Forward Search Added Variable t Tests and the Effect of Masked Outliers on Model Selection and Transformation

Anthony C. Atkinson*
The London School of Economics, London WC2A 2AE, UK and
Marco Riani†
Dipartimento di Economia, Università di Parma, Italy

February 12, 2002

Abstract

The "forward" search is a powerful general method for determining the impact of one or several observations on all aspects of a fitted model. However, monitoring the t tests for individual regression coefficients fails to identify the importance of observations to the significance of the individual regressors. We show that this failure is due to the ordering of the data by the search. We introduce an added variable test which has the desired properties since the projection leading to residuals destroys the effect of the ordering. An example illustrates the effect of several masked outliers on model selection.

A similar test for transformations does not have such good distributional properties. But the properties improve in the presence of "strong regression", when the projection destroys the correlation which is distorting the null distribution of the test statistic.

Keywords: Box-Cox transformation; projection; score statistic; t distribution; transformation to normality; very robust methods.

1 Introduction

The forward search is a powerful general method for detecting multiple masked outliers and for determining their effect on models fitted to data.

*e-mail: a.c.atkinson@lse.ac.uk
†e-mail: mriani@unipr.it

Here we develop a method for detecting the effect of outliers on the t tests for coefficients in a regression model and derive the distribution of the statistics. We show how our new procedure can aid model selection. We also discuss the distributional properties of a related test for transformations.

Atkinson and Riani (2000) describe the use of the forward search in linear and nonlinear regression, response transformation and in generalized linear models. One example of its use in multivariate analysis is Riani and Atkinson (2001), which is concerned with transformations and discriminant analysis. The method starts by fitting a small, robustly chosen, subset of the observations to the data. The subset is then repeatedly increased in size and a series of parameter estimates is obtained, from which, for example, residuals and t tests can be calculated. The observations are so chosen that outliers enter towards the end of the search. As a result, plots of parameter estimates and of residuals are stable until the outliers start to be included. The change in plots of estimates and residuals reveals the effect of the outliers. However, similar "forward" plots of the t statistics in regression are not informative, as the t tests start large and gradually decrease to their final values in the full sample (Atkinson and Riani 2000, p.72). In this paper we use added variable plots to exhibit the effect of outliers on the t statistics in the regression model. We show examples in which a few outliers can either make, or destroy the significance of a variable. Such information is of major importance where t tests are used in variable selection.

The forward search is described in more detail in the next section and an example given using data on the length of stay in hospital in which the plot of t statistics is impossible to interpret. In §3 the forward plot for added variable t statistics is defined and its properties derived. Theoretical results and simulations show that the statistics have the correct t distribution, independently of the ordering of the observations. Examples with outliers are in §4: the effect of the outliers is clearly revealed. The next section considers briefly the related plot for transformations using a constructed variable. Here the null distribution is no longer exactly t.

2 The Forward Search and Forward Plots

2.1 General Principles

For all n observations the standard regression model is written as

$$y = Q\theta + \epsilon \tag{1}$$

where Q is $n \times p$ and the errors ϵ satisfy the second-order assumptions with variances σ^2. We estimate the parameter θ by least squares from carefully

selected subsets of m observations

To start the forward search we take $m = p$ and sample a large number of subsets, here 1,000, to each of which a regression is fitted by least squares. We take as the starting subset that which yields the smallest median squared residual. We move forward in the search, incrementing m by one by taking as the next subset the observations with the $m+1$ smallest least squares residuals. We now give a concise description of the algorithm and its consequences for the analysis of data.

2.2 Step 1. Choice of the Initial Subset

Let $Z = (Q, y)$, so that Z is $n \times (p+1)$ and let $S^{(m)}_{i_1,\ldots,i_m} \equiv \{z_{i_1}, \ldots, z_{i_m}\}$, be a unique m-tuple where $z^T_{i_1}$ is the i_1-th row of Z, for $1 \leq i_1, \ldots, i_m \leq n$ and $i_j \neq i_{j'}$. Specifically, let $\boldsymbol{\iota}^T = [i_1, \ldots, i_m]$ and let $e_{i,S^{(m)}_{\boldsymbol{\iota}}}$ be the least squares residual for unit i given observations in $S^{(m)}_{\boldsymbol{\iota}}$. We take as our initial subset the p-tuple $S^{(p)}_*$ which satisfies

$$\tilde{e}^2_{[med],S^{(p)}_*} = min_{\boldsymbol{\iota}}[\tilde{e}^2_{[med],S^{(p)}_{\boldsymbol{\iota}}}], \tag{2}$$

where $\tilde{e}^2_{[l],S^{(p)}_{\boldsymbol{\iota}}}$ is the l-th ordered squared residual among $\tilde{e}^2_{i,S^{(p)}_{\boldsymbol{\iota}}}$, $i = 1, \ldots, n$,

$$\text{med} = [(n+p+1)/2], \tag{3}$$

and $[(n+p+1)/2]$ denotes the integer value of $(n+p+1)/2$. Criterion (2) provides a least median of squares method for regression models with independent errors (Rousseeuw 1984; Hawkins 1993). The breakdown point of this estimator is asymptotically 50%.

2.3 Step 2: Adding Observations During the Forward Search

Given a subset of dimension $m \geq p$, say $S^{(m)}_*$, the forward search moves to dimension $m+1$ by selecting the $m+1$ units with the smallest squared least squares residuals, the units being chosen by ordering all squared residuals $\tilde{e}^2_{i,S^{(m)}_*}$, $i = 1, \ldots, n$.

In most moves from m to $m+1$ just one new unit joins the subset. However, two or more units may join $S^{(m)}_*$ as one or more leave, usually as the search includes units which belong to a cluster of outliers. Step 2 of the forward search is repeated until all units are included in the subset.

