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 outlying 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 $C_p$, (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-models to be considered. Often, the evaluation of all possible sub-models 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 outlying 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 $C_p$, $AIC$ 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
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 $\theta$ 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 $\hat{F}_n(t) = \sum_{i=1}^n 1_{x_i < t}/n$; then estimators for the parameter vector $\theta$ are obtained as solutions to the set of estimating equations:

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

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 $\delta(x; \theta, \hat{F}_n) = \frac{f^*(x; \theta)}{m^*(x; \theta)} - 1$, 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 $\psi$-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 equa-
tions.

To calculate the Pearson residuals we need to select the smoothing pa-
rameter $h$. 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 $\{y_1, \ldots, y_n\}$ a sample of dimension $n$ from an unknown dis-
btribution and $\{x_1, \ldots, x_n\}$ a sample of vector from $p$ explanatory variables
which could included the intercept term. Considering the regression model
\[ y = x \beta + \epsilon \] and assuming a parametric family \( M = \{ m(\epsilon; \sigma); \sigma \in \Sigma \} \) we let \( z(\beta) = y - x \beta \) the residuals for a specific value of the parameter vector \( \beta \) and \( \hat{F}_n(t; \beta) = \sum_{i=1}^{n} 1_{z_i(\beta) < t}/n \) 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 \( \beta \) is a solution of the estimating equation:

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

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

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

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

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

\[
\begin{align*}
\sum_{i=1}^{n} w(z_i(\beta); \sigma, \hat{F}_n(\beta)) \left( y_i - x_i \beta \right) x_i &= 0 \\
\sum_{i=1}^{n} w(z_i(\beta); \sigma, \hat{F}_n(\beta)) \left( (y_i - x_i \beta)^2 - \sigma^2 \right) &= 0
\end{align*}
\] (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:

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

$$y = x_A \beta_A + \varepsilon$$

where $A$ is a subset of $d_A$ distinct positive integers that are less or equal to $p$ and $\beta_A$ (or $x_A$) is the $d_A$ vector containing the components of $\beta$ (or $x$) that are indexed by the integers in $A$.

Further, let $\mathcal{M}_I$ the set of models $A$ such that at least one nonzero component of $\beta$ is not in $\beta_A$ and $\mathcal{M}_{II}$ the set of models $A$ such that $\beta_A$ contains all nonzero components of $\beta$. Hence $\mathcal{M}_I$ is the set of all models that are a proper subset of the “true model”, while $\mathcal{M}_{II}$ 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 $\hat{\beta}_A$.

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:

$$\hat{\sigma}_A^2 = \frac{1}{\sum_{i=1}^{n} \hat{w}_i} \sum_{i=1}^{n} \hat{w}_i z_i \left( \hat{\beta}_A \right)^2$$  \hspace{1cm} (3)

Let now consider the weighted likelihood ratio test $L_{RT_w}$:

$$L_{RT_w} = \frac{\Pi_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\hat{\sigma}_A^2}} \exp \left[ -\frac{1}{2} \frac{z_i^2}{\hat{\sigma}_A^2} \right] \right\} \hat{w}_i}{\Pi_{i=1}^{n} \left\{ \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp \left[ -\frac{1}{2} \frac{z_i^2}{\hat{\sigma}^2} \right] \right\} \hat{w}_i}$$  \hspace{1cm} (4)
\[
\begin{align*}
&= \left( \frac{\sigma^2}{\hat{\sigma}^2} \right) - \sum_{i=1}^{n} \frac{\psi_i}{\hat{\sigma}^2} \exp \left\{ \frac{1}{2} \frac{z_i (\hat{\beta}_A)^2}{\hat{\sigma}^2_A} - \frac{1}{2} \frac{z_i}{\hat{\sigma}^2} \right\} \\
&= \left( \frac{\sigma^2}{\hat{\sigma}^2} \right) - \sum_{i=1}^{n} \frac{\psi_i}{\hat{\sigma}^2} \exp \left\{ -\frac{1}{2} \frac{\sum_{i=1}^{n} \hat{\psi}_i z_i (\hat{\beta}_A)^2}{\hat{\sigma}^2_A} + \frac{1}{2} \frac{\sum_{i=1}^{n} \hat{\psi}_i z_i (\hat{\beta})^2}{\hat{\sigma}^2} \right\} \\
&= \left( \frac{\sigma^2}{\hat{\sigma}^2} \right) - \sum_{i=1}^{n} \frac{\psi_i}{\hat{\sigma}^2}
\end{align*}
\]

Further let consider the follows transformation:

\[
LRT^*_{w} = LRT_{w} - \frac{\sum_{i=1}^{n} \psi_i}{\hat{\sigma}^2} - 1
= \frac{\sigma^2 - \hat{\sigma}^2}{\hat{\sigma}^2}
\]

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. \(A(0) = 0, A'(0) = 1\) and \(A''(\delta)\) is a bounded continuous function of \(\delta\)

A3. the observations \(z_i(\hat{\beta}_0)\) are from the model \(M(\cdot; \sigma_0)\) with density \(m(\cdot; \sigma_0)\) and \(\sigma_0\) belong to the parametric space \(\Sigma\)

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. \(\sup_z \left| \tilde{F}_n(z; \hat{\beta}) - M(z; \hat{\sigma}) \right| \overset{P}{\to} 0\)
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 $A$.

