Regression Diagnostics and the Forward Search 2

A. C. Atkinson, London School of Economics

March 2 2009

Evidence for transformation of the response in regression often depends on observations that are ill-fitted by the model for untransformed data. Such observations appear to be outliers when the wrong model is fitted. We start by comparing analyses of the same data transformed and not. We then show how the FS can provide evidence about whether the data should be transformed.

Choosing whether to transform the response is only one aspect of building a statistical model. Section 4 describes the combination of the FS with added variable t tests to determine which terms to include in a regression model. The FS is then extended to determine the influence of individual observations in the more general case that Mallows' C_p is used to choose the model.

3 Transformations to Normality in Regression

3.1 Wool Data

The wool data, taken from Box and Cox (1964), give the number of cycles to failure of a worsted yarn under cycles of repeated loading. The number of cycles to failure (a non-negative response) ranges from 90, for the shortest specimen subject to the most severe conditions, to 3,636 for observation 19 which comes from the longest specimen subjected to the mildest conditions. In their analysis Box and Cox (1964) recommend that the data be fitted after the log transformation of y. We compare analyses of the transformed and untransformed data, to show some of the information provided by the forward search.

Figure 20(a) shows, for the untransformed data, the plot of least squares residuals e against fitted values $\hat{y}$. There is appreciable structure in this plot,

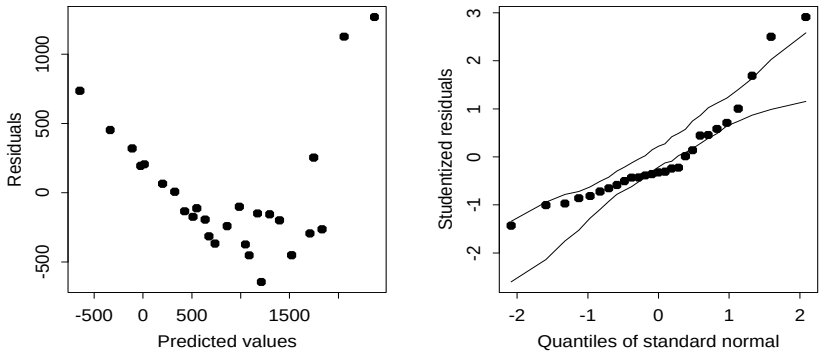

Figure 20: Wool data: (a) least squares residuals e against fitted values $\hat{y}$; (b) normal QQ plot of studentized residuals

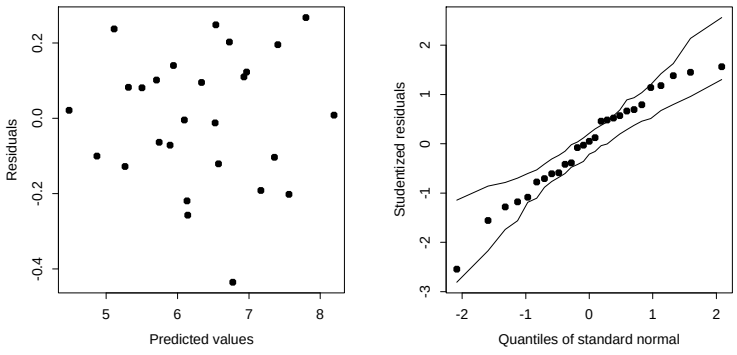

Figure 21: Transformed wool data: residual plots for log y: (a) least squares residuals against fitted values; (b) normal QQ plot of studentized residuals

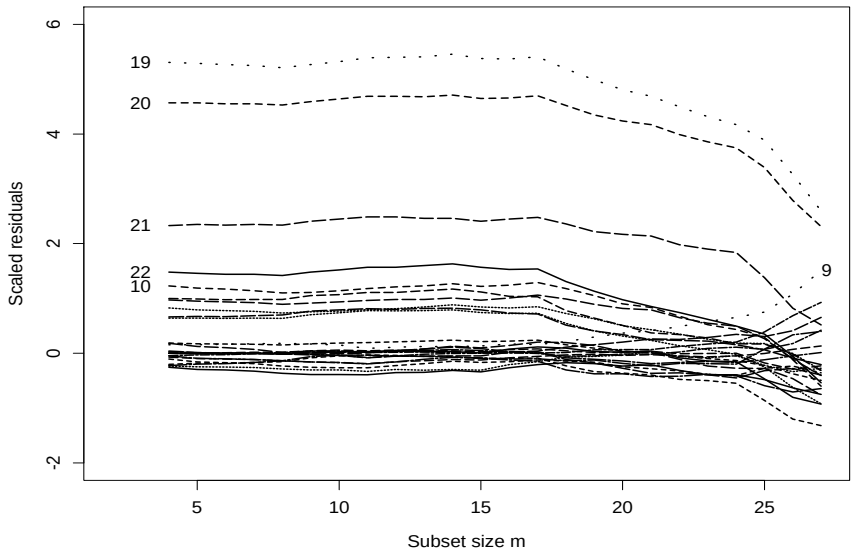

Figure 22: Wool data: forward plot of least squares residuals scaled by the final estimate of σ. The four largest residuals can be directly related to the levels of the factors

unlike Figure 21(a) for the log transformed data which is without structure, as it should be if model and data agree. The right-hand panels of the figures are normal QQ plots. That for the transformed data is an improvement, although there is perhaps one too large negative residual, which however lies within the simulation envelope of the studentized residuals in panel (b). This plot is also much better behaved than its counterpart being much more nearly a straight line. We now consider the results of our forward searches for these data.

The forward plot of scaled residuals for the untransformed data is in Figure 22 with that for the transformed data in Figure 23. We have already noted the four large residuals in the plot for the untransformed data and the activity towards the end of the search. The plot for the transformed data seems both more stable and more symmetrical, although observations 24 and 27 initially have large residuals. Do these observations have any effect on the selection of the logarithmic transformation?

3.2 Transformation of the Response

The logarithmic is just one possible transformation of the data. Might the square root or the reciprocal be better? We describe the parametric family of power transformations introduced by Box and Cox that combines such

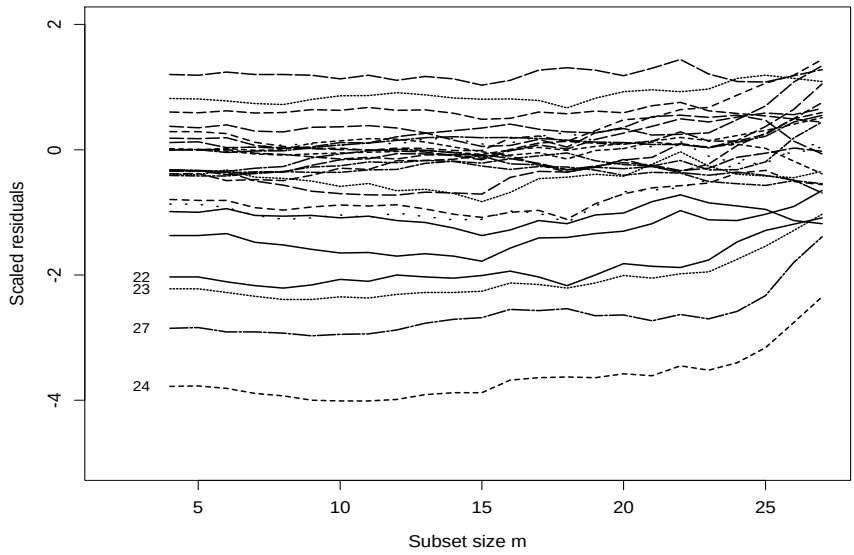

Figure 23: Transformed wool data: forward plot of least squares residuals for $\log y$ scaled by the final estimate of σ. Are observations 24 and 27 important in the choice of transformation?

transformations in a single family.

For transformation of just the response y in the linear regression model, Box and Cox (1964) analyze the normalized power transformation

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

where the geometric mean of the observations is written as $\dot{y} = \exp(\Sigma \log y_i / n)$. The model fitted is multiple regression with response $z(\lambda)$; that is,

$$z(\lambda) = X\beta + \epsilon. \tag{2}$$

When $\lambda = 1$, there is no transformation: $\lambda = 1/2$ is the square root transformation, $\lambda = 0$ gives the log transformation and $\lambda = -1$ the reciprocal. For this form of transformation to be applicable, all observations need to be positive. For it to be possible to detect the need for a transformation the ratio of largest to smallest observation should not be too close to one.

The intention is to find a value of λ for which the errors in the $z(\lambda)$ (2) are, at least approximately, normally distributed with constant variance and for which a simple linear model adequately describes the data. This is attempted by finding the maximum likelihood estimate of λ, assuming a normal theory linear regression model.

Once a value of λ has been decided upon, the analysis is the same as that using the simple power transformation

$$y(\lambda) = \begin{cases} (y^\lambda - 1)/\lambda & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases} . \tag{3}$$

However the difference between the two transformations is vital when a value of λ is being found to maximize the likelihood, since allowance has to be made for the effect of transformation on the magnitude of the observations.