2.4 Consequences of the Forward Search

The forward search estimator $\hat{\theta}_{FS}$ is the collection of least squares estimators in each step of the forward search, that is:

$$\hat{\theta}_{FS} = \left(\hat{\theta}^*_p, \ldots, \hat{\theta}^*_n\right), \tag{4}$$

with $\hat{\theta}^*_m$ the least squares estimator from subset $S^{(m)}_*$. In the absence of outliers and systematic departures from the model, the parameter estimates for each m are unbiased estimators of the same quantity. So both parameter estimates and residuals should remain stable and approximately constant during the forward search.

The method combines a highly robust procedure with least squares estimators. The zero breakdown point of least squares estimators renders $\hat{\theta}_{FS}$ sensitive to the introduction of atypical units into the subset used for fitting the model. In particular, if there are k outliers, the forward procedure will include these towards the end of the search, usually in the last k steps. Until then, residual plots and parameter estimates will remain approximately constant.

2.5 Monitoring the Search

Despite the stability of parameter estimates and residuals, the estimate of σ^2 does not remain constant during the forward search since observations are sequentially selected which have small residuals. Thus, even in the absence of outliers, the residual mean square estimate $s^2_{S^{(m)}_*} < s^2_{S^{(n)}_*} = s^2$ for $m < n$. A consequence of the decrease of $s^2_{S^{(m)}_*}$ with m is that the values of the t statistics for the parameters θ_j are initially very large, decreasing as m increases. The added variable tests of §3 avoid this problem.

2.6 Surgical Unit Data

Neter, Kutner, Nachtsheim, and Wasserman (1996, pp.334 & 438) analyse 108 observations on the time of survival of patients who had a particular kind of liver surgery. The four explanatory variables, $x_1 - x_4$, are: a blood clotting score; a prognostic index which includes the age of the patient; an enzyme function test score and a score for liver function. The response is survival time. For the moment we follow Neter et al. (1996) and use the log of time as the response. The properties of a test for this transformation are described in §5.

Variable	constant	x_1	x_2	x_3	x_4
t	12.45	14.49	25.24	27.31	0.114
$\hat{\beta}_j$	0.521	0.0690	0.00964	0.00917	0.000977

Table 1: Transformed Surgical Unit data: t tests and parameter estimates for all 108 observations

Table 1 gives the t tests and parameter estimates when all 108 observations are fitted. It seems clear that the constant and the first three explanatory variables are all highly significant, but that x_4 need not be included in the model. We now investigate how this conclusion depends on individual observations.

Figure 1 shows the result of the forward search starting with $m = 5$ from 1,000 starting points. The upper panel gives the forward plot of the estimated coefficients. Since, as Table 1 shows, these are of differing magnitude, they have been standardized by their values in the middle of the search. The three significant regression coefficients are extremely stable: only the non-significant $\hat{\beta}_4$ changes appreciably during the search.

This impression of stability is echoed in the lower panel of Figure 1 which shows the least squares residuals at each step of the forward search, standardised by the estimate of σ^2 at the end of the search. There is a stable, but asymmetrical, pattern at the beginning of the search with a few negative outliers, particularly observations 28 and 43. Further on in the search the positive residuals begin to become smaller as most residuals trend downwards. It is interesting to note the evidence of masking in these data: at the end of the search the two most negative outliers are no longer especially extreme: observations 28 and 43 provide the 5th and 6th most negative residuals.

A question is whether this structure has an effect on inferences about the model. The asymmetric distribution of errors in the earlier part of the search may indicate something wrong with the log transformation. We return to this question in §5. Here we look at the estimation of σ^2 and then at the t tests.

The upper panel of Figure 2 shows how the estimate of σ^2 behaves during the search. Since the search proceeds by adding observations which have small residuals, the estimate of σ^2 is initially very small. As the figure shows, in this case, the values initially grow slowly, but later on more quickly, in a roughly parabolic way. The smooth curve suggests that there are no important outliers, such as those indicated by the jumps in Figure 1.8 of Atkinson and Riani (2000). The effect of this sequence of estimates is clear in the

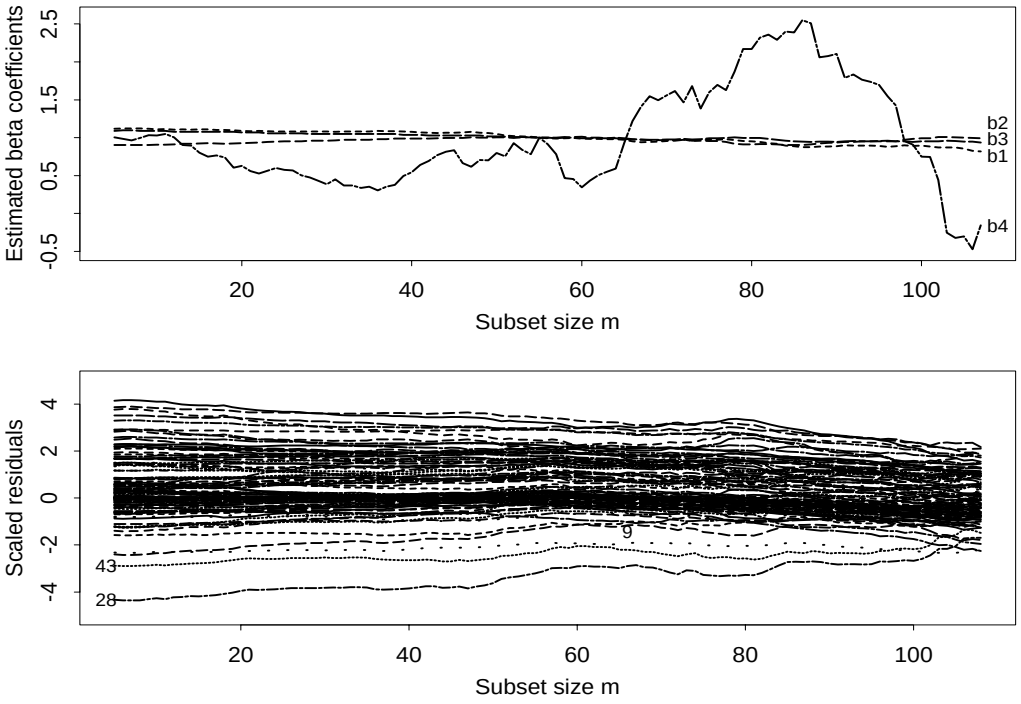

Figure 1: Transformed Surgical Unit data: upper panel, forward plot of parameter estimates and, lower panel, forward plot of scaled residuals

lower panel of Figure 2, which shows the behaviour of the t statistics for the four regression coefficients and the intercept. The decreasing values for all curves echo the increase in the values of s^2. Initially even the coefficient of x_4 is significant, although it soon decreases to only marginal significance. All four curves, of course, finish at the values in Table 1. But such a plot is quite uninformative about any outliers and their effect. We now use added variables to find an informative and easily interpreted plot of the t statistics.