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

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

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

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

$$+ \sup_{i} |\hat{w}_i - 1| \frac{1}{n} \sum_{i=1}^{n} z_i(\hat{\beta}_L)^2$$

$$\xrightarrow{p} 0$$

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

$$|\hat{\sigma}^2 - \hat{\sigma}^2_L| = \left| \frac{1}{\sum_{i=1}^{n} \hat{w}_i} \sum_{i=1}^{n} \hat{w}_i z_i(\hat{\beta})^2 - \frac{1}{n} \sum_{i=1}^{n} z_i(\hat{\beta}_L)^2 \right| \xrightarrow{p} 0$$
Similarly we have $|\hat{\sigma}_A^2 - \hat{\sigma}_A^2| \xrightarrow{p} 0$ for all $A \in \mathcal{M}_{II}$ and finally $|L_{RT}^* - L_{RT}^*| \xrightarrow{p} 0$.

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

On the other hand when the data is contaminated it is preferable to compare the $L_{RT}^* (\sum_{i=1}^{n} \hat{w}_i - p)/q$ with an $F_{q;\tilde{n}-p}$ where $\tilde{n}$ 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 $\lambda_w$ defined in Agostinelli (1998c) and Agostinelli and Markatou (2000) since, with the normal model we have

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

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 $\hat{\beta}_A$ with a weighted least square estimator $\tilde{\beta}_A$ 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 $W_{i_{LF}}$-test the test based on the weighted least square, it can be shown that the $WF$-test and the $W_{i_{LF}}$-test are asymptotic equivalent for all models in $\mathcal{M}_{II}$.
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_τF-test. We now introduce the Forward selection algorithm. First of all we estimate the weights \( \hat{\omega}_i \) from the full model. The residual sum of squares of a particular set \( \mathcal{A} \) of variables is then

\[
RSS_A = \sum_{i=1}^{n} \hat{\omega}_i \, z_i (\hat{\beta}_A)^2.
\]

The same definitions and results hold when we define \( \tilde{RSS}_A \) based on weighted least square estimator as

\[
\tilde{RSS}_A = \sum_{i=1}^{n} \tilde{\omega}_i \, z_i (\hat{\beta}_A)^2
\]

where \( \hat{\beta}_A \) is the solution of the weighted least square estimator based on \( \hat{\omega}_i \).

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 \( RSS_{A+1} \). The ratio

\[
R_e = \frac{RSS_A - RSS_{A+1}}{RSS_{A+1}/(\sum_{i=1}^{n} \hat{\omega}_i - d_A - 1)}
\]

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 \( RSS_{A-1} \) be the smallest \( RSS \) which can be obtained after deleting any variable from the previously selected variables. The ratio

\[
R_d = \frac{RSS_{A-1} - RSS_A}{RSS_A/(\sum_{i=1}^{n} \hat{\omega}_i - d_A)}
\]
is calculated and compared with a threshold 'F-to-delete' value, say $F_d$. If $R_d < F_d$ the variable is deleted from $A$.

The Stepwise selection algorithm is a variation on the Forward selection method. After each variable is added to $A$ 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 of 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

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

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

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

Hence when an addition is followed by a deletion, the new $RSS$, say $RSS^*_A$, is such that

$$RSS^*_A \leq RSS_A \frac{1 + F_d/\left(\sum_{i=1}^{n} \hat{w}_i - d_A\right)}{1 + F_e/\left(\sum_{i=1}^{n} \hat{w}_i - d_A - 1\right)} \quad (7)$$

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

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

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

11
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 run 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
following two hypothesis sets:

\[
\begin{align*}
H_{0a} : & \quad \beta_1 = 0, \beta_2 = 0, \beta_3 \neq 0, \beta_4 \neq 0 \\
H_{1a} : & \quad \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0
\end{align*}
\]

\[
\begin{align*}
H_{0b} : & \quad \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 = 0, \beta_4 = 0 \\
H_{1b} : & \quad \beta_1 \neq 0, \beta_2 \neq 0, \beta_3 \neq 0, \beta_4 \neq 0
\end{align*}
\]

In figure 1 we report the results for the $H_{0a}, H_{1a}$ set, in its left part the value of the $WF_{ib}$-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 $WF_{ib}$-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 $WF_{t_a}$-test and F-test, and the weights associated to $WF_{t_a}$-test c.

