The Statistician (2001)
50, *Part* 1, *pp.* 63–78

Regression diagnostics for binomial data from the forward search

Anthony C. Atkinson

London School of Economics and Political Science, UK

and Marco Riani

Università di Parma, Italy

[Received September 1998. Final revision September 2000]

Summary. We suggest a simple robust method for the detection of atypical and influential observations in binomial data. Our technique is based on a forward search procedure which orders the observations from those most in agreement with a specified generalized linear model to those least in agreement with it. The effectiveness of the forward search estimator in detecting masked multiple outliers, and more generally in ordering binomial data, is shown by means of three data sets. Plots of diagnostic quantities during the forward search clearly show the effect of individual observations on residuals and test statistics. These examples reveal the strength of our method in showing the structure of the data in a way which is more simple and effective than it would be by using standard deletion diagnostic procedures.

Keywords: Generalized linear models; Graphical methods; Influential observations; Outliers; Robust methods

1. Introduction

Multiple outliers may strongly affect the generalized linear model fitted to data, as may unidentified distinct subsets. But such important observations may be difficult to identify. In regression the single-deletion diagnostics described in Cook and Weisberg (1982) and Atkinson (1985) may fail owing to 'masking' if there is more than one outlier. More recent regression methods using multiple-deletion diagnostics, such as those of Barrett and Gray (1997) and Haslett (1999), may likewise fail either owing to masking or computational requirements and interpretability if there are too many outliers. For generalized linear models single-deletion methods are summarized in chapter 12 of McCullagh and Nelder (1989). These may all be thought of as 'backward' methods: they start from a fit to all the data and then study the effects of deletion. If there is just one outlier, it will be found by the deletion of single cases, but, for the general analysis of data, a method is needed which can establish the number of outliers and their effect on inferences drawn from the data.

Our approach, based on a robust and at the same time efficient estimator, uses a forward search through the data. Related algorithms have recently been suggested in regression models with independent errors, in multivariate analysis and for data transformations (Atkinson, 1994; Atkinson and Riani, 1997; Riani and Atkinson, 2000). Our proposal extends this field of research

Address for correspondence: Anthony C. Atkinson, Department of Statistics, London School of Economics and Political Science, London, WC2A 2AE, UK.
E-mail: a.c.atkinson@lse.ac.uk

0039–0526/01/50063

to the case of generalized linear models. In this paper we demonstrate how a greater insight into the structure of data can be obtained by starting from a very robust fit to a carefully chosen subset of the data and then working forwards, adding observations to the subset used for fitting.

The results are presented through plots which are both easy to interpret and powerful in revealing the structure of the data. In the examples to be presented, it is found that features of forward plots can be related to structure in the scatterplot matrices of the data.

The outline of the paper is as follows. In Section 2 we state our notation and introduce the forward search estimator. In Section 3 we apply our method to three examples of binomial data. In Section 4 we compare our method with the traditional backward approach. Section 5 contains additional comments and conclusions.

2. Details of the forward search

2.1. Generalized linear models

Diagnostic methods for generalized linear models use functions of residuals, as they do for regression. The residual sum of squares in regression is replaced by the deviance

$$D(\hat{\beta}) = \sum_{i=1}^{n} d_i^2, \tag{1}$$

a sum of n non-negative components. The deviance residuals are the signed square roots of these components:

$$d_i = \operatorname{sgn}(y_i - \hat{\mu}_i)\sqrt{d_i^2},$$

where $\hat{\mu}_i$ is the fitted value of the ith observation when $\hat{\beta}$ is the estimate of the vector of p parameters in the linear predictor $\eta_i = x_i^{\mathrm{T}}\beta$. Since the dispersion parameter does not have to be estimated for binomial data, the value of the deviance can be compared with the χ^2-distribution on $n - p$ degrees of freedom to provide a test of the goodness of fit. The standard work on these models is McCullagh and Nelder (1989). A description of diagnostic methods for generalized linear models is in chapter 6 of Atkinson and Riani (2000).

The forward search algorithm described in this paper has three steps: the first concerns the choice of an initial subset, the second refers to the way in which we progress in the forward search and the third relates to the monitoring of the statistics during the progress of the search. In the following subsections we consider these three aspects separately.

