Building Regression Models with the Forward Search

Anthony C. Atkinson
Department of Statistics, The London School of Economics
Houghton St, London WC2A 2AE, UK
E-mail: a.c.atkinson@lse.ac.uk

Marco Riani
Dipartimento di Economia, Università di Parma,
Via J. Kennedy 6, 43100 Parma, Italy
E-mail: mriani@unipr.it

Abstract. *We give an example of the use of the forward search in building a regression model. The standard backwards elimination of variables is supplemented by forward plots of added variable t statistics that exhibit the effect of each observation on the process of model building. Attention is also paid to the effect of individual observations on selection of a transformation. Variable selection using AIC is mentioned, as is the analysis of multivariate data.*

Keywords. Added variable; Box-Cox transformation; constructed variable; masking; outliers; robust estimation.

1. Introduction

The forward search is a powerful general method for detecting multiple masked outliers and for determining their effect on inferences about models fitted to data. This paper presents a canonical example of the use of the forward search in building a regression model. Both the selection of variables and transformations of the response are considered. In addition, some mention is made of methods for the analysis of multivariate data, including clustering.

Standard statistical techniques are based on aggregate statistics; in normal theory regression these would be the sufficient statistics for the parameters of the linear model and for estimation of the error variance. The books [12] and [3] describe the use of single observation deletion diagnostics to explore the contribution of each observation to inference. If there are several outliers, these can sometimes be deleted in turn and their importance revealed. However, particularly if there are several similar outliers, or if the data contain unsuspected clusters, the outliers may be masked; they may not become apparent until several observations have been deleted. Unfortunately, the combinatorial explosion of the number of combinations of observations that have to be deleted renders such backwards deletion procedures impractical. An example for regression in which deletion procedures fail to reveal a complicated structure of outliers is in [4].

The book [18] explores the use of very robust methods in the detection of outliers. These methods use estimators based on a single carefully chosen subset of the data. On the contrary, in the forward search we base estimators on a series of subsets of the data and so obtain multiple views of any hidden structure. We monitor the evolution of residuals, parameters estimates and inferences as the subset size increases, presenting our results as "forward plots" which show the evolution of the quantities of interest as a function of sample size. Use of the forward search is described by [4] for linear and nonlinear regression, response transformation and in generalized linear models. Related forward techniques for multivariate data are given in [8].

The forward search is described in more detail in the next section. In §3 we define added-variable t statistics for regression coef-

ficients. Theoretical results and simulations show that the statistics have the correct t distribution, independently of the ordering of the observations in the search. The resulting procedure for the selection of regression variables is exemplified in §3.4 with the analysis of data on ozone concentration in California.

The next section briefly considers the related plot for transformations using a constructed variable. Here the null distribution is no longer exactly t. The paper concludes with brief mentions of applications of the forward search in model selection and, in §6, the analysis of multivariate data.

2. The Forward Search

Details of the forward search for regression are given in Chapter 2 of [4]. The method typically starts by fitting a small, robustly chosen, subset of m_0 of the n observations to the data. In the example in this paper we take $m_0 = p$, the number of parameters in the regression model. We sample 1,000 subsets to each of which a regression is fitted by least squares and the median of the n squared residuals is calculated. We take as the starting subset $S^*(m_0)$ that which yields the smallest median squared residual. This is the algorithm for least median of squares introduced by [17]. For general m we move forward to a larger subset by ordering the n squared residuals from the least squares fit to the subset $S^*(m)$ of m observations and using the $m+1$ observations with the smallest squared residuals to form $S^*(m+1)$. In this way we obtain a series of parameter estimates for $p \leq m \leq n$, 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 remain stable as m grows. 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 and Cox poison data in §4.4 of [4] illustrates the dependence of the order in which observations enter $S^*(m)$ on the particular transformation of the response.

There is nothing special about the starting point of the search. For small problems we can search over all subsets of size p of the n observations. Another possibility is to use the Least Trimmed Squares estimator, for which again see [17].

During the search we can monitor quantities indicative of model quality or inadequacy, such as residuals or the score test for transformations. In regression we can also monitor the evolution of s^2, the estimate of the error variance. Because 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 tests for the parameters in the model decrease during the search and are hard to interpret. An example is on p. 72 of [4]. In [5] the method of added variables is used to provide plots of t tests which are orthogonal to the search. We exemplify and extend this procedure both for testing for explanatory variables and for tests of transformations of the response.

3. An Added Variable t Test

3.1. Added Variables

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 the observations in $S^*(m)$.