The $WF_{t_a}$-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 $WF_{t_b}$-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 $WF_{t_a}$-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 WFₜₐ-test and F-test and their p-value for H₀₀ₐ.

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 (x₁ = time to run 1.5 miles (minutes), x₂ = age (year), x₃ = weight (kg), x₄ = heart rate while running (same time oxygen rate measured), x₅ = maximum heart rate recorded while running and x₆ = heart rate while resting). Using stepwise procedure with Fₑ = 4 and Fₜ = 2 (or forward with Fₑ = 4, or backward, Fₜ = 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 respect 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

<table>
<thead>
<tr>
<th></th>
<th>e1 Step</th>
<th>e1 For</th>
<th>e1 Back</th>
<th>e2 Step</th>
<th>e2 For</th>
<th>e2 Back</th>
<th>e3 Step</th>
<th>e3 For</th>
<th>e3 Back</th>
<th>e4 Step</th>
<th>e4 For</th>
<th>e4 Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>152</td>
<td>158</td>
<td>140</td>
<td>148</td>
<td>149</td>
<td>132</td>
<td>25</td>
<td>26</td>
<td>36</td>
<td>140</td>
<td>145</td>
<td>127</td>
</tr>
<tr>
<td>Extra 1</td>
<td>38</td>
<td>37</td>
<td>41</td>
<td>37</td>
<td>41</td>
<td>41</td>
<td>4</td>
<td>8</td>
<td>12</td>
<td>40</td>
<td>51</td>
<td>50</td>
</tr>
<tr>
<td>Extra 2</td>
<td>8</td>
<td>4</td>
<td>17</td>
<td>9</td>
<td>5</td>
<td>19</td>
<td>5</td>
<td>0</td>
<td>6</td>
<td>10</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Extra 3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Extra 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>9</td>
<td>57</td>
<td>8</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>56</td>
<td>55</td>
<td>59</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>5</td>
<td>99</td>
<td>54</td>
<td>77</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

\( n = 30 \), model p1, WF-test

<table>
<thead>
<tr>
<th></th>
<th>e1 Step</th>
<th>e1 For</th>
<th>e1 Back</th>
<th>e2 Step</th>
<th>e2 For</th>
<th>e2 Back</th>
<th>e3 Step</th>
<th>e3 For</th>
<th>e3 Back</th>
<th>e4 Step</th>
<th>e4 For</th>
<th>e4 Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>155</td>
<td>156</td>
<td>139</td>
<td>145</td>
<td>150</td>
<td>132</td>
<td>28</td>
<td>31</td>
<td>42</td>
<td>139</td>
<td>145</td>
<td>126</td>
</tr>
<tr>
<td>Extra 1</td>
<td>33</td>
<td>37</td>
<td>40</td>
<td>39</td>
<td>38</td>
<td>40</td>
<td>2</td>
<td>7</td>
<td>12</td>
<td>48</td>
<td>49</td>
<td>51</td>
</tr>
<tr>
<td>Extra 2</td>
<td>9</td>
<td>5</td>
<td>19</td>
<td>8</td>
<td>5</td>
<td>20</td>
<td>4</td>
<td>1</td>
<td>7</td>
<td>10</td>
<td>3</td>
<td>19</td>
</tr>
<tr>
<td>Extra 3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Extra 4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>8</td>
<td>56</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>58</td>
<td>56</td>
<td>61</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>3</td>
<td>101</td>
<td>49</td>
<td>67</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>

\( n = 30 \), model p1, WF-test

<table>
<thead>
<tr>
<th></th>
<th>e1 Step</th>
<th>e1 For</th>
<th>e1 Back</th>
<th>e2 Step</th>
<th>e2 For</th>
<th>e2 Back</th>
<th>e3 Step</th>
<th>e3 For</th>
<th>e3 Back</th>
<th>e4 Step</th>
<th>e4 For</th>
<th>e4 Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>167</td>
<td>174</td>
<td>168</td>
<td>160</td>
<td>160</td>
<td>157</td>
<td>10</td>
<td>10</td>
<td>21</td>
<td>153</td>
<td>167</td>
<td>158</td>
</tr>
<tr>
<td>Extra 1</td>
<td>25</td>
<td>21</td>
<td>28</td>
<td>28</td>
<td>27</td>
<td>30</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>34</td>
<td>22</td>
<td>35</td>
</tr>
<tr>
<td>Extra 2</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>8</td>
<td>1</td>
<td>50</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>44</td>
<td>13</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>31</td>
<td>49</td>
<td>39</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>25</td>
<td>24</td>
<td>46</td>
<td>1</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>7</td>
<td>3</td>
<td>115</td>
<td>18</td>
<td>60</td>
<td>9</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