2.2. Step 1: choice of the initial subset

If the model contains p parameters, our forward search algorithm starts with the selection of a subset S_*^p of p units. Observations in this subset are intended to be outlier free. If n is moderate and $p \ll n$, the choice can be performed by exhaustive enumeration of all $\binom{n}{p}$ distinct p-tuples; otherwise we evaluate the properties of some large number of p-tuples, often 1000. If d_{i,S_j^p} is the deviance residual for unit i given that observations in S_j^p are used in fitting the model, the initial subset is such that

$$d^2_{[\mathrm{med}],S_*^p} = \min_j(d^2_{[\mathrm{med}],S_j^p}), \tag{2}$$

where $d^2_{[l],S_j^p}$ is the lth ordered squared residual among d^2_{i,S_j^p}, $i = 1, \ldots, n$,

$$\mathrm{med} = p + \left[\frac{n-p}{2}\right] \tag{3}$$

and $[(n-p)/2]$ denotes the integer part of $(n-p)/2$. Criterion (2) provides a robust fit to the data and is similar to the least median of squares method for regression models with independent errors (Rousseeuw, 1984; Hawkins, 1993). Although the breakdown point of this estimator for binomial data is unknown, we need only to initialize the algorithm with a subset which does not contain atypical observations.

Remark 1. For binomial data, and other generalized linear models, there is a choice of residuals (e.g. McCullagh and Nelder (1989), page 397). In this paper we use deviance residuals because they are more stable than Pearson residuals when the fitted probabilities are close to 0 or to 1. We found that standardizing the residuals to allow for the effect of different leverages on the variance had no observable effect.

Remark 2. The method is not sensitive to the method used to select an initial subset. For example, the least median of squares criterion (2) can be replaced by the least trimmed squares criterion in which the sum of the squares of the smallest med residuals is minimized.

2.3. Step 2: adding observations during the forward search

Given a subset S_*^m of $m \geqslant p$ observations, the forward search selects the $m+1$ units with the smallest squared deviance residuals, the units being chosen by ordering all squared deviance residuals $d_{i,S_*^m}^2$, $i = 1, \ldots, n$.

The forward search estimator $\hat{\beta}_{\mathrm{FS}}$ is the collection of maximum likelihood estimators at each step of the forward search,

$$\hat{\beta}_{\mathrm{FS}} = (\hat{\beta}_{S_*^p}, \ldots, \hat{\beta}_{S_*^m}, \ldots, \hat{\beta}_{S_*^n}), \tag{4}$$

where $\hat{\beta}_{S_*^m}$ is the maximum likelihood estimator using the observations in S_*^m.

In most moves from m to $m+1$ just one new unit joins the subset. But sometimes two or more units join S_*^m as one or more leave. Such an event is unusual, only occurring when the search includes one unit which belongs to a cluster of outliers. At the next step the remaining outliers in the cluster seem less outlying and so several may be included at once. Of course, several other units then must leave the subset.

The search which we use avoids, in the first steps, the inclusion of outliers and provides a natural ordering of the data according to the specified null model. In this approach we use a highly robust method and at the same time maximum likelihood (i.e. fully efficient) estimators. The zero breakdown point of maximum likelihood estimators, in the context of the forward search, is not a disadvantage. The introduction of atypical influential observations is signalled by sharp changes in the curves which monitor parameter estimates, t-tests or any other statistic at every step. In this context, the robustness of the method does not derive from the choice of a particular estimator with a high breakdown point, but from the progressive inclusion of the units into a subset which, in the first steps, is outlier free. As a bonus of the suggested procedure, the observations can be naturally ordered according to the specified null model and it is possible to know how many of them are compatible with a particular specification. Furthermore, the approach suggested enables us to analyse the inferential effect of the atypical units (outliers) on the results of statistical analyses. For example, the problem of overdispersion in generalized linear models is often discussed. In our approach (as we show in one of our examples) we can quantify how many units are responsible for this phenomenon.

Remark 3. The search can often recover from a start which is not very robust. For example, Atkinson and Riani (1997) used bivariate box plots to provide an initial subset for multivariate data from which the grossest outliers are removed. The first few steps of the search are very active, as other outliers are identified and removed. But the final, informative, third of the forward search is insensitive to the precise selection of the initial subset.

Remark 4. Parameter estimation in generalized linear models is iterative. The search can be made faster by using $\hat{\beta}_{S_*^{m-1}}$ as the initial estimate for the iterative calculation of $\hat{\beta}_{S_*^m}$.

2.4. Step 3: ordering of the data

Step 2 of the forward search is repeated until all units are included in the subset. If just one observation enters S_*^m at each move, the algorithm provides an ordering of the data according to the specified null model, with observations furthest to it joining the subset at the last stages of the procedure. Through the joint examination of simple plots, which monitor the effect on the statistics of the sequential inclusion of the units, we can obtain great insight into the structure of the data.

One of the most important plots monitors all residuals at each step of the forward search. Large values of the residuals among cases that are not in the subset indicate the presence of outliers, as do non-smooth changes in the value of the residual deviance.