The likelihood of the transformed observations relative to the original observations y is

$$(2\pi\sigma^2)^{-n/2} \exp\{-(y(\lambda) - X\beta)^T(y(\lambda) - X\beta)/2\sigma^2\} J,$$

where the Jacobian

$$J = \prod_{i=1}^{n} \left| \frac{\partial y_i(\lambda)}{\partial y_i} \right| \tag{4}$$

allows for the change of scale of the response due to transformation

For the power transformation (3),

$$\frac{\partial y_i(\lambda)}{\partial y_i} = y_i^{\lambda - 1}$$

so that

$$\log J = (\lambda - 1) \sum \log y_i = n(\lambda - 1) \log \dot{y}.$$

The maximum likelihood estimates of the parameters are found in two stages. For fixed λ the likelihood (4) is maximized by the least squares estimates

$$\hat{\beta}(\lambda) = (X^T X)^{-1} X^T z(\lambda),$$

with the residual sum of squares of the $z(\lambda)$,

$$R(\lambda) = z(\lambda)^T (I - H) z(\lambda) = z(\lambda)^T A z(\lambda). \tag{5}$$

Division of (5) by n yields the maximum likelihood estimator of σ^2 as

$$\hat{\sigma}^2(\lambda) = R(\lambda)/n.$$

For fixed λ we find the loglikelihood maximized over both β and λ by substitution of $\hat{\beta}(\lambda)$ and $s^2(\lambda)$ into (4) and taking logs. If an additive constant is ignored this partially maximized, or profile, loglikelihood of the observations is

$$L_{\max}(\lambda) = -(n/2) \log\{R(\lambda)/(n - p)\} \tag{6}$$

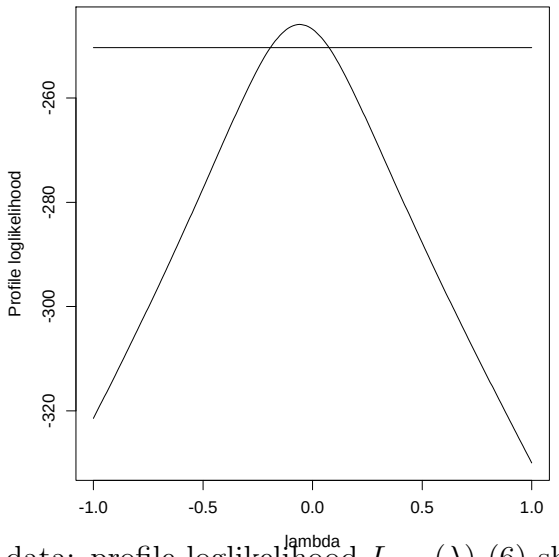

Figure 24: Wool data: profile loglikelihood $L_{\max}(\lambda)$ (6) showing the narrow 95% confidence interval for λ

so that $\hat{\lambda}$ minimizes $R(\lambda)$. For inference about the transformation parameter λ, Box and Cox suggest likelihood ratio tests using (6), that is, the statistic

$$T_{LR} = 2\{L_{\max}(\hat{\lambda}) - L_{\max}(\lambda_o)\} = n \log\{R(\lambda_0)/R(\hat{\lambda})\}. \tag{7}$$

Figure 24 is a plot of the profile loglikelihood $L_{\max}(\lambda)$, (6). It provides very strong evidence for the log transformation, with the maximum likelihood estimate $\hat{\lambda}$ equal to -0.059. The horizontal line on the plot at a value of $L_{\max}(\hat{\lambda})$ $-3.84/2$ cuts the curve of the profile loglikelihood at -0.183 and 0.064, providing an approximate 95% confidence region for λ. This plot, depending as it does solely on the value of the residual sum of squares $R(\lambda)$, is of course totally uninformative about the contribution of individual observations to the transformation.

To find a test statistic that can readily reflect the contribution of individual observations, we first require some theory from regression on the effect of the addition of an extra variable to a regression model.

3.3 Added Variables

The added-variable plot provides a method, in some circumstances, of assessing the impact of individual observations on estimates $\hat{\beta}_k$ of single parameters in a multiple regression model. The starting point is to fit a model including all variables except the one of interest, the "added" variable. The plot is based on residuals of the response and of the added variable. To test for

transformations the added variable is replaced by a "constructed" variable derived from the data.

We extend the regression model to include an extra explanatory variable, the added variable w, so that

$$E(Y) = X\beta + w\gamma, \tag{8}$$

where γ is a scalar. The least squares estimate $\hat{\gamma}$ can be found explicitly from the normal equations for this partitioned model

$$X^T X\hat{\beta} + X^T w\hat{\gamma} = X^T y \tag{9}$$

and

$$w^T X\hat{\beta} + w^T w\hat{\gamma} = w^T y. \tag{10}$$

If the model without γ can be fitted, $(X^T X)^{-1}$ exists and (9) yields

$$\hat{\beta} = (X^T X)^{-1} X^T y - (X^T X)^{-1} X^T w\hat{\gamma}. \tag{11}$$

Substitution of this value into (10) leads, after rearrangement, to

$$\hat{\gamma} = \frac{w^T(I-H)y}{w^T(I-H)w} = \frac{w^T Ay}{w^T Aw}. \tag{12}$$

Since $A = (I-H)$ is idempotent, $\hat{\gamma}$ can be expressed in terms of the two sets of residuals

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

and

$$\overset{*}{w} = (I-H)w = Aw \tag{13}$$

as

$$\hat{\gamma} = \overset{*}{w}^T e/(\overset{*}{w}^T \overset{*}{w}). \tag{14}$$

Thus $\hat{\gamma}$ is the coefficient of linear regression through the origin of the residuals e on the residuals of the new variable w, both after regression on the variables in X.

Because the slope of this regression is $\hat{\gamma}$, a plot of e against $\overset{*}{w}$ is often used as a visual assessment of the evidence for a regression and for the assessment of the contribution of individual observations to the relationship. Such a plot is called an added variable plot.

3.4 Constructed Variables for Transformations

In the likelihood ratio test (7) a numerical maximization is required to find the value of $\hat{\lambda}$. This is cumbersome when calculating deletion diagnostics or when using the FS, since a maximization is required for each subset of interest. For regression models a computationally simpler alternative test is the extension of the added variable method to yield an approximate score statistic derived by Taylor series expansion of (1) as

$$\begin{aligned} z(\lambda) &\doteq z(\lambda_0) + (\lambda - \lambda_0) \left. \frac{\partial z(\lambda)}{\partial \lambda} \right|_{\lambda=\lambda_0} \\ &= z(\lambda_0) + (\lambda - \lambda_0) w(\lambda_0), \end{aligned} \tag{15}$$

which only requires calculations at the hypothesized value λ_0. In (15) $w(\lambda_0)$ is the constructed variable for the transformation. Differentiation of $z(\lambda)$ for the normalized power transformation yields

$$\begin{aligned} w(\lambda) &= \frac{\partial z(\lambda)}{\partial \lambda} \\ &= \frac{y^\lambda \log y}{\lambda \dot{y}^{\lambda-1}} - \frac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} (1/\lambda + \log \dot{y}). \end{aligned} \tag{16}$$

The combination of (15) and the regression model $y = x^T\beta + \epsilon$ leads to the model

$$\begin{aligned} z(\lambda_0) &= x^T\beta - (\lambda - \lambda_0) w(\lambda_0) + \epsilon \\ &= x^T\beta + \gamma\ w(\lambda_0) + \epsilon, \end{aligned} \tag{17}$$

where $\gamma = -(\lambda - \lambda_0)$, which is of the form of (8). The two sets of residuals in the constructed variable plot, analogously to (17) are

$$\overset{*}{z}(\lambda) = (I - H) z(\lambda) = A z(\lambda)$$

and

$$\overset{*}{w}(\lambda) = (I - H) w(\lambda) = A w(\lambda). \tag{18}$$

If $\hat{\lambda}$ is close to λ_0, γ will be close to zero and there will be no significant slope in the constructed variable plot.

As an example, Figure 25 is the constructed variable plot for the wool data when $\lambda = 1$. With its positive slope, the plot shows clear evidence of the need for a transformation, evidence which seems to be supported by all the data. The most influential points seem to be observations 20 and 19, which are the two largest observations and 9, 8, 7 and 6, which are the four smallest. The sequential nature of these sets of numbers reflects that the

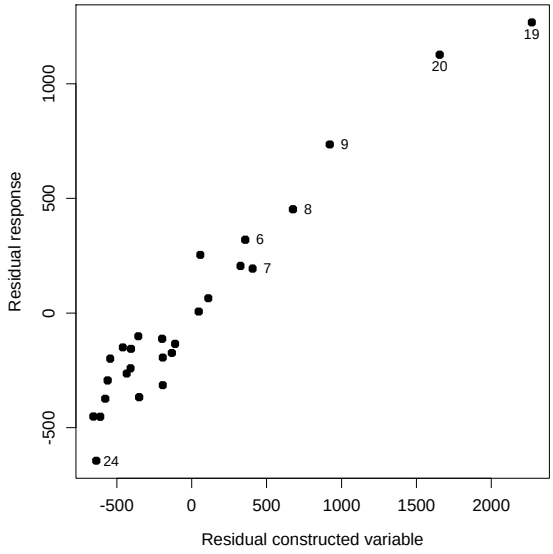

Figure 25: Wool data: constructed variable plot for $\lambda = 1$. The clear slope in the plot indicates that a transformation is needed. The largest observations are 19 and 20: the labelled points in the centre of the plot have the four smallest values of y

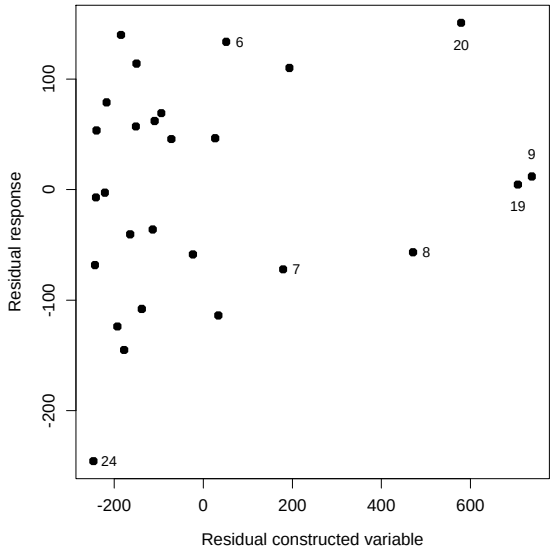

Figure 26: Wool data: constructed variable plot for $\lambda = 0$. The absence of trend indicates that the log transformation is satisfactory

data are from a factorial experiment and are presented in standard order. The contrasting constructed variable plot for $\lambda = 0$ is in Figure 26. There is no trend in the plot and the transformation seems entirely acceptable. The residuals from the six observations that were extreme in the previous plot now lie within the general cloud of points.

However the plot is one of residuals against residuals. As we have already argued, points of high leverage tend to have small residuals. Thus, if something important to the regression happens at a leverage point, it will often not show on the plot. Examples, for the constructed variable for transformation of the response, are given by Cook and Wang (1983) and by Atkinson (1985, §12.3). Instead of the plot, these authors suggest looking at the effect of individual observations on the t test for γ.