3 An Added Variable t Test

3.1 Added Variables

In order to obtain useful forward plots of t tests we rewrite the regression model (1) as

$$y = Q\theta + \epsilon = X\beta + w\gamma + \epsilon, \tag{5}$$

where γ is a scalar. We in turn take each of the columns of Q as the vector w (except the column corresponding to the constant term in the model). The

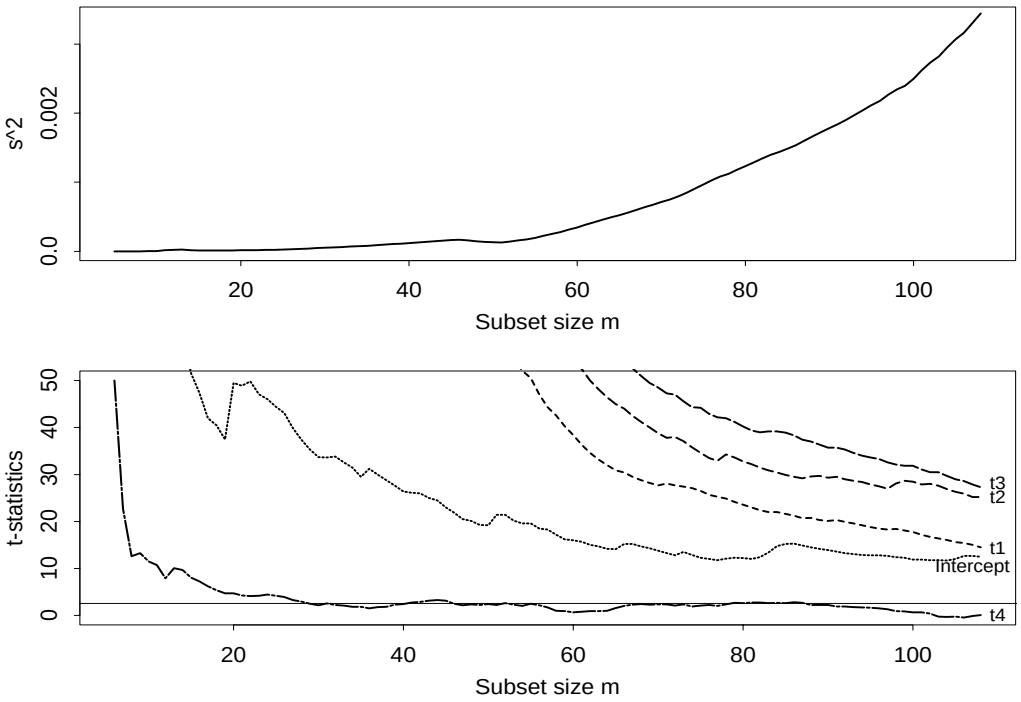

Figure 2: Transformed Surgical Unit data: upper panel, forward plot of s^2 and, lower panel, forward plot of t statistics

well-established approach of added variables - for example §2.2 of Atkinson and Riani (2000) - leads to an expression for the least squares estimate $\hat{\gamma}$ as a function of residuals from the regression of y and w on X. This representation also leads to added variable plots (Cook and Weisberg 1982, p.44; Atkinson 1985, p.67) which can be used to detect an influential observation. Here we use it for the derivation and properties of t tests in the forward search.

Let the least squares estimate

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{6}$$

when the fitted values from this regression are

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy. \tag{7}$$

Then the least squares residuals of y are

$$e = \overset{*}{y} = y - \hat{y} = (I - H)y = Ay \tag{8}$$

and the residuals of w

$$\overset{*}{w} = (I - H)w = Aw. \tag{9}$$

The least squares estimator of γ in (5) is

$$\hat{\gamma} = \overset{*}{w}^T e/(\overset{*}{w}^T \overset{*}{w}) = w^T Ay/(w^T Aw), \tag{10}$$

with variance

$$\text{var } \hat{\gamma} = \sigma^2/(\overset{*}{w}^T \overset{*}{w}) = \sigma^2/(w^T Aw). \tag{11}$$

Calculation of the t test for γ also requires s_w^2, the residual mean square estimate of σ^2 from regression on X and w, which can be written as

$$(n-p)s_w^2 = y^T Ay - (y^T Aw)^2/(w^T Aw). \tag{12}$$

The t statistic for testing that $\gamma = 0$ is thus

$$t_\gamma = \hat{\gamma}/\{s_w^2/(w^T Aw)\}^{1/2}. \tag{13}$$

3.2 Added Variables for Testing $\gamma = \gamma_0$

In model building interest is usually in whether $\gamma = 0$, that is whether a variable should be included in the model. An added variable formulation can also be used for testing that γ has the non-zero value γ_0, when the test is

$$t_{\gamma_0} = (\hat{\gamma} - \gamma_0)/\{s_w^2/(w^T Aw)\}^{1/2}. \tag{14}$$

Under this hypothesis (5) is

$$y = X\beta + w\gamma_0 + \epsilon. \tag{15}$$

Subtraction of the vector of offsets $w\gamma_0$ from both sides of (5) yields the general model

$$y(\gamma_0) = y - w\gamma_0 = X\beta + w(\gamma - \gamma_0) + \epsilon = X\beta + w\gamma'. \tag{16}$$

If $\gamma = \gamma_0$, γ' will be zero and there should be no evidence of regression of $y(\gamma_0)$ on w. The added variable calculations of the preceding section go through with the residuals e replaced by the residuals

$$e(\gamma_0) = (I - H)y(\gamma_0) = Ay(\gamma_0). \tag{17}$$

3.3 Adding an Observation in the Forward Search

In the forward search the quantities of §3.1 are calculated for a subset of size m. We now derive the effect on the values of $\hat{\gamma}$ and of t_γ of adding observation $m+1$.

