Robust goodness-of-fit tests of the classical linear regression model

Oleksandr Movshuk
Assistant Research Professor, ICSEAD
and
Visiting Assistant Professor, Graduate School of Economics,
Kyushu University

Working Paper Series Vol. 2002-01
March 2002

The International Centre for the Study of East Asian Development, Kitakyushu

Robust goodness-of-fit tests of the classical linear regression model

Oleksandr Movshuk*
International Centre for the Study of East Asian Development
11-4 Otemachi, Kokurakita, Kitakyushu, 803-0814, Japan (email: movshuk@icsead.or.jp)

ABSTRACT

This paper develops two goodness-of-fit tests to verify the joint null hypothesis that regression disturbances have zero mean and constant variance, and are generated from the Gaussian normal distribution. Initially, these tests use a high-breakdown regression estimator to identify a subset of regular observations, and then calculate standardized prediction residuals and studentized prediction residuals, from which the final test statistics are derived. A Monte Carlo study demonstrates that the first test is particularly sensitive to a small number of regression outliers with non-zero mean or unusually large variance, and in general to regression misspecifications that produce regression disturbances with longer tails than could be justified by the normality assumption. In contrast, the second test detects a substantial number of regression outliers, specifications with incorrect functional forms, omissions of relevant variables, and short tails in the distribution of the error term. While most specification tests are designed for a particular alternative, the joint application of the proposed tests has a high power to detect a wide range of breakdowns of the linear regression model. The omnibus property of the suggested tests makes redundant the current practice of running the battery of various specification tests.

* The financial support from ICSEAD and Kyushu University is gratefully acknowledged.

1. INTRODUCTION

Although the classical regression model $Y = X\beta + \varepsilon$ postulates that the error term ε has zero mean, constant variance σ^2 and normal distribution for all its elements, conventional econometric measures of goodness of fit do not directly examine these distributional assumptions. For example, the traditional measure of goodness of fit – the R^2 statistic – compares the variation of estimated OLS residuals with the variation of dependent variable *Y*. However, this ratio does not verify any distributional properties of ε. As a result, researchers that examine the goodness-of-fit of their models by R^2 may fail to notice substantial deviations from the assumed distribution of ε.

Distributional assumptions about ε can be also verified by univariate tests for normality, such as the Jarque-Bera, Shapiro-Wilk, or Shapiro-Francia tests. Since the vector of regression disturbances ε is not observable, the univariate tests for normality are usually applied to available estimates of ε, most often to OLS residuals $\hat{\varepsilon} = (I - V)\varepsilon$, where V is the projection matrix $X(X'X)^{-1}X'$. The distribution of $\hat{\varepsilon}$ does converge to ε asymptotically (Theil, 1971, p. 378-379), but in finite samples their correspondence may be poor[1]. In particular, if some rows of X are unusually large, then the corresponding diagonal elements of matrix V (denoted hereafter by v_{ii}) can also become large, producing so-called 'high-leverage points'. Since the variance of i[th] element of OLS residual vector $\hat{\varepsilon}$ is $\sigma^2(1 - v_{ii})$, high leverage points may substantially reduce the variance of OLS residuals[2]. In general, the projection matrix V can modify OLS residuals so much that the distribution of $\hat{\varepsilon}_i$ may have little in common with the distribution of ε. This may greatly diminish the power of univariate tests of normality. For

[1] This is because the distribution of ε' depends on the configuration of the matrix *V*.

[2] The drop in variance is especially conspicuous when a dummy variable for i^{th} observation is used. In this case $v_{ii} = 1$, so that the variance, as well as the value of $\hat{\varepsilon}_i$, would become zero even though the corresponding regression disturbance ε_i may have unusually large variance, or non-zero mean

example, if deviations from the assumed normality of ε are located at high-leverage points, univariate tests for normality may fail to spot these deviations in estimated $\hat{\varepsilon}_i$ [3].

Another drawback of univariate tests for normality is that they usually examine the null hypothesis of normal distribution *regardless of* any specific parametric value for mean and variance of ε. In contrast, the full ideal conditions of the linear regression model explicitly postulate that each element of vector ε has *zero* mean. Thus, most of the reported normality tests are in fact examining a broader null hypothesis than the one postulated by the linear regression model.

In this paper I introduce two goodness of fit tests of the linear regression model that avoid these pitfalls of mechanical application of univariate goodness-of-fit tests to OLS residuals. A Monte Carlo study demonstrates a high degree of complimentarity in the power of these tests. In particular, the first test is sensitive to a small number of regression outliers, defined as observations with non-zero mean or unusually large variance of ε. More generally, the first test is sensitive to regression misspecification that results in ε with *longer* tails than could be justified by the normality assumption. In contrast, the second test has a good power to detect a large number of regression outliers, specifications with incorrect functional forms, omissions of relevant variables, and short tails in the distribution of ε.

While the vast majority of specification tests in econometrics are designed for a particular alternative[4], the joint application of the proposed tests has a high power to detect most major breakdowns of the classical linear regression model[5]. The omnibus property of the

---

[3] This was demonstrated by Weisberg (1980), who compared the power of the univariate Shapiro-Wilk test with respect to unobservable ε and to corresponding regression residuals $\hat{\varepsilon}$, resulting from different configurations of matrix V. When, for example, ε had log-normal distribution, the Shapiro-Wilk easily detected the non-normality of ε (the power was 99 per cent), but depending on V, the test power with OLS residuals varied from 41 to 91 per cent.

[4] Even including a few 'general' tests that are ostensibly designed to guard against unspecified violations of full ideal conditions, as demonstrated by Thursby (1989).

[5] With the exception of the independence of ε and the full rank of X.

suggested tests makes redundant the current practice of running the battery of various specification tests.

The paper is organized as follows. Section 2 describes test statistics. Their non-standard distribution under the null is discussed in section 3, followed by an illustrative example in section 4. Section 5 reports the power of suggested and conventional tests in detecting major violations of the linear regression model, and section 6 offers conclusions.

2. TEST ALGORITHM

Consider the standard regression model $y = X\beta + \varepsilon$, where y is $(n \times 1)$ vector of observations on a dependent variable, X is $(n \times k)$ matrix of n observations on k independent variables (including the intercept), β is $(k \times 1)$ vector of unknown regression coefficients, and ε is $(n \times 1)$ vector of unobservable regression disturbances, assumed to be *i.i.d.* $N(0, \sigma^2 I_n)$. The matrix X is assumed to be fixed and of full rank.

To validate the distributional properties of ε, the test algorithm uses the sequence of recursive residuals w_i $(i = k+1, ..., n)$

$$w_i = (y_i - x_i \hat{\beta}_{i-1}) \Big/ \sqrt{1 + x_i^{'} (X_{i-1}^{'} X_{i-1})^{-1} x_i} \tag{1}$$

where $\hat{\beta}_{i-1} = (X'_{i-1} X_{i-1})^{-1} X'_{i-1} Y_{i-1}$ is the OLS estimate of β, calculated from preceding $i-1$ observations. As shown by Brown *et al.* (1975), if the null hypothesis $\varepsilon \sim N(0, \sigma^2)$, then $w \sim N(0, \sigma^2)$. In other words, if regression disturbances ε have zero mean and constant variance σ^2 and are normally distributed, then estimated recursive residuals w have the same property.

The values of recursive residuals depend on data ordering, so that even a minor data permutation can produce a completely different set of recursive residuals. In particular, the exact normality of recursive residuals holds only if they are calculated on randomly-ordered

observations, or by any variable that is statistically independent of w_i (such as any independent variable in X, or predicted values $X\hat{\beta}$). If, in contrast, the data ordering is determined by analyzed data (*i.e.*, determined endogenously), the normality of w is only approximate.

A. ENDOGENOUS DATA ORDERING

Starting from Brown *et al.* (1975), most application of recursive residuals assumed that the ordering of observations is either fixed or known a priori (with frequent references to a 'natural' order, such as time or by some independent variable). In contrast, the objective of this paper is to select an endogenous order of observations in a way that maximizes the power of recursive residuals to detect departures from the null hypothesis. The endogenous ordering makes the distribution of w nonstandard, but this problem can be solved by Barnard's (1963) suggestion to approximate the distribution of any non-standard statistic with simulated random data, generated under the null hypothesis.

The task of ordering observations endogenously is to ensure that the estimation subset of preceding observations contains only regular observations, while discordant observations penetrate into the estimation subset as late as possible. In statistical literature, this approach was applied for testing regression outliers by Hadi and Simonoff (1993). They suggested to assemble the most regular observations in the initial subset of k observations (which is the minimum size to calculate the first non-zero recursive residual w_{k+1}), and then enlarge the estimation subsets recursively by the least deviant observation outside the estimation subset.

If the initial subset of k observations contains the most regular observations, the Hadi-Simonoff algorithm will move discordant observations to the very end of the ordered sample. Essentially, their algorithm precludes the harmful impact of regression outliers on the estimated regression parameters as long as regression outliers are kept outside the estimation subset.

