efficient and robust estimators, in the sense of breakdown, provided that one selects a root based on using the parallel disparity measure (Markatou *et al.*, 1998).

An algorithm based on re-sampling techniques is used to identified the roots of the estimating equation 1. Sub-samples of fixed dimension and without replications are sampled from the dataset. From each of these sub-samples a maximum likelihood estimator are evaluated and used to start the re-weighted algorithm for solving the weighted likelihood estimating equations.

To calculate the Pearson residuals we need to select the smoothing parameter h. Markatou *et al.*, (1995) select $h^2 = g\sigma^2$, where g is a constant that is independent of the scale of the model and it is selected so that it assigns a very small weight to an outlying observation.

Agostinelli (1998a, 1998b) extended the methodology to the regression model. Let $\{y_1, \cdots, y_n\}$ a sample of dimension n from an unknown distribution and $\{x_1, \cdots, x_n\}$ a sample of vector from p explanatory variables which could included the intercept term. Considering the regression model

$y = x\beta + \varepsilon$ and assuming a parametric family $\mathcal{M} = \{m(\varepsilon;\sigma); \sigma \in \Sigma\}$ we let $z(\beta) = y - x\beta$ the residuals for a specific value of the parameter vector β and $\hat{F}_n(t;\beta) = \sum_{i=1}^n \mathbf{1}_{z_i(\beta)<t}/n$ the empirical cumulative distribution. Hence the Pearson residual would be on the shape $\delta(z;\sigma,\hat{F}_n(\beta)) = f^*(z;\beta)/m^*(z;\sigma) - 1$ where $f^*(z;\beta) = \int k(z;t,h)\ d\hat{F}_n(t;\beta)$ and $m^*(z;\sigma) = \int k(z;t,h)\ dM(t;\sigma)$. Therefore the weighted likelihood estimator of the parameter vector β is a solution of the estimating equation:

$$\sum_{i=1}^{n} w\left(z_i(\beta);\sigma,\hat{F}_n(\beta)\right) u\left(z_i(\beta);\sigma\right) = 0$$

where $u\left(z_i(\beta);\sigma\right) = \frac{\partial}{\partial\beta}\log m\left(z_i(\beta);\sigma\right)$, while an estimator of the nuisance parameter could be find as a solution of the following estimating equation:

$$\sum_{i=1}^{n} w\left(z_i(\beta);\sigma,\hat{F}_n(\beta)\right) u_\sigma\left(z_i(\beta);\sigma\right) = 0$$

where $u_\sigma\left(z_i(\beta);\sigma\right) = \frac{\partial}{\partial\beta}\log m\left(z_i(\beta);\sigma\right)$.

When $m(z;\sigma)$ belong to a normal scale family on the form $\mathcal{M} = \{\mathcal{N}(0,\sigma^2); \sigma^2 \in \mathcal{R}^+ /\{0\}\}$, the estimating equations are:

$$\begin{cases} \sum_{i=1}^n w\left(z_i(\beta);\sigma,\hat{F}_n(\beta)\right)\left(y_i - x_i\beta\right) x_i &= 0 \\ \sum_{i=1}^n w\left(z_i(\beta);\sigma,\hat{F}_n(\beta)\right)\left(\left(y_i - x_i\beta\right)^2 - \sigma^2\right) &= 0 \end{cases} \tag{2}$$

When the presence of leverage points is suspected an extended version of the weight function can be used (Agostinelli, 1998a).

3 The Weighted F Test

In this section we introduce a robust version of the F test based on the weighted likelihood. Let us consider the following linear regression model:

$$y = x\beta + \varepsilon$$

where y is a response variable, x is a vector of p possible explanatory variables, β is a vector of p unknown parameters and ε is a random error, with normal model $\mathcal{N}(0, \sigma)$, where σ is a scale parameter. Because some of the components of β may be 0, a reduced model might be used:

$$y = x_{\mathcal{A}}\beta_{\mathcal{A}} + \varepsilon$$

where $\mathcal{A}$ is a subset of $d_{\mathcal{A}}$ distinct positive integers that are less or equal to p and $\beta_{\mathcal{A}}$ (or $x_{\mathcal{A}}$) is the $d_{\mathcal{A}}$ vector containing the components of β (or x) that are indexed by the integers in $\mathcal{A}$.

Further, let $\mathcal{M}_{\mathcal{I}}$ the set of models $\mathcal{A}$ such that at least one nonzero component of β is not in $\beta_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{II}}$ the set of models $\mathcal{A}$ such that $\beta_{\mathcal{A}}$ contains all nonzero components of β. Hence $\mathcal{M}_{\mathcal{I}}$ is the set of all models that are a proper subset of the "true model", while $\mathcal{M}_{\mathcal{II}}$ is the set of all models such that the "true" model is a subset of them.

From the dataset we can estimate using the weighted likelihood estimating equations 2 the value of the parameters that best fit the majority of the data for the full model including all the p explanatory variables, namely $\hat{\beta}$, and the reduced model, namely $\hat{\beta}_{\mathcal{A}}$.