From (12) the residual sum of squares of regression of m observations only on X can be written

$$R(y,y) = y^T A y. \tag{18}$$

Let the new observation be y_+, with explanatory variables x_+ and w_+. The leverage of the new observation is

$$h_+ = x_+^T (X^T X)^{-1} x_+, \tag{19}$$

which is ≥ 0, but unlike leverages for deletion, need not be ≤ 1. The residual for the new observation is

$$e_+ = y_+^* = y_+ - x_+^T \hat{\beta}, \tag{20}$$

where $\hat{\beta}$ is given by (6).

Let $R_+(y,y)$ be the residual sum of squares of the $m+1$ observations after regression on X and x_+. The (Bartlett)-Sherman-Morrison-Woodbury formula is customarily used as one way of deriving deletion diagnostics (Cook and Weisberg 1982, p.210 and §2.7 of Atkinson and Riani (2000) give references). With a change of sign it can be used for the addition of observations when it follows that

$$R_+(y,y) = R(y,y) + e_+^2/(1+h_+) = y^T A y + y_+^{*2}/(1+h_+). \tag{21}$$

The expressions for $\hat{\gamma}$, s_w^2 and t_γ are all functions of the form $R(a,b)$ where the vectors a and b are either w or y. It then follows from (21) that these residual sums and products after the addition of one more observation become

$$R_+(a,b) = R(a,b) + a_+^* b_+^*/(1+h_+) = a^T A b + a_+^* b_+^*/(1+h_+), \tag{22}$$

where a_+^* and b_+^* are the residuals of a_+ and b_+, as in (20), after regression on X.

As a result of these relationships we can, for example, write the t test (13) for $m+1$ observations as

$$t_\gamma^+ = \frac{(m+1-p)^{1/2} R_+(w,y)}{\{R_+(y,y) R_+(w,w) - R_+^2(w,y)\}^{1/2}}, \tag{23}$$

with the quadratic forms given by (22)

3.4 Orthogonality and the Non-centrality Parameter

The results of §2.6 showed that the usual t tests in the forward search do not have a null t distribution. The search orders the data using all the variables in Q, that is X and w. The observations in the subset are the $m+1$ smallest order statistics of the residuals from the parameter estimate $\hat{\theta}^*_m$. These observations yield small estimates of σ^2 and over-large values for the t statistics, especially at the beginning of the search.

We now show that the added variable test is not affected by the ordering of the data and so has the required distribution.

Result. The added variable test (13) follows the t distribution under the customary normal conditions for regression models.

Proof. In searches using the added variable test, we fit the reduced model $E(Y) = X\beta$, the residuals from which are used to determine the progress of the search. We do not include w in the model. The choice of observations to include in the subset thus depends only on y and X. But the results of §3.1 show that the added variable test (13) is a function solely of the residuals $\overset{*}{w}$ and $\overset{*}{y}$, which by definition, are in a space orthogonal to X. The ordering of observations using X therefore does not affect the null distribution of the test statistic. Since, for normally distributed errors, the estimates $\hat{\gamma}$ and s^2 are independent, it follows that the null distribution of the statistic is Student's t.

Although the null distribution of the test statistic is unaffected by the forward search, the value of the non-centrality parameter is dependent on the search, since the values of the w_i in the subset will depend upon the ordering of the observations by the search. We now find an expression for the mean of the test statistic, which shows its dependence on X.

Provided that σ^2 is estimated consistently, for example using s^2_w (12), it follows from the results in §3.1 for a subset of size $m+1$, that, asymptotically,

$$E(t_\gamma) = E(\hat{\gamma}-\gamma_0)\{R_+(w,w)/\sigma^2\}^{1/2} = \{(\gamma-\gamma_0)/\sigma\}\{w^T A w + w^{*2}_+/(1+h_+)\}^{1/2}. \tag{24}$$

This expression makes explicit the dependence of the value of t_γ not only on the difference between γ and γ_0 but on the residuals of w after regression on X. If w lies almost in the space spanned by X, the test will have low power even if the difference between γ and γ_0 is not negligible. In such cases, care needs to be taken with procedures, such as backward selection, which automatically exclude variables with small t values. In such cases, although w adds little in explanatory power to a model already including X, it may provide a good model in combination with some of the other variables.