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

$$y = Q\theta + \epsilon = X\beta + w\gamma + \epsilon, \tag{2}$$

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). Thus if the columns of Q are the $p-1$ regression variables x_2 to x_p, we exclude each in turn and reinclude it as w. We

perform a forward search using only the variables in X and then use the well-established approach of added variables - for example §2.2 of [4] - to calculate the t test for the inclusion of w in a manner orthogonal to the search.

This methodology 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 ([12], p. 44; [3], 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 estimator be

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{3}$$

when the fitted values from this regression are

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy \tag{4}$$

and the least squares residuals of y and w are

$$e = \overset{*}{y} = y - \hat{y} = (I - H)y = Ay \tag{5}$$

and

$$\overset{*}{w} = (I - H)w = Aw. \tag{6}$$

The least squares estimator of γ in (2) is

$$\hat{\gamma} = \overset{*}{w}^T e / (\overset{*}{w}^T \overset{*}{w}) = w^T Ay / (w^T Aw), \tag{7}$$

with variance

$$\text{var } \hat{\gamma} = \sigma^2 / (\overset{*}{w}^T \overset{*}{w}) = \sigma^2 / (w^T Aw). \tag{8}$$

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

$$(n-p)s_w^2 = y^T Ay - (y^T Aw)^2 / (w^T Aw). \tag{9}$$

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

$$t_\gamma = \hat{\gamma} / \{s_w^2 / (w^T Aw)\}^{1/2}. \tag{10}$$

3.2. 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

$$t_{\gamma_0} = (\hat{\gamma} - \gamma_0) / \{s_w^2 / (w^T Aw)\}^{1/2}. \tag{11}$$

Under this hypothesis (2) is

$$y = X\beta + w\gamma_0 + \epsilon. \tag{12}$$

Subtraction of the vector of offsets $w\gamma_0$ from both sides of (2) yields the general model

$$\begin{aligned} y(\gamma_0) &= y - w\gamma_0 \\ &= X\beta + w(\gamma - \gamma_0) + \epsilon \\ &= X\beta + w\gamma'. \end{aligned} \tag{13}$$

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

$$e(\gamma_0) = (I - H)y(\gamma_0) = Ay(\gamma_0). \tag{14}$$

3.3. Forward Plots of Added-Variable Tests

We perform one forward search for each of the $p-1$ choices of w in (2), or p choices if the decision to include the intercept also needs to be examined. Since we are regressing on a different set of variables in each search, we may expect that the observations will enter $S^*(m)$ in a different order for each choice of w. However, observations with outlying values of y will enter at the end of many or all searches. To summarise all this information we plot these $p-1$ values of the added-variable t statistic in a "forward plot" as a function of m. In this way we can see whether individual observations or groups of observations are causing changes in our assessment of the significance of the explanatory variables.

As an example, [4] analyse data from [15] (pp. 334 & 438) on the time of survival of 108 patients who had a particular kind of liver surgery. There are four explanatory variables. The forward plot of added-variable t statistics in Fig. 1 of [5] shows that evidence for the significance of three out of the four variables grows steadily during the search, with

the fourth remaining non-significant. There is no evidence of the egregious effect of any observations, either singly or in groups. However, modifications of the data show how outliers can cause a variable to be significant only when they are included or how the significance of a variable can be destroyed, as viewed by the t statistic calculated at the end of the search. Despite this masking of the effect of the altered observations, the forward plot of added-variable t statistics in their Fig. 3 reveals how these conclusions are caused by the presence of outliers. We now consider a more complicated regression example.

3.4. Ozone Data

Section 3.4 of [4] present a forward analysis of data on ozone concentration in which there are eight potential explanatory variables. The regression model is chosen using a standard analysis based on t statistics when all observations are fitted. A forward search is then used to explore the properties of the chosen model. We now supplement this analysis by use of forward plots of added-variable t tests.

The data, given by [4] in their Table A.7, are the first 80 observations on a series of daily measurements, from the beginning of the year, of ozone concentration and meteorological variables in California. The full set of 300 observations were used by [11] when introducing the ACE algorithm. The data are given in the supporting material for [13], with a scatter plot of two variables on p. 25 of that book, in which it is clear that ozone concentration is related to daily temperature, x_1.

