Choosing between two parametric models robustly

VICTORIA-FESER, Maria-Pia

Abstract

In this paper we propose a robust version of Cox-type test statistics for the choice between two non-nested hypotheses. We first show that the influence of small amounts of contamination in the data on the test decision can be very large. Secondly we build a robust test statistic by using the results on robust parametric tests available in the literature and show that the level of the robust test is stable. Finally, we show numerically not only the good property of robustness of this new test statistic, but also that its asymptotic distribution is a good approximation of its sample distribution, unlike for the classical test statistic.

Reference

VICTORIA-FESER, Maria-Pia. Choosing between two parametric models robustly. London School of Economics, 1995
Choosing between two parametric models robustly

by Maria-Pia Victoria-Feser

London School of Economics
Houghton Street
London WC2A 2AE, UK

Statistics Research Report LSERR21, LSE

June 1995

Abstract

In this paper we propose a robust version of Cox-type test statistics for the choice between two non-nested hypotheses. We first show that the influence of small amounts of contamination in the data on the test decision can be very large. Secondly we build a robust test statistic by using the results on robust parametric tests available in the literature and show that the level of the robust test is stable. Finally, we show numerically not only the good property of robustness of this new test statistic, but also that its asymptotic distribution is a good approximation of its sample distribution, unlike for the classical test statistic.

Key words: M-estimators, model choice, robust tests.
1 Introduction

In this paper we study the problem of choosing a model which according to certain criteria, best represents the data. We focus on tests between two non-nested hypotheses as defined by Cox (1961) and Cox (1962).

The areas of application are wide (see Loh 1985). In the last decade several Cox type statistics have been developed mainly in order to simplify the procedure when dealing with particular models (see among others Davidson and MacKinnon 1981, Fisher and McAleer 1981, MacKinnon, White, and Davidson 1983, Gourieroux, Monfort, and Trognon 1983). These ‘Cox-type’ statistics are actually parametric tests based on an artificial compound model in which the models under the null hypothesis and under the alternative hypothesis are represented (as a special case for the latter). The Cox statistic can as well be seen as a Lagrange multiplier test based on a compound model (see Atkinson 1970, Breusch and Pagan 1980 and Dastoor 1985).

Although it is widely accepted that these statistics are very useful in a statistical analysis, they have been often criticised for several reasons. The most studied one is the lack of accuracy of the approximation of the sample distribution of the statistic by its asymptotic distribution (see e.g. Atkinson 1970, Williams 1970, Godfrey and Pesaran 1983 and Loh 1985). Another (less studied) reason but at least as important is the lack of robustness of
Cox type statistics. Hall (1985), in the context of regression models, noting that the full information maximum likelihood estimator under normality is extremely sensitive to misspecifications of the error distribution, stated that ‘... the development of non-nested selection techniques based on more robust estimators would appear extremely desirable...’. Aguirre-Torres and Gallant (1983) propose a generalisation of the Cox statistic based on M-estimators for the parameters. The same idea can be found in Hampel, Ronchetti, Rousseeuw, and Stahel (1986), chap 7. However, they leave open the question of the choice of the $\rho$-function defining the M-estimators.

The aim of this paper is to propose a robust procedure based on optimal bounded influence parametric tests developed recently by Heritier and Ronchetti (1994). We begin by showing numerically how much a Cox-type test can be badly influenced by a very small amount of contamination in the data. We then show theoretically this effect by computing the level influence function (LIF) (Rousseeuw and Ronchetti 1979, Ronchetti 1982) of the Cox test. Finally, we show that we can bound this LIF with a new test obtained by considering the Cox test as a Lagrange multiplier test and applying the results on robust parametric tests. We end with the application of this new procedure to well known models and compare its power with the classical Cox statistic, by means of simulations. In particular, we will see that the new procedure is not only robust to small model deviations or contamina-
tions but also that for at least the chosen particular cases, the asymptotic
distribution of the robust test statistic is a better approximation of its sample
distribution than in the classical case.

2 Robustness properties of Cox-type statistics

In this section, we first briefly present the Cox and Cox-type test statistics.
Second, we explain the robustness concepts in the case of tests. To illustrate
the problem of non robustness of tests, we then do a simulation study which
shows that for one example (the quantal response problem) only a few outliers
can lead the test to a false decision. In a fourth subsection, we compute the
\( LIF \) of the Cox test and show that it is unbounded, meaning that for any
model under the hypothesis, only one outlier can lead the test to a false
decision.