To judge the importance of individual variables we find it useful to monitor t-statistics. Non-smooth changes in parameter estimates are indications of influential observations, which can also be detected through the monitoring of a 'forward version' of the Cook statistic (Cook and Weisberg (1982), page 116). When the first influential observation enters the subset, the statistic exhibits an upward movement. If the atypical observation belongs to a cluster, the value of the statistic is likely to decrease after the inclusion of the first outlier, because the introduction of the remaining units in the cluster reinforces the change in parameter estimates. We can also look at the values of the leverages for all units in the subset.

Several link functions $g(\mu)$ are available for modelling the relationship between the expected proportion of successes $\mu = E(Y)/n$ and the linear predictor η. We use the four given in Table 1, where Φ is the distribution function of the standard normal distribution. For binary data, i.e. data where the response is either 0 or 1, there are advantages in using a fifth link, the inverse sine, the properties of which are discussed on page 253 of Atkinson and Riani (2000).

To analyse the adequacy of the link we can monitor, at each step, the t-statistic for the constructed variable $\hat{\eta}^2_{S_*^m}$, where $\hat{\eta}_{S_*^m}$ is the estimated linear predictor and the statistic is calculated without changing the iterative weights. In what follows we shall call this statistic the 'goodness-of-link test'.

As our examples show, all these plots must be considered as different instruments which give complementary information about the structure of the data. For example, in our third example, a considerable change in deviance residuals produces big changes in the curves of the t-statistics. This is usually accompanied by a peak in the forward plot of Cook's distance and high leverage for the unit which joins the subset.

Table 1. Link functions used in the analysis

Name	*Link* $g(\mu) = \eta$
Logit or logistic	$\log\{\mu/(1-\mu)\}$
Probit	$\Phi^{-1}(\mu)$
Complementary log–log	$\log\{-\log(1-\mu)\}$
Log–log	$\log\{-\log(\mu)\}$

3. Examples

In this section, to show the power of the forward search in showing the structure of the data, we analyse three data sets. In the first example attention is focused on the goodness-of-link test for three links. The introduction of the last two units causes the value of the statistic to go outside the significance bounds for two of the three. In the second example we show the relationship between model building and diagnostics: starting from a simple model we are forced to include quadratic terms. We can then identify strange observations and lack of structure in part of the data. In the third example the focus is on a group of observations which strongly influence the linear predictor and suggest the presence of overdispersion.

3.1. Bliss's beetle data

Eight groups of around 60 beetles were subjected to eight different doses of insecticide and the number of beetles killed was recorded. (The data were originally given by Bliss (1935) and are reported in many text-books, e.g. Flury (1997), page 526.) The resulting data are binomial and interest is in modelling the relationship between the dose and the probability of death. Fig. 1 shows plots of deviance residuals (in absolute value) from forward searches for three models in which the explanatory variable is log(dose). Although the residuals were calculated only for each value of m we have joined them by a continuous line, a visual fiction which improves the interpretability of this and the remaining plots. The three links used were the logit, probit and complementary log–log. For the logit link observations 1 and 2 are the last two to be included in the forward search. The crossing of the lines at the end of the plot in Fig. 1(a) shows that the inclusion of observations 1 and 2 seems noticeably to affect the ordering of the residuals. With the probit link units 3 and 4 (the last two to be included) seem to be different from the rest of the data—they are badly predicted by models in which they are not included. However, the residuals from the forward search with the complementary log–log-link show no such behaviour—all residuals are smaller than 2 throughout, and relatively constant.

Fig. 2 shows a plot of the score test for the link from the forward search, the order of introduction of the observations again being different for the three links. For the logit and probit links these plots show evidence of a lack of fit at the 5% level, which is indicated by the statistic going outside the bounds in the plot. Although it is the inclusion of the last two observations which causes the values of the statistic to become significant, it is clear from the steady upward trend of the plots that the lack of fit is due to all observations. On the contrary, the plot for the complementary log–log-link shows no evidence of any departure from this model. This plot also shows that unit 5, which is the one with the biggest residual for the complementary log–log-link and the last to be included in this forward search, has no effect on the t-test for the constructed variable.

This example shows how our procedures can quickly lead to the identification of unsatisfactory models and provide checks of models which seem good—there is no evidence of a strong effect of any individual observation on any aspect of the fitted complementary log–log-model. Finally, if reference is made back to the data, observations 1–4 are those with the lower dose levels. Plots of the proportion dead against log(dose) show that the dose–response curve is not symmetrical, a feature modelled by the complementary log–log-link, but not by the other two, symmetrical, links.