3.5 Approximate Score Test for Transformations

The approximate score statistic $T_p(\lambda_0)$ for testing the transformation is the t statistic for regression on $w(\lambda_0)$ in (17). This can either be calculated directly from the regression in (17), or from the formulae for added variables in §3.3 in which multiple regression on x is adjusted for the inclusion of an additional variable. The t test for $\gamma = 0$ is then the test of the hypothesis $\lambda = \lambda_0$. To make explicit the dependence of both numerator and denominator of the test statistic on λ we can write our special case of (14) as

$$\hat{\gamma}(\lambda) = \overset{*}{w}^T(\lambda)\,\overset{*}{z}(\lambda)/\{\overset{*}{w}^T(\lambda)\,\overset{*}{w}(\lambda)\}.$$

The approximate score test for transformations is thus

$$\begin{aligned} T_p(\lambda) &= -\frac{\hat{\gamma}(\lambda)}{\sqrt{s_w^2(\lambda)/\{w^T(\lambda)Aw(\lambda)\}}} \\ &= -\frac{\hat{\gamma}(\lambda)}{\sqrt{s_w^2(\lambda)/\{\overset{*}{w}^T(\lambda)\,\overset{*}{w}(\lambda)\}}}. \end{aligned} \tag{19}$$

The negative sign arises because in (17) $\gamma = -(\lambda - \lambda_0)$. The mean square estimate of σ^2 can be written in the form

$$(n-p-1)s_w^2(\lambda) = \overset{*}{z}^T(\lambda)\,\overset{*}{z}(\lambda) - \{\overset{*}{z}^T(\lambda)\,\overset{*}{w}(\lambda)\}^2/\{\overset{*}{w}^T(\lambda)\,\overset{*}{w}(\lambda)\}.$$

These formulae show how $\hat{\gamma}$ is the coefficient for regression of the residuals $\overset{*}{z}$ on the residuals $\overset{*}{w}$, both being the residuals from regression on X. If, as is usually the case, X contains a constant, any constant in $w(\lambda)$ can be disregarded in the construction of $\overset{*}{w}$.

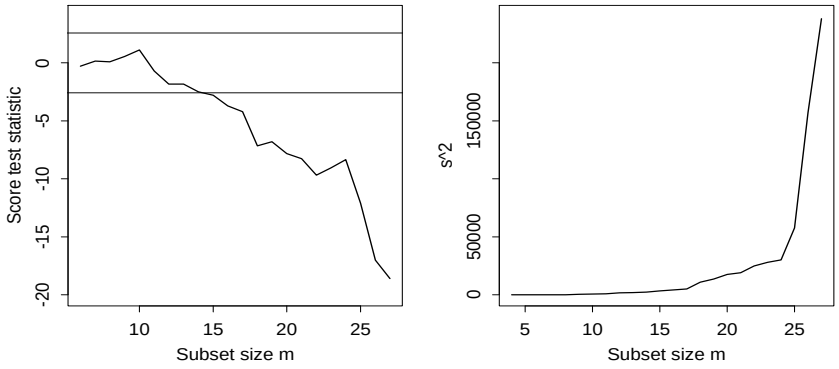

Figure 27: Wool data: (a) score test for transformation during the forward search and (b) the increasing value of the estimate s^2

The two most frequently occurring values of λ in the analysis of data are one and zero: either no transformation, the starting point for most analyses, or the log transformation.

3.6 The Fan Plot in the Forward Search

We monitor the value of $T_p(\lambda)$ during the forward search. Figure 27(a) is a plot for the untransformed wool data of the value of $T_p(1)$ during the forward search. The null distribution is approximately normal. If the data do not need transformation the values should lie within the 99% limits of ± 2.58 shown on the plot. However, the value of the statistic trends steadily downward, indicating that the evidence for a transformation is not confined to just the last few large observations, but that there are contributions from all observations. The negative value of the statistic indicates that a transformation such as the log or the reciprocal should be considered.

In contrast, Figure 28(a), is the forward plot of the approximate score statistic$T_p(1)$, that is for the log transformation, when the data are log transformed. The observations giving rise to large residuals, which enter at the end of the search, have no effect whatsoever on the value of the statistic. The plot of the parameter estimates in Figure 28(b) shows how stable the estimates of the parameters are during this forward search.

For data sets of this size we find it satisfactory to base our analyses on five values of $\lambda : -1, -0.5, 0, 0.5$ and 1. We perform five separate searches. The data are transformed and a starting point is found for each forward search,

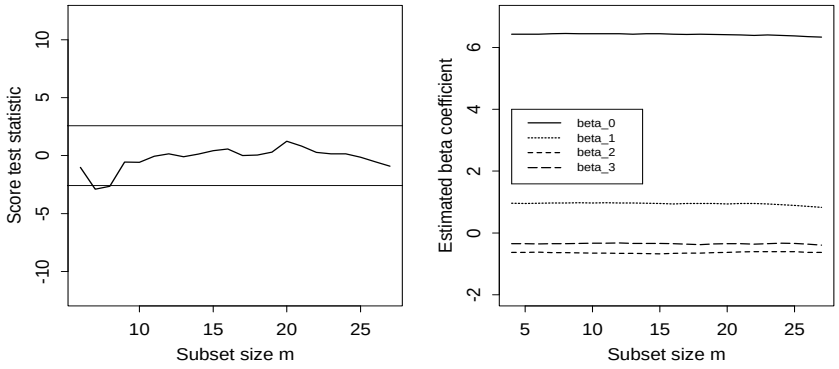

Figure 28: Transformed wool data: (a) score test for transformation during the forward search, showing that the log transformation is satisfactory and (b) the extremely stable values of the parameter estimates

which then proceeds independently for each λ using the transformed data. For the wool data we found the five initial subsets by exhaustive search of all subsets. Figure 29 shows the values of the approximate score statistic $T_p(\lambda)$ as the subset size m increases. The central horizontal bands on the figure are at ± 2.58, containing 99% of a standard normal distribution. For obvious reasons, we refer to this kind of forward plot as a fan plot.

Initially, apart from the very beginning when results may be unstable, there is no evidence against any transformation. When the subset size m equals 15 (56% of the data), $\lambda = 1$ is rejected. The next rejections are $\lambda = 0.5$ at 67% and -1 at 74%. The value of $\lambda = 0$ is supported not only by all the data, but also by our sequence of subsets. The observations added during the search depend on the transformation. In general, if the data require transformation and are not transformed, or are insufficiently transformed, large observations will appear as outliers. Conversely, if the data are overtransformed, small observations will appear as outliers. This is exactly what happens here. For $\lambda = 1$ and $\lambda = 0.5$, working back from $m = 27$, the last cases to enter the subset are 19, 20 and 21, which are the three largest observations. Conversely, for $\lambda = -1$ and $\lambda = -0.5$ case 9 is the last to enter, preceded by 8 and 7, which are the three smallest observations. For the log transformation, which produces normal errors, there is no particular pattern to the order in which the observations enter the forward search.

Similar results are obtained if $T_p(\lambda)$ is replaced by the signed square root

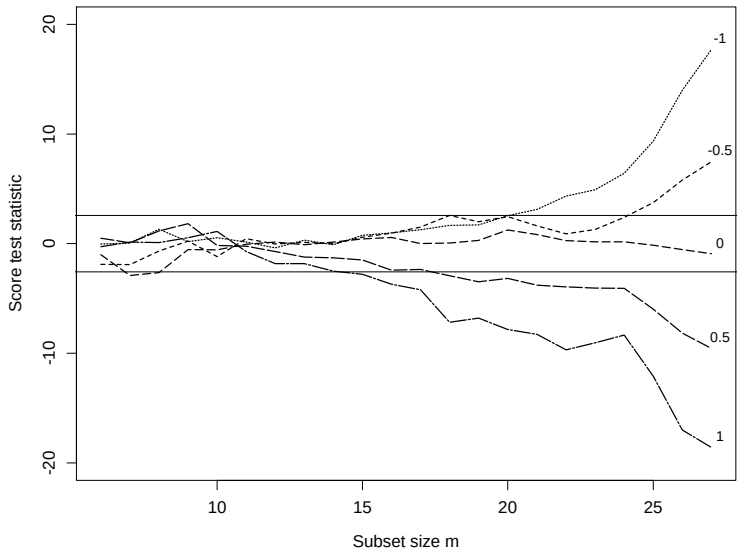

Figure 29: Wool data: fan plot – forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost; log y is indicated

of the likelihood ratio test (7).

3.7 Poison Data

The poison data from Box and Cox (1964) are like the wool data, well behaved: there are no outliers or influential observations that cannot be reconciled with the greater part of the data by a suitable transformation. Our fan plot and the other graphical procedures all clearly indicate the reciprocal transformation. We then consider a series of modifications of the data in which an increasing number of outliers is introduced. We show that the fan plot reveals the structure.

The data are the times to death of animals in a 3×4 factorial experiment with four observations at each factor combination. All our analyses use an additive model, that is, without interactions, so that $p = 6$, as did Box and Cox (1964) when finding the reciprocal transformation. The implication is that the model should be additive in death rate, not in time to death.

Our analysis is again based on five values of $\lambda : -1, -0.5, 0, 0.5$ and 1. The fan plot of the values of the approximate score statistic $T_p(\lambda)$ for each search as the subset size m increases is given in Fig 30 and shows that the log transformation is acceptable as is the inverse square root transformation ($\lambda = -0.5$). Initially, for small subset sizes, there is no evidence against

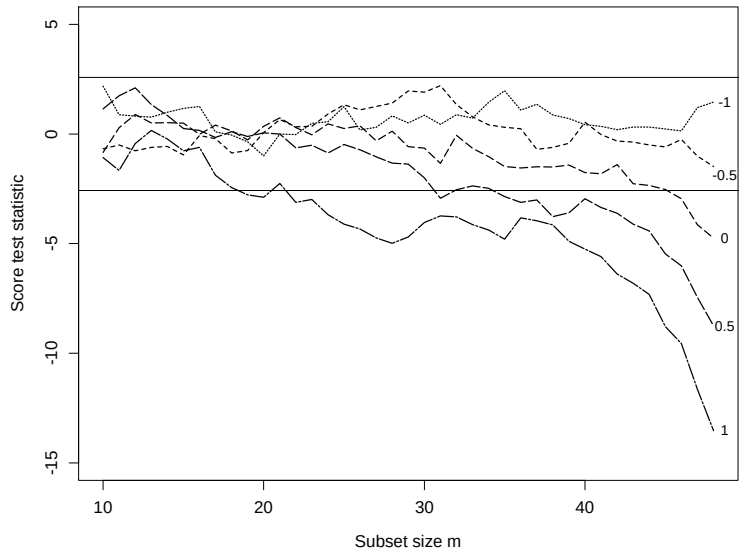

Figure 30: Poison data: fan plot – forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost: both $\lambda = -1$ and $\lambda = -0.5$ are acceptable

any transformation. During the whole forward search there is never any evidence against either $\lambda = -1$ or $\lambda = -0.5$ (for all the data $\hat{\lambda} = -0.75$). The log transformation is also acceptable until the last four observations are included by the forward search. The plot shows how evidence against the log transformation depends critically on this last 8% of the data. However, evidence that some transformation is needed is spread throughout the data, less than half of the observations being sufficient to reject the hypothesis that $\lambda = 1$.