\( n = 30 \), model p2, WF-test

<table>
<thead>
<tr>
<th></th>
<th>e1 Step</th>
<th>e1 For</th>
<th>e1 Back</th>
<th>e2 Step</th>
<th>e2 For</th>
<th>e2 Back</th>
<th>e3 Step</th>
<th>e3 For</th>
<th>e3 Back</th>
<th>e4 Step</th>
<th>e4 For</th>
<th>e4 Back</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>167</td>
<td>176</td>
<td>167</td>
<td>160</td>
<td>169</td>
<td>159</td>
<td>8</td>
<td>14</td>
<td>22</td>
<td>154</td>
<td>168</td>
<td>158</td>
</tr>
<tr>
<td>Extra 1</td>
<td>25</td>
<td>20</td>
<td>29</td>
<td>27</td>
<td>21</td>
<td>31</td>
<td>7</td>
<td>4</td>
<td>10</td>
<td>32</td>
<td>21</td>
<td>36</td>
</tr>
<tr>
<td>Extra 2</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>50</td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>10</td>
<td>41</td>
<td>12</td>
<td>0</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>30</td>
<td>45</td>
<td>40</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Missing 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>28</td>
<td>24</td>
<td>52</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Other</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>7</td>
<td>1</td>
<td>3</td>
<td>113</td>
<td>22</td>
<td>61</td>
<td>11</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 1: Results from the 200 Monte Carlo run for robust stepwise (Step), forward (For) and backward (Back) selection procedure for \( n = 30 \).
<table>
<thead>
<tr>
<th></th>
<th>e1</th>
<th></th>
<th>e2</th>
<th></th>
<th>e3</th>
<th></th>
<th>e4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
</tr>
<tr>
<td>Correct</td>
<td>164</td>
<td>173</td>
<td>160</td>
<td>161</td>
<td>168</td>
<td>154</td>
<td>32</td>
</tr>
<tr>
<td>Extra 1</td>
<td>36</td>
<td>27</td>
<td>36</td>
<td>34</td>
<td>28</td>
<td>36</td>
<td>15</td>
</tr>
<tr>
<td>Extra 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Extra 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Extra 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>7</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>77</td>
</tr>
</tbody>
</table>

n = 60, model p1, WF-test

<table>
<thead>
<tr>
<th></th>
<th>e1</th>
<th></th>
<th>e2</th>
<th></th>
<th>e3</th>
<th></th>
<th>e4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
</tr>
<tr>
<td>Correct</td>
<td>164</td>
<td>174</td>
<td>160</td>
<td>162</td>
<td>169</td>
<td>156</td>
<td>33</td>
</tr>
<tr>
<td>Extra 1</td>
<td>36</td>
<td>26</td>
<td>36</td>
<td>34</td>
<td>28</td>
<td>35</td>
<td>14</td>
</tr>
<tr>
<td>Extra 2</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td>Extra 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Extra 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>62</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>79</td>
</tr>
</tbody>
</table>

n = 60, model p2, WF-test

<table>
<thead>
<tr>
<th></th>
<th>e1</th>
<th></th>
<th>e2</th>
<th></th>
<th>e3</th>
<th></th>
<th>e4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
</tr>
<tr>
<td>Correct</td>
<td>182</td>
<td>193</td>
<td>180</td>
<td>177</td>
<td>188</td>
<td>177</td>
<td>16</td>
</tr>
<tr>
<td>Extra 1</td>
<td>17</td>
<td>7</td>
<td>18</td>
<td>19</td>
<td>9</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Extra 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>32</td>
</tr>
<tr>
<td>Missing 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>51</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>90</td>
</tr>
</tbody>
</table>

n = 60, model p2, WF_{it}-test

<table>
<thead>
<tr>
<th></th>
<th>e1</th>
<th></th>
<th>e2</th>
<th></th>
<th>e3</th>
<th></th>
<th>e4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
<td>For</td>
<td>Back</td>
<td>Step</td>
</tr>
<tr>
<td>Correct</td>
<td>182</td>
<td>193</td>
<td>180</td>
<td>177</td>
<td>189</td>
<td>177</td>
<td>17</td>
</tr>
<tr>
<td>Extra 1</td>
<td>17</td>
<td>7</td>
<td>18</td>
<td>20</td>
<td>9</td>
<td>20</td>
<td>5</td>
</tr>
<tr>
<td>Extra 2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Missing 1</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Missing 2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>6</td>
</tr>
<tr>
<td>Missing 3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>34</td>
</tr>
<tr>
<td>Missing 4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>Other</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>88</td>
</tr>
</tbody>
</table>

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.