3.2. Liver cancer

Zelterman (1999) quoted data on the incidence of liver cancer in mice, which we reproduce in Table 2, giving the number of mice developing cancer and the total number tested, which forms

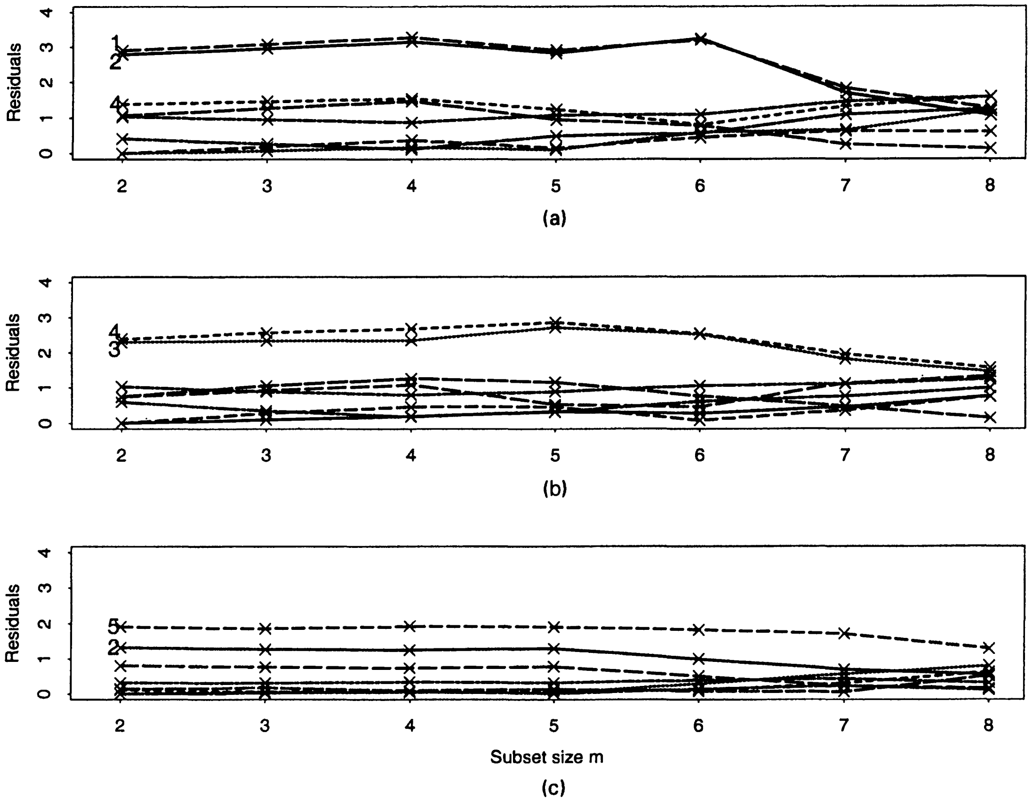

Fig. 1. Bliss's beetle data—absolute values of deviance residuals as the subset size increases from 2 to 8: (a) logit link; (b) probit link; (c) complementary log–log-link

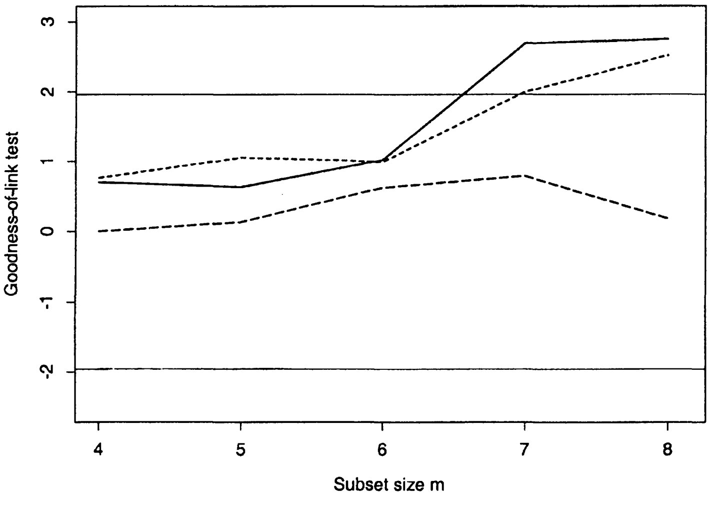

Fig. 2. Bliss's beetle data—goodness-of-link test from the three forward searches: ———, logit link; --------, probit link; – – – –, complementary log–log-link