3.8 Modified Poison Data

For the introduction of a single outlier into the poison data we change observation 8, one of the readings for Poison II, group A, from 0.23 to 0.13. This is not one of the larger observations so the change does not create an outlier in the scale of the original data. The effect on the estimated transformation of all the data is however to replace the reciprocal with the logarithmic transformation: $\hat{\lambda} = -0.15$. And, indeed, the fan plot of the score statistics from the forward searches in Figure 31 shows that, at the end of the forward search, the final acceptable value of λ is 0, with -0.5 on the boundary of the acceptance region.

But, much more importantly, Figure 31 clearly reveals the altered obser-

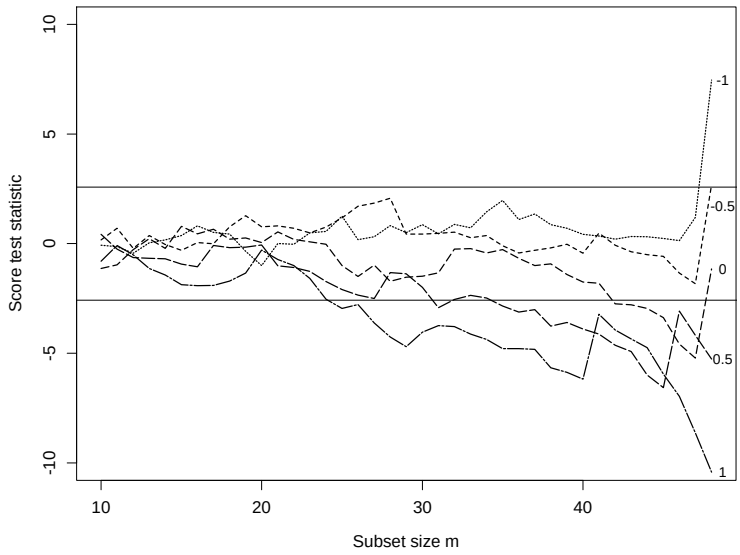

Figure 31: Modified poison data: fan plot – forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost: the effect of the outlier is evident in making $\lambda = 0$ appear acceptable at the end of the search

vation and the differing effect it has on the five searches. Initially the curves are the same as those of Figure 30. But for $\lambda = 1$ there is a jump due to the introduction of the outlier when $m = 41$ (85% of the data), which provides evidence for higher values of λ. For other values of λ the outlier is included further on in the search. When $\lambda = 0.5$ the outlier comes in at $m = 46$, giving a jump to the score statistic in favour of this value of λ. For the other values of λ the outlier is the last value to be included. Inclusion of the outlier has the largest effect on the inverse transformation. It is clear from the figure how this one observation is causing an appreciable change in the evidence for a transformation.

3.9 Doubly Modified Poison Data: An Example of Masking

The simplest example of masking is when one outlier hides the effect of another, so that neither is evident, even when single deletion diagnostics are used. As an example we further modify the poison data. In addition to the previous modification, we also change observation 38 (Poison I, group D) from 0.71 to 0.14.

For the five values of λ used in the fan plot the five values of the approximate score test for the transformation are:

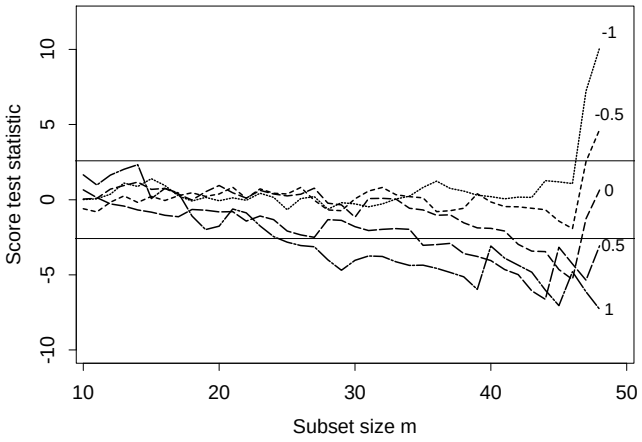

Figure 32: Doubly modified poison data: fan plot – forward plot of $T_p(\lambda)$ for five values of λ. The curve for $\lambda = -1$ is uppermost; the effect of the two outliers is clear

λ	-1	-0.5	0	0.5	1
$T_p(\lambda)$	10.11	4.66	0.64	-3.06	-7.27

It seems clear that the data support the log transformation and that all other transformations are firmly rejected. However, diagnostics based on the deletion of single observations fail to break the masking of the two outliers (see Chapter 4 of Atkinson and Riani 2000).

The effect of the two outliers is clearly seen in the fan plot, Figure 32. The plot also reveals the differing effect the two altered observations have on the five searches. Initially the curves are similar to those of the original data shown in Figure 30. The difference is greatest for $\lambda = -1$ where addition of the two outliers at the end of the search causes the statistic to jump from an acceptable 1.08 to 10.11. The effect is similar, although smaller, for $\lambda = -0.5$. It is most interesting however for the log transformation. Towards the end of the search this statistic is trending downwards, below the acceptable region. But addition of the last two observations causes a jump in the value of the statistic to a nonsignificant value. The incorrect log transformation is now acceptable.

For these three values of λ the outliers are the last two observations to be included in the search. They were created by introducing values that are too near zero when compared with the model fitted to the rest of the data. For the log transformation, and more so for the reciprocal, such values become

extreme and so have an appreciable effect on the fitted model. For the other values of λ the outliers are included earlier in the search. The effect is most clearly seen when $\lambda = 1$; the outliers come in at $m = 40$ and 46, giving upward jumps to the score statistic in favour of this value of λ. For the remaining value of 0.5 one of the outliers is the last value to be included.

3.10 Distributions in the Fan Plot: Wool Data

We now investigate the null distribution of $T_p(0)$. The score test is a t test for regression on a constructed variable which is however a function of the response. If this relationship between y and w is ignored, we would expect the score statistic to have a t distribution, apart from any effect of the ordering of observations due to the forward search. Fig.33 shows, for the logtransformed wool data, the forward plot of $T_p(0)$ during the forward search together with the results of 1,000 simulations when the data are generated with $\lambda = 0$ using the parameter estimates for this value of λ and all n observations. The simulated 90, 95 and 99 percentage points of the distribution of the statistic show that the distribution of the statistic starts with longer tails than the normal but that, by half-way through this search, the distribution is close to the asymptotic standard normal distribution. The agreement is very good until the end of the search when there is a slight spreading of the distribution, recalling the bell of a trumpet. This slightly larger variance when $m = n$ is in line with the simulation results of Atkinson and Lawrance (1989).

3.11 Trumpets and Constructed Variables

In this section we look at the relationship between the simulation envelopes and the parameter values used in the simulations. In envelopes of residuals in regression (Atkinson 1985, p.35) the linear model used for simulation does not matter and a sample of standard normal variables is used. What does matter is the hat matrix of the model fitted to the data, which affects the variances and covariances of the residuals. However, in transformation, the parameters of the linear model can also have an effect. We now study the effect of the dependence of the trumpet at the end of the search on the values of these parameters.

The effect of the trumpet was small in Fig.33, for which the squared multiple correlation coefficient R^2 had the value 0.97 at the end of the search. Analyses of other data sets with smaller values of R^2 gave plots that were more spread out at the end of the search.

Fig.34 shows, for the log transformed wool data, simulation envelopes in which the linear model has the same constant and value of s^2 as those in

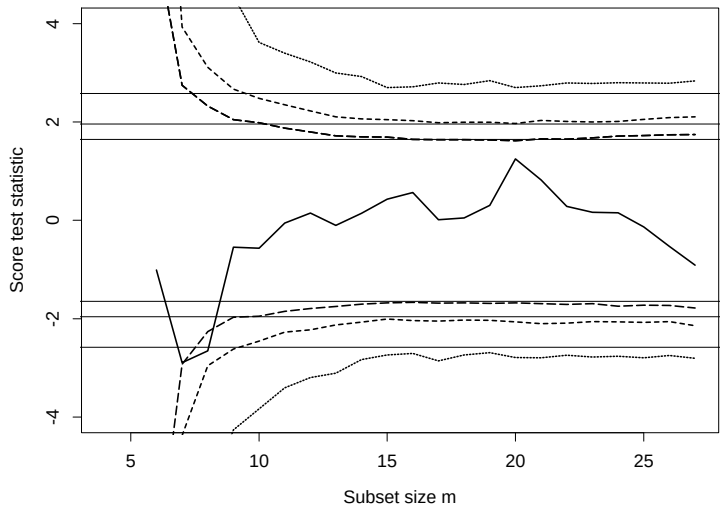

Figure 33: Logtransformed wool data: forward plot of $T_p(0)$ with 90%, 95% and 99% theoretical confidence bands and simulation envelopes using parameter estimates $\hat{\beta}(n)$ from the end of the search when $\lambda = 0$.

the data, but in which the values of the remaining three parameters b in the linear model have been divided by ten, that is, $b = \hat{\beta}(n)/10$. As a result, the average value of R^2 in the simulated data sets is reduced to 0.28. The effect on the simulation envelopes, compared to those in Fig.33 is clear. Although symmetrical, the envelopes are now too large throughout, especially towards the end of the search, where there is an appreciable trumpet.