We use the final weights $\hat{w} = w(z(\hat{\beta}); \hat{\sigma}, \hat{F}_n(\hat{\beta}))$ from the full model to obtain an estimators of the scale parameter in the reduced models:

$$\hat{\sigma}^2_{\mathcal{A}} = \frac{1}{\sum_{i=1}^{n} \hat{w}_i} \sum_{i=1}^{n} \hat{w}_i z_i \left(\hat{\beta}_{\mathcal{A}}\right)^2 \tag{3}$$

Let now consider the weighted likelihood ratio test LRT_w:

$$LRT_w = \frac{\prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\hat{\sigma}^2_{\mathcal{A}}}} \exp\left[-\frac{1}{2} \frac{z_i \left(\hat{\beta}_{\mathcal{A}}\right)^2}{\hat{\sigma}^2_{\mathcal{A}}} \right] \right\}^{\hat{w}_i}}{\prod_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left[-\frac{1}{2} \frac{z_i \left(\hat{\beta}\right)^2}{\hat{\sigma}^2} \right] \right\}^{\hat{w}_i}} \tag{4}$$

$$
\begin{aligned}
&= \frac{(\hat{\sigma}^2_{\mathcal{A}})^{-\frac{\sum_{i=1}^n \hat{w}_i}{2}}}{(\hat{\sigma}^2)^{-\frac{\sum_{i=1}^n \hat{w}_i}{2}}} \prod_{i=1}^n \exp\left\{-\frac{1}{2}\frac{z_i\left(\hat{\beta}_{\mathcal{A}}\right)^2}{\hat{\sigma}^2_{\mathcal{A}}} + \frac{1}{2}\frac{z_i\left(\hat{\beta}\right)^2}{\hat{\sigma}^2}\right\}^{\hat{w}_i} \\
&= \frac{(\hat{\sigma}^2_{\mathcal{A}})^{-\frac{\sum_{i=1}^n \hat{w}_i}{2}}}{(\hat{\sigma}^2)^{-\frac{\sum_{i=1}^n \hat{w}_i}{2}}} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^n \hat{w}_i z_i\left(\hat{\beta}_{\mathcal{A}}\right)^2}{\hat{\sigma}^2_{\mathcal{A}}} + \frac{1}{2}\frac{\sum_{i=1}^n \hat{w}_i z_i\left(\hat{\beta}\right)^2}{\hat{\sigma}^2}\right\} \\
&= \left(\frac{\hat{\sigma}^2_{\mathcal{A}}}{\hat{\sigma}^2}\right)^{-\frac{\sum_{i=1}^n \hat{w}_i}{2}}
\end{aligned}
$$

Further let consider the follows transformation:

$$
\begin{aligned}
LRT^*_w &= LRT_w^{-\frac{2}{\sum_{i=1}^n \hat{w}_i}} - 1 \\
&= \frac{\hat{\sigma}^2_{\mathcal{A}} - \hat{\sigma}^2}{\hat{\sigma}^2}
\end{aligned}
$$

In the following we stated the Theorem for the asymptotic distribution of the LRT^*_w.

Theorem 3.1 *Under the following conditions:*

A1. $k(z,t,h)$ *is a bounded variation density kernel and the smoothing parameter* h *is a positive constant*

A2. $A(0)=0$, $A'(0)=1$ *e* $A''(\delta)$ *is a bounded continuous function of* δ

A3. the observations $z_i(\beta_0)$ *are from the model* $M(\cdot;\sigma_0)$ *with density* $m(\cdot;\sigma_0)$ *and* σ_0 *belong to the parametric space* Σ

A4. $\hat{\theta} = \{\hat{\beta};\hat{\sigma}\}$ *is a consistent estimator of* $\{\beta_0;\sigma_0\}$

A5. $\sup_z \left|\frac{\partial}{\partial\sigma} M(z;\sigma)\right| < \infty$

A6. $\sup_z \left|\hat{F}_n(z;\hat{\beta}) - M(z;\hat{\sigma})\right| \xrightarrow{p} 0$

we have the following asymptotic result

$$LRT_w^* \ \frac{\sum_{i=1}^n \hat{w}_i - p}{q} \sim F_{q;n-p}$$

for all the models in $\mathcal{M}_{II}$ where q is the number of parameters that do not belong to $\mathcal{A}$.

Proof: Note that, the condition *A3.* ensure that the stochastic model is correctly specified and hence no contamination is present. Under this fact it could be shown that (Agostinelli, 1998a)

$$\sup_i |\hat{w}_i - 1| \xrightarrow{p} 0.$$

Let $\hat{\beta}_L$ and $\hat{\sigma}_L^2$ the maximum likelihood estimator of the coefficients and scale parameters and let $LTR^* = (\sigma^2_{\mathcal{A}\ L} - \sigma_L^2)/\sigma_L^2$ the classical likelihood ratio test, then

$$\begin{aligned}
\frac{1}{n}\left|\sum_{i=1}^n \hat{w}_i z_i(\hat{\beta})^2 - \sum_{i=1}^n z_i(\hat{\beta}_L)^2\right| &\leq \frac{1}{n}\left|\sum_{i=1}^n \hat{w}_i z_i(\hat{\beta})^2 - \sum_{i=1}^n \hat{w}_i z_i(\hat{\beta}_L)^2\right| \\
&+ \frac{1}{n}\left|\sum_{i=1}^n \hat{w}_i z_i(\hat{\beta}_L)^2 - \sum_{i=1}^n z_i(\hat{\beta}_L)^2\right| \\
&\leq \sup_i \hat{w}_i \ \frac{1}{n}\sum_{i=1}^n \left|z_i(\hat{\beta})^2 - z_i(\hat{\beta}_L)^2\right| \\
&+ \sup_i |\hat{w}_i - 1| \frac{1}{n}\sum_{i=1}^n z_i(\hat{\beta}_L)^2 \\
&\xrightarrow{p} 0
\end{aligned}$$

