Biometrika (2002), **89,** 4, *pp.* 939–946

Printed in Great Britain

Forward search added-variable *t*-tests and the effect of masked outliers on model selection

BY ANTHONY C. ATKINSON

Department of Statistics, London School of Economics, London WC2A 2AE, U.K.

a.c.atkinson@lse.ac.uk

AND MARCO RIANI

Dipartimento di Economia, Università di Parma, 43100 Parma, Italy

mriani@unipr.it

SUMMARY

Monitoring the *t*-tests for individual regression coefficients in 'forward' search fails to identify the importance of observations to the significance of the individual regressors. 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. Comments are given on the related test for response transformations.

Some key words: Box-Cox transformation; Projection; Score statistic; *t*-distribution; Transformation to normality; Very robust methods.

1. INTRODUCTION

Forward search is a powerful general method for detecting multiple masked outliers, for determining their effects on models fitted to data and for detecting systematic model inadequacy. Here we develop a method for detecting the effects of individual observations 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 mention the distributional properties of a related test for transformations.

Forward search is briefly described in the next section. In § 3 the forward plot for added-variable *t*-statistics is defined and its properties are 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. Brief comments on the related plot for transformations conclude the paper.

2. FORWARD SEARCH AND FORWARD PLOTS

Details of forward search for regression are given in Atkinson & Riani (2000, Ch. 2). The method starts by fitting a small, robustly chosen, subset of m of the n observations to the data. In the example in this paper we start with $m = p$, the number of parameters in the regression model. We sample 1000 subsets to each of which a regression is fitted by least squares and take as the starting subset that which yields the smallest median squared residual. This is the algorithm for least median of squares introduced by Rousseeuw (1984). We move forward to a larger subset by ordering the n squared residuals from the least squares fit to the subset of m observations and using the $m + 1$ observations with the smallest squared residuals as our new larger subset. Usually one observation

is added to the subset at each step, but sometimes two or more are added as one or more leave, which is often an indication of the introduction of some of a cluster of outliers. In this way we obtain a series of parameter estimates for $p \leqslant m \leqslant n$, which progresses from very robust at the beginning of the search to least squares at the end. In the absence of outliers, the parameter estimates and plots of all n residuals remain stable as m grows; an example is Atkinson & Riani (2000, p. 71). The search is such that observations which are far from the fitted model enter at the end of the search; these may be outliers, or an unidentified subset, or they may indicate a systematic failure of the model. The analysis of the Box & Cox poison data in Atkinson & Riani (2000, § 4.4) shows the dependence of the order in which observations enter the subset used in fitting on the transformation employed.

During the search we monitor quantities indicative of model quality or inadequacy. In addition to residuals we can look at score tests for transformations or for link functions, or measures of curvature in nonlinear models. We can also study the evolution of plots, for example of profile likelihoods. In regression we can monitor the evolution of s^2, the estimate of the error variance. Since the search orders the observations by the magnitude of their residuals from the fitted subsets, the value of s^2 increases during the search, although not necessarily monotonically. As a consequence, even in the absence of outliers and model inadequacies, the values of the t-statistics for the parameters in the model decrease during the search and are hard to interpret. An example is Atkinson & Riani (2000, p. 72). Here we use the method of added variables to provide plots of t-tests which are orthogonal to the search.

3. AN ADDED-VARIABLE t-TEST

3·1. *Added variables*

In order to obtain useful forward plots of t-tests we write the regression model for all n observations as

$$y = Q\theta + \varepsilon = X\beta + w\gamma + \varepsilon, \tag{1}$$

where Q is $n \times p$, the errors ε satisfy the second-order assumptions with variances σ^2 and γ is a scalar. In turn we take each of the columns of Q as the vector w, except for the column corresponding to the constant term in the model. The well-established approach of added variables (Atkinson & Riani, 2000, § 2.2) leads to an expression for the least squares estimator $\hat{\gamma}$ as a function of residuals from the regression of y and w on X. This representation also leads to added variable plots (Cook & Weisberg, 1982, p. 44; Atkinson, 1985, p. 67) which can be used to detect an influential observation, although deletion methods are preferable (Atkinson, 1985, § 12.3).