Table 2. Liver data

Observation	*Total number tested*	*Number with cancer*	*Dose*	*Months on study*	*Observation*	*Total number tested*	*Number with cancer*	*Dose*	*Months on study*
1	199	0	0.00	9	37	287	6	0.60	16
2	147	1	0.30	9	38	193	7	0.75	16
3	76	1	0.35	9	39	100	2	1.00	16
4	52	0	0.45	9	40	110	7	1.50	16
5	345	0	0.60	9	41	153	0	0.00	17
6	186	0	0.75	9	42	69	4	0.30	17
7	168	1	1.00	9	43	443	5	0.35	17
8	169	1	1.50	9	44	302	6	0.45	17
9	164	0	0.00	12	45	230	8	0.60	17
10	151	1	0.30	12	46	166	9	0.75	17
11	27	2	0.35	12	47	85	3	1.00	17
12	14	1	0.45	12	48	82	1	1.50	17
13	283	2	0.60	12	49	555	6	0.00	18
14	153	0	0.75	12	50	2014	34	0.30	18
15	149	3	1.00	12	51	1102	20	0.35	18
16	152	2	1.50	12	52	550	15	0.45	18
17	133	1	0.00	14	53	411	13	0.60	18
18	42	1	0.30	14	54	382	17	0.75	18
19	25	0	0.35	14	55	213	19	1.00	18
20	14	2	0.45	14	56	211	24	1.50	18
21	243	1	0.60	14	57	762	20	0.00	24
22	124	0	0.75	14	58	2109	164	0.30	24
23	127	1	1.00	14	59	1361	128	0.35	24
24	127	1	1.50	14	60	888	98	0.45	24
25	115	0	0.00	15	61	758	118	0.60	24
26	75	1	0.30	15	62	587	118	0.75	24
27	35	1	0.35	15	63	297	76	1.00	24
28	20	0	0.45	15	64	314	126	1.50	24
29	203	3	0.60	15	65	100	17	0.00	⩾33
30	109	1	0.75	15	66	445	135	0.30	⩾33
31	99	5	1.00	15	67	100	72	0.35	⩾33
32	100	1	1.50	15	68	103	42	0.45	⩾33
33	205	1	0.00	16	69	67	30	0.60	⩾33
34	66	2	0.30	16	70	75	37	0.75	⩾33
35	61	2	0.35	16	71	31	22	1.00	⩾33
36	304	3	0.45	16	72	11	9	1.50	⩾33

the binomial denominator. There are eight doses, units unspecified, and observations are taken at nine unequally spaced times, making 72 observations in all. We use our analysis to demonstrate the relationship between model building and the information on individual observations provided by the forward search.

Fig. 3 is a forward plot of the goodness-of-link test for three links when a first-order model is fitted, i.e. a model in which the linear predictor includes terms in dose and time. Clearly, none of the models is adequate: the log–log-link seems to be generally misspecified; the other two links deteriorate markedly at the end of the search. The results are confirmed by the residual deviances in Table 3: all are over 200, to be compared with the χ^2-distribution on 69 degrees of freedom.

We next try a full second-order model. The forward plot of the goodness-of-link tests for three links is in Fig. 4. Now there is no systematic evidence of inadequacy, although there is an abrupt change at the end of the plot for the logit link, caused by the inclusion of observation 67, the last to enter in all searches. Table 3 shows that the deviances with all observations are still somewhat

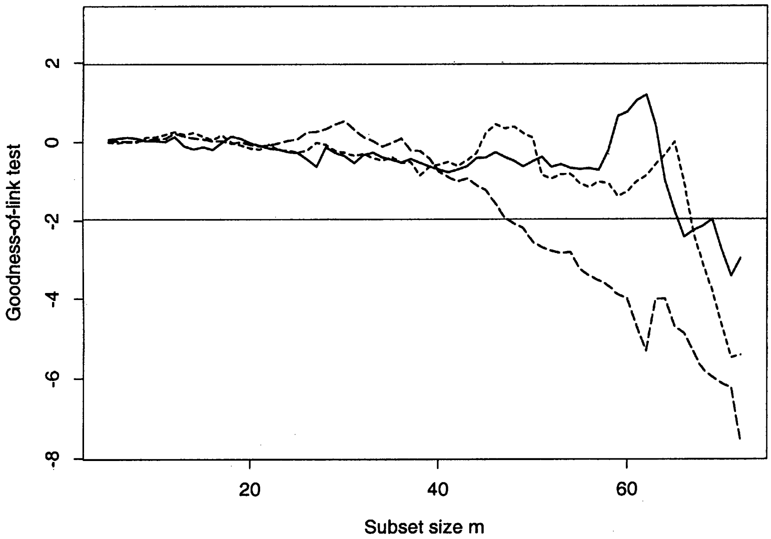

Fig. 3. Liver cancer data, first-order model—goodness-of-link tests for three link functions (observation 67 is the last to enter in all three searches): ———, logit; --------, complementary log–log; – – – –, log–log

Table 3. Liver cancer data: residual deviances and associated degrees of freedom for various models and link functions in the last two steps of the forward search

	Results for the following models:					
	First order		*Full second order*		*Second order without interaction*	
	$m = n$	$m = n - 1$	$m = n$	$m = n - 1$	$m = n$	$m = n - 1$
Degrees of freedom	69	68	66	65	67	66
Logit	207.0	167.3	140.9	84.74	144.2	89.52
Complementary log–log	228.7	186.7	148.5	89.21	148.8	89.27
Log–log	241.9	185.2	133.6	81.98	—	—

high. However, the deletion of observation 67 reduces all residual deviances to around the 95% point of χ^2, with the log–log-model fitting best.