The composition of the initial subset of k observation is the Achilles' heel of the Hadi-Simonoff algorithm. They suggested to use observations with the smallest absolute 'adjusted residuals', defined by $\hat{\varepsilon}_i \big/ (1 - v_{ii})^{1/2}$. Basically, this is OLS residual $\hat{\varepsilon}_i$, divided by its standard error. The normalization solves the problem of heteroskedasticity of $\hat{\varepsilon}_i$. However, adjusted residuals are still immune to so-called 'masking effect', when regression outliers with large ε are 'masked' by small OLS residuals $\hat{\varepsilon}$ [6] and remain unidentified.

Some estimators are not susceptible to the masking effect, including, among others, the least trimmed squares (LTS) estimator, introduced by Rousseeuw (1984). Like the OLS, the LTS minimizes the sum of squared residuals, but disregards the impact of the most discordant part of data. Define *[q]* the integer part of q, and let $h = \left[(n + k + 1)/2 \right]$. While the OLS criteria minimizes the sum of squared residuals for all observations, the LTS criteria minimizes the sum of only h smallest squared residuals, and ignores *(n-h)* observations with larger residuals.

Rousseeuw and Hubert (1997, p. 5) demonstrated that parameter estimates by the LTS remain bounded if there are as many $s = n - h = [(n - k)/2]$ or fewer regression outliers. Thus, the LTS estimator becomes unbounded only with $(s + 1)$ regression outliers. In contrast, the OLS breaks down with even a single outlying observation.

One drawback of the LTS estimator is that, unlike OLS, its objective function does not have a closed form solution. As a result, LTS estimates are calculated by various algorithms, most often by the original *PROGRESS* code from Rousseeuw (1984), which is based on so-called elemental set algorithm. In essence, this algorithm picks up a subset of k observations (called 'elemental set'), and computes exact parameter estimates for k independent variables. Then it calculates residuals for the remaining *n-k* observations. These residuals are squared and sorted,

---

[6] In essence, the masking effect is due to the fitting criteria of the OLS estimator. The OLS minimizes the sum of squared deviations of *all* observations, even if some observations are regression outliers. When the proportion of outliers is large, the OLS fit tries to accommodate the outlying observations as well, producing OLS residuals that do not clearly distinguish the presence of outlying observations.

and the sum of h smallest residuals is calculated as the LTS fitting criteria. If the fitting criteria is smaller than the one from previous 'elemental set' trials, the criteria is stored (together with parameter estimates). The procedure is repeated with different subsets of size k, and the final parameter estimates are ones that eventually produced the smallest LTS fitting criteria.

Clearly, the algorithm is certain to find the global LTS minimum only if it evaluates all $\binom{n}{k}$ combinations of the data. This is feasible for small models, but the computational cost rapidly becomes prohibitive even for moderate n and k. As shown in table 1, for $n = 50$ and $k = 7$, the algorithm requires evaluation of as many as 99,884,400 elemental subsets. With $n = 100$ and $k = 7$ the number of elemental sets soars to 16,007,560,800!

Recently, Rousseeuw and Van Driessen (1999) developed a much faster LTS algorithm which finds a close approximation to the LTS global minimum with much less computational cost. The modified LTS algorithm is based on the idea that instead of evaluating all $\binom{n}{k}$ elemental subsets of data, it is sufficient to minimize the probability that elemental sets contain at least a single discordant observation. Let m and ϖ be the number of elemental sets and the maximum share of outlying observations, respectively. Since elemental sets are picked up at random, the probability ρ that at least one elemental subset contains at least one discordant observation is $\rho = (1-(1-\varpi)^k)^m$. Setting ρ to a sufficiently small number (say, 0.01), and fixing ϖ at its maximum[7] asymptotic level 0.50, the lower limit of 'elemental set' regressions to keep ρ below 0.01 can be quite feasible, such as just 590 for $n = 50$ and $k = 7$ (see column 4 of table 1)[8].

---

[7] The share of discordant observations could not exceed 0.5, since otherwise the distinction between regular and discordant observations is not meaningful.

[8] Note that ρ does not involve the sample size n, so that the same number of 590 'elemental set' regressions is sufficient to for models with $n = 100$ and $k = 7$ and so on.

In fact, the default number of evaluated elemental sets m_d in the *PROGRESS* code often ensures that ρ is very low. Consider a case where $n = 50$ and $k = 7$, for which the default m_d is set to 3,000. The probability that at least one discordant observations affects the LTS parameter estimates is, in fact, as low as $(1-(1-0.5)^7)^{3000} \cong 6\times 10^{-11}$. Similar probabilities for other default cases are given in column 7 of table 1.

This feature of the *PROGRESS* code is used in the following LTS-OLS algorithm. Initially, the algorithm ranks the data endogenously from the most regular to more discordant observations. It produces the first basic subset B_1 of regular observations. However, in contrast to the Hadi-Simonoff algorithm, the subset contains h (rather than k) observations.

TEST ALGORITHM (PART 1)

Step 1. Apply the LTS estimator to all n observations, and evaluate LTS minimizing criteria $\sum_{i=1}^{h}(y_i - x_i\hat{\boldsymbol{\beta}}_{LTS})_{(i)}^2$ with at least the default number of elemental sets m_d in the *PROGRESS* code.

Step 2. Using LTS parameter estimates after m_d elemental set trials, calculate LTS residuals.

Step 3. Sort all n observations by absolute values of their LTS residuals, and include in the initial basic subset B_1 h observations with the smallest absolute LTS residuals.

Step 4. Apply OLS estimator to observations in subset B_1, and calculate OLS absolute residuals for these h observations.

Step 5. Sort observations in subset B_1 by absolute values of their OLS residuals.

It will be shown shortly that the addition of OLS in Steps 4 and 5 improves the precision of parameter estimates after the LTS fit in Step 1. This efficiency gain can also be explained intuitively. Parameter estimates by the elemental sets algorithm are *exactly* determined by k observations. In contrast, the subsequent OLS estimate in Step 4 takes into account the information from h observations in B_1 (with $h \geq k$).

To verify the robustness, efficiency, and computational cost of the combined LTS-OLS algorithm, I ran a small Monte Carlo experiment with k set to 9 (so that the identification of regular observations becomes fairly complicated), while sample size n set to 20, 60, and 100. These values of k and n required the evaluation of 1.68×10^5, 1.48×10^{10} and 1.90×10^{12} elemental sets. The regular data-generating process was $Y=\sum_{j=1}^{k}\beta_j X_j+\varepsilon$, with $X_1=1$, $\beta_j=1$ for all j and $\varepsilon\sim N(0,1)$.

In total, 5 patterns of data were considered. In the first experiment all regression disturbances are generated as $\varepsilon\sim N(0,1)$ with no regression outliers (replicating the ideal specification of the linear regression model). In other experiments, regular data were replaced by discordant observations, generated as $N(100,10)$.

As mentioned before, parameter estimates of LTS fit remain bounded if the number of regression outliers is $s=n-h$ or fewer. Utilizing this property, the robustness of LTS with respect to the robustness bound s was verified. Specifically, the efficiency of OLS and LTS estimators with just a single outlier (second experiment), $s/2$ outliers (third experiment), s outliers (fourth experiment), and $s+1$ outliers (fifth experiment) was considered.

As discussed before, leverage points can have sizeable impact on the distribution of OLS residuals. Thus, whether the leverage points may affect the LTS estimator is also investigated. First, outliers at low leverage points, when each row of matrix X was generated from uniform distribution $U(0,15)$ were generated. Secondly, outliers at high leverage points, with corresponding rows of matrix X distributed as $N(15,10)$ and other rows of X generated as low leverage points were planted.

To obtain random numbers, a combined multiple recursive generator *MRG32k5a*, suggested by L'Ecuyer (1998, p. 13, 15)[9] was used. Each experiment consisted of 500 replications.

[9]The generator has two components of order five, and is implemented in TSP 4.5. The maximal period length of the generator is about 1.07×10^{96}, compared with the conventional period length of 2.15×10^9.

The efficiency of OLS and LTS was evaluated by the mean squared error (MSE) around the true parameter values. Table 2 reports major results for OLS, LTS, and LTS-OLS. The table also reports timing (in minutes) for all $500 \times 5 = 2{,}500$ replications of OLS, LTS, and LTS-OLS.

In experiment 1, the data-generating process coincides with the full ideal conditions of the linear regression model. No wonder that the OLS had a clear advantage in efficiency, yielding the lowest MSE. Interestingly, even with no regression outliers, the combined LTS-OLS algorithm achieved some improvement over LTS. The gain in efficiency was about 10 per cent, most noticeably for large *n*. Thus, the introduction of Steps 4 and 5 does result in improved efficiency of estimating regression parameters.

In the second experiment with a single regression outlier, the OLS estimator broke down spectacularly, especially with high leverage outliers. On the other hand, parameter estimates by both LTS and LTS-OLS were always bounded as long as the number of outliers did not exceed the robustness limit *s*. Overall, this theoretical property of high-breakdown robust estimators was verified for all sample sizes. However, both LTS and LTS-OLS broke down when the number of outliers surpassed *s*. Finally, there was the same efficiency of LTS and LTS-OLS for both low and high leverage points, indicating that, unlike OLS, these robust estimators are not sensitive to outliers in both Y- and X-directions.