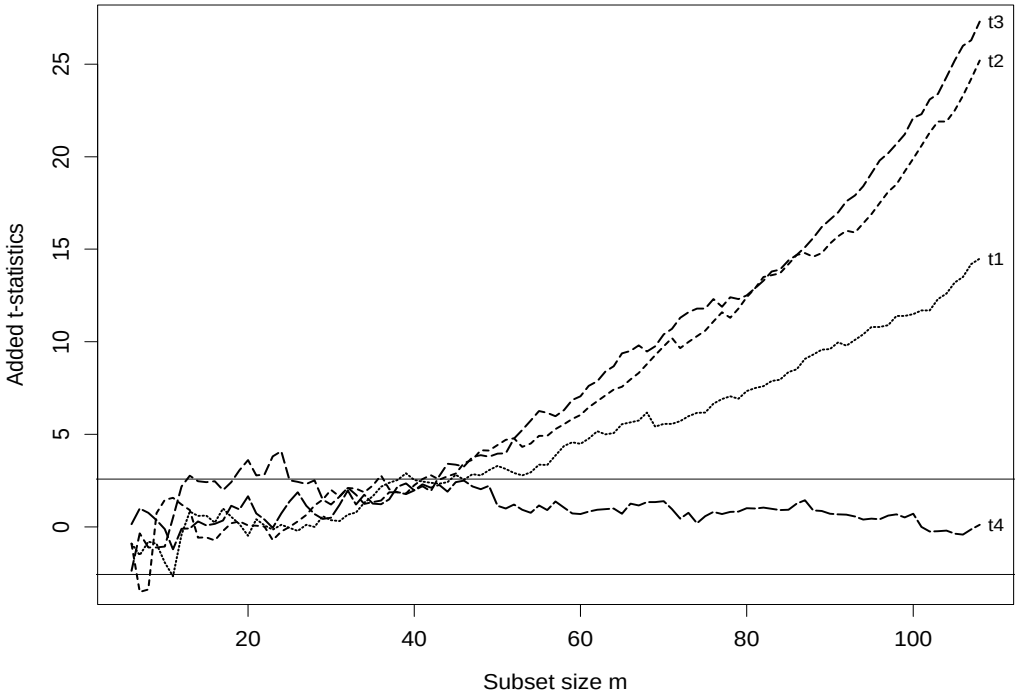

Figure 3: Transformed Surgical Unit data: forward plot of the four added-variable t statistics. To be compared with the lower panel of Figure 2

3.5 Surgical Unit Data

There are four explanatory variables in the surgical data. In order to use the method of added variables, each has to be omitted in turn and be treated as the added variable w. Four forward searches are therefore used, each using three of the four variables. The resulting plot of the four forward t statistics is in Figure 3. These curves behave as we would hope: initially no variables are significant, although x_3 is briefly significant at the 1% level around $m = 20$. The curves then rise smoothly to their values at the end of the search given in Table 1. For any value of m these are the kind of values we would expect to obtain from randomly sampling the observations and calculating the t statistics from the m observations, that is without an ordering effect.

There are two further points about the forward added variable plot of Figure 3. One is that we have included horizontal lines to indicate significance levels. These are based on the normal distribution. The upper panel of Figure 4 repeats the curve for t_4 in Figure 3 but with confidence limits calculated from the percentage points of the t distribution and found by

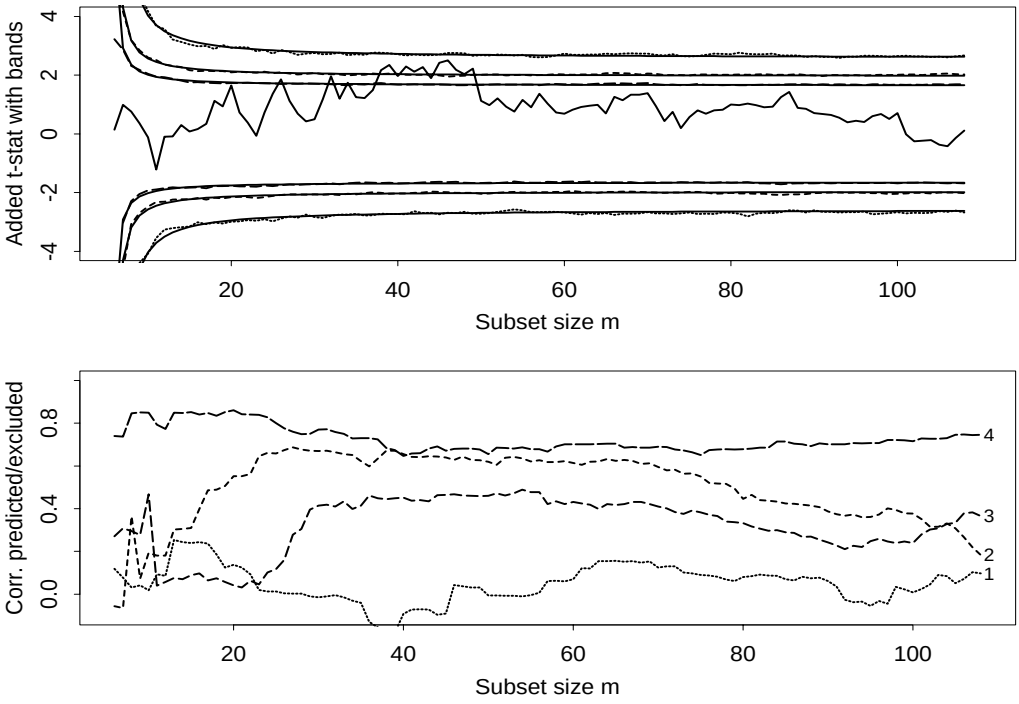

Figure 4: Transformed Surgical Unit data: upper panel, forward plot of added-variable t statistic for x_4, percentage points of the t distribution and averages of 10,000 simulations; lower panel, correlation between predictions from fitting X and the excluded variable

simulation of 10,000 samples. Theory and simulation agree.

Looking at Figure 3 would suggest that x_4 should be dropped from the model, a conclusion also reached, for example, by Hoeting et al. (1996) who were investigating a Bayesian method of variable selection in the presence of outliers. Calculation of the C_p statistics for all the data also shows that x_4 should be dropped. The lower panel of Figure 4, based on the searches which produced Figure 3, shows the correlation between each residual added variable $\overset{*}{w}$ and the prediction $\hat{y}$ (7) from regression on X. Variables 1, 2 and 3 have low correlations with the predictions when they are excluded from the fit. However x_4 is highly correlated with the predictions from the other variables. This suggests it may be important under the kind of data perturbations considered by Breiman (1996).

4 Multiple Outliers

4.1 Theory

Multiple outliers can both be hard to detect and can completely alter inferences about the correctness of individual models. We now suppose that the data are contaminated by k mean shift outliers, which will enter the search after the good observations (§2.4). The model for these observations is

$$E(Y_+) = X_+\beta + w_+\gamma + \Delta, \tag{25}$$

with X_+ a $k \times (p-1)$ matrix and the other vectors $k \times 1$: Δ is a vector of arbitrary shift parameters.