Next we consider whether all the terms in the full second-order model are needed and whether observation 67 is important to the significance of the terms. Fig. 5 is a forward plot of the t-statistics for the three links. Figs 5(a) and 5(b), for the logit and the complementary log–log-links, show that the interaction term is not significant. For the log–log-link all higher order terms are significant: the non-significant first-order terms therefore must be included. We therefore give in Table 3 the residual deviances for the logit and complementary log–log-links without the interaction term. Even when observation 67 is dropped, both deviances are slightly larger than the 95% point of the χ^2-distribution on 66 degrees of freedom.

We now try to find out why the logit and complementary log–log-links are showing some sign of a lack of fit. Fig. 6 shows the goodness-of-link test for these two models. The effect of observation 67 at the end of the search is obvious. But, before this, there is a clear increase in the

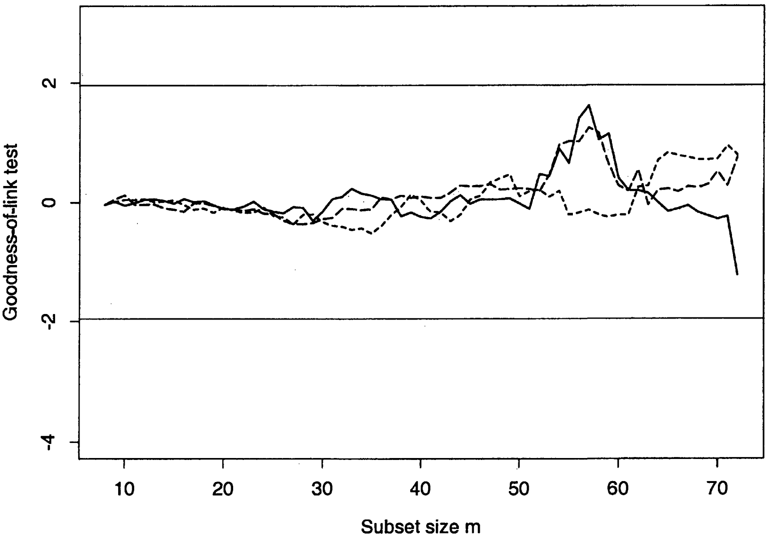

Fig. 4. Liver cancer data, full second-order model—goodness-of-link tests for three link functions: ———, logit; ---------, complementary log–log; – – – –, log–log

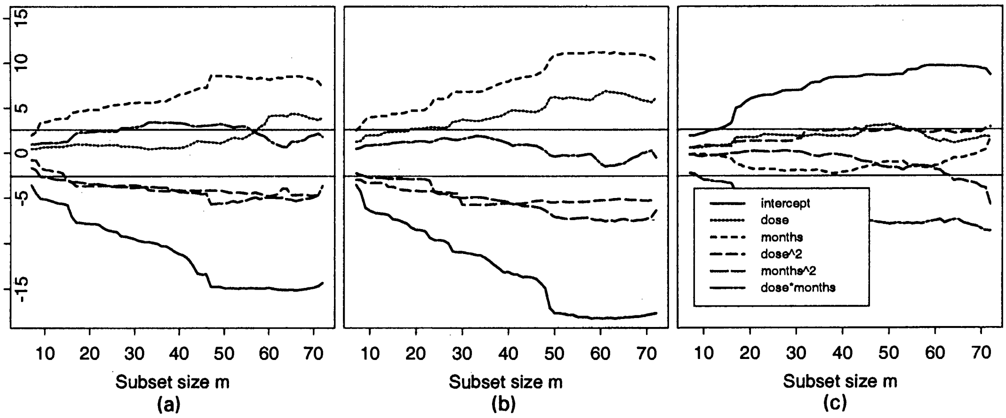

Fig. 5. Liver cancer data, full second-order model—forward plots of *t*-statistics for three link functions: (a) logit; (b) complementary log–log; (c) log–log

statistics, caused by the inclusion of a group of observations, units 11, 20, 48 and 42 working backwards in order for the complementary log–log-link. The order is 11, 48, 20 and 42 for the logit link.