2.1 The test statistics

In general, it is assumed under \( H_0 \) (the hypothesis under test) that the model
is \( F_\alpha^0 \) (with density \( f^0(\cdot; \alpha) \)) and that under \( H_1 \) (the alternative hypothesis)
the model is \( F_\beta^1 \) (with density \( f^1(\cdot; \beta) \)), where \( \alpha \) and \( \beta \) are parameter vectors.
For simplicity and without loss of generality, we suppose that the observations are univariate, i.e. we observe $x_1, \ldots, x_n, x_i \in \mathcal{R}$. Let $L_0(x; \hat{\alpha}) = \log f^0(x; \hat{\alpha})$ and $L_1(x; \hat{\beta}) = \log f^1(x; \hat{\beta})$ be the log-likelihood functions, where $\hat{\alpha}$ and $\hat{\beta}$ are the corresponding maximum likelihood estimators and define $L(x; \hat{\alpha}, \hat{\beta}) = L_0(x; \hat{\alpha}) - L_1(x; \hat{\beta})$. Cox (1961) and Cox (1962) proposed the following test statistic

$$U_{\text{Cox}} = n^{-1} \sum_{i=1}^n L(x_i; \hat{\alpha}, \hat{\beta}) - \int L(x; \hat{\alpha}, \hat{\beta}) f^0(x; \hat{\alpha}) dx$$

where $\sum$ stands for $\sum_{i=1}^n$ and $\hat{\beta}_\alpha$ is the pseudo maximum likelihood estimator defined as the solution in $\beta$ of $\int \partial/\partial \beta \log f^1(x; \beta) f^0(x; \hat{\alpha}) dx = 0$. Two straightforward modifications of $U_{\text{Cox}}$ 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, the asymptotic null distribution of $n^{1/2}U_{\text{Cox}}$ is the normal distribution with mean 0 and variance $\text{var}(F^0_\alpha) = E[L^2] - E[L]^2 - E[(s^0)^T L E[(s^0)^T s^0]^{-1} E[s^0 L]]$, where $L = L(x; \alpha, \beta_\alpha)$, $s^0 = s^0(x; \alpha) = \partial/\partial \alpha \log f^0(x; \alpha)$ and $E[\cdot]$ is the expectation with respect to $F^0_\alpha$, the argument of $\text{var}$. In practice one needs a consistent estimator of $\text{var}(F^0_\alpha)$, e.g. when $\alpha$ is replaced by $\hat{\alpha}$.
2.2 Robustness and tests

Developments in the field of robust statistics have mainly concerned parameter estimation but much less has been done for testing procedures. However, robust tests are important and the reason is quite obvious: if we estimate the parameters of a model robustly and then use a classical procedure to test hypotheses about the parameters or about the estimated model, the test is not necessarily robust. There are two fundamental goals in robust testing. The level of a test should be stable under small, arbitrary departures from the null hypothesis (robustness of validity), and the test should retain a good power under small, arbitrary departures from specified alternatives (robustness of efficiency).

In this paper, we follow Hampel (1968) infinitesimal approach based on M-estimators. Ronchetti (1979), Ronchetti (1982), Rousseeuw and Ronchetti (1979) and Rousseeuw and Ronchetti (1981) were the first to adapt Hampel’s optimality problem for estimators to testing procedures, in the case of testing a null hypothesis about a one-dimensional parameter. Hampel’s optimality problem for testing procedures can be stated as: Under a bound on the influence of small contamination on the test’s level and power (robustness requirement), the power of the test at the ideal model is maximised (efficiency requirement). More recently, Heritier and Ronchetti (1994) have extended
the existing theory on robust parametric tests to general parametric models.

However, to our knowledge nothing has been done for model choice tests.

Our first step is to study the robustness properties of Cox type statistics. After a numerical simulation, we show analytically that the Cox statistic is not robust. By means of the $LIF$, we find the influence of an infinitesimal amount of contamination in the data on the true level of the test, i.e. the level of the test at the model. Actually, this level is often approximated by the asymptotic level. The general technique consists of assuming a distribution in the neighbourhood of the model under the null hypothesis and then studying the effect on the asymptotic level of the test. To compute the $LIF$, the neighbourhood of the model at $H_0$ is given by

$$F_{\varepsilon,n}^0 = (1 - \varepsilon_n)F_0^0 + \varepsilon_n \Delta_z$$

where $\Delta_z$ is the distribution which gives a probability of 1 to an arbitrary point $z$. Note that a more general neighbourhood is defined when instead of $\Delta_z$ we choose any distribution. However, to determine the bias on the level of the test, by choosing 2, one actually has the worst situation, i.e. the $LIF$ determines the worst bias on the level (see Hampel, Ronchetti, Rousseeuw, and Stahel 1986, p. 175). To study the effect of this model deviation, it is necessary to choose a contamination $\varepsilon_n$ that tends to zero at the rate $n^{-\frac{1}{2}}$, 

7
or its effects will soon dominate everything and give divergence (see Hampel, Ronchetti, Rousseuw, and Stahel 1986, chap 3). Therefore, \( \varepsilon_n = n^{-1/2} \varepsilon \).

With parametric tests, one has to choose the rate of convergence of the test statistic to avoid overlapping neighbourhoods between the hypotheses under \( H_0 \) and \( H_1 \) (see Heritier and Ronchetti 1994).

In subsection 2.3, we compute the asymptotic level of the Cox test under \( F_{\varepsilon,n}^0 \) and then derive the \( LIF \) by taking the derivative of this level with respect to \( \varepsilon \) at \( \varepsilon = 0 \) when \( n \to \infty \). But first we present a simulation study.