Atkinson and Riani (2002b) show that the fact that a low value of R^2 accompanies wide simulation envelopes at the end of the search can be explained by considering the structure of the constructed variable plots, which are scatter plots of residual transformed response against the residual constructed variable. The score statistic $T_p(\lambda)$ is the t test for interceptless regression in this plot. In the absence of evidence for a transformation, this plot often looks like a random scatter of points. However, in simple cases, the plots can have a near parabolic structure, even when there is no evidence for a transformation. If there are several explanatory variables and strong regression the parabolic pattern disappears. However if there is weak regression, giving a low value of R^2, the parabolic structure remains, although in a distorted form.

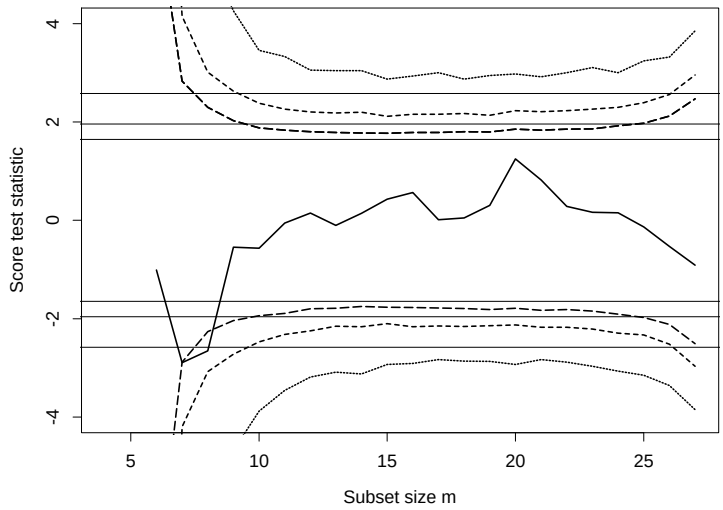

Figure 34: Logtransformed wool data: forward plot of $T_p(0)$ with 90%, 95% and 99% simulation envelopes using parameter estimates $b = \hat{\beta}(n)/10$. There is now a trumpet towards the end of the envelopes

4 Model Building

Monitoring the t tests for individual regression coefficients in the "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 which results in an increase of s^2 as m increases and a decrease in the values of the t statistics. We use an added variable test which has the desired properties since the projection leading to residuals destroys the effect of the ordering. The distribution of the test statistic is investigated and an example illustrates the effect of several masked outliers on model selection.

4.1 An Added Variable t Test

We now extend the added variable model of (8) from one variable to all the variables in the model in turn. We write the regression model for all n observations as

$$y = Q\theta + \epsilon = X\beta + w\gamma + \epsilon, \tag{20}$$

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.

4.2 Orthogonality and the Properties of the t Statistic

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.

On the contrary, 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. However, the results of §3.3 show that the added variable test 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. If the errors were normally distributed, the estimates $\hat{\gamma}$ and s^2 would be independent, and the null distribution of the statistic would be Student's t. Because, in the search, we are fitting truncated samples, the errors have slightly shorter tails than normal, albeit with no noticeable effect on the distribution of the statistic (§4.10).

4.3 Surgical Unit Data

Neter et al. (1996, pp.334 & 438) analyse 108 observations on the time of survival of patients who had a particular kind of liver surgery. There are four explanatory variables. The response is survival time. We follow Neter et al. (1996) and use the logarithm to base ten of time as the response.

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. 35. 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.

In the figure we have included horizontal lines to indicate significance levels. These are based on the normal distribution. Figure 36(a) repeats the curve for t_4 in Fig. 35 but with confidence limits calculated from the

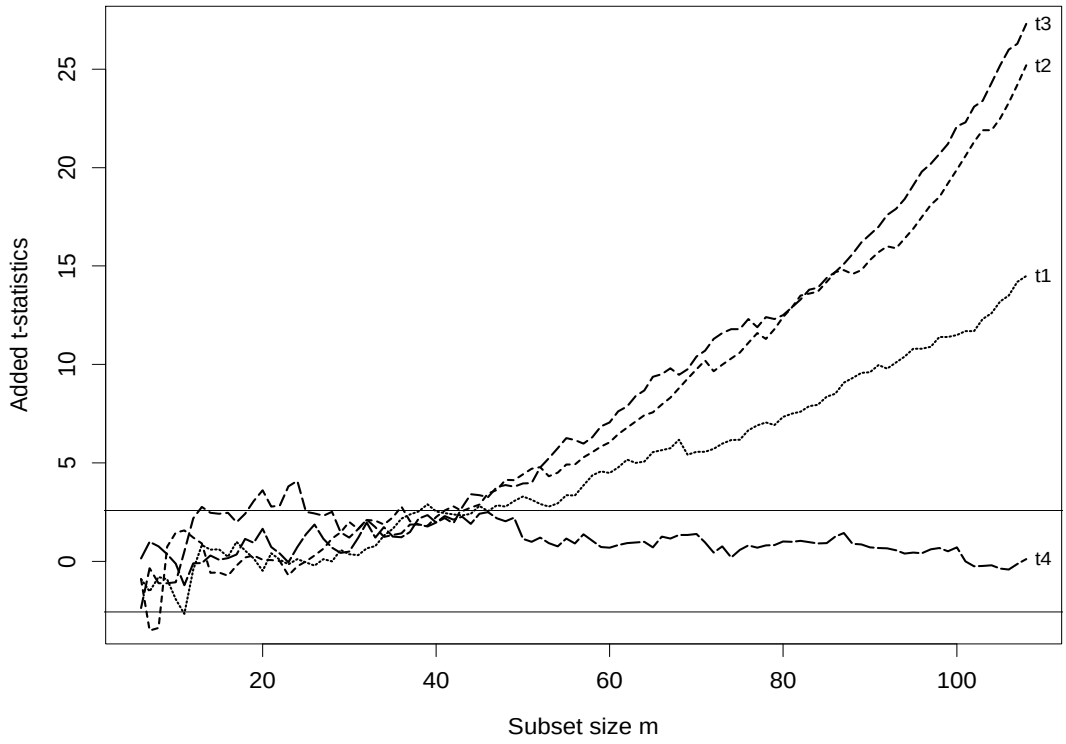

Figure 35: Transformed Surgical Unit data: forward plot of the four added-variable t statistics, t_1, t_2, t_3 and t_4.

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. The conclusion is that x_4 should be dropped from the model.

4.4 Multiple Outliers: 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. The model for these observations is

$$E(Y_+) = X_+\beta + w_+\gamma + \Delta, \tag{21}$$

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

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

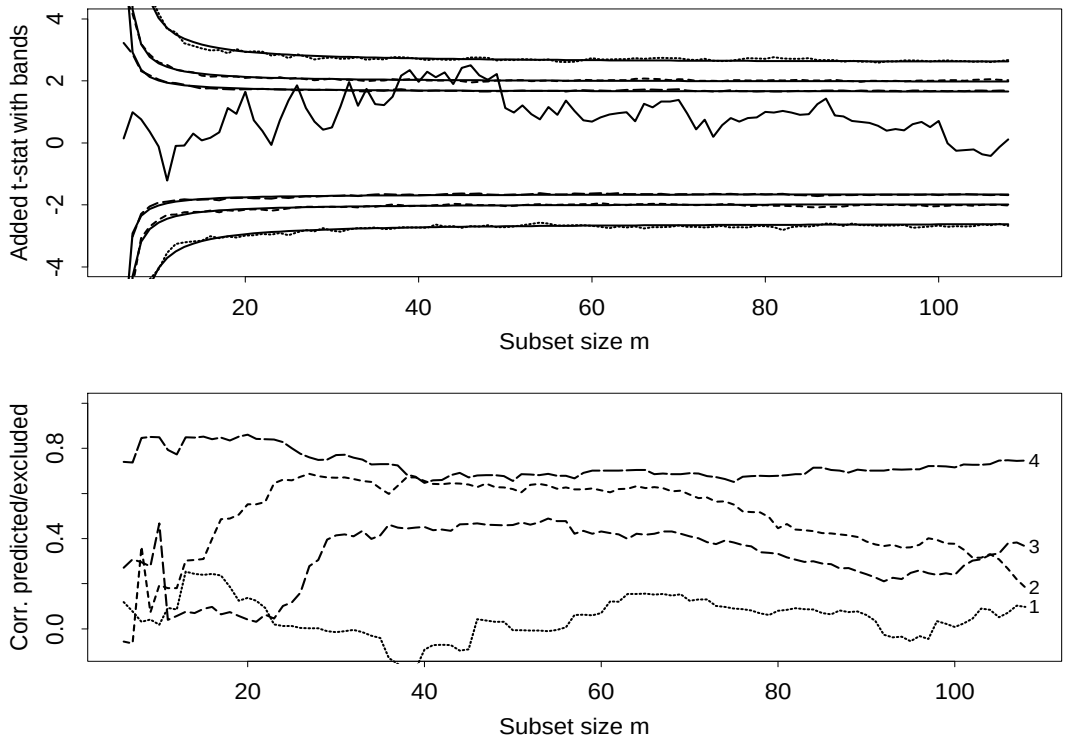

Figure 36: Transformed Surgical Unit data: (a) forward plot of added-variable t statistic for x_4, percentage points of the t distribution and averages of 10,000 simulations; (b) correlation between predictions from fitting X and the excluded variable **ignore**

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. Fig. 37 shows evidence of this decrease.