as $n \to \infty$. Hence $\left|\sum_{i=1}^n \hat{w}_i z_i(\hat{\beta})^2 - \sum_{i=1}^n z_i(\hat{\beta}_L)^2\right| = o_p(n)$ but $\sum_{i=1}^n \hat{w}_i = O_p(n)$. Then as $n \to \infty$ we have

$$\left|\hat{\sigma}^2 - \hat{\sigma}_L^2\right| = \left|\frac{1}{\sum_{i=1}^n \hat{w}_i}\sum_{i=1}^n \hat{w}_i z_i(\hat{\beta})^2 - \frac{1}{n}\sum_{i=1}^n z_i(\hat{\beta}_L)^2\right| \xrightarrow{p} 0$$

Similarly we have $|\hat{\sigma}^2_{\mathcal{A}} - \hat{\sigma}^2_{\mathcal{A}\ L}| \xrightarrow{p} 0$ for all $\mathcal{A} \in \mathcal{M}_{\mathcal{I}\mathcal{I}}$ and finally $|LTR^*_w - LTR^*| \xrightarrow{p} 0$.

This means that when no contamination is present in the data the LRT^*_w $(\sum_{i=1}^n \hat{w}_i - p)/q$ has the same asymptotic distribution of the classical LRT^* $(n-p)/q$ based on the maximum likelihood, that is, an F with q and $n-p$ degree of freedom for all $\mathcal{A} \in \mathcal{M}_{\mathcal{I}\mathcal{I}}$.

On the other hand when the data is contaminated it is preferable to compare the LRT^*_w $(\sum_{i=1}^n \hat{w}_i - p)/q$ with an $F_{q;\hat{n}-p}$ where $\hat{n}$ is the integer number closed to $\sum_{i=1}^n \hat{w}_i$.

Note, that the definition 4 agree with the definition of the weighted likelihood ratio test function λ_w defined in Agostinelli (1998c) and Agostinelli and Markatou (2000) since, with the normal model we have

$$\lambda_w = \log\left(\frac{\hat{\sigma}^2_{\mathcal{A}}}{\hat{\sigma}^2}\right) \sum_{i=1}^n \hat{w}_i$$

In order to find, under the assumptions stated in Theorem 3.1, an WF-test function that is asymptotic equivalent to the classical one regardless of the considered model we need to replace the WLEE $\hat{\beta}_{\mathcal{A}}$ with a weighted least square estimator $\tilde{\beta}_{\mathcal{A}}$ with weights based on the full model, i.e. $\hat{w}_i$. The proof of this fact is similar to that of the Theorem 3.1 and will be omitted. Further, let W_{ls}F-test the test based on the weighted least square, it can be shown that the WF-test and the W_{ls}F-test are asymptotic equivalent for all models in $\mathcal{M}_{\mathcal{I}\mathcal{I}}$.

4 The Robust Stepwise Regression

In this section we introduce a robust version of the Forward, Backward and Stepwise regression methods based on the WF-test or on the W_{ls}F-test. We now introduce the Forward selection algorithm. First of all we estimate the weights $\hat{w}_i$ from the full model. The residual sum of squares of a particular set $\mathcal{A}$ of variables is then

$$RSS_{\mathcal{A}} = \sum_{i=1}^{n} \hat{w}_i \; z_i(\hat{\beta}_{\mathcal{A}})^2.$$

The same definitions and results hold when we define $RSS_{\mathcal{A}}$ based on weighted least square estimator as

$$\tilde{R}SS_{\mathcal{A}} = \sum_{i=1}^{n} \hat{w}_i \; z_i(\tilde{\beta}_{\mathcal{A}})^2$$

where $\tilde{\beta}_{\mathcal{A}}$ is the solution of the weighted least square estimator based on $\hat{w}_i$.

We start with the intercept variable in $\mathcal{A}$. Suppose the smallest RSS which can be obtained by adding another variable to the present set is $RSS_{\mathcal{A}+1}$. The ratio

$$R_e = \frac{RSS_{\mathcal{A}} - RSS_{\mathcal{A}+1}}{RSS_{\mathcal{A}+1}/(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}} - 1)} \tag{5}$$

is calculated and compared with a threshold 'F-to-enter' value, say F_e. If $R_e > F_e$ the variable is added to $\mathcal{A}$ and the algorithm follows until no other variables can be added.

Differently, in the Backward selection algorithm we start with all the explanatory variables in $\mathcal{A}$. Let $RSS_{\mathcal{A}-1}$ be the smallest RSS wich can be obtained after deleting any variable from the previously selected variables. The ratio

$$R_d = \frac{RSS_{\mathcal{A}-1} - RSS_{\mathcal{A}}}{RSS_{\mathcal{A}}/(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}})} \tag{6}$$

is calculated and compared with a threshold 'F-to-delete' value, say F_d. If $R_d < F_d$ the variable is deleted from $\mathcal{A}$.