### 2.3 Simulation study

In this subsection we illustrate numerically, through one example, the non-robustness properties of Cox type statistics. We consider here the quantal responses problem treated among others by Cox (1962), Atkinson (1970) and Loh (1985). At \( k \) levels of a variable \( x_i \), called the dose level, \( n_i \) experiments are performed. The number \( z_i \) of successes is distributed binomially with index \( n_i \) and probability \( \pi_0^i \) under \( H_0 \) and \( \pi_1^i \) under \( H_1 \). The purpose of the experiment is to determine the relationship between the dose level and the parameter of the binomial distribution. The two common models are the one- and two-hit models, defined respectively by \( \pi_0^i = 1 - e^{-\alpha x_i} \) and \( \pi_1^i = 1 - e^{-\beta x_i} - \beta x_i e^{-\beta x_i} \). We chose five dose levels \( x_1 = 0.5, x_2 = 1, x_3 = 2, x_4 = 4 \)
and $x_5 = 8$ (see Cox 1962), and for each of them we simulated 30 binary data with probability of success $\pi_i^0$, with parameter $\alpha = 2$. We computed the maximum likelihood estimator by means of a Newton-Raphson iteration (see Thomas 1972). We computed the (standardised) White statistic and found it was equal to 0.8677 with corresponding $p$-value of 19.3% (with $\hat{\alpha} = 2.18$ and $\hat{\beta} = 3.99$) leading then to the acceptance of $H_0$.

But what happens if some data are changed (the Bernouilli trial is changed to the value of 0 when it is equal to 1, or to 1 when it is equal to 0)? Intuitively, by looking at (1) one can see that unfortunately the classical tests are not robust, because the maximum likelihood estimator of $\alpha$ is not robust and $\sum L(x_i; \alpha, \beta)$ can be determined by only one extreme observation. To show that, we changed the value of two binary data (the first two corresponding to the level $x_4 = 4$) and again computed the (standardised) White statistic. As the amount of contamination is very small, one would expect the decision not to be influenced by it. However, this time we found a value of 4.6311 with corresponding $p$-value of less than $5 \cdot 10^{-4}$% (with $\hat{\alpha} = 1.63$ and $\hat{\beta} = 2.98$) leading to the rejection of $H_0$.

We will see later that it is possible to build a robust version of Cox-type statistics that is not badly influenced by misleading observations and moreover can automatically detect them.
2.4 Level influence function

In this subsection, we compute the LIF of the Cox test. A similar result can be obtained for the Atkinson and White tests.

Under $H_0$ we have seen that

$$n^{1/2}U(F^{(n)}) = n^{1/2} \left( \int L(x; \hat{\alpha}, \hat{\beta})dF^{(n)}(x) - \int L(x; \hat{\alpha}, \beta_0) f(x; \hat{\alpha})dx \right)$$

has an asymptotic normal distribution with zero mean and variance $\text{var}(F_0^\alpha)$, where $F^{(n)}$ denotes the empirical distribution. Note that $\hat{\alpha}$ and $\hat{\beta}$ can be written as functionals of the empirical distribution, i.e. $\hat{\alpha}(F^{(n)})$ and $\hat{\beta}(F^{(n)})$.

The test decision will be to reject $H_0$ if $|n^{1/2}U(F^{(n)})| > \kappa_\omega^*$, $(\kappa_\omega^* = \text{var}(F_0^\alpha)^{1/2} \Phi^{-1}(1 - \omega/2)$, $\omega$ is the nominal level and $\Phi$ is the cumulative standard normal distribution). Since the sample distribution of $n^{1/2}U(F^{(n)})$ is not known, the decision is taken by using its asymptotic distribution.

Therefore, we define the true asymptotic level $\omega(0)$ (the argument 0 stands for no model contamination) at the model $F_0^\alpha$ of the test as the probability that $|n^{1/2}U(F^{(n)})|$ exceeds the critical value $\kappa_\omega^*$. Let $\kappa_\omega = \kappa_\omega^*/\text{var}(F_0^\alpha)^{1/2}$, then we have

$$2^{-1}\omega(0) = 1 - \Phi(\kappa_\omega)$$

Since the model is not always exactly true, the actual asymptotic level will
be in general biased. This can lead the test to a rejection of \( H_0 \) only because of an infinitesimal proportion of outliers. As a neighbourhood distribution, we consider the \( \varepsilon \)-contamination distribution given by (2), with \( \varepsilon_n = \varepsilon n^{-1/2} \).

Assuming \( F^0_{\varepsilon,n} \) under \( H_0 \) we then have that

\[
n^{1/2}U(F^0_{\varepsilon,n}) = n^{1/2} \left( \int L(x; \hat{\alpha}(F^0_{\varepsilon,n}), \hat{\beta}(F^0_{\varepsilon,n}))dF^0_{\varepsilon,n}(x) - \int L(x; \alpha, \beta_\alpha)f(x; \alpha)dx \bigg|_{\alpha=\hat{\alpha}(F^0_{\varepsilon,n})} \right)
\]

If we assume Fréchet differentiability (see Heritier and Ronchetti 1994, appendix 2, Clarke 1983 and Clarke 1986), we have that

\[
n^{1/2} \left( U(F^{(n)}) - U(F^0_{\varepsilon,n}) \right) \xrightarrow{\text{d}} N \left[ 0, var(F^0_\alpha) \right]
\]

where \( \xrightarrow{\text{d}} \) stands for convergence in distribution as \( n \to \infty \). This result enables us to derive the LIF of the test.