4.5 Modified Surgical Unit Data

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

We contaminate up to 12 observations in two different ways in order to produce two different effects. The resulting forward plots of the t tests are in Fig. 37. In Fig. 37(a) the effect of the modification has been to make x_1 non-significant; previously it was the most important variable. Since x_1 is the added variable, the search orders the observations using the regression

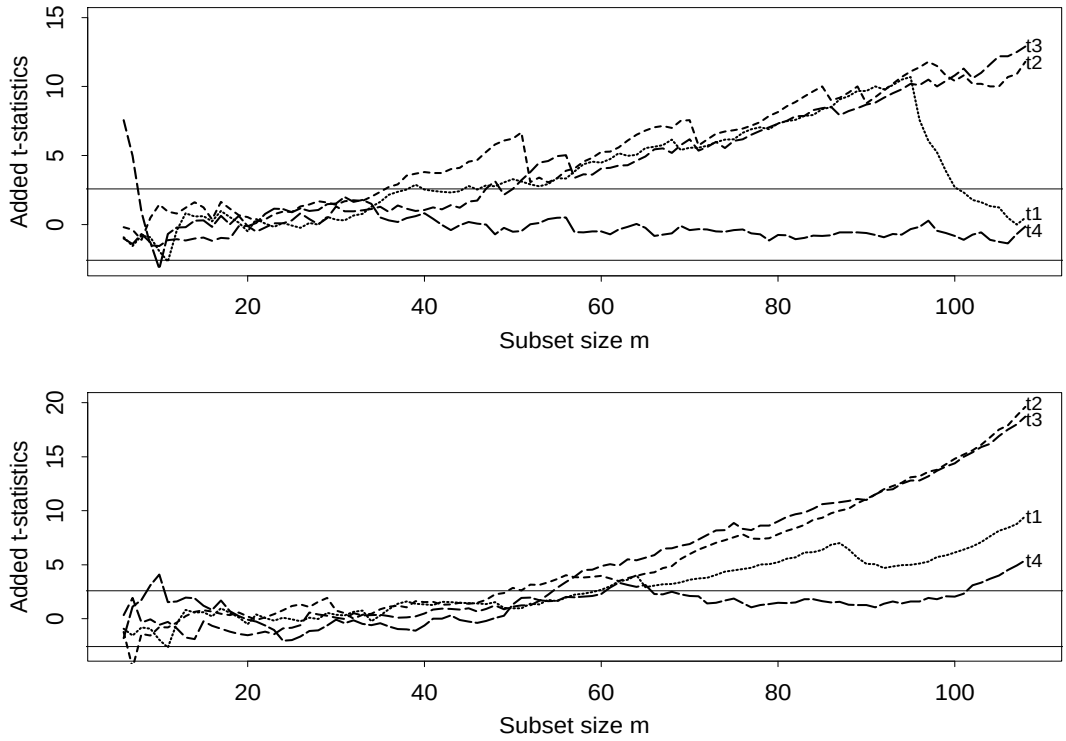

Figure 37: Modified Transformed Surgical Unit data: both panels show forward plots of added-variable t statistics, t_1, t_2, t_3 and t_4. (a) outliers render x_1 non-significant; (b) now the outliers make x_4 significant.

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.

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 x_1; in the second case it would lead to the incorrect inclusion of x_4 in the model.

The outliers are easily found using the forward plots of statistics, parameter estimates, Cook distances and the other diagnostic measures exemplified in Atkinson and 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. 37 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. For further discussion see Atkinson and Riani (2002a).

4.6 Ozone Data

With four explanatory variables the forward plot of added variable t statistics provided a clear indication of the model. However, with more variables, the situation can be less clear. One particular difficulty is that, with correlated explanatory variables, deletion of one variable can cause large changes in the values of the other t statistics.

As an illustration of this point, §3.4 of Atkinson and Riani (2000) presents 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 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 values of the non-negative response range from 2 to 24 and techniques like those of §3 indicate the log transformation. In addition, there is an upwards trend in the residuals from the fitted model with $\log y$ as response, so that 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. Figure 38 is the forward plot of added-variable t statistics for this model. The trend and x_5 are 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 virtually all 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 Figure 38 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

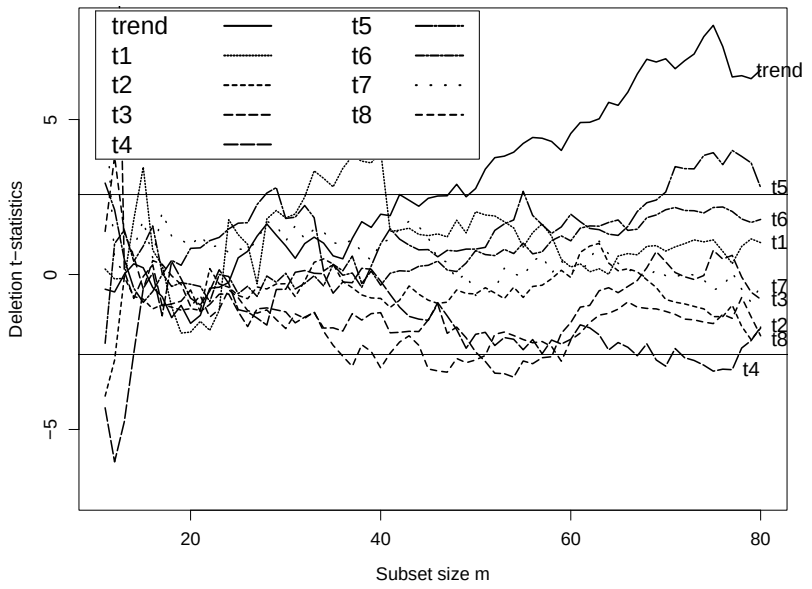

Figure 38: Logged ozone data: forward plot of added-variable $\boldsymbol{t}$ statistics; horizontal band contains 99% of the normal distribution. The trend and x_5 are most significant. The plot reflects overfitting

analysis parallels that on p. 70 of Atkinson and Riani (2000), 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 Figure 39.

At this point x_4 has the smallest t statistic, -1.64 and Atkinson and Riani (2000) next delete this variable. However, Figure 39 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.

Figure 40 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 not been included on the plots.

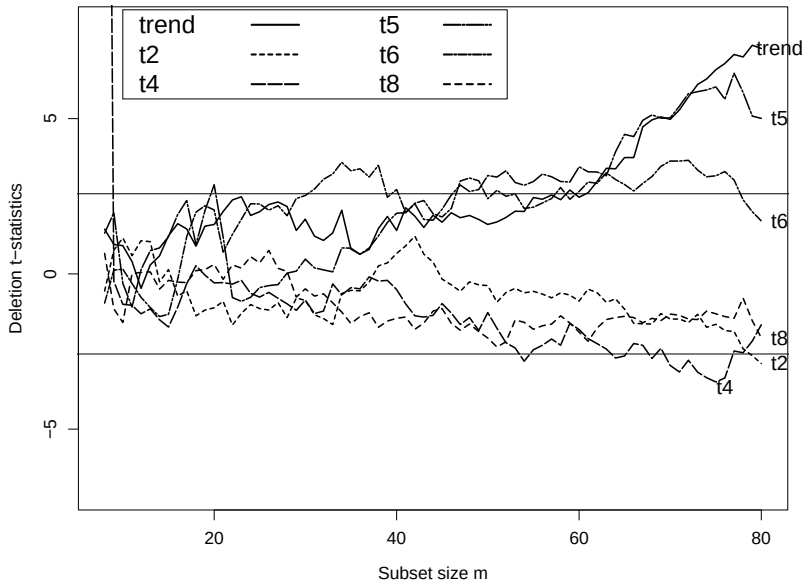

Figure 39: Logged ozone data: forward plot of added-variable t statistics; horizontal band contains 99% of the normal distribution. The least significant variable at the end of the search is x_4, but it is appreciably more significant than x_8 for most of the search.

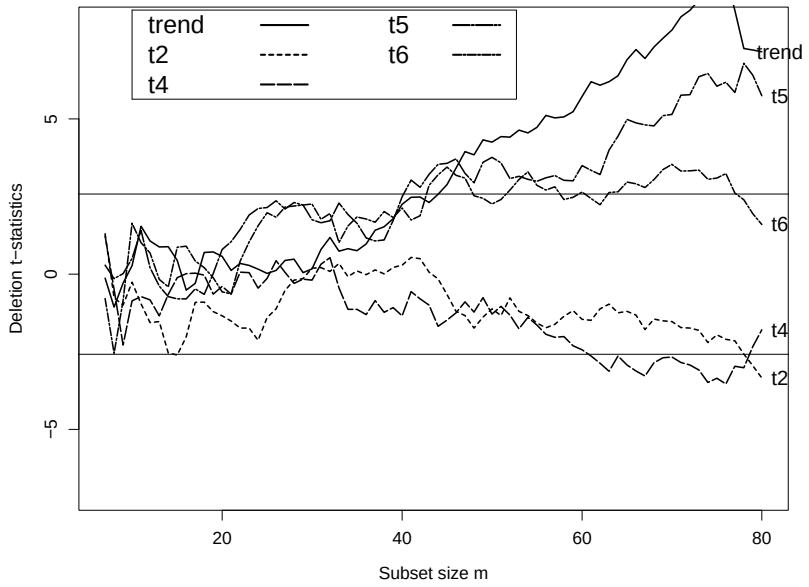

Figure 40: Logged ozone data: forward plot of added-variable t statistics; horizontal band contains 99% of the normal distribution. All five terms are either significant at the 1% level at the end of the search or have been so earlier

4.7 Aggregate Statistics: C_p

The analysis above augmented the standard procedure 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 Mallows C_p, 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 are interested in the linear multiple regression model $y = X\beta + \epsilon$, in which X is an $n \times p$ full-rank matrix of known constants, with ith row x_i^T. As before, the normal theory assumptions are that the errors ϵ_i are i.i.d. $N(0, \sigma^2)$. The residual sum of squares from fitting this model to the data is $R_p(n)$. The purpose is to compare various sets of explanatory variables and so various forms of the matrix X, over a range of values of p.

In the selection of regression variables using C_p, σ^2 is estimated from a large regression model with $n \times p^+$ matrix X^+, $p+ > p$, of which X is submatrix. The unbiased estimator of σ^2 comes from regression on all p^+ columns of X^+ and can be written

$$s^2 = R_{p^+}(n)/(n - p^+). \tag{22}$$

That model is chosen which minimizes

$$C_p = R_p(n)/s^2 - n + 2p = (n - p^+)R_p(n)/R_{p^+}(n) - n + 2p. \tag{23}$$

One derivation of C_p (Mallows 1973) is that it provides an estimate of the mean squared error of prediction at the n observational points from the model with p parameters provided the full model with p^+ parameters yields an unbiased estimate of σ^2. Then $\mathrm{E}\{R_p(n)\} = (n-p)\sigma^2$, $\mathrm{E}(s^2) = \sigma^2$ and $\mathrm{E}(C_p)$ is approximately p. An intuitive interpretation of (23) is that when comparing models the reduction in the residual sum of squares from the addition of extra parameters is penalized by twice the number of extra parameters.