The Stepwise selection algorithm is a variation on the Forward selection method. After each variable is added to $\mathcal{A}$ by R_e a test is made to see if any of the previously selected variables can be deleted by R_d.

While it is easy to see that the Forward and Backward selection algorithms will stop in a finite number di steps for the convergence of the Stepwise selection algorithm we follow Miller (1990).

From 5 it follows that when the criterion for adding a variable is satisfied

$$RSS_{\mathcal{A}+1} \leq \frac{RSS_{\mathcal{A}}}{1 + F_e / \left(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}} - 1\right)}$$

while from 6 it follows that when the criterion for deletion of a variable is satisfied

$$RSS_{\mathcal{A}} \leq RSS_{\mathcal{A}+1} \left\{1 + F_d / \left(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}}\right)\right\}$$

Hence when an addition is followed by a deletion, the new RSS, say $RSS^*_{\mathcal{A}}$, is such that

$$RSS^*_{\mathcal{A}} \leq RSS_{\mathcal{A}} \frac{1 + F_d / \left(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}}\right)}{1 + F_e / \left(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}} - 1\right)} \tag{7}$$

The procedure stops when no further additions or deletions are possible which satisfy the criteria. As each $RSS_{\mathcal{A}}$ is bounded below by the smallest RSS for any subset of $d_{\mathcal{A}}$ variables, by ensuring that the RSS is reduced each time that a new subset of $d_{\mathcal{A}}$ variables is found, convergence is guaranteed. From 7 it follows that

$$\frac{1 + F_d / \left(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}}\right)}{1 + F_e / \left(\sum_{i=1}^{n} \hat{w}_i - d_{\mathcal{A}} - 1\right)} < 1$$

and hence a sufficient condition is $F_d < F_e$.

5 Examples and Simulations

In this section we present two examples and a Monte Carlo simulations. All functions related to the weighted versions have been written in Fortran 77 and they have been interfaced with R (CRAN, Ihaka and Gentleman, 1996). They are available in any CRAN mirror. All examples and simulations are ran on a PC under Linux OS.

For the weights we have used a normal kernel with the smoothing parameter equal to $g\ \hat{\sigma}^2$ and $g = 0.032$. A Hellinger Residual Adjustment Function is used. To identify the roots of the estimating equations a bootstrap approach is used with 100 bootstrap sub-samples with dimension 10 and 20 for the two sample size.

Example: WF_{ls}-test. In this example we show the performance of the WF_{ls}-test in a particular context. We have simulated 50 observations from the normal regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ where $\epsilon \sim \mathcal{N}(0,1)$, $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\} = \{5, 0, 0, 0.25, 0.18\}$ and $x_1 \sim \mathcal{U}(-4,4)$, $x_2 \sim \mathcal{U}(-6,6)$, $x_3 \sim \mathcal{U}(-8,8)$, $x_4 \sim \mathcal{U}(-10,10)$. Other 50 observations were added, progressively, to the first 50 observations to get 51 different dataset, from size 50 to 100. These last observations are generated with the same model with parameters: $\epsilon \sim \mathcal{N}(0, 0.5)$, $\{\beta_0, \beta_1, \beta_2, \beta_3, \beta_4\} = \{4, 0, 0, -0.25, -0.18\}$ and $x_1 \sim \mathcal{U}(3.5, 4.5)$, $x_2 \sim \mathcal{U}(5.5, 6.5)$, $x_3 \sim \mathcal{U}(-8.5, -7.5)$, $x_4 \sim \mathcal{U}(-10.5, -9.5)$. Further, the contamination level of each dataset is equal to $i/(50+i)$ where i is the number of observations from the second model added to the first one. For each of the 51 dataset we have test the