Let \( \mu(\varepsilon) = \lim_{n \to \infty} n^{1/2}U(F^0_{\varepsilon,n}) \). The actual asymptotic level under the hypothetical model \( F^0_{\varepsilon,n} \) is then given by

\[
2^{-1}\omega(\varepsilon) = 1 - \Phi \left( \frac{\kappa^*_\omega - \mu(\varepsilon)}{var(F^0_\alpha)} \right) = 1 - \Phi(k(\varepsilon))
\]

The bias on the asymptotic level can be approximated by a first order Taylor expansion of \( \omega(\varepsilon) \) around \( \omega(0) \), i.e.

\[
\omega(\varepsilon) - \omega(0) = \varepsilon \cdot \frac{\partial \omega(\varepsilon)}{\partial \varepsilon} \bigg|_{\varepsilon=0} + O(\varepsilon^2).
\]
where $\frac{\partial \omega(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0} = LIF(z; \omega, F_{\alpha}^0)$. Therefore, the $LIF$ carries all the information about the first order approximation of the bias on the asymptotic level of the test. If the $LIF$ can take arbitrarily large values, it means the test is not robust, or in other words the decision can be determined by only a few observations. We have that

$$LIF(z; \omega, F_{\alpha}^0) = 2 \cdot \text{var}(F_{\alpha}^0)^{-1/2} \frac{\partial}{\partial k} \Phi(k(\varepsilon)) \frac{\partial \mu(\varepsilon)}{\partial \varepsilon}\bigg|_{\varepsilon=0}$$

If the asymptotic variance $\text{var}(F_{\alpha}^0)$ is estimated consistently (by $\text{var}(\hat{F}_\alpha)$) then it can be shown that the effect of the contamination on the bias is (asymptotically) null. This is because $\varepsilon n^{-1/2}$ converges more rapidly to 0 than the empirical variance to the asymptotic variance.

For consistent estimators of $\beta_\alpha$, i.e. $\hat{\beta}(F_{\alpha}^0) = \beta_\alpha$, we have that

$$\frac{\partial}{\partial \varepsilon} \mu(\varepsilon) \bigg|_{\varepsilon=0} = - \left\{ \int L(x; \alpha, \beta_\alpha) dF_{\alpha}^0(x) - L(z; \alpha, \beta_\alpha) + \int L(x; \alpha, \beta_\alpha) s_0(x; \alpha) dF_{\alpha}^0(x) \cdot IF^*(z, \hat{\alpha}, F_{\alpha}^0) \right\}$$

where $IF^*(z, \hat{\alpha}, F_{\alpha}^0) = \lim_{n \to \infty} \left[ \{\partial/\partial \varepsilon_n\} \hat{\alpha}(F_{\varepsilon,n}^0)\bigg|_{\varepsilon=0} \right] = - \left[ \int (\partial/\partial \alpha^T) s_0(x; \alpha) dF_{\alpha}^0(x) \right]^{-1}$ $s_0(z; \alpha)$ is the influence function (IF) of the estimator $\hat{\alpha}$. The bias on the
asymptotic level is then given by

\[ \omega(\varepsilon) - \omega(0) = -2 \cdot \varepsilon \cdot \text{var}(F_0^\alpha)^{-1/2} \int_{-\infty}^{\infty} y d\Phi(y) \cdot \left\{ \int L(x; \alpha, \beta_\alpha) dF_0^\alpha(x) - L(z; \alpha, \beta_\alpha) + \int L(x; \alpha, \beta_\alpha) s^0(x; \alpha) dF_0^\alpha(x) \cdot IF^*(z, \hat{\alpha}, F_0^\alpha) \right\} + O(\varepsilon^2) \]  

That means that a single observation \( z \) such that \( L(z; \alpha, \beta_\alpha) = \log f^0(z; \alpha) - \log f^1(z; \beta_\alpha) \) or \( s^0(z; \alpha) \) is large can make the bias on the asymptotic level very large. Indeed, the non-robustness of the test, i.e. the bias on the asymptotic level, is due simultaneously to

- the non-robustness of the estimator of the parameter \( \alpha \)
- the non-robustness of the test statistic

While \( s^0(z; \alpha) \) is up to a multiplicative constant the IF of the maximum likelihood estimator of the parameter under the null hypothesis, \( L(z; \alpha, \beta_\alpha) \) is directly related to the influence on the test statistic. Therefore, it is not sufficient to base a test on robust estimators for the parameters only. Indeed, a robust estimator for \( \alpha \) guaranties a bounded value for \( IF^*(z, \hat{\alpha}, F_0^\alpha) \) but not for \( L(z; \alpha, \beta_\alpha) \). For example, if we want to test the Gamma \((F_{\alpha_1, \alpha_2}, \alpha_1 \) is the shape parameter and \( \alpha_2 \) is the scale parameter) against the Lognormal \((F_{\beta_1, \beta_2}, \beta_1 = \mu \) and \( \beta_2 = \sigma^2 \)), the difference between the log-likelihood
functions evaluated at any point $z$ is given by

$$
\left( \alpha_1 - \frac{\beta_1}{\beta_2} \right) \log(z) + \frac{1}{2\beta_2^2} \log(z)^2 - \alpha_2 z + Q(\alpha_1, \alpha_2, \beta_1, \beta_2)
$$

which can be large when $z$ is large or small. Finally, note that (3) doesn’t depend on the $IF$ of $\hat{\beta}$, such that it is not necessary to choose a robust estimator for $\beta$ to build a robust test statistic.