In order to show the effect of these outliers we extend (22) to the effect of adding several observations and obtain

$$R_+(a,b) = R(a,b) + a_+^{*T}(I_k + H_+^k)^{-1}b_+^* = a^TAb + a_+^{*T}(I_k + H_+^k)^{-1}b_+^*. \tag{26}$$

Now a_+^* and b_+^* are $k \times 1$, I_k is the $k \times k$ identity matrix and H_+^k the $k \times k$ hat matrix for the extra observations. Then the estimate of γ can be written

$$\hat{\gamma} = \frac{w^TAy + w_+^{*T}(I_k + H_+^k)^{-1}y_+^*}{w^TAw + w_+^{*T}(I_k + H_+^k)^{-1}w_+^*}. \tag{27}$$

In (27) $w_+^* = w_+ - X_+(X^TX)^{-1}X^Tw$. Since, from (5) and (6), $\mathrm{E}(\hat{\beta}) = \beta + (X^TX)^{-1}w\gamma$,

$$E(\hat{\gamma}) = \gamma + \frac{w_+^{*T}(I_k + H_+^k)^{-1}\Delta}{w^TAw + w_+^{*T}(I_k + H_+^k)^{-1}w_+^*}. \tag{28}$$

The effect of the vector of shift parameters may be either to increase or decrease $E(\hat{\gamma})$ depending on the signs of γ, Δ and of w_+^*. As different variables are selected to be the added variable, the effect of Δ will change depending on the various vectors w_+^*. However the effect of Δ is always modified by projection into the space orthogonal to X.

The effect of the outliers on the estimate of σ^2 is to cause it to increase. There will thus be a tendency for the t statistics to decrease after the introduction of the outliers even if $\hat{\gamma}$ increases. We show evidence of this decrease in Figure 5.

4.2 Surgical Unit Data

We now modify the surgical unit data to show the effect of masked outliers on the forward plot of t statistics. The effect of the outliers is clear.

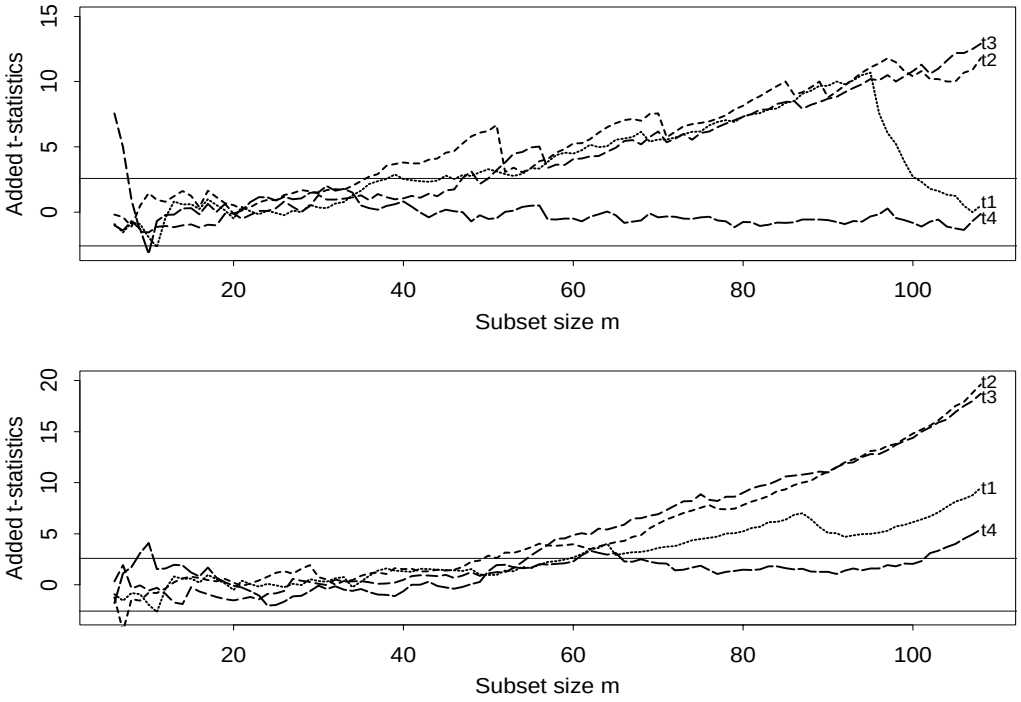

Figure 5: Modified Transformed Surgical Unit data: both panels show forward plots of added-variable t statistics. Upper panel, outliers render x_1 non-significant: lower panel; now the outliers make x_4 significant

The data were modified by Hoeting et al. (1996) who added five outliers to the first 54 observations. We contaminate up to 12 observations in two different ways in order to produce two different effects. The actual changes in the data are recorded in Table 2, with the forward plots of the t tests in Figure 5. In the upper panel the effect of the modification has been to make variable one non-significant: previously it was the most important variable. The plot very dramatically shows that, for this search without x_1, the observations have been ordered with the outliers at the end and that this group of observations has a dramatic effect on the added variable t test for variable one.

The plots of the forward t tests for variables 2 and 3 in the upper panel of Figure 5 show some peaks, followed by local declines. These come from the inclusion of the outliers, which are coming in at different points in the different forward searches and having less dramatic effects on the values of the t statistics.

The lower panel of Figure 5 shows the effect of a different kind of contamination, which, in this case, makes variable four significant at the end of the

	Units contaminated								Contamination
First	8	10	18	31	42	45	50	95	Added log(3)
contamination					28	73	75	103	Subtracted log(3)
Second									
contamination	43	50	98	33	29	17	28	18	Added log(2)

Table 2: The two contaminations of the data leading to the behaviour shown in Figure 5

search. The other variables remain significant, but the effect of the outliers, entering earlier in the search is again evident, particularly for x_1.

At the end of §4.2 it was argued that one effect of outliers was to inflate the estimate of σ^2 and so to shrink the values of the t statistics. Comparison of the values of t_2 and t_3 in Figure 3 with those in the two panels of Figure 5 illustrates this argument.

These plots very clearly show the effect of the outliers on the t tests for regression. Variable selection using t tests in the first example would lead to the incorrect dropping of variable one: in the second case it would lead to the incorrect inclusion of variable 4 in the model.