Most importantly, the Monte Carlo experiment showed that the most substantial gains in efficiency of the LTS-OLS algorithm can be expected with large *n* and substantial number of planted outliers. For example, for $n = 60$ and $s = 60 - [(60 - 9)/2] = 25$ regression outliers the MSE of LTS was 30.8, while for LTS-OLS it was just 4.9. Similarly, for $n = 100$ and $s = 100 - [(100 - 9)/2] = 45$, regression outliers the MSE of LTS increased to 33.0, while for LTS-OLS the MSE, in fact, dropped to 2.4.

Therefore, the combined use of LTS and OLS estimators appears to provide both robust and fast solution for setting apart the regular part of analyzed data. As the last row of table 2 shows, on the whole 2,500 repetitions of the LTS-OLS algorithm took about 42 minutes (for

$n = 100$) on a standard personal computer, or slightly more than 1 second per repetition of LTS-OLS algorithm.

Once LTS-OLS algorithm differentiates the initial basic subset B_1, the recursive ordering of remaining observations proceeds as follows.

TEST ALGORITHM (PART 2)

Step 6. For $r \in B_1$, calculate $\hat{\boldsymbol{\beta}}_r = (X'_r X_r)^{-1} X'_r Y_r$ and $\hat{s}_r = SSR_r / (n_r - k)$.

Step 7. For $d \notin B_1$, calculate standardized prediction residuals $w_d = (y_d - x_d \hat{\boldsymbol{\beta}}_r) \Big/ \sqrt{1 + x'_d (X'_r X_r)^{-1} x_d}$ and studentized prediction residuals $t_d = w_d / \hat{s}_r$.

Step 8. For $d \notin B_1$, sort observations by absolute values of w_d and t_t, and store the smallest ones among $d \notin B_1$ as $w_{(h+1)}$ and $t_{(h+1)}$.

Step 9. Extend subset B_1 with an observation, for which w_d or t_t are the smallest. Define the augmented subset as B_2, and go to step 6.

Step 10. Continue until the estimation subset contains $n-1$ observations.

Step 11. Calculate $w_{(n)}$ and $t_{(n)}$ for the n^{th} observation, and stop.

B. GOODNESS-OF-FIT TEST, BASED ON STUDENTIZED PREDICTION RESIDUALS

The test statistic is calculated from the sequence of studentized prediction residuals $t_{(h+1)}$, $t_{(h+2)}, \ldots, t_{(n-1)}, t_{(n)}$. Under the null hypothesis $\varepsilon \sim N(0, \sigma^2)$, the sequence of studentized prediction residuals approximately follows the Student's t-distribution with $h-k, h-k+1, \ldots, n-k-2, n-k-1$ degrees of freedom (Cook, Weisberg, 1982). In order to achieve the highest power, the test searches for a particular studentized prediction residual that violates the null hypothesis most significantly.

Note that the null distribution in the sequence of test statistics $t_{(h+1)}, t_{(h+2)}, \ldots, t_{(n-1)}, t_{(n)}$ depends on the *varying* degrees of freedom, which precludes their direct comparison. To remove the impact of varying degrees of freedom, one solution is to calculate two-tail probabilities of

absolute test statistics $|t_{(h+1)}|$, $|t_{(h+2)}|$,…, $|t_{(n-1)}|$, $|t_{(n)}|$, and then identify the most substantial break in the series by the smallest two-tail probability.

Another alternative, which is better suited for tabulating critical values of the test, is to convert each studentized prediction residual t_i into a normal deviate *with an equivalent two-tail probability*. Then such normalized statistics can be compared directly. Hawkins (1991, p. 223) recommended to use normalizing transformation, developed by Wallace (1959, p.1125):

$$z^{*} = \frac{8v+1}{8v+3}\sqrt{v\log_e(1+t^2/v)} \tag{2}$$

where t denotes studentized residual with corresponding v degrees of freedom.

Another transformation, also due to Wallace (*ibid.*) is more precise for small degrees of freedom:

$$\begin{aligned} z^{**} &= \left[1-\frac{2}{8v+3}\sqrt{1-e^{-s^2}}\right]\{v\log_e(1+t^2/v)\}\sqrt{v\log_e(1+t^2/v)} \\ s &= \frac{0.184(8v+3)}{v}\{\log_e(1+t^2/v)\}^{-1/2} \end{aligned} \tag{3}$$

Finally, the following normalizing transformation from the ACM algorithm 395 (Hill, 1970) is especially accurate. For tail probabilities as small as 10^{-11}, the transformation has absolute error of about 0.0001 for $v \cong 10$. Then it rapidly decreases to less than 0.000001 for $v \geq 100$:

$$\begin{aligned} z^{***} &= w + \frac{(w^3+3w)}{b} - \frac{4w^7+33w^5+240w^3+855w}{10b(b+0.8w^4+100)} \\ b &= 48(v-0.5)^2 \qquad w = \sqrt{(v-0.5)\log_e(1+t^2/v)} \end{aligned} \tag{4}$$

Using any of these transformations, the sequence of $t_{(h+1)}$, $t_{(h+2)}$,…, $t_{(n-1)}$, $t_{(n)}$ is transformed into the sequence of normalized test statistics $z_{(h+1)}$, $z_{(h+2)}$,…, $z_{(n-1)}$, $z_{(n)}$, and the final test statistic for the first goodness-of-fit test is

$$\tilde{z} = \sup|z_i| \tag{5}$$

for $i = h+1, h+2, ..., n-1, n$.

How the statistical significance of $\tilde{\chi}$ statistic (5) can be evaluated? Though repeated data permutations by the size of various residuals increase the test power, they simultaneously introduce complicated interdependencies in the data. As a result, the analytical distribution of $\tilde{z}$ becomes intractable. Yet, the distribution of $\tilde{\chi}$ statistic under the null can be approximated by the Barnard's (1963) Monte Carlo inference, which is specifically designed for cases when the distribution of a test statistic is not known (Gentle, 1998, p. 141-142). The method proceeds as follows:

- Generate many subsets of artificial data according to the null hypothesis $H_0 : \varepsilon \sim \mathrm{N}(0, \sigma^2)$, where σ^2 does not have to be explicitly specified. This is because studentized prediction residuals belong to the class of pivotal statistics, with distribution of $\tilde{\chi}$ independent of unknown parameters β and σ^2. Thus, without loss of generality, the null distribution of $\tilde{\chi}$ can be obtained with arbitrary values of β and σ^2. For example, one can simply fix all unknown parameters β to zero and set $\sigma^2 = 1$, generating artificial data as $y_B = \varepsilon$, $\varepsilon \sim \mathrm{N}(0,1)$.
- The test statistic $\tilde{\chi}$ is calculated from $(n \times 1)$ vector y_B and $(n \times k)$ actual matrix X. The calculated bootstrap test statistic $\tilde{\chi}_B$ is stored. The procedure is repeated B times. Upon completing, all test statistics $\tilde{\chi}_B$ are sorted in absolute values.
- Count how many times the actual test statistic $\tilde{\chi}$ exceeds $\tilde{\chi}_B$, with approximate p-value equal to $\hat{p}(\tilde{\chi}) = \dfrac{1}{B+1} \sum_{s=1}^{B} I\left(\tilde{\chi}_s > \tilde{\chi}\right)$, where $I(\cdot)$ is the indicator function.

Under the mild regularity conditions, it can be shown that as $B \to \infty$, the estimated p-value $\hat{p}(\tilde{\chi})$ will tend to the true p-value $p(\tilde{\chi})$ (Horowitz, 1997). Moreover, the Monte Carlo approximation for pivotal and two-sided test statistics (like $\tilde{\chi}$) makes error of order $O(n^{-2})$. In contrast, the traditional asymptotic approximations make errors of size $O(n^{-1})$, thus supporting

the advantage of using the Bernard's procedure to approximate the p-value of highly non-standard (but pivotal) $\tilde{\tau}$.

2C. GOODNESS-OF-FIT TEST BASED ON RECURSIVE RESIDUALS.

The second test essentially applies the Shapiro-Wilk and Shapiro-Francia test statistics to recursive residuals w_i, but takes into account the assumed zero mean of each w_i.

After the endogenous sorting of all *n* observations is complete, recursive residuals w_i are calculated for $i = k+1, k+2, ..., n-1, n$ by (1). Let $\vec{w}' = (w_{(1)}, w_{(2)}, ..., w_{(n-k)})$ be a vector of ordered $(n-k)$ recursive residuals, and let $x_{(1)} \le x_{(2)} \le ... \le x_{(n-k)}$ be an ordered sample from a standard normal distribution $N(0,1)$. Define the expected values of order statistics from $N(0,1)$ by vector $m' = (m_1, m_2, ..., m_{n-k})$.

A well known property of the normal distribution is that it is invariant under linear transformations. Let $\vec{w}$ be the ordered sequence of recursive residuals $w_{(1)}, w_{(2)}, ..., w_{(n-k)}$. If vector $\vec{w}$ comes from a normal distribution $N(\mu, \sigma^2)$, then $E(\vec{w}_{(i)}) = \mu + \sigma m_i$, where m_i is i^{th} element of vector *m*. Accordingly, the plot of $\vec{w}$ on the vector *m* should tend to be linear, while normalized deviations from this linear relationship can indicate to what extent the vector of recursive residuals approaches the hypothesized normal distribution.

The postulated zero mean of regression disturbances implies $\mu = 0$, with the plot of $\vec{w}$ on *m* passing through the origin. Then the extent of linear relationship between $\vec{w}$ and *m* can be evaluated by the *uncentered* R^2 statistic after estimating regression specification $\vec{w}_{(i)} = \sigma m_i$ by the OLS estimator.