## 3 Robust model choice tests

Since the Cox statistic can be viewed as a parametric Lagrange multiplier test, we present an optimal robust version by making use of the work of Heritier and Ronchetti (1994) on robust bounded-influence tests in general parametric models. In particular we show that when we apply the results on robust bounded-influence Lagrange multiplier test to the model choice test, we bound exactly the right quantity. We therefore obtain a robust version of Cox and Cox type statistics. We end this section with some simulation results.
3.1 Robust bounded-influence Lagrange multiplier test

In this subsection, we briefly present the robust version of the parametric Lagrange multiplier test developed by Heritier and Ronchetti (1994).

Consider a general parametric model \( \{F_\theta\} \) with density \( f(\cdot; \theta) \) and scores function \( s(\cdot; \theta) \). We are interested in testing the null hypothesis \( H_0 : \theta(2) = 0 \) and \( \theta(1) \) unspecified, against the alternative \( H_1 : \theta(2), \theta(1) \) unspecified, where \( \theta(2) (r_2 \times 1) \) is the parameter of interest. Heritier and Ronchetti (1994) consider tests which rely on M-estimators \( T_\psi \) of \( \theta \) defined by \( \sum \psi(x_i, T_\psi) = 0 \).

They generalise the Lagrange multiplier (or scores) statistic with \( R^2 = U_{GLM}^T C^{-1} U_{GLM} \), where \( U_{GLM} = n^{-1} \sum \psi(x_i; T^*_\psi(2)) \), \( \psi(\cdot; \cdot)(2) \) is the second part of dimension \( r_2 \times 1 \) of the vector \( \psi(\cdot; \cdot) \), \( T^*_\psi \) is the M-estimator in the reduced model, i.e the solution of \( \sum \psi(x_i; T^*_\psi(1)) = 0 \), with \( T^*_\psi(2) = 0 \) and where the matrix \( C \) depends on the asymptotic covariance matrix of the M-estimator (see Heritier and Ronchetti 1994). Heritier and Ronchetti (1994) show that under the null hypothesis \( nR^2 \) is asymptotically distributed according to a \( \chi^2_{r_2} \) distribution. They compute the \( LIF \) and find that it is proportional to the square of the \( IF \) of the statistic \( U_{GLM} \). The latter is equal to the self-standardised \( IF \) of the estimator \( T_{\psi(2)} \) (see Hampel, Ronchetti, Rousseeuw, and Stahel 1986, p. 252-257). An optimal test is then one which maximises the power of the test at the model, given a bound on the \( LIF \).
Since the $LIF$ of the Lagrange multiplier test is proportional to the self-standardised $IF$ of the partitioned estimator $T_{\psi(2)}$, then the robust version of the Lagrange multiplier test is the optimal robust self-standardised estimator $T_{\psi(2)}$ with partitioned parameters given in Hampel, Ronchetti, Rousseeuw, and Stahel (1986), p. 252-257. For a given bound $c \geq r_2^{1/2}$, it is based on the following $\psi$-function

$$
\psi_c(x; \theta)_{(1)} = A_{(11)}s(x; \theta)_{(1)}
$$

$$
\psi_c(x; \theta)_{(2)} = h_c \left( [A[s(x; \theta) - a]]_{(2)} \right)
$$

where $h_c(x)$ is the Huber function $h_c(x) = x \min(1; c/\|x\|)$, $a_{(1)} = 0$, the $r_2$-dimensional vector $a_{(2)}$ and the lower triangular matrix $A$ are determined by the equations

$$
\int \psi_c(x; \theta)_{(2)}dF_\theta(x) = 0 \quad (4)
$$

$$
\int \psi_c(x; \theta)\psi_c^T(x; \theta)dF_\theta(x) = I \quad (5)
$$

The solution for $\psi$ leads to a simplification in the expression of the statistic. Indeed, we have $C = I$. The robust Lagrange multiplier statistic is then given by $R^2 = U_{GLM}^T U_{GLM} = \{n^{-1} \sum \psi_c(x_i; (T_{MLE(1)}, 0))_{(2)} \}^T \cdot \{n^{-1} \sum \psi_c(x_i; (T_{MLE(1)}, 0))_{(2)} \}$. We note that depending on the choice of the
bound $c$, one can have a more or less robust test statistic. The lower the bound $c$ the more robust the test statistic. At $c = \infty$, we have the classical non robust statistic.