We begin by regressing ozone concentration on the eight explanatory variables. The forward plot of added-variable t statistics is in Fig. 1. A surprising feature of the fitted model is that none of the t tests for the coefficients are significant at the end of the search, the most extreme value being -1.32, although the value of R^2 is 0.430. One reason for this seemingly poor fit may be that some of the variables are highly correlated, leading

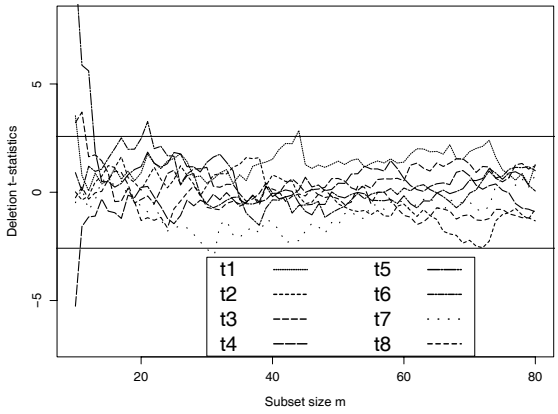

Figure 1: Ozone data: forward plot of added-variable t statistics; horizontal band contains 99% of the normal distribution. There are no significant variables

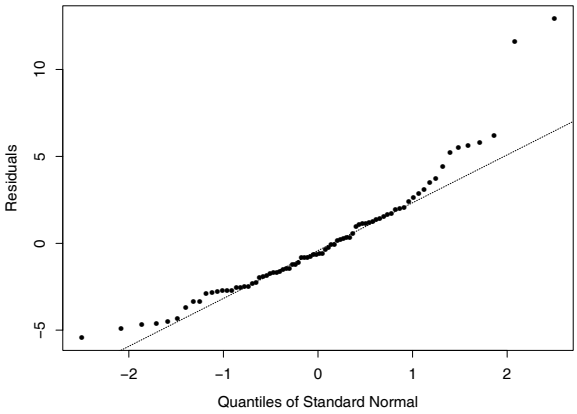

Figure 2: Ozone data: QQ-plot of residuals. Some lack of normality is evident

to small t values.

The QQ plot of raw residuals in Fig. 2 shows that the assumptions of constant error variance for regression do not hold. In fact, [4] find evidence for a log transformation of the data. We scrutinise the evidence for this in §4. However, such a transformation is *a priori* likely; regression assumes at least approximately constant error variance, but here the non-negative responses range from 2 to 24. In addition, Fig. 3.36 of [4] shows evidence of an upwards trend in the residuals from the fitted model with $\log y$ as response, so they, and we, include a linear term in time in our model. The observations that lie furthest from this trend are 65, 56, 53 and 31.

There are now nine explanatory variables including the trend. Fig. 3 is the forward plot of added-variable t statistics for this model.

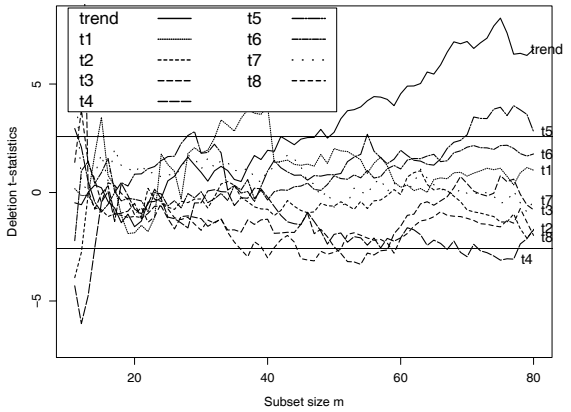

Figure 3: Logged ozone data: forward plot of added-variable t statistics; horizontal band contains 99% of the normal distribution. The trend and x_5 are most significant. The plot reflects overfitting.

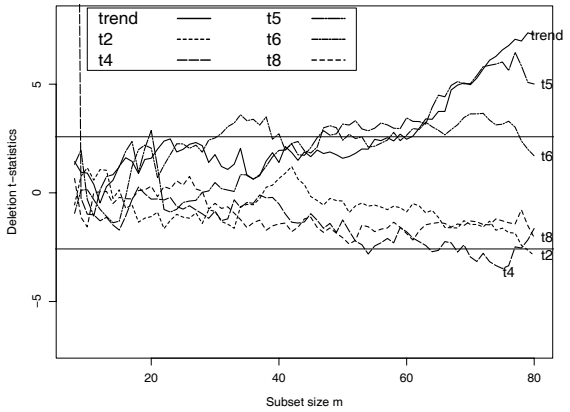

Figure 4: Logged ozone data: forward plot of added-variable t statistics; horizontal band contains 99% of the normal distribution. The least significant variable at the end of the search is x_4, but it is appreciably more significant than x_8 for most of the search.