The contribution of individual units to the deviance can be assessed from forward plots of deviance residuals. Fig. 7 shows such a plot for the complementary log–log-link without the interaction term, since the removal of this term has a negligible effect on the deviance. The most obvious feature of the plot is the large residual for observation 67, which is never well explained. Observation 66 has a large leverage and large residual until it enters, when the residual becomes small. Other observations with large residuals are 11, 20, 42, 12 and 48, all of which enter at the end of the search.

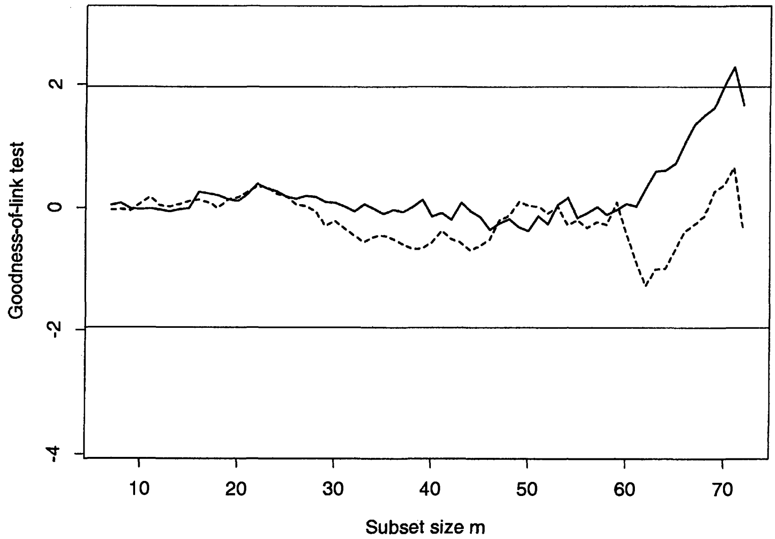

Fig. 6. Liver cancer data, second-order model without interaction—goodness-of-link tests for two link functions showing the effect of outliers at the end of the search (observation 67 is the last to enter): ———, logit; ---------, complementary log–log

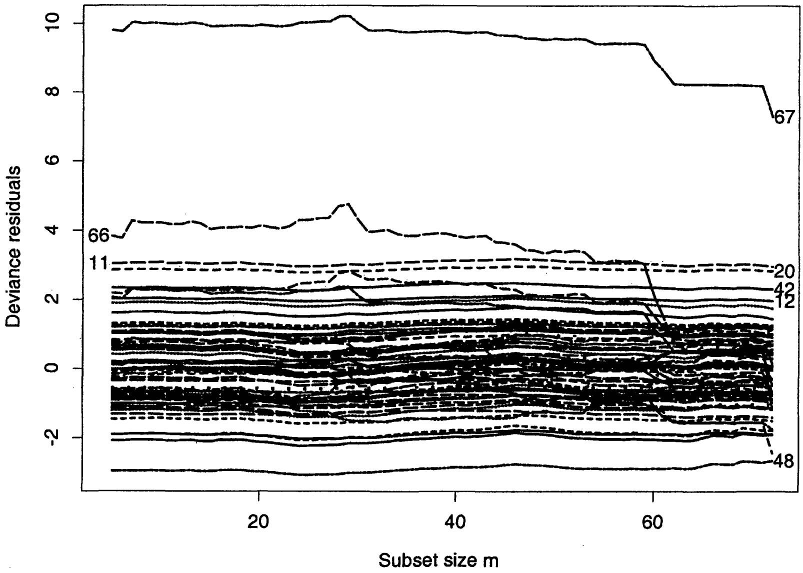

Fig. 7. Liver cancer data, second-order model without interaction—forward plot of deviance residuals for the complementary log–log-link

To interpret these results we show, in Fig. 8, two plots of the data against dose. In Fig. 8(a) there is a clear structure of linear increase of proportion affected with dose for the higher months on study. Observation 67 does not fit this structure. The adjacent unit, 66, has over four times as many mice as any other group tested at over 33 months (see Table 2): hence the effect of high leverage exhibited in Fig. 7. However, it lies on the general response curve. The other larger

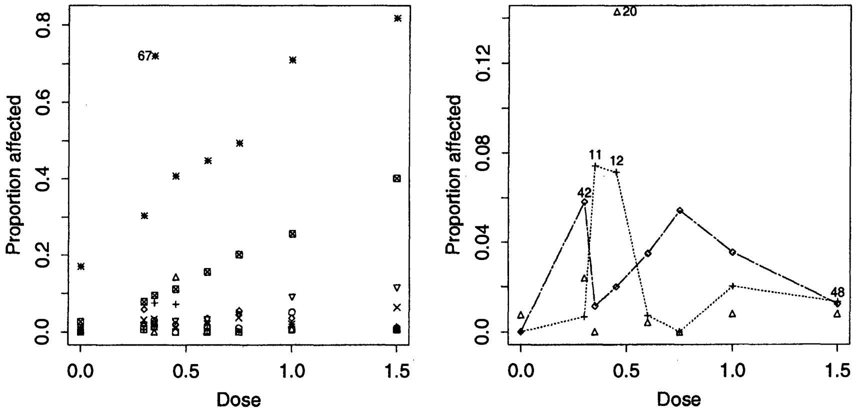

Fig. 8. Liver cancer data—scatterplots of response against dose: □, 9 months; +, 12 months; △, 14 months; ○, 15 months; ×, 16 months; ◇, 17 months; ▽, 18 months; ☒, 24 months; $*$, 33 months