### 3.2 Robust Cox-type statistic

In this subsection we apply the results of Heritier and Ronchetti (1994) to Cox-type statistics when they are interpreted as a Lagrange multiplier test. If we construct the comprehensive model

$$f^c(x; \theta) = \left\{ f^0(x; \alpha) \right\}^\lambda \left\{ f^1(x; \beta) \right\}^{(1-\lambda)} \left[ \int \left\{ f^0(y; \alpha) \right\}^\lambda \left\{ f^1(y; \beta) \right\}^{(1-\lambda)} dy \right]^{-1}$$

(6)

where $\theta = (\alpha, \lambda)^T$, then the Lagrange multiplier test statistic corresponding to the hypothesis $H_0 : \lambda = 1$ against the alternative $H_1 : \lambda \neq 1$ leads to the Cox, Atkinson or White statistic, depending on the choice of the estimator $\tilde{\beta}$ of $\beta$. Note that we could reparameterise the problem by defining $\gamma = \lambda - 1$, $H_0 : \gamma = 0$, $H_1 : \gamma \neq 0$. This would lead to the same results.

Under $H_0$ the scores function corresponding to $f^c$ is given by

$$s^c(x; \theta) = \frac{\partial}{\partial \theta} \log f^c(x; \theta) \bigg|_{\lambda = 1} = \begin{bmatrix} s^c(x; \theta)(1) \\ s^c(x; \theta)(2) \end{bmatrix}$$
where

\[ s^c(x; \theta)_{(1)} = \frac{\partial}{\partial \alpha} \log f^c(x; \theta) \bigg|_{\lambda=1} = \frac{\partial}{\partial \alpha} \log f^0(x; \alpha) = s^0(x; \alpha) \]

and

\[ s^c(x; \theta)_{(2)} = s_{Cox}(x; \alpha, \beta) \]

\[ = \frac{\partial}{\partial \lambda} \log f^c(x; \theta) \bigg|_{\lambda=1} = L(x; \alpha, \beta) - \int L(y; \alpha, \beta)f^0(y; \alpha)dy \]

The corresponding \( \psi_c \)-function is then given by

\[
\psi_c(x; \theta) = \begin{bmatrix}
A_{(11)}s^0(x; \alpha) \\
[A_{(21)}s^0(x; \alpha) + A_{(22)}[s_{Cox}(x; \alpha, \beta) - a_{(2)}]] W_c(x; \alpha, \beta)
\end{bmatrix}
\]

where

\[
W_c(x; \alpha, \beta) = \min \left\{ 1; c \cdot \left| A_{(21)}s^0(x; \alpha) + A_{(22)}[s_{Cox}(x; \alpha, \beta) - a_{(2)}] \right|^{-1} \right\} .
\]

The robust Cox-type statistic is finally given by

\[
U_{GLM} = \frac{1}{n} \sum_{i=1}^{n} \left[ A_{(21)}s^0(x_i; \hat{\alpha}) + A_{(22)}[s_{Cox}(x_i; \hat{\alpha}, \beta) - a_{(2)}] \right] W_c(x_i; \hat{\alpha}, \beta)
\]

where \( \hat{\alpha} \) is the maximum likelihood estimator of \( \alpha \) and \( A_{(21)}, A_{(22)}, a_{(2)} \) are
determined implicitly by (4) and (5), with $F_{\theta} = F_{\hat{\theta}}^c$ is the comprehensive model and at $\theta = \hat{\theta} = (\hat{\alpha}, 1)$.

The asymptotic normality of $n^{1/2}U_{GLM}$ under $H_0$ is proven in Heritier and Ronchetti (1994) (see proof of proposition 2). In practice however, $\beta$ needs to be estimated. If $\tilde{\beta}$ is a consistent estimator, then the asymptotic normality still holds. Indeed, to prove it, one would follow the proof given by Heritier and Ronchetti (1994) and add a development of the statistic around $\tilde{\beta}$. The added term would disappear for consistent estimators of $\beta$.

Knowing that the bias on the asymptotic level (see (3)) is proportional to $s^0(z; \alpha)$ and to $L(z; \alpha, \beta)$, we see that by using the robust version of the Lagrange multiplier test with the comprehensive model (6), we bound exactly the right quantity. Therefore, the use of (8) avoids that the decision is influenced by a small amount of outliers.

To compute the robust Lagrange multiplier statistic, we propose to use the same algorithm as for optimal B-robust estimators (see e.g. Victoria-Feser and Ronchetti 1994). For a given bound $c$, it is given by the following 4 steps.

$Step 1$: Compute the maximum likelihood estimator for $\alpha$. $\hat{\theta} = (\hat{\alpha}, 1)^T$. 

Step 2: Solve for $A_{(21)}$, $A_{(22)}$ and $a_{(2)}$, the following implicit equations

\[
\int \psi_c(x; \hat{\theta}) f^0(x; \hat{\alpha}) dx = 0 \quad (9)
\]
\[
\int \psi_c(x; \hat{\theta}) \psi_c(x; \hat{\theta})^T f^0(x; \hat{\alpha}) dx = 0 \quad (10)
\]
\[
\int \psi_c(x; \hat{\theta}) \psi_c(x; \hat{\theta})^T f^0(x; \hat{\alpha}) dx = 1 \quad (11)
\]

where $A_{(12)} = a_{(1)} = 0$. As starting values, one could choose $a = 0$ and $A$ such that $A^{-1}A^{-T} = \int s^c(x; \hat{\theta})s^c(x; \hat{\theta})^T f^0(x; \hat{\alpha}) dx$.