In the standard application of model selection procedures both n and s^2 are fixed, the variable factors being the value of p and the regressors that are being considered.

Models with small values of C_p are preferred. Statements are often made that those models with values of C_p near p are acceptable. In §4.10 we consider the distribution of values of C_p and try to make this statement more precise.

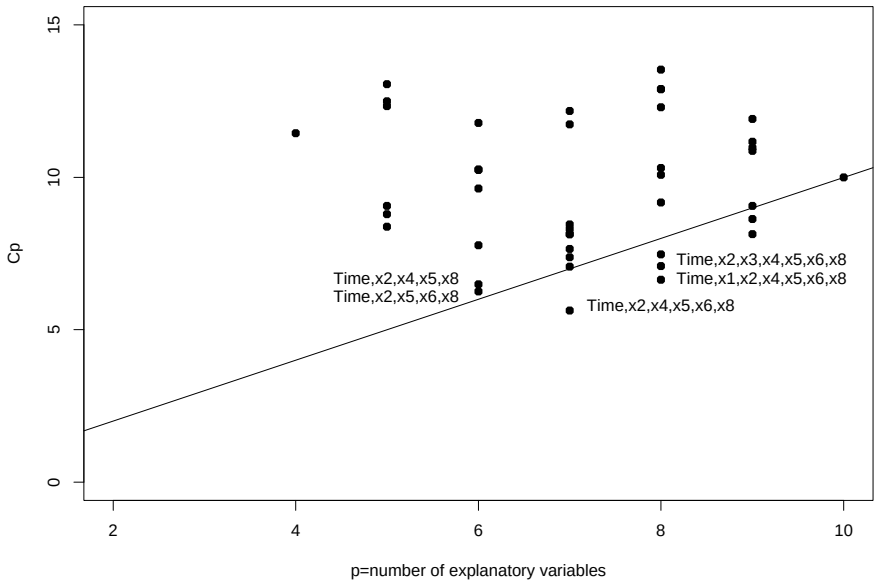

Figure 41: C_p plot for the ozone data. The combination of the two best models for $p = 6$ yields the best model for $p = 7$

4.8 The Ozone Data

Figure 41 is a C_p plot for the ozone data in which the smaller values of C_p for subset models are plotted against p. It shows a typical shape. Initially, for small p, all models have values of C_p much greater than p, and so these small models are not satisfactory. The best relatively small models are for $p = 6$, and 7. All models include a constant and the time trend. The model with smallest C_p for $p = 6$ also includes variables 2, 5, 6 and 8. This is the model selected by Atkinson and Riani (2000, p. 70). In the second-best model for $p = 6$, variable 4 replaces variable 6, giving the model including variables 2, 4, 5 and 8. The best model for $p = 7$ includes both these variables. Good models for larger values of p add further variables to the model for $p = 7$, giving rise to larger values of C_p.

Above we argued for variables 2, 4, 5 and 6. The model with minimum C_p in Figure 41 is for $p = 7$ and includes the constant, the trend and variables 2, 4, 5, 6 and 8. However, this model may be too large, since the t values for x_4 and x_6 are respectively -1.64 and 1.71. Our purpose is to determine how the choice of model is influenced by outliers or other unsuspected structure.

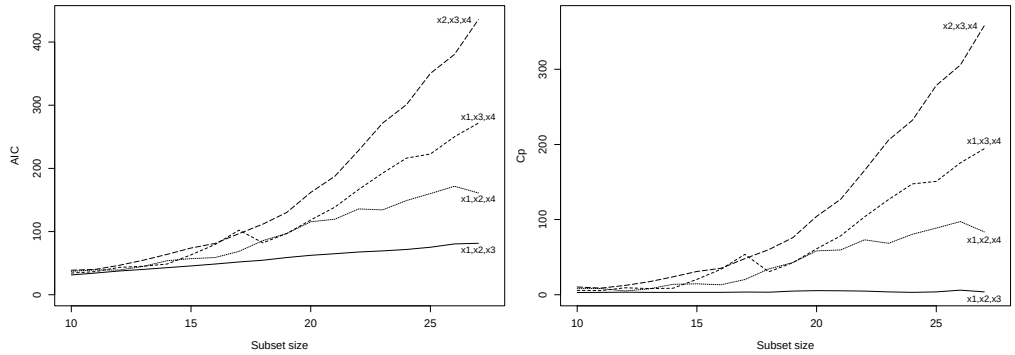

Figure 42: Wool data: three explanatory variables plus 1 noise variable. Forward plots of AIC(m) and $C_p(m)$ for $p = 4$

4.9 Forward C_p

The information criterion (23) for all observations is a function of the residual sums of squares $S_p(n)$ and $S_{p^+}(n)$. For a subset of m observations we can then define the forward value of this criterion as

$$C_p(m) = (m - p^+)R_p(m)/R_{p^+}(m) - m + 2p. \tag{24}$$

For each m we calculate $C_p(m)$ for all models of interest. However, some care is needed in interpreting this definition. For each of the models with p parameters, the search may be different, so that the subset $S_*(m)$ will depend on which model is being fitted. This same subset is used to calculate $R_{p^+}(m)$, so that the estimate s^2 in (22) may also depend on the particular model being evaluated as well as on m.

4.10 The Distribution of C_p in the Forward Search

The distribution of C_p is given, for example, by Mallows (1973) and by Gilmour (1996). From (23) we require the distribution of the ratio of two nested residual sums of squares. It is straightforward to show that the required distribution is

$$C_p \sim (p^+ - p)F + 2p - p^+, \qquad \text{where} \qquad F \sim F_{p^+-p,n-p^+}. \tag{25}$$

Gilmour comments that when $n - p^+$ is small, E(C_p) can be appreciably greater than p. In our example, with $n = 80$, this is not the case.

These results apply to C_p which is calculated from the full sample. However, in the forward search with $m < n$ we take the central m residuals to calculate the sums of squares $R_{p^+}(m)$ and $R_p(m)$. These sums of squares

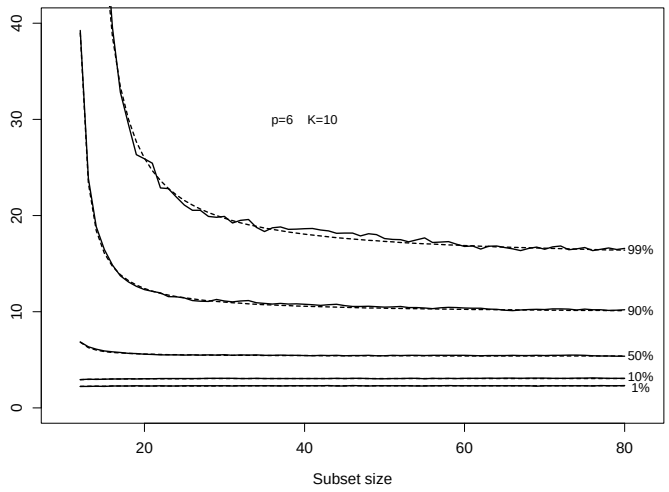

Figure 43: Comparison between empirical and theoretical envelopes for $C_p(m)$ based on the F distribution (25) when $n = 80$, $p = 6$ and $p^+ = 10$: 1%, 10%, 50%, 90% and 99% quantiles

are accordingly based on truncated samples and will have smaller expectations than those based on a full sample of m observations. Specifically $\mathrm{E}\{s^2(m)\} < \sigma^2$. We conducted a small simulation study to check the effect of this truncation on the distribution of $C_p(m)$.

Figure 43 shows a forward plot of the empirical distribution from 10,000 simulations of 80 observations with $p = 6$ and $p^+ = 10$. We give the empirical 1%, 10%, 50%, 90% and 99% points as n varies from 12 to 80, together with those calculated from the full sample distribution of C_p defined in (25). Amazingly, the distribution of $C_p(m)$ during the search is indistinguishable from that of the full sample statistic for sample size m. Accordingly, we can use (25) directly to provide envelopes for our forward plots.

4.11 Forward C_p Plots for the Ozone Data

We examine model selection by a forward plot for each plausible value of p. From Figure 41 it seems that $p = 6$ is a good choice, that is a constant, the trend and four explanatory variables. We also check other values of p.

Figure 44 shows the forward plots of $C_p(m)$ from $m = 59$ for p from 4 to 7, including only those models that have small values of $C_p(m)$ in this region of the search. These plots confirm our earlier choice of $p = 6$. However, a feature for all values of p is that many of the curves increase in the last two steps. The plot for $p = 6$ shows that, when $m = 78$, minimising the value of C_p leads to the choice of model with terms in x_2, x_4, x_5 and x_6, although

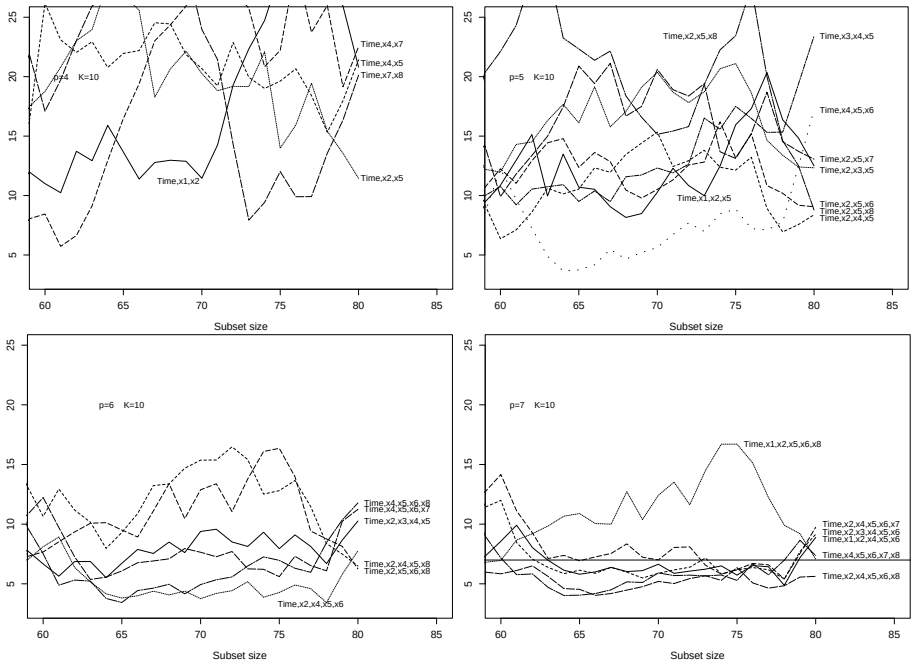

Figure 44: Ozone data: forward plots of $C_p(m)$ when $p = 4$, 5, 6 and 7. The last two observations to enter the subset have a clear effect on model choice

this is only the third best model of this size when $m = n$. This plot clearly and elegantly shows how the choice of model is being influenced by the last two observations to enter the forward search.