The trend and x_5 are now significant at the 1% level. In most cases there is an appreciable decrease in significance in the last few steps of the search; t_4 is the most extreme example, changing from significant to not so. Each of these curves corresponds to a forward search in which X is different, so the units may enter in a different order. However, working backwards, the units that enter in the last few steps in all, or the majority, of searches are 65, 56, 31 and 53. These are precisely the units that were found to be outlying from the time trend. Our forward plot makes clear their influence on inferences drawn from the data.

A second feature of Fig. 3 is the jagged nature of the curves. This is a symptom of overfitting; there are so many explanatory variables that the values of the coefficients are responding to slight fluctuations in the data.

Initially we used a backwards procedure to select variables, based on the t statistics at the end of the search, but augmented by plots of the added-variable t statistics to ensure that this summary value was representative of behaviour for all $S^*(m)$. Proceeding in this way, always dropping the least significant variable, led, in turn, to the removal of x_7, x_3 and x_1. This analysis parallels that on p. 70 of [4], who however do not plot the t statistics. As the result of this process we obtain a model with a logged response, that includes a trend and terms in x_2, x_4, x_5, x_6, and x_8. The forward plot of the added-variable t statistics is in Fig. 4.

At this point x_4 has the smallest t statistic, -1.64 and [4] next delete this variable. However, Fig. 4 shows that there are rapid changes in the values of the t statistics in the last few steps of the search as the four observations we identified as potential outliers enter $S^*(m)$. In particular, the significance of x_8 is highest at the end of the search, but still remains within the 99% band as it has for the whole search. On the contrary, the statistic for x_4 increases steadily in significance throughout much of the search, lying outside the 99% region for several values of m just before inclusion of the final observations appreciably reduces its significance. We accordingly remove x_8 from the model.

Fig. 5 is the forward plot of added-variable t statistics for this model including four explanatory variables and the trend. As the figure shows, all variables and the trend are either significant at the end of the search or have been so for a part of the search just before the inclusion of the last observations. This then is our final model, with a logged response, the five variables shown in the plot and, of course, a constant term. This has been highly significant throughout and so has

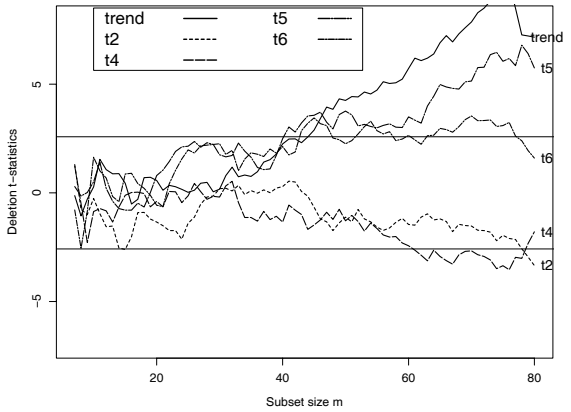

Figure 5: Logged ozone data: forward plot of added-variable t statistics; horizontal band contains 99% of the normal distribution. All five terms are either significant at the 1% level at the end of the search or have been so earlier

not been included on the plots. The final value of R^2 for this fitted model is 0.67, an appreciable improvement on the value of 0.43 for the first model we fitted.

3.5 Null Distribution of t Statistics

We now consider the exact distribution of the added-variable t statistics that we have judged in our plots against the normal distribution. In our searches we fit the reduced model $\mathrm{E}(y) = X\beta$, 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 $S^*(m)$ depending only on y and X. But the added-variable test (10) is a function solely of the residuals $\overset{*}{w}$ and $\overset{*}{y}$ which 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 is Student's t on $m-p$ degrees of freedom. Consequently, for small values of m, the percentage points of the null distribution forming the envelope will be slightly greater than those in our figures. Fig. 1 of [5] shows envelopes for an example with $n = 108$ and $p = 5$. This figure shows what is known from the properties of the t distribution, that for degrees of freedom ν greater than around 20, there is no practical difference between the normal and t distributions. The distribution is confirmed by simulation envelopes from 10,000 forward searches that are indistinguishable from those from the t distribution. The conclusion is that the parallel sided bands in our figures provide an excellent guide to significance.