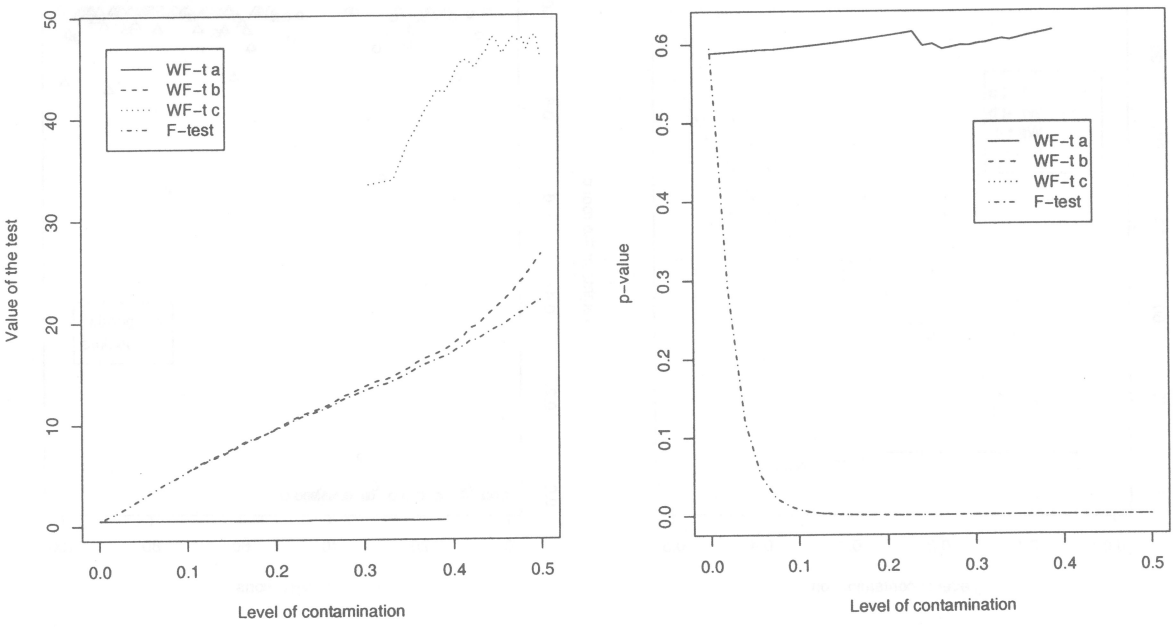

Figure 1: Value of the WF-test and F-test and their p-value for H_{0a}.

following two hypothesis sets:

$$\begin{cases} H_{0a} & : \beta_1 = 0, \beta_2 = 0, \beta_3 \neq 0, \beta_4 \neq 0 \\ H_{1a} & : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0 \end{cases} \quad \begin{cases} H_{0b} & : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = 0, \beta_4 = 0 \\ H_{1b} & : \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0 \end{cases}$$

In figure 1 we report the results for the H_{0a}, H_{1a} set, in its left part the value of the WF_{ls}-test and F-test, on the right their corresponding p-value. Since for certain levels of contamination the weighted likelihood estimating equation have more than one root, for illustrative purpose, we report the value of WF_{ls}-test corresponding to each of them.

To pick up one root it is possible to use the parallel disparity measure approach as describe in Markatou *et al.* (1998). However in a real situation it would be very important to consider the results from each roots in order to have a complete analysis.

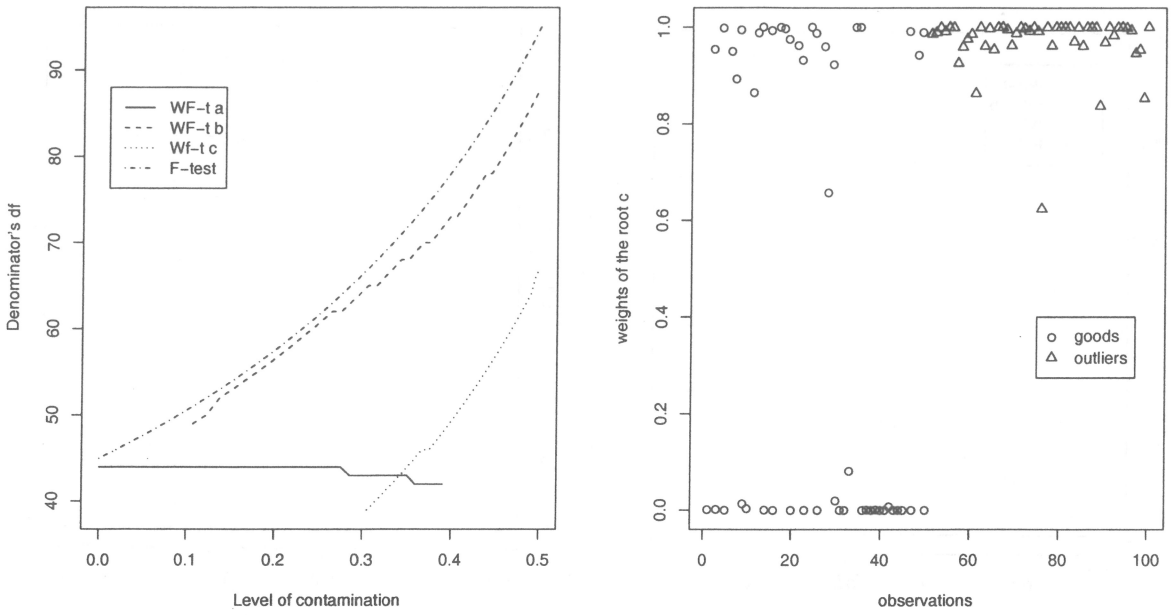

Figure 2: Denominator's degree of freedom for WF_{ls}-test and F-test, and the weights associated to WF_{ls}-test c.

The WF_{ls}-test associated to the root a, that is the robust one, (WF-t a) perform very well until 40% of contamination level, after that the root disappear. The WF_{ls}-test associated to the root b (WF-t b) behave like the classical F-test, while the WF-t c appear only with high levels of contamination and corresponding to a root c that downweight only some of the good points as show in the right part of the figure 2 for the dataset with 100 observations. In the left part of the same figure, we report the degree of freedom of the denominator in the WF_{ls}-test and F-test for each root.

Finally, in figure 3 we report the same results for H_{0b} and H_{1b} set. The same good performance of the WF-t a is illustrated.