The least squares estimator of β is $\hat{\beta}$, with hat matrix $H = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. The least squares residuals of y and w are

$$e = \overset{*}{y} = y - \hat{y} = (I - H)y = Ay, \tag{2}$$

$$\overset{*}{w} = (I - H)w = Aw. \tag{3}$$

The least squares estimator γ in (1) is

$$\hat{\gamma} = \overset{*}{w}{}^{\mathrm{T}}e/(\overset{*}{w}{}^{\mathrm{T}}\overset{*}{w}) = w^{\mathrm{T}}Ay/(w^{\mathrm{T}}Aw), \tag{4}$$

with variance

$$\mathrm{var}(\hat{\gamma}) = \sigma^2/(\overset{*}{w}{}^{\mathrm{T}}\overset{*}{w}) = \sigma^2/(w^{T}Aw). \tag{5}$$

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

$$(n - p)s_w^2 = y^{\mathrm{T}}Ay - (y^{\mathrm{T}}Aw)^2/(w^{\mathrm{T}}Aw). \tag{6}$$

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

$$t_\gamma = \hat{\gamma}/\{s_w^2/(w^{\mathrm{T}}Aw)\}^{\frac{1}{2}}. \tag{7}$$

3·2. *Adding an observation in the forward search*

In a 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 (6) the residual sum of squares of regression of m observations only on X can be written

$$R(y, y) = y^{\mathrm{T}} A y. \tag{8}$$

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

$$h_+ = x_+^{\mathrm{T}} (X^{\mathrm{T}} X)^{-1} x_+, \tag{9}$$

which is nonnegative, but, unlike leverages for deletion, may exceed one. The residual for the new observation is

$$e_+ = y_+^* = y_+ - x_+^{\mathrm{T}} \hat{\beta}. \tag{10}$$

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 & Weisberg (1982, p. 210) and Atkinson & Riani (2000, § 2.7) give references. With a change of sign it can be used for the addition of observations when it follows that

$$R_+(y, y) = R(y, y) + e_+^2/(1 + h_+) = y^{\mathrm{T}} A y + y_+^{*2}/(1 + h_+). \tag{11}$$

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 (11) that these residual sums and products after the addition of one more observation become

$$R_+(a, b) = R(a, b) + a_+^* b_+^*/(1 + h_+) = a^{\mathrm{T}} A b + a_+^* b_+^*/(1 + h_+), \tag{12}$$

where a_+^* and b_+^* are the residuals of a_+ and b_+, as in (10), after regression on X.

As a result of these relationships we can, for example, write the t-test (7) for $m+1$ observations as

$$t_\gamma^+ = \frac{(m + 1 - p)^{1/2} R_+(w, y)}{\{R_+(y, y) R_+(w, w) - R_+^2(w, y)\}^{\frac{1}{2}}}, \tag{13}$$

with the quadratic forms given by (12).

3·3. *Orthogonality and the noncentrality parameter*

Since 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* (7) *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 be included in the subset thus depends only on y and X. However, the results of § 3·1 show that the added-variable test (7) 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 noncentrality 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_w^2 from (6), it follows from the results in § 3·1 for a subset of size $m+1$ that, asymptotically,

$$E(t_\gamma) = E(\hat{\gamma} - \gamma_0) \left\{ \frac{R_+(w, w)}{\sigma^2} \right\}^{\frac{1}{2}} = \left(\frac{\gamma - \gamma_0}{\sigma} \right) \left(w^{\mathrm{T}} A w + \frac{w_+^{*2}}{1 + h_+} \right)^{\frac{1}{2}}. \tag{14}$$

This expression makes explicit the dependence of the value of t_γ not only on the difference between γ and γ_0 but also 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. Although in such cases 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.