The contaminations have been designed so that the outliers are masked and are not readily recognisable from statistics calculated from all the data. It is perhaps possible that they could be found by a careful study of QQ and other plots of residuals and from the scatterplots of the data. Certainly they are easily found using the forward plots of statistics, parameter estimates, Cook distances and the other diagnostic measures exemplified in Chapter 3 of Atkinson and Riani (2000). But this is not the point. The purpose of our method is to discover precisely the effect of individual observations on the t tests for the variables included in the model. The plots in Figure 5 do exactly that. It is clear that a subset of observations are indicating a different model from the majority of the data. Which these observations are follows from the order in which the observations enter the search. In both examples the contaminated observations were the last to enter the searches in which inferences were changed.

5 Transformations of the Univariate Response in Regression

5.1 A Constructed Variable Score Test

The surgical unit data have been analysed using the logarithm of time. We now test whether this transformation is appropriate. The constructed-variable test we use is similar in form to the added-variable test of §3.1 but, as we see, has different distributional properties.

The test was introduced by Atkinson (1973) to test the value of the transformation parameter λ in the Box and Cox (1964) family of normalized power transformations

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \dot{y}^{\lambda-1} & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0, \end{cases} \tag{29}$$

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i/n)$. In this test the variable w in (13)is replaced by a constructed variable which, provided X includes a constant, can be written

$$w(\lambda) = \begin{cases} y^\lambda\{\log(y/\dot{y}) - 1/\lambda\}/(\lambda \dot{y}^{\lambda-1}) & \lambda \neq 0 \\ \dot{y} \log y(0.5 \log y - \log \dot{y}) & \lambda = 0. \end{cases} \tag{30}$$

Chapter 3 of Atkinson and Riani (2000) gives examples of the use of forward plots of this added variable test for transformations. Because outliers in one scale may not appear to be outlying after a different transformation, the forward search is performed for five different values of λ. Although forward plots of the test statistic are easily interpreted, the statistic cannot have exactly a t distribution: the constructed variable (30) is a function of the response. Thus the response and the constructed variable are not independent and so the conditions for the t distribution of t_γ (13)do not hold. However, the statistic depends on the properties of the residuals of y and $w(\lambda)$, the correlation between which depends on the projection matrix A. We use plots and simulation to investigate the effect of this projection on the distribution of the test statistic in the forward search.

5.2 Examples

Two examples are considered: the unmodified surgical unit data before transformation and a random sample of 108 observations from a standard normal observation with mean ten, which are then exponentiated. The log transformation ($\lambda = 0$) is to be expected.

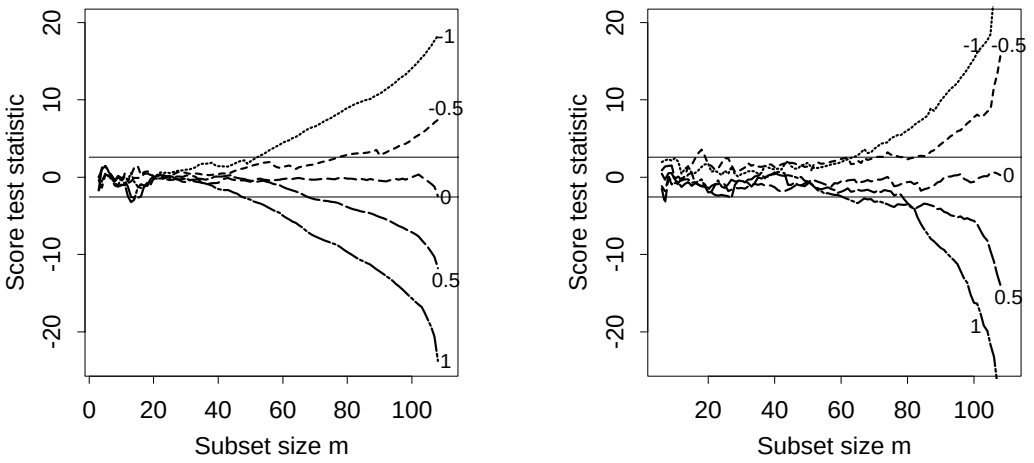

Figure 6: Fan plots (forward plots of constructed variable tests for transformations). Left-hand panel, exponentiated random sample; right-hand panel, untransformed Surgical Unit data. The log transformation is indicated for both examples

The left-hand panel of Figure 6 shows a "fan" plot for the exponentiated normal sample, that is a forward plot of the score statistics for five values of λ between -1 and 1. The log transformation is apparently indicated, except at the very end of the search where one or two observations suggest rejection of $\lambda = 0$. The right-hand panel shows a similar plot for the untransformed Surgical Unit data. Again the log transformation is indicated, but this time by all 108 observations.

In Figure 6 the confidence bands drawn are for the normal distribution. We have argued, and seen in Figure 4, that the t distribution is appropriate for added variable tests. However the results of 10,000 simulations in the left-hand panel of Figure 7 show that the null distribution for the constructed-variable test is not t: in the centre of the search the bands are too narrow and, at the end of the search, there is a "trumpet" as the distribution broadens out. The downwards steps at the end of the search which were noted in Figure 6 are now seen not to take the test statistic outside the central 95% of the simulated distribution. The log transformation is indeed acceptable.