4. Transformation of the Response

The ozone 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 [2] for the value of the transformation parameter λ in the Box-Cox [10] family of normalized power transformations

$$z(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda \dot{y}^{\lambda-1} & \lambda \neq 0 \\ \dot{y} \log y & \lambda = 0, \end{cases} \qquad (15)$$

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 (10)is replaced by a constructed variable which, provided X includes a constant, can be written

$$w(\lambda) =$$

$$\begin{cases} y^\lambda\{\log(y/\dot{y}) - 1/\lambda\}/(\lambda \dot{y}^{\lambda-1}) & \lambda \neq 0 \\ \dot{y} \log y(0.5 \log y - \log \dot{y}) & \lambda = 0. \end{cases} \qquad (16)$$

In addition, the response y in the regression model (2) is replaced by the normalized transformed response $z(\lambda)$. The test of a particular value λ_0 of the transformation parameter is the t test for the significance of the constructed variable $w(\lambda_0)$ in a regression in which the explanatory variables X are also included. In this application the added variable formulation of the regression model arises naturally, with no need to select successive columns of Q in (1).

Chapter 4 of [4] gives examples of the use of forward plots of this constructed variable test for transformations. Although forward

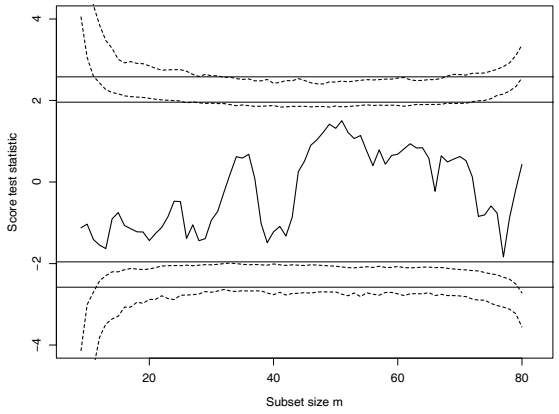

Figure 6: Logged ozone data: forward plot of constructed-variable t statistics. Horizontal bands contains 95 and 99% of the normal distribution. The curved simulation envelopes come from 20,000 simulations

plots of the test statistic are easily interpreted, the statistic cannot have exactly a t distribution; the constructed variable (16) 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 (10) do not hold. The statistic depends on the properties of the residuals of these varailes, that is of $z(\lambda_0)$ and $w(\lambda_0)$, the correlation between which depends on the projection matrix A. Plots and simulation are used by [6] to investigate the effect of this projection on the distribution of the test statistic in the forward search.

Fig. 6 shows the forward plot of t test for the constructed variable for the logarithmic transformation of the ozone data. At the end of the search this lies in the centre of the confidence region with a value of 0.43. The figure also shows that the last observations to enter the search, particularly the last two, do have a noticeable effect on the values of the test statistic. However, the significance of the value is not changed. Also given in the figure are bootstrap bounds from 20,000 simulations. The figure shows that the normal approximation to the distribution hold well in the centre of the region, but less well at the end, where what [6] call a "trumpet effect" is evident.

The simulation of these bands is not completely straightforward. In regression, to check the bands of the earlier figures, the observations can be simulated as coming from a standard normal distribution, in effect taking the linear parameters β as zero and the error variance σ^2 as one, the plots depending only on least squares residuals that are invariant to these values. However, in simulating data for checking the transformation of a particular set of data, we use a model fitted to the particular transformation that we want to test. In addition, if outliers are present, we exclude them from the data used for parameter estimation. Here, since the last few observations were not strongly outlying, we did not exclude them. The simulated envelopes in Fig. 6 show the typical breadth at the beginning of the search which results from estimation of σ^2 from few observations and so to the statistic having a t distribution on few degree of freedom. The trumpet effect at the end of the search is shown by [6] to depend on the value of R^2 in the regression. If this is high, the values of the residuals $\overset{*}{z}(\lambda_0)$ and $\overset{*}{w}(\lambda_0)$ are almost independent and the normal distribution provides a good approximation at the end of the search. If R^2 is low, the distribution is affected by the presence of y in both z and w. Here, with an R^2 value of 0.67 there is some broadening of the envelopes that makes the value of the statistic less significant. However, the effect is not enough to change our inferences based on the normal theory envelopes.

This concludes our analysis of the ozone data. Since the data are in time order, it is not surprising that we found it necessary to include a time trend. A function of day of the year for all 300 observations was found by [11] which rises sharply to a peak near day 120 and then declines more slowly to reach the initial value towards the end of the year. Initially their function is almost linear and so matches our linear trend. Over a longer time period the set of functions chosen for the effect of time should be cyclical. Sines and cosines are a natural first choice.