Step 3: Compute $U_{GLM}$ given in (8).

Step 4: At the nominal level $\omega$, accept $H_0$ if $\left| n^{1/2}U_{GLM} \right| < \kappa_\omega$, where $\kappa_\omega = \Phi^{-1}(1 - \omega/2)$.

It should be stressed that the third step is not really obvious to compute. Unlike the case of parameter estimation, one cannot solve equations (9), (10) and (11) by means of an iterative process. One has instead to use a routine to find the zero roots of a system of nonlinear equations.

3.3 Simulation study

In order to study the robustness properties of $U_{GLM}$ when compared to the classical Cox-type statistics, we used the test when the samples are contaminated and non contaminated. We choose to simulate Pareto samples and test
the Pareto distribution against the Exponential distribution by means of the
Atkinson statistic. Note that we obtain similar results with other Cox-type
statistics.

The Pareto density is given by $f_0(x; \alpha) = \alpha x^{-(\alpha+1)}x_0^\alpha$ with $0 < x_0 \leq x < \infty$, so that as alternative we considered the truncated exponential dis-
tribution given by $f_1(x; \beta) = \beta e^{-\beta(x-x_0)}$. We simulated 1000 samples of
200 observations from a Pareto distribution with parameter $\alpha = 3.0$ ($x_0 = 0.5$), samples that we contaminated by means of
$(1 - \epsilon \cdot 200^{-1/2}) F_{\alpha,x_0} + \epsilon \cdot
200^{-1/2} F_{\alpha,10^{-x_0}}$. For amounts of contaminations from $\epsilon = 0\%$ to $\epsilon = 20\%$,
Table 1 gives the actual levels of the classical and robust ($c = 2.0$) Atkinson
statistic when testing the Pareto against the Exponential distribution.

Table 1 here

The actual levels are the probabilities that ($n^{1/2}$ times) the test statistic
(absolute value) computed from the simulated samples exceed the critical
value $\kappa_\omega$, where $\omega$ is the nominal level. We can observe that the classical
statistic has a very strange behaviour since when there is no contamination
the null hypothesis is underrejected and with only small amounts of contamina-
tion, the null hypothesis is overrejected. The first phenomenon is probably
due to the fact that the approximation of the actual distribution of the Cox-
type statistics by means of their asymptotic distribution is not accurate (see
e.g. Williams 1970, Atkinson 1970 and Loh 1985). The second phenomenon
is the lack of robustness as shown in subsection 2.4. On the other hand,
we find that with the robustified Atkinson statistic not only the asymptotic
distribution is a good approximation of its sample distribution, but also that
the small departures from the model under the null hypothesis do not influ-
ence the level of the test at least for amounts of contamination up to about
\[ \varepsilon = 10\% \]. With more contamination (15\% and 20\%), the null hypothesis
tends to be slightly overrejected at the 5\% and 10\% levels, but this is not
to drastic compared to the classical case. In other words, the robust test is
very stable.

The fact that the level of the robust test is not very much influenced by
contamination is due to the structure of the test itself (see (8)). However,
that the asymptotic distribution of the robust test statistic is a good ap-
proximation of its sample distribution (compared to the classical test) can
seem at first rather surprising. This can be understood intuitively by re-
membering the probable causes of the problems in the classic test: Atkinson
(1970) remarked that some rather small (legitimate) observations have a
large influence on the value of the test statistic because one often takes their logarithm. With robust techniques the influence of such ‘extreme but legitimate’ observations is bounded, such that the null hypothesis is not under- or overrejected.

3.4 The quantal responses example

As a second example, we computed the robust White statistic for the quantal responses model presented in subsection 2.3. With the non contaminated sample, the robust test statistic \( c = 2.0 \) has a value of 0.2388 corresponding to a \( p \)-value of 40.5% leading to the acceptance of \( H_0 \). With the contaminated sample, the robust test statistic \( c = 2.0 \) has a value of 0.8132 corresponding to a \( p \)-value of 20.9% leading again and contrary to the classical White statistic to the acceptance of \( H_0 \). Moreover, with the robust test statistic we can look at the weights given by the robust statistic to the observations (see (7)) such that we can immediately point out the extreme observations.

3.5 Power comparison

In this subsection, we study the power properties of the robust Cox-type statistics by means of a simulation. Our aim is not to give a general answer but to give an idea of the loss of power with the robust statistic when the data
are not contaminated and the gain of power when the data are contaminated.