Since elements of vector $\vec{w} = w_{(1)}, w_{(2)}, ..., w_{(n-k)}$ have been sorted by OLS residuals, they are not independent. Thus, the OLS estimate of $\hat{\sigma}_{OLS} = (m'm)^{-1} m'\vec{w}$ will be unbiased, but inefficient. The inefficiency can be avoided by using the Aitken's GLS estimator, which yields the

fully efficient and unbiased estimate $\hat{\sigma}_{GLS} = (m' Q^{-1} m)^{-1} m' Q^{-1} \vec{w}$, where Q is the $(n-k)\times(n-k)$ variance-covariance matrix of the normal order statistic.

The GLS is most attractive from the efficiency standpoint. However, the GLS requires not only *m*, but also *Q*. Tietjen Kahaner and Beckman (1977) calculated exact values of Q by a highly cumbersome double numerical integration, and only for *n*=2(1)50. In contrast, exact values of m are less computationally-intensive, and were calculated by Harter (1961) for *n*=2(1)100(25)250(50)400. Fortunately, a number of approximations for m and Q have been proposed in the literature.

Approximations of m

- Royston (1982) developed two algorithms (NSCOR1 and NSCOR2) which currently are the most precise. NSCOR1 calculates m in the same way as was done by Harter (1961), with absolute error as small as 0.0000001 for $2 \le n \le 2000$. NSCOR2 is a rational approximation for *m*. This algorithm does not require numerical integration. As a result, NSCOR2 is significantly faster than NSCOR1, with accuracy no worse than 0.0001 for $2 \le n \le 2000$. I found that it took NSCOR1 about 9 minutes to calculate matrix m with $n = 500$, while NSCOR2 did this in just 3 seconds. The largest absolute difference between the alternative estimates of m was only 0.00004.
- Harter (1961, p. 155) suggested to use $\alpha_{1,n}$ to approximate m_1, m_n, and $\alpha_{(1,n)}$ for the rest of expected order statistics, with $\alpha_{1,n} = 0.315065 + 0.057974\,g - 0.009776\,g^2$ for $i = 1, n$ and $\alpha_{(1,n)} = 0.327511 + 0.058212\,g - 0.007909\,g^2$ for $i \neq 1, n$ (with $g = log_{10} n$). The formula achieves accuracy of 0.002 for $2 \le n \le 400$.

Approximations of Q

- Algorithm of Davis and Stephens (1978) is currently the best available approximation with no limit on the dimension of Q.

- Gupta (1952) noted that if the variance-covariance matrix Q is replaced by the identity matrix, the loss of efficiency would be negligible (in fact, setting $Q = I$ yields the OLS estimator $\hat{\sigma}_{OLS} = (m'm)^{-1} m'\vec{w}$). Later it was found that inefficiency of $\hat{\sigma}_{OLS}$ not only diminishes asymptotically (Ali, Chan, 1964), but also already becomes negligible for small *n*. For example, Barnett (1976, p. 49) showed that for $n = 8$ the relative inefficiency of $\hat{\sigma}_{OLS}$ (compared with $\hat{\sigma}_{GLS}$) is 0.9989, and it gradually rises to 0.9991 for larger *n*. Therefore, the Gupta's estimator $\hat{\sigma}_{OLS}$ is not only convenient, but also sufficiently accurate for most practical purposes.

To evaluate the postulated linearity between vectors $\vec{w}$ and *m*, I will consider three goodness-of-fit statistics, from which I subsequently pick up the best one after studying their null distribution and relative power.

The first statistic is basically the uncentered version of the Shapiro-Wilk (1965) test statistic:

$$W_0 = \frac{(a'\vec{w})^2}{\vec{w}'\vec{w}} \tag{6}$$

$$a' = m'Q^{-1} \Big/ \sqrt{m'Q^{-1}Q^{-1}m} \tag{6'}$$

The second test statistic results from the application of Gupta's estimator to (6'), with $Q = I$. This produces $b' = m'[m'm]^{-1/2}$, producing the uncentered version of Shapiro-Francia (1972) statistic:

$$W_0' = \frac{(b'\vec{w})^2}{\vec{w}'\vec{w}} = \frac{(m'\vec{w})^2 (m'm)^{-1}}{\vec{w}'\vec{w}} = \frac{(m'\vec{w})^2}{(m'm)(\vec{w}'\vec{w})} \tag{7}$$

Clearly, W_0' are uncentered correlation statistic $R^2(\vec{w}, m)$, or squared correlation coefficient between $\vec{w}$ and *m*. Consequently, (7) can also be alternatively calculated as

$$W_0' = 1 - \frac{\sum (\vec{w}_i - m_i \hat{\sigma}_{OLS})^2}{\sum \vec{w}_i^2} \tag{8}$$

Finally, the third test statistic results from using the Aitken estimator $\hat{\sigma}_{GLS}$ in (8) instead of

slightly less efficient $\hat{\sigma}_{OLS}$:

$$W_0'' = 1 - \frac{\sum(\vec{w}_i - m_i\hat{\sigma}_{GLS})^2}{\sum \vec{w}_i^2} \tag{9}$$

Test statistics (6)-(9) are, like $\tilde{z}$ statistic in (5), scale and origin invariant. Therefore, the Barnard's Monte Carlo inference can be used to calculate the null distribution of these test statistics.

3. APPROXIMATE DISTRIBUTION OF GOODNESS-OF-FIT STATISTICS UNDER THE NULL.

The null distribution of each test statistics (6)-(9) is dependent on a particular configuration of matrix *X*, the sample size *n* and the number of independent variables *k*. Similarly to the null distribution of Durbin-Watson statistic, this precludes the unique tabulation of exact percentage points that can be applied to any regression model. Though in this section I do report selected quintiles for test statistics, they should be considered only as rough benchmarks. The application of Barnard's procedure to actual model at hand should be always a preferred approach.

Principally, the percentage points in Tables 3 and 4 are presented to compare the sensitivity of test statistics to alternative ways of their computation. First, I investigated how different normalization to the $\tilde{z}$ statistic (such as z^*, z^{**} and z^{***}) affect quintiles of $\tilde{z}$. Second, I examined differences in calculated values of W_0' and W_0'' statistics that apply GLS and OLS estimates of $\hat{\sigma}$, respectively[10].

The percentage points were simulated for $n = 20(20)100$ and $k = 4$, with 2000 replications. Using econometric software TSP (version 4.5), I generated matrix *X* from the uniform distribution $U(0,1)$, while regression disturbances were generated as standard normal deviates $N(0,1)$.

---

[10] In the univariate case Weisberg (1974) detected only minor differences in percentage points of related Shapiro-Wilk and Shapiro-Francia statistics. If the same close agreement occurs with respect to estimated regression

Table 3 reports percentage points of $\tilde{z}^{***}$, and its difference with arguably less accurate normalizing transformation $\tilde{z}^{*}$. The table shows that the two versions of $\tilde{z}$ statistic turned out fairly close even for $n = 20$, with the maximum absolute difference just 0.0013. The discrepancy further disappears with increasing sample size. For $n = 100$ the difference does not exceed 0.0001.

Table 4 contains percentage points for W_0, W_0' and absolute difference between W_0' and W_0'' statistics. The following features are noteworthy:

1. For a specific quintile, each test statistic approaches unity as the sample size *n* increases. This property is typical for the original univariate Shapiro-Wilk and Shapiro-Francia tests.
2. The maximal difference between related W_0' and W_0'' test statistics is very insignificant even for $n = 20$ (0.00014), and drops rapidly to 0.00003 for $n = 40$, and to 0.00002 for $n = 100$. Thus, there seems to be little justification for the extra effort in calculating the GLS-based W_0'' statistic.
3. Critical values for W_0, W_0' and W_0'' statistics are much smaller compared with corresponding univariate tests for normality for the same sample size *n*. For example, with $n = 20$, 5%, the critical values of W_0 and Shapiro-Wilk statistics are 0.576 and 0.983, respectively.

4. ILLUSTRATIVE EXAMPLE

To illustrate the calculation of suggested test statistics, I will use a regression model and data from Mankiw *et al.* (1992, table II, p. 420). Dependent variable was GDP per working-age person in 1985, while independent variables included (*i*) investment/GDP ratio, (*ii*) the sum of growth rates of labor, technology, and depreciation rate, (*iii*) the percentage of working-age population

---

residuals, then there is hardly any gain in calculating a more computer-intensive test statistic W_0''

in secondary school. All variables were in logs[11]. Though Mankiw *et al.* estimated the specification with three international cross-sections (with 98, 75, and 22 countries), I will consider the smallest cross-section with OECD countries. Table 5 tracks major steps in the test algorithm.

To find the global LTS minimum, the elemental set algorithm requires evaluation of $\binom{22}{4} = 7{,}315$ subsets, which is feasible. After finding the exact LTS minimum, all n observations were sorted by the absolute values of their LTS residuals. After selecting subset B_1 with $h = [(22+4+1)]/2 = 13$ smallest LTS residuals, OLS fit was applied to these 13 observations. Then observations in B_1 were resorted by absolute values of their OLS residuals. Though rankings by LTS and OLS residuals are quite similar, Australian moves from rank 6 (when ranked by LTS residuals) to rank 1 (when ranked by OLS residuals in Steps 4-5). Japan also changed its rank from 8 (with LTS) to 7 (with OLS).