5. AIC and Model Selection

We have augmented the standard proce-

dure of backward elimination of regression variables with a forward search for each considered model. This backward procedure leaves unexplored the vast majority of models found by dropping each variable in turn. The comparison of this large number of models often uses a model selection criterion such as AIC [1] in which the increase in the loglikelihood of the model which comes from fitting extra terms is penalized by a multiple of the number of extra parameters. For regression models this procedure is identical to use of Mallows C_p [14], a function solely of an aggregate statistic for each model, in this case the residual sum of squares. The extension of our forward procedure to determine the effect of individual observations on model selection raises appreciable problems in the cogent presentation of the large amount of information that can be generated.

We have treated determination of the response transformation as a separate matter from model building. However, there is often a relationship between the fitted model and the best transformation. One possibility is, for each fitted model, to include the score test for transformations on the same plot as the added-variable t tests, in effect combining Fig. 6 with, for instance, Fig. 5. We would want the explanatory variables to be significant, but not the constructed variable. In model selection inclusion of the constructed variable for the transformation could be penalised in the same way as the inclusion of any other explanatory variable.

6. Envelopes and Multivariate Data

Our analysis of the ozone data relied heavily on the assessment of observed values of statistics in the forward search by reference to normal or t envelopes. For multivariate data [8] use forward plots of a variety of Mahalanobis distances to detect outliers and clusters in data. In some of their examples the structure of the data is revealed without the need for reference distributions. If a reference distribution is necessary, for example in formal testing for outliers or clusters, the best procedure depends on the dimensions of the problem. For small datasets we can use envelopes from bootstrap simulations to determine the threshold of statistics during the forward search. For moderate sized datasets we can use instead the polynomial approximations of [7]. For large samples [9] rescale a paradigmatic curve obtained by simulation to have the correct sample size and number of variables. A further possibility is use of the bounds of [16] derived from arguments using properties of order statistics and trimmed estimators.

References

[1] Akaike, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974; 19: 716-723.

[2] Atkinson, A. C. Testing transformations to normality. Journal of the Royal Statistical Society, Series B 1973; 35: 473-479.

[3] Atkinson, A. C. Plots, Transformations, and Regression. Oxford: Oxford University Press; 1985.

[4] Atkinson, A. C. and Riani, M. Robust Diagnostic Regression Analysis. New York: Springer–Verlag; 2000.

[5] Atkinson, A. C. and Riani, M. Forward search added variable t tests and the effect of masked outliers on model selection. Biometrika 2002; 89: 939-946.

[6] Atkinson, A. C. and Riani, M. Tests in the fan plot for robust, diagnostic transformations in regression. Chemometrics and Intelligent Laboratory Systems 2002; 60: 87-100.

[7] Atkinson, A. C. and Riani, M. Exploratory tools for clustering multivariate data. Computational Statistics and Data Analysis 2007; doi:10.1016/j.csda.2006.12.034.

[8] Atkinson, A. C., Riani, M. and Cerioli, A. Exploring Multivariate Data with the Forward Search. New York: Springer–Verlag; 2004.

[9] Atkinson, A. C., Riani, M. and Laurini, F. Approximate envelopes for finding an unknown number of multivariate outliers in large data sets. 2007. (Submitted).

[10] Box, G. E. P. and Cox, D. R. An analysis of transformations (with discussion). Journal of the Royal Statistical Society, Series B 1964; 26: 211-246.

[11] Breiman, L. and Friedman, J. H. Estimating optimal transformations for multiple regression and transformation (with discussion). Journal of the American Statistical Association 1985; 80: 580-619.

[12] Cook, R. D. and Weisberg, S. Residuals and Influence in Regression. London: Chapman and Hall; 1982.

[13] Cook, R. D. and Weisberg, S. An Introduction to Regression Graphics. New York: Wiley; 1994)

[14] Mallows, C. L. Some comments on C_p. Technometrics 1973; 15: 661-675.

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

[16] Riani, M., Atkinson, A. C. and Cerioli, A. Finding an unknown number of multivariate outliers in large data sets. 2008. (Submitted).

[17] Rousseeuw, P. J. Least median of squares regression. Journal of the American Statistical Association 1984; 79, 871-880.

[18] Rousseeuw, P. J. and Leroy, A. M. Robust Regression and Outlier Detection. New York: Wiley; 1987.