3·4. *Surgical unit data*

Neter et al. (1996, pp. 334, 438) analyse 108 observations on the times of survival of patients who had a particular kind of liver surgery. The four explanatory variables are as follows: x_1 is a blood clotting score; x_2 is a prognostic index, which includes the age of the patient; x_3 is an enzyme function test score; and x_4 is a score for liver function. The response is survival time. We follow Neter et al. (1996) and use the logarithm to base ten of time as the response. The properties of a test for this transformation are mentioned in § 5.

It seems clear when all 108 observations are fitted 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.

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 Fig. 1. These curves behave as we would hope: initially no variable is significant, although x_3 is briefly significant at the 1% level around $m = 20$. The curves then rise smoothly to their values when $m = n$, with the nonsignificant value of t_4 showing seemingly random fluctuations.

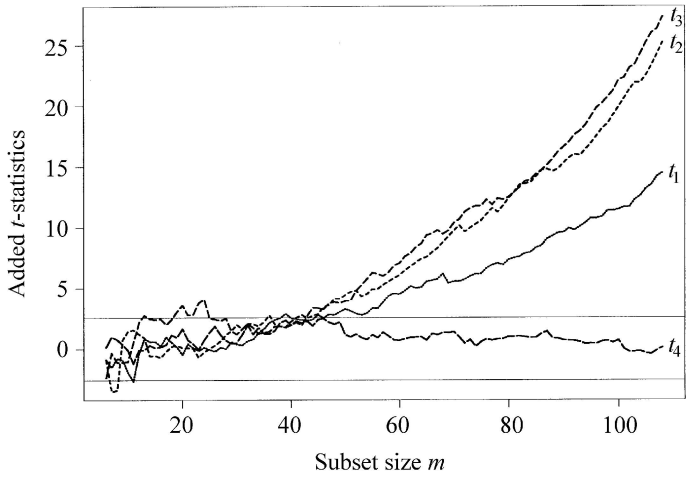

Fig. 1. Transformed surgical unit data: forward plot of the four added-variable t-statistics, t_1, t_2, t_3 and t_4.

In Fig. 1 we have included horizontal lines to indicate significance levels. These are based on the normal distribution. Figure 2(a) repeats the curve for t_4 in Fig. 1 but with confidence limits calculated from the percentage points of the t-distribution and found by simulation of 10 000 samples. Theory and simulation agree: despite the ordering of observations by the searches, the statistics follow the t-distribution.

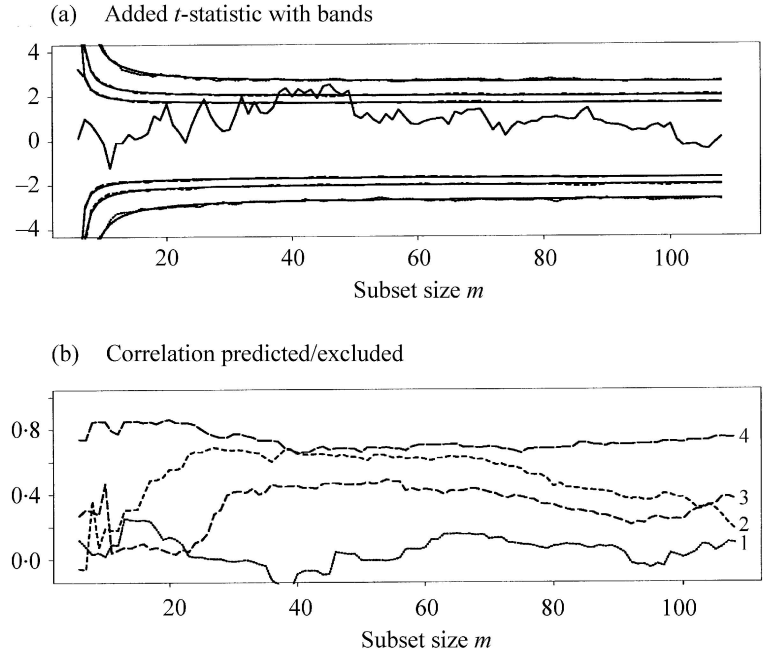

Fig. 2. Transformed surgical unit data: (a) forward plot of added-variable t-statistic for x_4, percentage points of the t-distribution and averages of 10 000 simulations; (b) correlation between predictions from fitting X and the excluded variable.