Once the first subset B_1 with h regular observations is formed, the recursive extension of estimation subset begins. During the first recursion, the least deviant observation among $(n-h) = 9$ potential outliers is Spain, with $w_{(h)} = -0.2347$, and $|t_{(h)}| = 3.3008$.

After storing these test statistics, Spain is included in the estimation subset B_1, OLS parameters are calculated with $h+1$ observations, and w_i and $|t_i|$ are calculated for the remaining 8 potential outliers. During the second recursion, Italy has the smallest absolute value of standardized (and studentized) test statistics. These test statistics ($w_{(h+1)} = -0.2403$ and $|t_{(h+1)}| = 2.3949$) are stored, Italy is included in the estimation subset B_2, and the algorithm seeks the next least deviant observation. Eventually, the estimation subset contains 21 observations, with Turkey being the last observation in the endogenous ranking of countries.

[11] The paper contains the original data used in regression analysis, and I was able to replicate exactly reported results.

At the next step test statistics are calculated. First, consider the calculation of $\tilde{\tilde{z}} = \sup|z_i|$ statistic. The first studentized prediction residual is $t_b = 3.3008$. Under the null, the statistic approximately follows Student's t-distribution with $v = (b+1) - 4 - 1 = 9$ degrees of freedom. The two-tail probability for t_b is 0.009218, which corresponds to the standard normal deviate 2.6039. Likewise, the next studentized prediction residual has two-tail probability 0.037643, which corresponds to the standard normal deviate 2.0787, and so on.

After completing the normalizing transformation, the largest among them yields $\tilde{\tilde{z}} = \sup|z_i| = 2.7221$. The test statistic separates Greece, as well as more extreme Portugal and Turkey, from the rest of the sample. These 3 countries form the subset of most probable outliers.

What is the probability that these deviant observations does not correspond to the postulated data-generating process of the linear regression model? To answer the question, the Barnard's Monte Carlo test is used, using the actual matrix X, and regression disturbances generated under the null $\varepsilon \sim N(0, \sigma^2)$. Without loss of generality, one can set $\sigma^2 = 1$ and $\beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

After calculating the test statistic $\tilde{\tilde{z}}$ with artificial data $B = 999$ times, approximate p-value of $\tilde{\tilde{z}}$ was $\hat{p}(2.7221) = \frac{1}{1000}\sum_{s=1}^{B} I(Z_s > Z) = 0.319$. Thus, under the null hypothesis of $\varepsilon \sim N(0, \sigma^2)$, it is quite likely that Greece, Portugal, and Turkey correspond to the postulated data-generating process[12].

Before calculating the second set of test statistics W_0, W_0', and W_0'', it is instructive to visually assess the linearity between sorted $w_{(i)}$ and m_i (see chart 1). The linear pattern is especially distorted by three negative observations (which, in fact, represent Greece, Portugal,

and Turkey, already identified by $\tilde{z}$ statistic), and by the largest positive $w_{(i)}$ which refers to Canada.

Using data from table 5, W_0, W_0', and W_0'' are calculated as follows:

$$W_0 = (a'\vec{w})^2 / \vec{w}'\vec{w} = 1.62999/1.96025 = 0.83152,$$

$$W_0' = 1 - \sum(\vec{w}_{(i)} - m_i\hat{\sigma}_{OLS})^2 / \sum \vec{w}_{(i)}^2 = 1 - 0.30941/1.96025 = 0.84216,$$

or alternatively,

$$W_0' = \left\{m_i'\vec{w}_{(i)} / \sqrt{m_i'm_i}\sqrt{\vec{w}_{(i)}'\vec{w}_{(i)}}\right\}^2 = \left(5.0929/\sqrt{15.7115}\sqrt{1.96025}\right)^2 = 0.84216,$$

$$W_0'' = 1 - \sum(\vec{w}_{(i)} - m_i\hat{\sigma}_{GLS})^2 / \sum \vec{w}_{(i)}^2 = 1 - 30945/1.96025 = 0.84214,$$

using the Aitken estimator $\hat{\sigma}_{GLS} = 0.32262$ (compared with $\hat{\sigma}_{OLS} = 0.32415$).

Approximate p-values for these W_0, W_0', W_0'' statistics were 0.320, 0.359, and 0.359 respectively. As before, I used 999 test runs with data, simulated under the null hypothesis[13].

Even though the difference between W_0' and W_0'' should be the most noticeable in small samples, note that the use of the more efficient Aitken estimator makes little difference between W_0' and W_0'', with essentially the same test statistics and p-values. Due to such little differences between W_0' and W_0'', I will hereafter concentrate on more dissimilar W_0 and W_0' statistics.

5. THE POWER OF GOODNESS-OF-FIT STATISTICS

How different is the power of suggested goodness-of-fit tests compared with available misspecification tests? To save space, I will consider $\tilde{z}$ statistic, calculated with the Hill's normalizing transformation (4), and also W_0 and W_0' statistics. Their power is compared with the following groups of standard specification tests.

---

[12] The p-value may also be checked in Table 3. For $n = 20$, the test statistic $\tilde{z} = 2.72$ corresponds to 0.300 percentage point.

[13] On the other hand, from Table 4 percentage points for W_0, W_0', W_0'' statistics were approximately 0.350, 0.380,

GENERAL MISSPECIFICATION TESTS AGAINST UNSPECIFIED ALTERNATIVES

1. RESET test of Ramsey (1969);
2. Durbin-Watson test;
3. ANOVA test for structural change of Chow (1960);

TESTS, BASED ON RECURSIVE RESIDUALS

4. CUSUM (Brown *et al.*, 1975);
5. CUSUM of Squares (*ibid.*);
6. ψ-test (Harvey and Collier, 1977);

TESTS FOR REGRESSION OUTLIERS

7. studentized residual test (Cook and Weisberg, 1982);
8. recursive test for multiple regression outliers (Hadi and Simonoff, 1993);

TESTS FOR HETEROSKEDASTICITY

9. White's test to detect heteroskedasticity of unknown form;
10. Godfrey, Breusch, and Pagan's test to detect heteroskedasticity of known form;

TESTS FOR UNIVARIATE NORMALITY

11. Jarque-Bera test;
12. Shapiro-Wilk test.

Unless otherwise indicated, matrix X contained the intercept, and logs of labor and capital inputs from table 7.1 in Green (1997, p. 345), with $n = 27$. Data-generating process was $Y_i = \varepsilon_i$, with $\varepsilon_i \sim \mathrm{N}(0,1)$.

I examined the following violations in the full ideal conditions:

GROUP 1. REGRESSION OUTLIERS WITH NON-ZERO MEAN OF REGRESSION DISTURBANCES

1-1. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i = 1$

1-2. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i = 1,...,5$

1-3. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i = 1,...,10$

---

and 0.380.

1-4. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i = 1,...,13$ [14]

1-5. One high leverage outlier $\varepsilon_i \sim \mathrm{N}(7,1)$, $i = 1$; $n = 30$, $k = 3$. X_1 and X_2 are distributed as U(0,15); but $X_1(1) = 20$, $X_2(1) = 20$

1-6. Five high leverage outliers. The same as 1-5, but with $i = 1,...,5$

GROUP 2. HETEROSKEDASTIC REGRESSION DISTURBANCES.

2-1. $\varepsilon_i \sim \mathrm{N}(0,10)$, $i = 1,...,5$

2-2. $\varepsilon_i \sim \mathrm{N}(0,10)$, $i = 1,...,10$

2-3. $\varepsilon_i \sim \mathrm{N}(0,10)$, $i = 1,...,13$

2-4. $Y = \varepsilon\left(1 + 10X_1^4 + 10X_2^4\right)$ for all observations

GROUP 3. OMITTED VARIABLES AND NON-LINEARITY

3-1. $Y = 4X_2^2 + \varepsilon$ for all observations.

3-2. $Y = 1.4^{\varepsilon}$

GROUP 4. NON-NORMALITY OF REGRESSION DISTURBANCES

4-1. Cauchy distribution

4-2. Log-normal distribution

4-3. Exponential distribution

4-4. Laplace distribution

4-5. Uniform distribution.

The total number of replications in each Monte Carlo experiment was 500. Some tests crucially depend on the ordering of observations. Since these Monte Carlo experiments mimic cross-sectional data, with no 'natural ordering' of data, I selected an ordering by values of OLS predicted values, which is often applied in practice.

The nominal level for individual tests was set at 5 per cent. I also studied the joint application of $\tilde{\tilde{\chi}}$ statistic together with either W_o or W_0' statistics. To keep the joint significant level for these two pairs of tests no larger than 5 per cent, the Bonferroni inequality was used[15].

Some of considered tests statistics had only asymptotic justification, which often resulted in actual size of test statistic under the null substantially less than nominal size of 5 per cent.

---

[14] Note that for n=27 and $k = 3$, $s = n - h = 13$.

The discrepancy was the most substantial with the Jarque-Bera test for normality, which is valid only asymptotically. To eliminate this bias in reported test power, I calculated for each test its exact critical values under the null. These finite-sample adjustments were used in all Monte Carlo experiments.