Example: Aerobic Fitness Prediction. To illustrate the new intro-

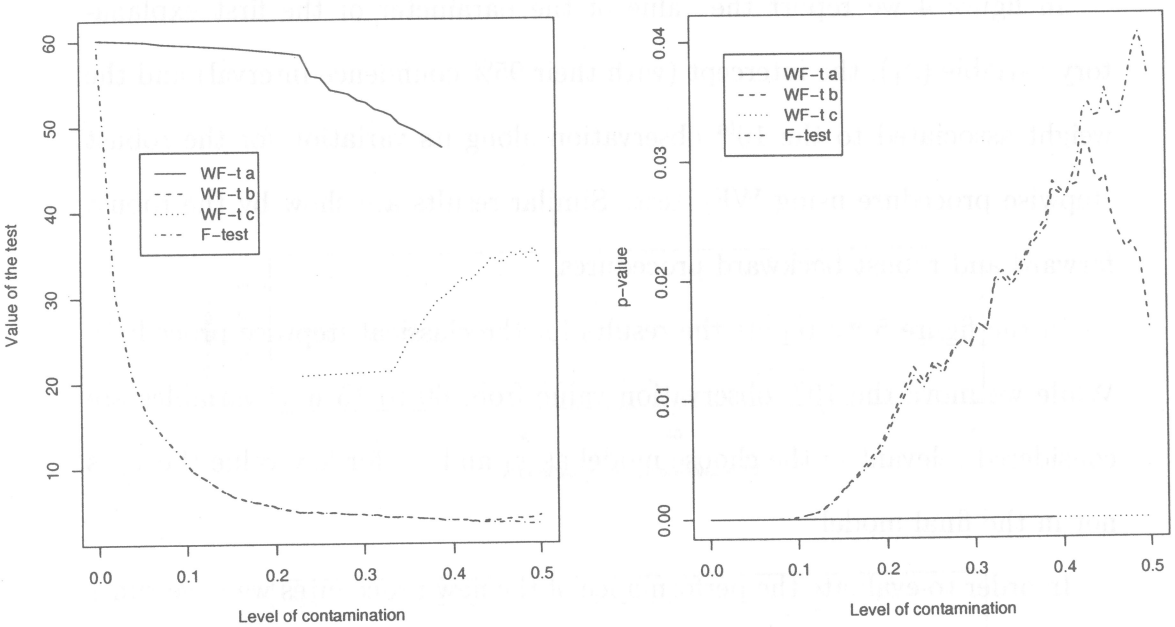

Figure 3: Value of the WF_{ls}-test and F-test and their p-value for H_{0b}.

duced methods, we have considered the dataset from the SAS/STAT User's Guide (1990, pag. 1443). This dataset has 31 observations, one dependent variable (y = oxygen intake rate, ml per kg of body weight per minute) and 6 explanatory variables (x_1 = time to run 1.5 miles (minutes), x_2 = age (year), x_3 = weight (kg), x_4 = heart rate while running (same time oxygen rate measured), x_5 = maximum heart rate recorded while running and x_6 = heart rate while resting). Using stepwise procedure with $F_e = 4$ and $F_d = 2$ (or forward with $F_e = 4$, or backward, $F_d = 8$) it turn out that only the first explanatory variable should be include in the model. We have the same result using the weighted version. To evaluated the stability of this result we run a sensitivity analysis moving observation 10 from 15 to 60 with step 2 (its original value was 60.055). We choose this observation since it is a

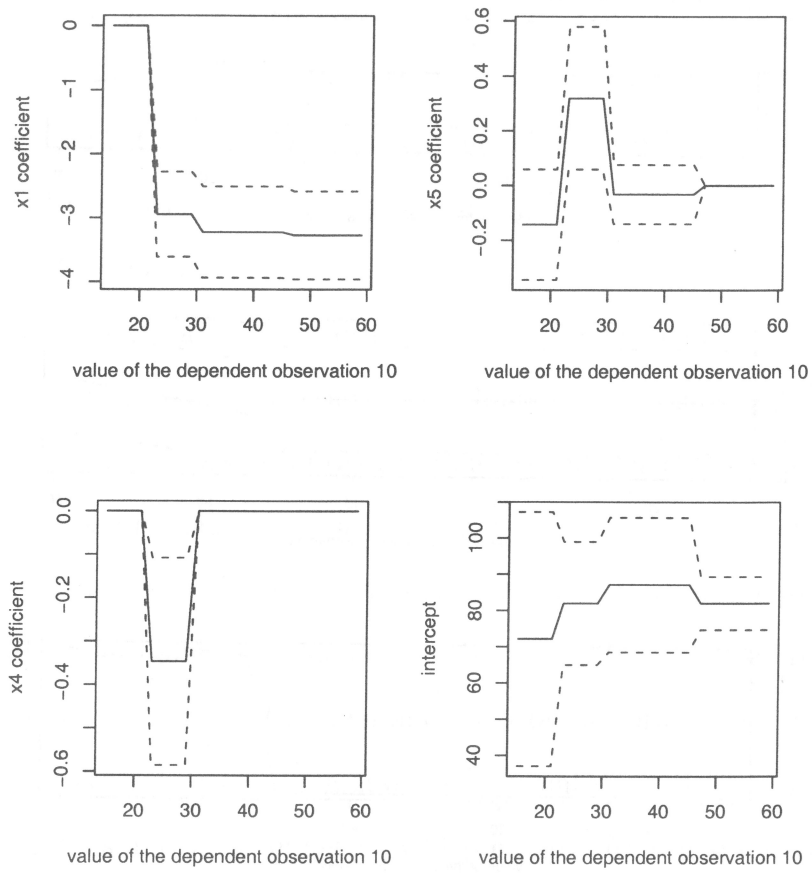

Figure 5: The variables considered by the classical stepwise procedure.

selection procedure based on WF-test and WF_{ls}-test. We let $F_e = 4$ and $F_d = 4$. For each situation we carried out 200 Monte Carlo simulation runs. In the case of multiple roots for the full model, the weights associated to the roots with smaller scale variance but total weights bigger then 0.6 are used in the evaluation in each Monte Carlo run.