When the models are non-nested, a useful concept of power is given by the probability of making the correct decision under the alternative, i.e. of accepting the true model and rejecting the false one (see Pesaran 1982). Actually, two tests are computed, one for the model under $H_1$ against the model under $H_0$ and another for the model under $H_0$ against the model under $H_1$. The power of the test is then estimated by the frequency of the samples that lead to the acceptance of $H_1$ and the rejection of $H_0$ by means of the two tests. We chose to test the Pareto against the Exponential distributions. We first simulated 1000 samples of 200 observations from the Exponential distribution with parameter $\beta = 4.0 \ (x_0 = 0.5)$, without contamination. In Table 2 we computed the power (corresponding to different levels) of the robust Atkinson statistic for different bounds $c$. We can first remark that the power of the classical test (i.e. $c = \infty$) could hardly be worse! The reason is probably that for this particular test (Pareto against Exponential), the approximation of the distribution of the test statistic by its asymptotic distribution is not accurate. With the robust version of the statistic this problem disappears and we find reasonably good powers. Moreover, the loss of power for using a more robust test statistic ($c = 2.0$) is not that large.
In a second step, we simulated 1000 samples of 200 observations from an Exponential distribution with parameter $\beta = 4.0 \ (x_0 = 0.5)$ and contaminated the samples by means of $(1 - \varepsilon \cdot 200^{-1/2}) F_{\beta, x_0} + \varepsilon \cdot 200^{-1/2} F_{\beta, 10 - x_0}$ from $\varepsilon = 0\%$ to $\varepsilon = 20\%$. We computed the power (corresponding to different levels) of the robust ($c = 2.0$) and classical Atkinson statistics in Table 3.

With the classical statistic, we can see that the test is just useless since no matter how little the data are contaminated, the power of the test is null. With the robust statistic, we can remark that not only the power is fairly high, but also it remains stable when the data are contaminated.
4 Conclusion

In this paper we focused on model choice procedures for separate models. We showed that the classical Cox-type statistics not only suffer from a lack of robustness but also their asymptotic distribution is not always an accurate approximation of their sample distribution. We therefore proposed a robust version of Cox-type statistics based on robust parametric tests for general parametric models. In particular we showed that with the new test statistic, the influence of small amounts of contamination in the observations is limited. We illustrated this result by means of simulations and found out that the asymptotic distribution of the robust test statistic is a good approximation of its sample distribution.

Acknowledgements

The author was partially supported by the ‘Fond National Suisse pour la Recherche Scientifique’. She would like to thank Prof. A. C. Atkinson for his helpful comments during the preparation of this paper, Prof. E. Ronchetti and Dr. M. Knott for their valuable comments on an earlier version of the paper.
References


Cox, D. R. (1962). Further results on tests of separate families of hypothe-


Statist. 21, 103–112.


Table 1: Actual levels (in %) of the classical and robust Atkinson statistic \((c = 2.0)\) with contamination (Pareto against Exponential)

<table>
<thead>
<tr>
<th>Amount of contamination</th>
<th>Classical statistic</th>
<th>Robust statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Nominal levels (in %)</td>
<td>Nominal levels (in %)</td>
</tr>
<tr>
<td></td>
<td>1% 3% 5% 10%</td>
<td>1% 3% 5% 10%</td>
</tr>
<tr>
<td>0%</td>
<td>2.1 3.1 3.5 5.2</td>
<td>1.3 3.5 5.5 10.2</td>
</tr>
<tr>
<td>3%</td>
<td>6.3 8.7 10.3 14.7</td>
<td>1.2 3.3 5.1 10.3</td>
</tr>
<tr>
<td>6%</td>
<td>13.1 18.5 22.5 27.6</td>
<td>1.4 3.6 5.4 10.7</td>
</tr>
<tr>
<td>10%</td>
<td>24.4 31.3 35.2 43.9</td>
<td>1.3 3.0 5.6 11.4</td>
</tr>
<tr>
<td>15%</td>
<td>35.6 44.6 49.9 58.1</td>
<td>1.4 4.1 7.9 14.5</td>
</tr>
<tr>
<td>20%</td>
<td>46.3 54.2 58.6 67.1</td>
<td>0.9 4.1 7.6 14.5</td>
</tr>
</tbody>
</table>

Table 2: Power of the robust Atkinson statistic (Pareto against Exponential)

<table>
<thead>
<tr>
<th>Bound c</th>
<th>Power at the levels</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1% 3% 5% 10%</td>
</tr>
<tr>
<td>2.0</td>
<td>0.74 0.86 0.88 0.90</td>
</tr>
<tr>
<td>5.0</td>
<td>0.88 0.94 0.94 0.89</td>
</tr>
<tr>
<td>10.0</td>
<td>0.78 0.94 0.96 0.90</td>
</tr>
<tr>
<td>15.0</td>
<td>0.53 0.87 0.93 0.91</td>
</tr>
<tr>
<td>∞</td>
<td>0.00 0.00 0.00 0.38</td>
</tr>
</tbody>
</table>

Table 3: Power of the classical and robust \((c = 2.0)\) Atkinson statistics (Pareto against Exponential), with contaminated samples

<table>
<thead>
<tr>
<th>Amount of contamination</th>
<th>Classical statistic</th>
<th>Robust statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Power at the levels</td>
<td>Power at the levels</td>
</tr>
<tr>
<td></td>
<td>1% 3% 5% 10%</td>
<td>1% 3% 5% 10%</td>
</tr>
<tr>
<td>0%</td>
<td>0.00 0.00 0.001 0.376</td>
<td>0.742 0.858 0.876 0.895</td>
</tr>
</tbody>
</table>