Table 6 contains major results of the Monte Carlo study. Initially, I will consider the power of individual $\tilde{z}$, W_0, and W_0' tests. Then I will proceed to the power of using $\tilde{z}$ - W_0 and $\tilde{z}$ - W_0' tests *jointly*.

The power of $\tilde{z}$ statistic is relatively high with one, five, and ten mean-shift regression outliers (experiments 1-1, 1-2, 1-3). The test power remains high in the following cases: (*i*) outliers, planted at both low-leverage and high-leverage points; (*ii*) abrupt shifts in the variance of ε; (*iii*) omitted variables; (*iv*) incorrect functional form; (*v*) non-normal distributions with long tails.

Compared with other tests, $\tilde{z}$ statistic demonstrated a wide-ranging sensitivity to misspecification errors, with just two cases of low power: first, with as many as 13 discordant observations (experiment 1-4), and when the distribution of ε had short tails (such as the uniform distribution in experiment 4-5).

The power of W_0 and W_0' test statistics was very similar. Both tests are especially sensitive to large number of outliers with non-zero mean, and to non-normal distributions with short tail.

However, a few times the power of W_0' was substantially higher (compare 0.266 of W_0 versus 0.340 of W_0' in experiment 1-2). W_0' also had advantage over W_0 in detecting the effect of omitted variables and non-linearity of regression function (experiments 3-1 and 3-2), among

---

[15] With individual significance levels for $\tilde{z}$ and W_o (or W_0') tests set to 2.5 per cent.

other cases, thus making W_0' as a preferable test statistic. W_0' also simpler to calculate, since it does not require the cumbersome calculation of the covariance-variance matrix $\underset{\sim}{Q}$.

There was an explicit complimentarity between $\tilde{\tilde{\chi}}$, on the one hand, and either W_0 or W_0' statistics, on the other hand. Not surprisingly, the joint use of $\tilde{\tilde{\chi}}$ - W_0 and $\tilde{\tilde{\chi}}$ - W_0' (denoted in table 6 as J-1, and J-2, respectively) detected basically the whole range of considered regression misspecifications. Compared with W_0, the higher power of W_0' became partially obscure in their joint application with $\tilde{z}$. Yet again, results for the combination of $\tilde{\tilde{\chi}}$ and W_0' were marginally better (as especially evident in experiment 3-1). Consequently, this pair of test statistics appears to be the best choice to detect a wide range of possible (and unspecified) specification errors, especially common to cross-sectional data.

In contrast, the power of most conventional specification tests was much less comprehensive. For example, the ostensibly "general" RESET test had non-trivial power only in cases when failures of full ideal conditions were related to the matrix X (experiments 1-5 and 1-6). However, when the correspondence was absent, RESET's power was consistently insignificant. Similarly disappointing power was shown by other ostensibly "general tests", including Durbin-Watson, ψ-test, and Chow structural stability test. The low power of these tests is due to their dependent on correct ordering of observations (such as the separation of data into two 'regimes', assumed to be knows in these tests. Though I used ordering by the predicted values, in many cases this ordering turned out unrelated to the true misspecification pattern, resulting in poor performance of conventional 'order-dependent' specification test.

The best power among standard tests was achieved by the Hadi-Simonoff test of multiple outliers. Though in general its power was close to the power of $\tilde{\tilde{\chi}}$ test, the superiority of the latter test was often substantial, especially in cases of 5 and 10 regression outliers. The case of 5 outliers, planted at high leverage points (experiment 1-6), shows the susceptibility of

Hadi-Simonoff test to the masking effect, especially compared with the power of $\tilde{\chi}$ test (0.150 and 0.674, respectively).

6. CONCLUSIONS

This paper introduced several tests to evaluate distributional assumptions of the classical regression model. In particular, the paper suggested using a high breakdown regression (LTS) to detect violations of the full ideal conditions, with a substantial increase in the power of the tests, particularly in the case of LTS-OLS algorithm. Another useful feature of the suggested tests is that they do not assume any specific alternative hypothesis. The paper found that similar omnibus test statistics included the Had-Simonoff test and Shapiro-Wilk normality test. However, they had a lower power than tests suggested in this paper, especially the joint use of $\tilde{\chi}$ and W_0' statistics. The omnibus property makes redundant the current practice of running the battery of various specification tests with often unknown joint significance level.

It remains unclear if the application of other high-breakdown estimators would lead to any further improvements in power. This seems to be a promising extension of the present work.

REFERENCES.

Ali, M. M and L. K. Chan (1964). "On Gupta's estimates of the parameters of the normal distribution," *Biometrica*, 51, 498-501.

Barnard, G. A. (1963). "Comment," *Journal of Royal Statistical Society, Series B,* 25, 294.

Barnett, V. (1976). "Convenient probability plotting positions for the normal distribution," *Applied Statistics*, 25, 47-50.

Brown, R. L., J. Durbin and J. M. Evans (1975). "Techniques for testing the constancy of regression relationships over time," *Journal of the Royal Statistical Society*, Series B, 37, 149-163.

Chow, G. C. (1960). "Tests of Equality between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 28, 591-605.

Cook, R. D. and S. Weisberg (1982). *Residuals and Influence in Regression.* New York: Chapman and Hall.

Davis, C. S. and M. A. Stephens (1978). "Algorithm AS128. Approximating the covariance matrix of normal order statistics," *Applied Statistics*, 27, 206-212.

Gentle, J. E. (1998). *Random Number Generation and Monte Carlo Methods.* New York: Springer-Verlag.

Greene, W. H. (1997). *Econometric Analysis*, 3rd edition, Prentice-Hall.

Gupta, A. K. (1952). "Estimation of the Mean and Standard Deviation of a Normal Population from a Censored Sample," *Biometrica*, 41, 296-301.

Hadi, A., and J. Simonoff (1993). "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, 88, 1264-1272.

Harter, H. L. (1961). "Expected Values of Normal Order Statistics," *Biometrica*, 48, 151-165.

Harvey, A. and P. Collier (1977). "Testing for Functional Misspecification in Regression Analysis," *Journal of Econometrics,* 6, 103–119.

Hawkins, D.M (1991). "Diagnostics for use with regression recursive residuals," *Technometrics*, 33, 221-234.

Hill, G. W. (1970). "Algorithm 395: Student t-Distribution," *Communications of ACM*, 13, 617-619.

Horowitz, J. L. (1997). "Bootstrap methods in econometrics: theory and numerical performance". In D.M. Kreps and K.F. Wallis, eds. *Advances in economics and econometrics: theory and application*s, Seventh World Congress, vol. 3. Cambridge, U.K.: Cambridge University Press.

L'Ecuyer, P. (1998). "Good Parameters and Implementations for Combined Multiple Recursive Random Number Generators," working paper, Department of Mathematics, University of Montreal, http://www.iro.umontreal.ca/~lecuyer/myftp/papers/combmrg2.ps

Mankiw. H. G., D. Romer and D. N. Weil (1992). "A contribution to the empirics of economic growth," *Quarterly Journal of Economics*, 107, 407-437.

Ramsey, J. B. (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis," *Journal of the Royal Statistical Society*, Series B, 31, 350-371

Rousseeuw, P. J. (1984). "Least Median of Squares Regression," *Journal of American Statistical Association,* 82, 851-857.

Rousseeuw, P. J. and A. M. Leroy (1987). *Robust Regressions and Outlier Detection.* New York: John Wiley.

Rousseeuw, P. J. and M. Hubert (1997). "Recent Developments in PROGRESS," manuscript, University of Antwerp , downloadable at http://win-www.uia.ac.be/u/statis/publications.html.

Rousseeuw, P. J. and K. Van Driessen (1999). "Computing LTS Regression for Large Data Sets," Technical Report, University of Antwerp, downloadable at http://win-www.uia.ac.be/u/statis/ /publications.html.

Royston, J. P. (1982). "Algorithm AS177, Expected Normal Order Statistics (Exact and Approximate)," *Applied Statistics*, 31, 161-165.

Shapiro, S. S. and R. S. Francia (1972). "An Analysis of Variance Test for Normality," *Journal of the American Statistical Association*, 67, 215-216.

Shapiro, S. S. and M. B. Wilk (1965). "An Analysis of Variance Test for Normality (Complete Samples)," *Biometrica*, 52, 591-611.

Theil, H. (1971). *Principles of Econometrics*. New York: John Wiley.

Thursby, J. G. (1989). "A Comparison of Several Specification Error Tests for A General Alternative," *International Economic Review*, 30, 217-230.

Tietjen, G. L., D. K. Kahaner and R. J. Beckman (1977). "Variance and Covariance of the Normal Order Statistic for Sample Sized 2 to 5". In: *Selected Tables in Mathematical Statistics*. Providence: American Mathematical Society.

Wallace, D. L. (1959). "Bounds on Normal Approximations to Student's and the Chi-Distributions," *Annals of Mathematical Statistics*, 30, 1121-1130.

Weisberg, S. (1974). "An Empirical Comparison of the Cumulative Distributions for W and W'," *Biometrica*, 61, 644-646.

Weisberg, S. (1980). "Comment on a Paper by White and MacDonald," *Journal of American Statistical Association*, 75, 28-31.