Figure 1 suggests that x_4 should be dropped from the model, a conclusion also reached by Hoeting et al. (1996). Figure 2(b), based on the searches which produced Fig. 1, shows the correlation between each residual added variable $\overset{*}{w}$ and the prediction $\hat{y}$ 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 both can 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; see § 2. The model for these observations is

$$E(Y_+) = X_+\beta + w_+\gamma + \Delta, \tag{15}$$

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 (12) to the effect of adding several

observations and obtain

$$R_+(a, b) = R(a, b) + a_+^{*\mathrm{T}}(I_k + H_+^k)^{-1}b_+^* = a^{\mathrm{T}}Ab + a_+^{*\mathrm{T}}(I_k + H_+^k)^{-1}b_+^*. \tag{16}$$

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 estimator of γ can be written

$$\hat{\gamma} = \frac{w^{\mathrm{T}}Ay + w_+^{*\mathrm{T}}(I_k + H_+^k)^{-1}y_+^*}{w^{\mathrm{T}}Aw + w_+^{*\mathrm{T}}(I_k + H_+^k)^{-1}w_+^*}. \tag{17}$$

In (17)

$$w_+^* = w_+ - X_+(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}w.$$

Since, from (1),

$$E(\hat{\beta}) = \beta + (X^{\mathrm{T}}X)^{-1}w\gamma,$$

$$E(\hat{\gamma}) = \gamma + \frac{w_+^{*\mathrm{T}}(I_k + H_+^k)^{-1}\Delta}{w^{\mathrm{T}}Aw + w_+^{*\mathrm{T}}(I_k + H_+^k)^{-1}w_+^*}. \tag{18}$$

The effect of the vector of shift parameters may be either to increase or to 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. Figure 3 below shows evidence of this decrease.

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.

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 1, with the forward plots of the t-tests in Fig. 3. In Fig. 3(a) the effect of the modification has been to make x_1 nonsignificant; previously it was the most important variable. Since x_1 is the added variable, the search orders the observations using the regression model in only x_2, x_3 and x_4. 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 x_1.

Table 1. *The two contaminations of the surgical unit data leading to the behaviour shown in Fig.* 3; *all logarithms are to base* 10

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

The plots of the forward t-tests for x_2 and x_3 in Fig. 3(a) show some peaks, followed by local declines. These come from the inclusion of the outliers, which enter at different points in the different forward searches and have less dramatic effects on the values of the t-statistics.

Figure 3(b) shows the effect of a different kind of contamination, which, in this case, makes x_4 significant at the end of the 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 § 4·1 it was argued that one effect of outliers was to inflate the estimate of σ^2 and