The constructed-variable plot for $m = 108$ in the right-hand panel of Figure 7 explains the behaviour of the simulated envelopes towards the end of the search. The shape of this plot is approximately parabolic - because there is no regression in this model, there is no projection into a space orthogonal to X to beak the relationship between y and $w(\lambda)$. The slight asymmetry in the plot is enough to cause some regression and a seemingly significant

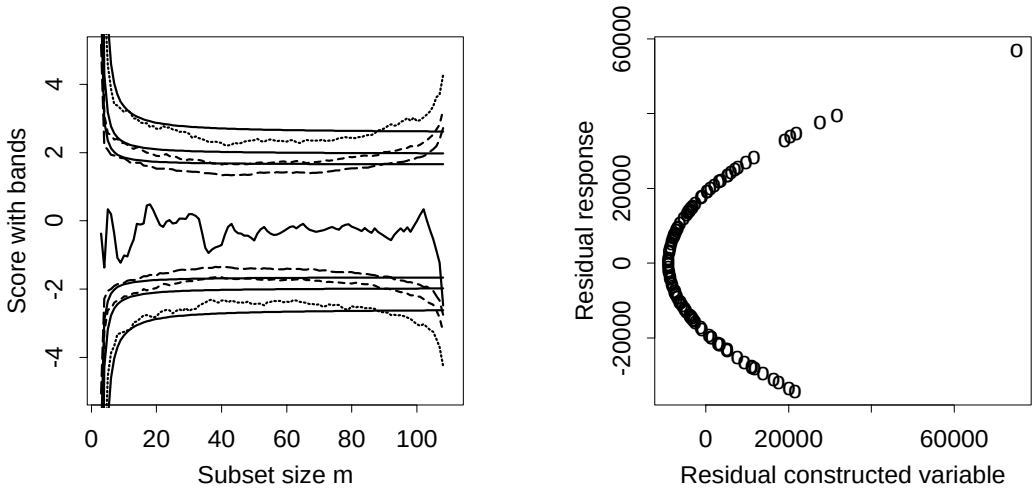

Figure 7: Exponentiated random sample: left-hand panel, fan plot; right-hand panel, constructed variable plot at m= 108

t value, although the value of R^2 is approximately zero. For $m = 107$, the observation in the top right hand corner of the plot is excluded from the subset, but there is still some asymmetry and some regression. Repetition of this structure over many simulations produces the trumpet effect.

The plots in Figure 8 for the untransformed Surgical Unit data are very different. Now there is strong regression : $R^2 = 0.947$. Although the simulated envelopes are slightly narrow in the centre of the search, the distribution of the statistic in the last steps is the hoped for Student's t. The right-hand panel of the figure shows the constructed variable plot, again at the end of the search. There is now sufficient regression on X that the projection destroys the relationship between y and $w(\lambda)$ and the distribution is as it is for the added-variable test of §3.

These results show the importance of strong regression in ensuring that the constructed-variable score test for transformations has a null t distribution. Plots like those of Figures 7 and 8 help interpret the results of Atkinson and Lawrance (1989) where the null distribution of the statistic was found to vary with data set. Another departure from the null t distribution sometimes arises because the data to be transformed must be non-negative. If a model is fitted for which there is a non-negligible probability of generating negative observations, these have to be rejected during simulation when the distribution of the data will be a truncated normal. Such data give rise to skew null distributions for the transformation statistic. Examples are in Atkinson and Riani (2002).

Although the distribution of the test for transformation is only approxi-

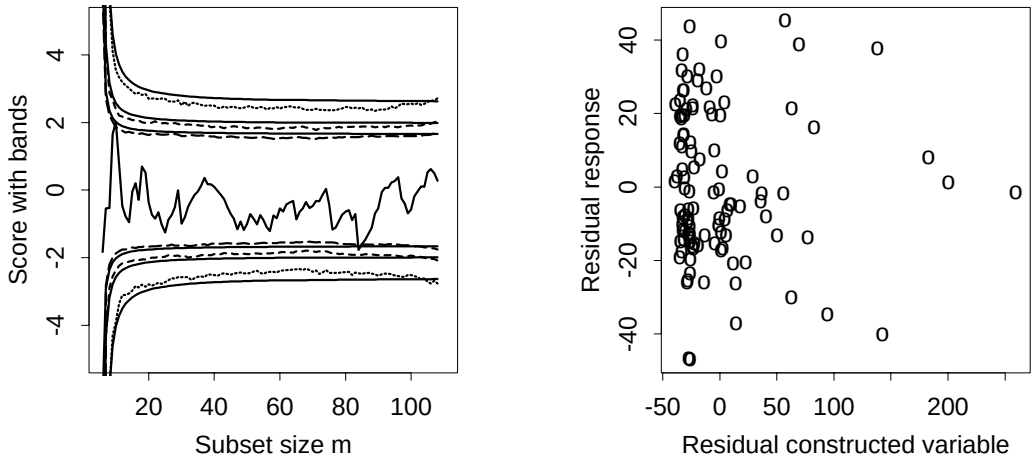

Figure 8: Untransformed Surgical unit data: left-hand panel, fan plot; right-hand panel, constructed variable plot at m= 108

mately t, the statistic for the added variable test for regressors, which is the main subject of this paper, has been shown to follow Student's t distribution. This behaviour is distinct from, and more useful for model identification than, the behaviour of the standard t tests during the forward search.

References

Atkinson, A. C. (1973). Testing transformations to normality. *Journal of the Royal Statistical Society, Series B 35*, 473–479.

Atkinson, A. C. (1985). *Plots, Transformations, and Regression*. Oxford: Oxford University Press.

Atkinson, A. C. and A. J. Lawrance (1989). A comparison of asymptotically equivalent tests of regression transformation. *Biometrika 76*, 223–229.

Atkinson, A. C. and M. Riani (2000). *Robust Diagnostic Regression Analysis*. New York: Springer–Verlag.

Atkinson, A. C. and M. Riani (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemometrics and Intelligent Laboratory Systems 60*, 87–100.

Box, G. E. P. and D. R. Cox (1964). An analysis of transformations (with discussion). *Journal of the Royal Statistical Society, Series B 26*, 211–246.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics 24*, 2350–2383.

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

Hawkins, D. M. (1993). The accuracy of elemental set approximations for regression. *Journal of the American Statistical Association 88*, 580–589.

Hoeting, J., A. E. Raftery, and D. Madigan (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics and Data Analysis 22*, 251–270.

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (1996). *Applied Linear Statistical Models, 4th edition*. New York: McGraw-Hill.

Riani, M. and A. C. Atkinson (2001). A unified approach to outliers, influence, and transformations in discriminant analysis. *Journal of Computational and Graphical Statistics 10*, 513–544.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association 79*, 871–880.