4.12 Outlier Detection

The last two observations to enter $S_*(m)$ are 56 and 65; these also seem to be outlying in the plot of residuals against trend in Figure 3.36 of Atkinson and Riani (2000). To detect outliers we calculate the deletion residual for the $n - m$ observations not in $S_*^{(m)}$. These residuals are

$$r_{i*}(m) = \frac{y_i - x_i^T \hat{\beta}_*(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}} = \frac{e_{i*}(m)}{\sqrt{s_*^2(m)\{1 + h_{i*}(m)\}}}, \qquad (26)$$

where $h_{i*}(m) = x_i^T \{X_*(m)^T X_*(m)\}^{-1} x_i$; the leverage of each observation depends on $S_*^{(m)}$. Let $i_{\min}$ denote the observation with the minimum absolute deletion residual among those not in $S_*^{(m)}$, that is

$$i_{\min} = \arg \min_{i \notin S_*^{(m)}} |r_{i*}(m)|.$$

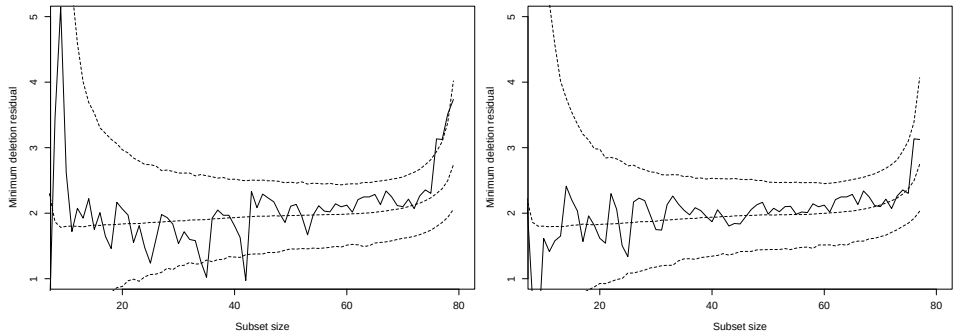

Figure 45: Ozone data: monitoring the minimum deletion residual (27). Left-hand panel, $n = 80$, right-hand panel $n = 78$. There are two outlying observations

To test whether observation $i_{\min}$ is an outlier we use the absolute value of the minimum deletion residual

$$r_{i_{\min}*}(m) = \frac{e_{i_{\min}*}(m)}{\sqrt{s^2_*(m)\{1 + h_{i_{\min}*}(m)\}}}, \tag{27}$$

as a test statistic. If the absolute value of (27) is too large, the observation $i_{\min}$ is considered to be an outlier, as well as all other observations not in $S^{(m)}_*$. Riani and Atkinson (2007) give further details and discuss the calculation of approximations to the distribution of the test statistic (27). We use simulation to find envelopes for the small value of n for the ozone data.

The left-hand panel of Figure 45 shows a forward plot of the minimum deletion residual for all 80 observations when the model contains variables 2,4 5 and 6, together with 1%, 50% and 99% simulation envelopes. The last two observations are clearly revealed as outlying. If they are removed and the envelopes recalculated for $n = 78$ we obtain the plot in the right-hand panel of Figure 5. There is no evidence of any further outlying observations.

We now return to model selection. Figure 46 gives the last part of the forward plot of $C_p(m)$ for $n = 78$ when $p = 6$, together with 2.5%, 50% and 97.5% quantiles calculated from (25). We give the curves only for those models that are one of the three best at some point for the last ten values of m. The model with variables 2, 4, 5 and 6 is clearly the best; unlike any other model its value of $C_p(m)$ lies in the lower half of the distribution for $m > 63$. There are many alternative six-parameter models with values of $C_p(78)$ lying below the 97.5% quantile. Plots for five such are shown in Figure 46. All however fall in the upper half of the distribution.

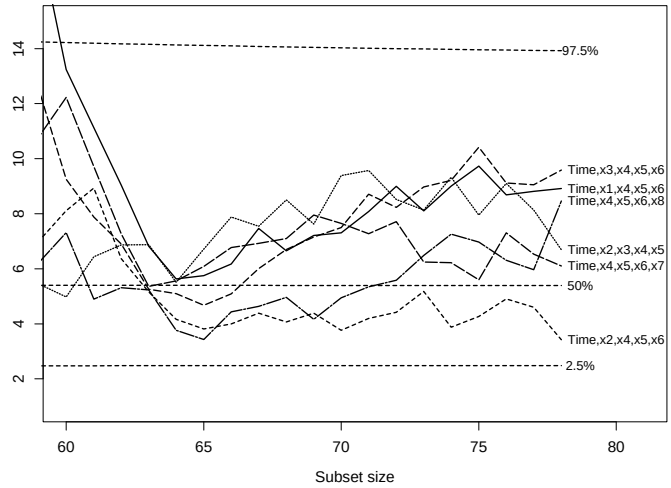

Figure 46: Ozone data without outliers: forward plots of $C_p(m)$ when $p = 6$, together with 2.5%, 50% and 97.5% quantiles from (25). The model including variables 2, 4, 5 and 6 is preferred

Table 1: Ozone data: effect of deletion of outliers on significance of terms in model with variables 2, 4, 5 and 6

	All 80 observations		$n = 78$	
Term	t	p-value	t	p-value
Constant	-4.83	0.000	-5.74	0.000
Time	7.16	0.000	8.99	0.000
x_2	- 3.34	0.001	-2.57	0.012
x_4	-1.79	0.077	-3.01	0.004
x_5	5.75	0.000	6.80	0.000
x_6	1.60	0.114	2.39	0.019
R^2	0.67		0.74	

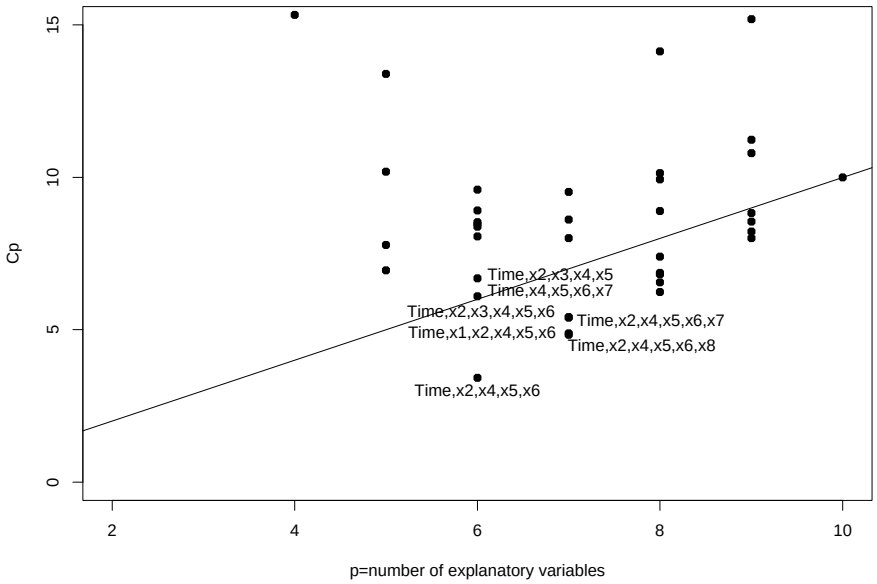

Figure 47: C_p plot for the ozone data after deletion of the two outliers. One model with $p = 6$ is now clearly best. In comparison, the best model in Figure 41, which had $p = 7$, was less sharply revealed

It is also interesting to consider the effect of deleting observations 56 and 65 on the properties of the final model. Table 1 lists the t-statistics for the six terms in the model and their significance both for all observations and for the 78 observations after deletion of the two outliers. When $n = 80$ neither x_4 nor x_6 are significant when they are both in the model. But deletion of the outliers causes the variables to be jointly significant, one at 2% and the other well past the 1% level.

We have based our argument on the plot for $p = 6$. So, finally we reproduce the C_p plot of Figure 41 for all values of p after the two outliers have been removed. The comparison is instructive. Now the model with variables 2, 4, 5 and 6 has an appreciably smaller value of C_p than the next best six-parameter model. In addition, this value is less than that for the best seven-parameter model. By detection and deletion of the outliers we have not only changed the selected model but have sharpened the choice of the best model

The distributional results in Figure 46 indicate some other potential models. Whether we need to be concerned to have more than one model depends on the purpose of model fitting. If the model is to be used to predict over the region over which the data have been collected and the system is unlikely to change, so that the correlations between the explanatory variables remain sensibly constant, then any of these models will give almost equally good pre-

dictions. If however the relationships between the variables may change, or predictions are needed in new regions where data are sparse or non-existent, then the outcomes of all satisfactory models, as selected here by $C_p(m)$, must be taken into account. The possible effects of climate change on ozone concentration in the Californian desert indicate that the consequences of several well-fitting models should be explored.

For further details of this analysis see Atkinson and Riani (2008).

References

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 (2002a). Forward search added variable t tests and the effect of masked outliers on model selection. *Biometrika 89*, 939–946.

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

Atkinson, A. C. and M. Riani (2008). A robust and diagnostic information criterion for selecting regression models. *Journal of the Japanese Statistical Society 38*, 3–14.

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.

Cook, R. D. and P. Wang (1983). Transformations and influential cases in regression. *Technometrics 25*, 337–343.

Gilmour, S. G. (1996). The interpretation of Mallows's C_p-statistic. *The Statistician 45*, 49–56.

Mallows, C. L. (1973). Some comments on C_p. *Technometrics 15*, 661–675.

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