The entries give the actual number of the runs falling into each category. The category "Correct" means that the correct model was chosen and is the key measure of the performance. "Extra 1" means that a model with one extra variable was chosen for which the true model is a proper subset. "Missing 1" indicates that the chosen model differed in one missing variable; "Extra 2", "Extra 3", "Extra 4" and "Missing 2", "Missing 3", "Missing 4" follow a similar pattern. "Other" means that the chosen model is not a subset or does not include the true model.

Table 1 gives the results for $n = 30$ and table 2 gives the results for $n = 60$. The performance of the procedures based on WF-test with respet to those based on WF_{ls}-test is very similar regardless of sample size and contamination type. The robust property are very goods for distribution e_2 (symmetric contamination case) and e_4 (asymmetric contamination case) while they are not performing very well for contamination type e_3. The efficiency is very similar to that of the classical procedure for distribution e_1 (the results for the classical procedures are not reported).

References

Agostinelli, C., (1998a). Inferenza statistica robusta basata sulla funzione di verosimiglianza pesata: alcuni sviluppi, Ph.D. Thesis, Dipartimento di

	e1			e2			e3			e4		
	Step	*For*	*Back*	*Step*	*For*	*Back*	*Step*	*For*	*Back*	*Step*	*For*	*Back*
$n = 30$, model p1, WF-test												
Correct	152	158	140	148	149	132	25	26	36	140	145	127
Extra 1	38	37	41	37	41	41	4	8	12	40	51	50
Extra 2	8	4	17	9	5	19	5	0	6	10	3	19
Extra 3	0	0	2	0	1	3	0	0	2	0	0	2
Extra 4	0	0	0	0	0	0	0	0	0	0	0	0
Missing 1	0	1	0	0	1	0	9	57	8	0	1	0
Missing 2	0	0	0	1	2	0	56	55	59	0	0	0
Other	2	0	0	5	1	5	99	54	77	2	0	2
$n = 30$, model p1, WF_{ls}-test												
Correct	155	156	139	145	150	132	28	31	42	139	145	126
Extra 1	33	37	40	39	38	40	2	7	12	48	49	51
Extra 2	9	5	19	8	5	20	4	1	7	10	3	19
Extra 3	0	1	1	1	3	5	0	0	2	0	0	2
Extra 4	1	0	0	1	0	0	0	0	0	0	0	0
Missing 1	0	1	0	0	1	0	8	56	9	0	1	0
Missing 2	0	0	0	1	2	0	58	56	61	0	0	0
Other	2	0	1	5	1	3	101	49	67	3	2	2
$n = 30$, model p2, WF-test												
Correct	167	174	168	160	160	157	10	10	21	153	167	158
Extra 1	25	21	28	28	27	30	4	5	6	34	22	35
Extra 2	2	0	0	4	5	0	2	0	0	3	0	0
Missing 1	0	0	4	0	0	8	1	50	15	0	0	5
Missing 2	0	2	0	0	0	0	12	44	13	0	3	0
Missing 3	0	2	0	0	0	0	31	49	39	0	5	0
Missing 4	0	0	0	1	1	2	25	24	46	1	2	0
Other	6	1	0	7	7	3	115	18	60	9	1	2
$n = 30$, model p2, WF_{ls}-test												
Correct	167	176	167	160	169	159	8	14	22	154	168	158
Extra 1	25	20	29	27	21	31	7	4	10	32	21	36
Extra 2	3	0	0	5	0	0	1	0	0	3	0	0
Missing 1	0	2	4	0	0	6	1	50	3	0	0	3
Missing 2	0	1	0	0	3	0	10	41	12	0	3	0
Missing 3	0	0	0	0	3	0	30	45	40	0	4	0
Missing 4	0	0	0	1	3	1	28	24	52	0	2	0
Other	5	1	0	7	1	3	113	22	61	11	2	3

Table 1: Results from the 200 Monte Carlo run for robust stepwise (*Step*), forward (*For*) and backward (*Back*) selection procedure for $n = 30$.