Table 1. Number of elemental sets with various versions of PROGRESS code.

n	k	Required m to find the global LTS minimum	ϖ	Sufficient m for $\rho < 0.01$	Default m_d	ρ for default m_d
50	2	1,225	0.500	16	1000	1×10^{-125}
50	3	19,600	0.500	35	1500	1×10^{-87}
50	4	230,300	0.500	72	2000	9×10^{-57}
50	5	2,118,760	0.500	145	2500	3×10^{-35}
50	6	15,890,700	0.500	293	3000	3×10^{-21}
50	7	99,884,400	0.500	590	3000	6×10^{-11}
50	8	536,878,650	0.500	1180	3000	8×10^{-6}
50	9	2,505,433,700	0.500	2353	3000	3×10^{-3}

Table 2. Efficiency comparison of OLS, LTS, and LTS-OLS estimators with different extent of outlier contamination.

	Low leverage outliers			*High leverage outliers*		
Sample	20	60	100	20	60	100
Outliers			*OLS*			
None	13.2	2.4	1.2	13.2	2.4	1.2
One	12,876.5	478.9	154.5	7,423.5	3,572.4	1,038.2
h/2	29,516.7	2,157.5	1,912.8	32,289.2	24,500.5	18,150.8
H	11,842.8	5,030.6	2,517.7	19,006.9	19,784.2	16,476.2
h+1	19,146.1	4,942.6	3,149.8	32,377.1	18,942.8	16,289.2
Timing	0.3	0.4	0.4	0.4	0.4	0.3
			LTS			
None	50.1	13.1	8.0	50.1	13.0	8.0
One	54.9	13.0	8.1	54.9	13.0	8.1
h/2	52.7	14.5	9.9	52.7	14.5	9.9
H	35.9	30.8	33.0	35.9	30.8	33.0
h+1	92,441.3	437.4	130.0	41,942.9	21,857.3	1,852.7
Timing	12.0	24.0	41.1	13.2	27.1	37.3
			LTS-OLS			
None	50.0	12.0	7.0	50.0	11.9	7.0
One	54.1	12.0	7.2	54.1	12.0	7.2
h/2	51.0	11.4	6.6	51.0	11.4	6.6
H	23.5	4.9	2.4	23.5	4.9	2.4
h+1	93,636.7	1,114.9	336.1	40,460.7	21,263.1	2,436.4
Timing	13.1	25.0	42.1	14.3	28.1	38.3

*Note*s: the number independent variables k was 9 (including intercept); matrix X with low leverage points was generated from *U(0,15*) distribution with seed 100; matrix X with high leverage points was generated from N(10,10) distribution with seed 1000; regression disturbances were generated as standard normal deviates with seed 200; each experiment included 500 replications. "Timing" is the total time required for $500\times5 = 2{,}500$ replications in 5 experiments (in minutes).

Table 3. Selected quintiles of $\tilde{z}$ statistics.

	$\tilde{z}^{***} = \sup\left\|z_i^{***}\right\|$					$\left\|\tilde{z}^{***} - \tilde{z}^{*}\right\|$				
n	*20*	*40*	*60*	*80*	*100*	*20*	*40*	*60*	*80*	*100*
0.900	2.21	2.27	2.31	2.34	2.40	.0007	.0001	.0000	.0000	.0000
0.800	2.31	2.36	2.39	2.44	2.51	.0012	.0002	.0001	.0000	.0000
0.700	2.40	2.43	2.47	2.53	2.60	.0010	.0001	.0001	.0000	.0000
0.650	2.44	2.47	2.50	2.57	2.64	.0008	.0002	.0001	.0000	.0000
0.600	2.47	2.51	2.54	2.61	2.67	.0010	.0002	.0001	.0000	.0000
0.550	2.50	2.55	2.58	2.65	2.71	.0012	.0004	.0001	.0000	.0000
0.500	2.54	2.58	2.61	2.69	2.76	.0011	.0002	.0001	.0000	.0000
0.450	2.58	2.61	2.66	2.74	2.81	.0010	.0004	.0001	.0001	.0000
0.400	2.62	2.66	2.71	2.79	2.86	.0013	.0002	.0001	.0001	.0000
0.350	2.67	2.70	2.75	2.84	2.91	.0011	.0006	.0001	.0001	.0000
0.300	2.72	2.76	2.80	2.89	2.98	.0012	.0002	.0001	.0001	.0000
0.250	2.79	2.80	2.86	2.95	3.05	.0012	.0002	.0001	.0001	.0000
0.200	2.86	2.86	2.93	3.03	3.11	.0013	.0003	.0001	.0001	.0000
0.150	2.94	2.95	3.03	3.11	3.20	.0012	.0003	.0001	.0001	.0001
0.100	3.07	3.07	3.14	3.23	3.29	.0011	.0004	.0002	.0001	.0001
0.050	3.26	3.28	3.33	3.41	3.48	.0003	.0004	.0002	.0001	.0001
0.025	3.42	3.48	3.51	3.59	3.66	.0013	.0005	.0002	.0001	.0001
0.010	3.64	3.71	3.72	3.84	3.87	.0011	.0007	.0003	.0002	.0001
0.005	3.78	3.84	3.89	3.91	3.94	.0010	.0008	.0003	.0002	.0001

Note: test statistic $\tilde{z}^{***} = \sup\left\|z_i^{***}\right\|$ was calculated by (4); test statistic $\tilde{z}^{*} = \sup\left\|z_i^{*}\right\|$ was calculated by (2).

Table 4. Selected quintiles of W_0, W_0' and W_0'' tests.

	W_o					W_o'					$\lvert W_0' - W_o'' \rvert$				
n	*20*	*40*	*60*	*80*	*100*	*20*	*40*	*60*	*80*	*100*	*20*	*40*	*60*	*80*	*100*
0.900	0.951	0.967	0.976	0.981	0.984	0.952	0.969	0.977	0.982	0.985	.0001	.0000	.0000	.0000	.0000
0.800	0.933	0.957	0.968	0.975	0.979	0.934	0.959	0.970	0.976	0.980	.0000	.0000	.0000	.0000	.0000
0.700	0.916	0.948	0.962	0.969	0.974	0.916	0.949	0.963	0.970	0.975	.0000	.0000	.0000	.0000	.0000
0.650	0.906	0.943	0.958	0.966	0.971	0.907	0.944	0.959	0.967	0.972	.0001	.0000	.0000	.0000	.0000
0.600	0.897	0.936	0.954	0.963	0.968	0.898	0.938	0.954	0.963	0.968	.0001	.0000	.0000	.0000	.0000
0.550	0.888	0.929	0.949	0.958	0.965	0.887	0.930	0.950	0.959	0.965	.0000	.0000	.0000	.0000	.0000
0.500	0.877	0.921	0.944	0.954	0.961	0.875	0.923	0.945	0.955	0.961	.0001	.0000	.0000	.0000	.0000
0.450	0.863	0.912	0.939	0.948	0.957	0.859	0.914	0.939	0.948	0.957	.0001	.0000	.0000	.0000	.0000
0.400	0.846	0.903	0.931	0.943	0.952	0.847	0.902	0.933	0.944	0.953	.0000	.0000	.0000	.0000	.0000
0.350	0.832	0.893	0.924	0.936	0.945	0.832	0.892	0.925	0.936	0.946	.0000	.0000	.0000	.0000	.0000
0.300	0.812	0.880	0.915	0.928	0.939	0.811	0.881	0.915	0.929	0.939	.0000	.0000	.0000	.0000	.0000
0.250	0.785	0.863	0.904	0.918	0.931	0.785	0.863	0.903	0.919	0.931	.0000	.0000	.0000	.0000	.0000
0.200	0.751	0.845	0.888	0.908	0.921	0.749	0.845	0.888	0.909	0.920	.0001	.0000	.0000	.0000	.0000
0.150	0.717	0.818	0.868	0.893	0.908	0.719	0.817	0.868	0.893	0.907	.0001	.0000	.0000	.0000	.0000
0.100	0.662	0.776	0.846	0.871	0.892	0.666	0.778	0.845	0.871	0.891	.0001	.0000	.0000	.0000	.0000
0.050	0.576	0.716	0.797	0.833	0.860	0.580	0.717	0.797	0.833	0.859	.0000	.0000	.0000	.0000	.0000
0.025	0.518	0.663	0.752	0.797	0.828	0.525	0.660	0.751	0.797	0.828	.0000	.0000	.0000	.0000	.0000
0.010	0.482	0.602	0.697	0.748	0.774	0.492	0.606	0.696	0.743	0.774	.0000	.0000	.0000	.0000	.0000
0.005	0.468	0.558	0.679	0.718	0.744	0.476	0.566	0.680	0.717	0.744	.0000	.0000	.0000	.0000	.0000

Note: test statistics W_0, W_0' and W_0'' were calculated by (6), (7)-(8), and (9), respectively.