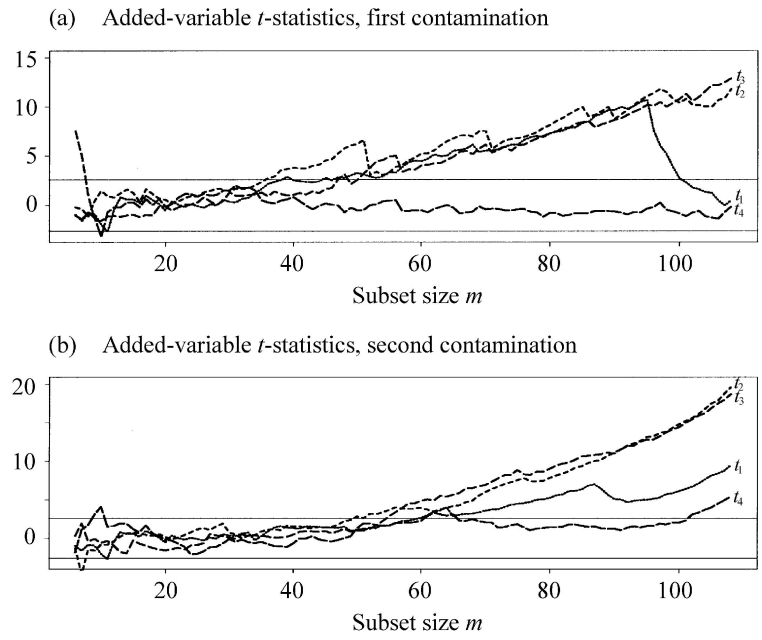

Fig. 3. Modified transformed surgical unit data: forward plots of added-variable t-statistics, t_1, t_2, t_3 and t_4, with contaminations given in Table 1. In (a) outliers render x_1 nonsignificant; in (b) outliers make x_4 significant.

so to shrink the values of the t-statistics. Comparison of the values of t_2 and t_3 in Fig. 1 with those in Fig. 3(a) and Fig. 3(b) illustrates this argument.

These plots very clearly show the effect of the outliers on the t-tests for regression. Variable selection using t-tests with the first contamination would lead to the incorrect dropping of x_1; with the second contamination it would lead to the incorrect inclusion of x_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 Q–Q 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 Atkinson & Riani (2000, Ch. 3), but this is not the point. The purpose of our method is to discover precisely the effects of individual observations on the t-tests for the variables included in the model. The plots in Fig. 3 do exactly that. It is clear that a subset of observations are indicating a different model from the majority of the data. The identities of these observations follow 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. Discussion

The surgical unit data were analysed using the logarithm of time. A London School of Economics research report by the authors uses the forward version of the constructed-variable test of Atkinson (1973) for response transformation to assess this transformation. Although the test is similar in form to the added-variable test of § 3·1, the statistic cannot have exactly a t-distribution as the constructed variable 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 in (7) do not hold. However, the results of Atkinson & Riani (2002) show that, provided there is appreciable regression, the projection matrix A ensures that the constructed-variable score test for transformations almost has

a null t-distribution. This conclusion helps to interpret the results of Atkinson & Lawrance (1989), where the null distribution of the statistic was found to vary with dataset.

Although the distribution of the test for transformation is only approximately t, the statistic for the added-variable test for regressors, which is the 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. *J. R. Statist. Soc.* B **35**, 473–9.
ATKINSON, A. C. (1985). *Plots, Transformations, and Regression*. Oxford: Oxford University Press.
ATKINSON, A. C. & LAWRANCE, A. J. (1989). A comparison of asymptotically equivalent tests of regression transformation. *Biometrika* **76**, 223–9.
ATKINSON, A. C. & RIANI, M. (2000). *Robust Diagnostic Regression Analysis*. New York: Springer-Verlag.
ATKINSON, A. C. & RIANI, M. (2002). Tests in the fan plot for robust, diagnostic transformations in regression. *Chemomet. Intel. Lab. Syst.* **60**, 87–100.
BREIMAN, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350–83.
COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.
HOETING, J., RAFTERY, A. E. & MADIGAN, D. (1996). A method for simultaneous variable selection and outlier identification in linear regression. *Comp. Statist. Data Anal.* **22**, 251–70.
NETER, J., KUTNER, M. H., NACHTSHEIM, C. J. & WASSERMAN, W. (1996). *Applied Linear Statistical Models*, 4th ed. New York: McGraw-Hill.
ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Am. Statist. Assoc.* **79**, 871–80.

[*Received February* 2002. *Revised April* 2002]