	e1			e2			e3			e4		
	Step	*For*	*Back*	*Step*	*For*	*Back*	*Step*	*For*	*Back*	*Step*	*For*	*Back*
$n = 60$, model p1, WF-test												
Correct	164	173	160	161	168	154	32	39	38	159	164	155
Extra 1	36	27	36	34	28	36	15	12	20	37	32	38
Extra 2	0	0	4	3	2	10	4	4	8	3	3	7
Extra 3	0	0	0	0	0	0	0	0	4	0	0	0
Extra 4	0	0	0	0	0	0	0	0	0	0	0	0
Missing 1	0	0	0	0	1	0	7	47	6	0	1	0
Missing 2	0	0	0	0	1	0	62	57	59	0	1	0
Other	0	0	0	2	0	0	77	39	63	1	0	0
$n = 60$, model p1, WF_{ls}-test												
Correct	164	174	160	162	169	156	33	39	39	158	166	155
Extra 1	36	26	36	34	28	35	14	12	21	38	31	38
Extra 2	0	0	4	3	2	9	5	3	7	3	2	7
Extra 3	0	0	0	0	0	0	1	0	2	0	0	0
Extra 4	0	0	0	0	0	0	0	0	0	0	0	0
Missing 1	0	0	0	0	0	0	6	50	5	0	0	0
Missing 2	0	0	0	0	1	0	62	58	62	0	1	0
Other	0	0	0	1	0	0	79	38	64	1	0	0
$n = 60$, model p2, WF-test												
Correct	182	193	180	177	188	177	16	22	21	153	167	158
Extra 1	17	7	18	19	9	20	5	2	6	34	22	35
Extra 2	1	0	0	1	0	0	0	0	0	3	0	0
Missing 1	0	0	2	0	0	2	0	26	15	0	0	5
Missing 2	0	0	0	0	0	0	6	42	13	0	3	0
Missing 3	0	0	0	0	1	0	32	47	39	0	5	0
Missing 4	0	0	0	0	2	0	51	45	46	1	2	0
Other	0	0	0	3	0	1	90	14	60	9	1	2
$n = 60$, model p2, WF_{ls}-test												
Correct	182	193	180	177	189	177	17	21	22	154	168	158
Extra 1	17	7	18	20	9	20	5	2	10	32	21	36
Extra 2	1	0	0	1	0	0	0	0	0	3	0	0
Missing 1	0	0	2	0	0	2	0	24	3	0	0	3
Missing 2	0	0	0	0	0	0	6	45	12	0	3	0
Missing 3	0	0	0	0	0	0	34	49	40	0	4	0
Missing 4	0	0	0	0	2	0	48	44	52	0	2	0
Other	0	0	0	2	0	1	88	13	61	11	2	3

Table 2: Results from the 200 Monte Carlo run for robust stepwise (*Step*), forward (*For*) and backward (*Back*) selection procedure for $n = 60$.

Scienze Statistiche, Università di Padova.

Agostinelli, C., (1998b). Verosimiglianza pesata nel modello di regressione lineare, XXXIX Riunione scientifica della Società Italiana di Statistica, Sorrento.

Agostinelli, C., (1998c). Un approccio alla verifica d'ipotesi robusta basato sulla funzione di verosimiglianza pesata, submitted to *Statistica.*

Agostinelli, C., and Markatou, M., (2000). Test of hypotheses based on the weighted likelihood methodology, *Statistica Sinica*, in press.

Agostinelli, C., (1999). Robust model selection in regression via weighted likelihood methodology, submitted to *Statistics & Probability Letters.*

Agostinelli, C., (2000). Robust model selection by Cross-Validation via weighted likelihood methodology, submitted to *Statistica Sinica.*

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

Beale, E.M.L., Kendall, M.G., Mann, D.W., (1967), The discarding of variables in multivariate analysis, *Biometrika*, 54, 357-366.

Efroymson, (1960). Multiple regression analysis, in *Mathematical Methods for Digital Computers*, eds. A. Ralston and H.S. Wilf, 191-203, Wiley, New York.

Garside, M.J., (1965). The best sub-set in multiple regression analysis, *Applied Statistics*, 14, 196-200.

Goldberger, A.S, and Jochems, D.B., (1961). Note on stepwise least squares, *Journal of the American Statistical Association*, **56**, 105-110.

Goldberger, A.S., (1961). Stepwise least squares: Residual analysis and specification error, *Journal of the American Statistical Association*, **56**, 998-1000.

Hadi, A.S., (1990). A stepwise procedure for identifying multiple outliers in linear regression, in *Proc. Statist. Comput. Sect.*, 137-142, Amer. Statist. Assoc. (Alexandria, VA).

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

Ihaka, R., Gentleman, R., (1996). R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics, 5, 299-314.

Lindsay, B.G., (1994). Efficiency versus robustness: The case for minimum Hellinger distance and related methods, *Annals of Statistics,* **22**, 1018-1114.

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

Markatou, M., Basu, A. and Lindsay, B.G., (1995). Weighted likelihood estimating equations: The continuous case, *Technical Report*, Department of Statistics, Columbia University, New York.

Markatou, M., Basu, A. and Lindsay, B.G., (1998). Weighted likelihood estimating equations with a bootstrap root search, *Journal of the American Statistical Association*, **93**, 740-750.

Miller, A.J., (1984). Selection of subsets of regression variables (with discussion), *Journal of the Royal Statistical Society*, Ser. A., **147**, 389-425.

Miller, A.J., (1990). *Subset selection in regression*, Chapman-Hall, New York.

Ronchetti, E., (1997). Robustness aspects of model choice, *Statistica Sinica*,

7, 327-338.

SAS Institute, (1990). *SAS/Stat User's Guide, volume 2, version 6, fourth edition*, SAS Institute Inc., Cary, NC.

Shao, J., (1993). Linear model selection by cross-validation, *Journal of the American Statistical Association*, **88**, 486-494.

Stone, M., (1974). Cross-validation choice and assessment of statistical predictors, *Journal of the Royal Statistical Society*, Ser. B., **36**, 111-147.