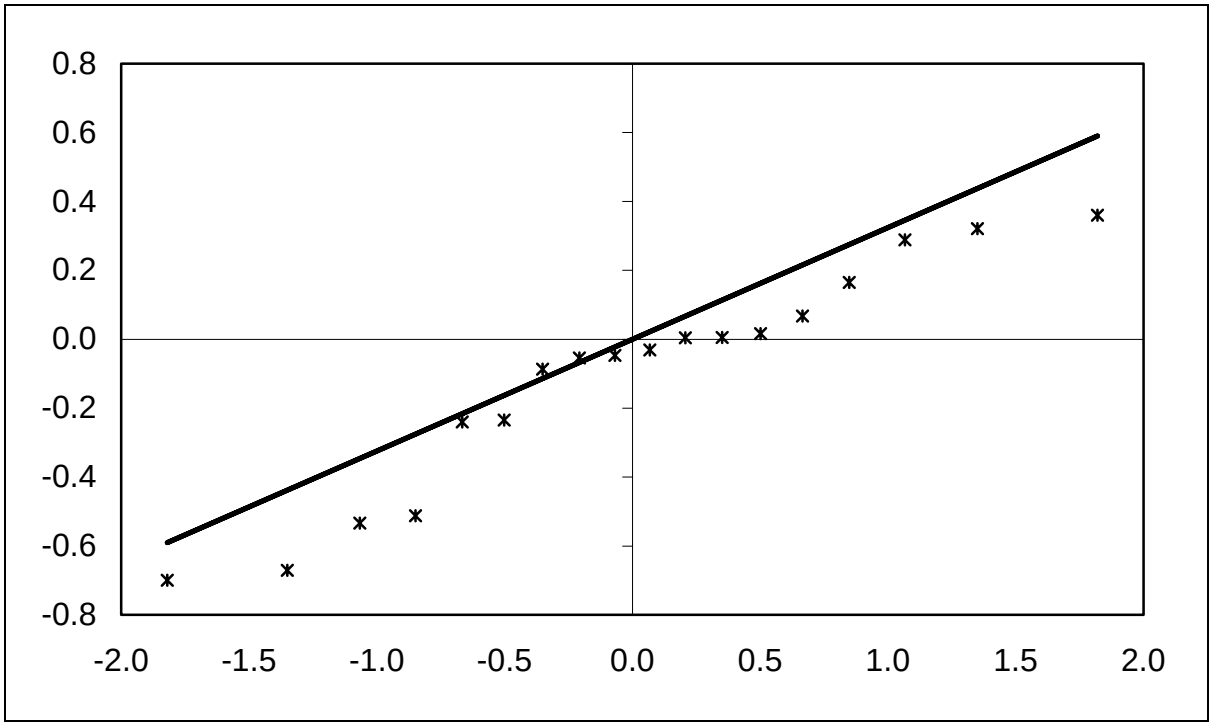
Reference: 5% critical values value for Shapiro-Wilk statistics - for *n=20:* 0.983; for *n=40:* 0.987.
5% critical values value for Shapiro-Francia statistics - for *n=80:* 0.994; for *n=100:* 0.995.

Table 5 . Illustrative example.

Countries	$\lvert\varepsilon_{LTS}\rvert$	Rank of $\lvert\varepsilon_{LTS}\rvert$	Unsorted w_i	Sorted $w_{(i)}$	Exact m_i	a_i	$\lvert t_i\rvert$	v	Two-tail prob. of $\lvert t_i\rvert$	z_i^{***}
1 Australia	0.013	6								
2 Belgium	0.000	1								
3 Switzerland	0.000	2								
4 Netherlands	0.000	3								
5 New Zealand	0.000	4	0.004798	-0.699774	-1.820032	-0.488522				
6 Germany	0.013	5	0.003823	-0.670700	-1.350414	-0.325853				
7 Japan	0.034	8	-0.030719	-0.533521	-1.065728	-0.254458				
8 Sweden	0.024	7	0.016617	-0.512288	-0.848125	-0.202592				
9 UK	0.044	9	-0.047092	-0.240255	-0.664795	-0.158851				
10 France	0.069	10	0.067700	-0.234764	-0.501582	-0.119878				
11 Finland	0.087	11	-0.053951	-0.086826	-0.350837	-0.083862				
12 Austria	0.099	12	-0.086826	-0.053951	-0.207735	-0.049660				
13 Denmark	0.144	13	0.164375	-0.047092	-0.068803	-0.016448				
14 Spain	0.278	14	-0.234764	-0.030719	0.068803	0.016448	3.300786	9	0.009218	2.603849
15 Italy	0.292	15	-0.240255	0.003823	0.207735	0.049660	2.394887	10	0.037643	2.078721
16 Norway	0.305	16	0.288183	0.004798	0.350837	0.083862	2.401803	11	0.035120	2.106968
17 Canada	0.379	17	0.359943	0.016617	0.501582	0.119878	2.537724	12	0.026047	2.225510
18 USA	0.431	18	0.320430	0.067700	0.664795	0.158851	1.896853	13	0.080286	1.749031
19 Ireland	0.437	19	-0.512288	0.164375	0.848125	0.202592	2.785164	14	0.014600	2.442144
20 Greece	0.782	20	-0.699774	0.288183	1.065728	0.254458	3.158925	15	0.006487	2.722100
21 Portugal	0.939	21	-0.670700	0.320430	1.350414	0.325853	2.423169	16	0.027614	2.202721
22 Turkey	1.031	22	-0.533521	0.359943	1.820032	0.488522	1.699376	17	0.107472	1.609662

Source: Mankiw *et al.* (1992), specification in the last column of Table II, p. 420.

Chart 1. Illustrative example: plot of ordered recursive residuals $\vec{w}_i$ on the expected values of order statistic m_i.



Step 6. Monte Carlo results

	$\tilde{\tilde{\tau}}$	W_0	W_0'	*J-1*	*J-2*	*RESET*	*DW*	*Chow*	*Cusum*	*Cusum squares*	*Ψ-test*	*STUD*	*HS*	*WHITE*	*GBP*	*SW*	*JB*
1.1. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i=1$	.992	.106	.216	.988	.988	.016	.014	.074	.062	.488	.030	.996	.992	.976	.116	.808	.942
1.2. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i=1,...,5$	.930	.266	.340	.916	.916	.000	.008	.000	.000	.030	.000	.014	.890	.150	.018	.134	.078
1.3. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i=1,...,10$	.854	.988	.992	.944	.940	.000	.012	.002	.000	.000	.000	.000	.806	.040	.022	.898	.000
1.4. $\varepsilon_i \sim \mathrm{N}(7,1)$, $i=1,...,13$	.010	.480	.488	.446	.446	.000	.014	.002	.056	.000	.002	.000	.104	.008	.002	.692	.000
1.5. High leverage, $\varepsilon_i \sim \mathrm{N}(7,1)$, $i=1$	.920	.114	.168	.892	.892	.976	.368	.732	.008	.930	.036	.934	.968	.972	.982	.288	.546
1.6. High leverage $\varepsilon_i \sim \mathrm{N}(7,1)$, $i=1,...5$	.674	.148	.144	.602	.604	1.000	.950	.982	.088	.982	.382	.040	.150	.098	.002	.036	.032
2.1. $\varepsilon_i \sim \mathrm{N}(0,10)$, $i=1,...,5$	.782	.056	.110	.708	.708	.080	.058	.034	.254	.728	.174	.660	.786	.148	.040	.634	.734
2.2. $\varepsilon_i \sim \mathrm{N}(0,10)$, $i=1,...10$	.556	.018	.028	.444	.444	.028	.034	.036	.132	.256	.084	.390	.566	.024	.006	.566	.528
2.3. $\varepsilon_i \sim \mathrm{N}(0,10)$, $i=1,...,13$	.370	.004	.014	.268	.268	.034	.042	.036	.116	.248	.058	.288	.388	.040	.002	.352	.344
2.4. $Y=\varepsilon\left(1+10X_1^4+10X_2^4\right)$	.422	.050	.076	.364	.366	.062	.096	.058	.152	.548	.114	.390	.422	.372	.588	.208	.306
3.1. $Y=4X_2^2+\varepsilon$	.368	.614	.734	.460	.482	1.000	1.000	1.000	1.000	1.000	1.000	.292	.340	1.000	.888	.126	.108
3.2. $Y=1.4^{\varepsilon}$	.624	.292	.372	.600	.604	.498	.270	.220	.006	.910	.038	.516	.610	.456	.682	.374	.462
4.1. Cauchy distribution	.946	.544	.628	.916	.916	.090	.078	.044	.108	.772	.070	.892	.948	.104	.088	.914	.922
4.2. Log-Normal distribution	.910	.910	.932	.964	.966	.066	.058	.034	.036	.578	.022	.826	.906	.134	.076	.962	.942
4.3. Exponential distribution	.564	.514	.608	.660	.664	.040	.036	.038	.068	.266	.028	.478	.562	.094	.078	.766	.672
4.4. Laplace distribution	.320	.022	.038	.244	.246	.044	.044	.046	.056	.194	.038	.304	.324	.072	.048	.266	.334
4.5. Uniform distribution	.016	.186	.192	.158	.152	.026	.066	.048	.054	.022	.030	.002	.008	.040	.056	.144	.000
Summary: cases with power above 0.200	*15*	*8*	*9*	*16*	*16*	*4*	*4*	*4*	*2*	*12*	*2*	*12*	*14*	*5*	*4*	*13*	*11*

The following abbreviations are used in the table: DW – Durbin-Watson test; STUD – test of maximum studentized residual; HS – Hadi-Simonoff test; WHITE – White's test; GBP – Godfrey, Breusch, and Pagan's test; JB – Jarque-Bera test; SW – Shapiro-Wilk test.