deviance residuals in Fig. 8 are caused by units for which the incidence of cancer is much lower. In interpreting Table 2 remember that there is wide variability in the numbers of mice tested.

The part of the plot for units which have 12, 14 or 17 months on study is shown enlarged in Fig. 8(b). These groups contain units having large residuals in Fig. 7. Fig. 8(b) shows that units 11, 12 and 20 have a proportion of affected mice that is much higher than that of their corresponding group. For the group with 17 months on study, units 42 and 48 show proportions of affected mice which are respectively much higher and much lower than would be expected from their dose levels. In Fig. 7 the residual for unit 42 is always larger than 2 in all steps of the forward search, whereas that for unit 48 is always less than -2.5. These are all units which enter at the end of the forward search.

Finally, if we go back to Table 2 we can see that the study seems to have been poorly designed, with too many observations at too low times and dose levels.

A conclusion from this analysis is that observation 67 is a clear outlier and should be omitted. The best model we have found is that for the log–log-link with the full second-order model. The structure of the outliers is obvious and there is no indication of masking.

3.3. Toxoplasmosis and rainfall

We now briefly consider an example in which masking is present. First we show how the forward search simply exhibits the structure of the data, then, in the next section, describe how backward deletion methods cannot elucidate the structure. The data are taken from Efron (1978) and are the proportion of subjects aged between 11 and 15 years testing positively for toxoplasmosis, as a function of rainfall (in millimetres), in 34 cities of El Salvador. He fitted a binomial model with a logit link to a cubic function in standardized rainfall z. This fit has some strange features: the cubic term in the model is significant at the 1% level, whereas some of the lower order terms are not significant at all. However, the relationship does not explain all the variation in the data—the residual deviance is 62.63 on 30 degrees of freedom, which is significant evidence of a lack of fit at a level of 0.043%, if asymptotic theory is an adequate guide.

Fig. 9 is the plot of residuals from the forward search: it is very different from Fig. 7 in which most residuals, apart from that for observation 66, were virtually constant throughout the forward search. Here there is a large amount of change towards the end of the search, which is where we

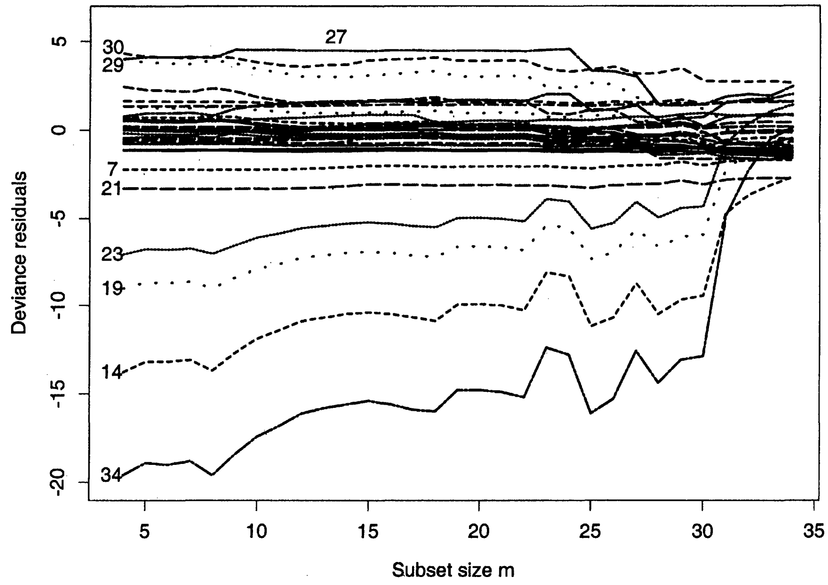

Fig. 9. Toxoplasmosis data—forward plot of deviance residuals

expect outliers to enter. Fig. 9 clearly shows that four units (34, 14, 19 and 23) have very large negative residuals until $m = 30$. But some of the signs of these residuals change when $m = n = 34$: units 23 and 34 have *positive* deviance residuals of 1.39 and 0.13, whereas unit 19 has a small *negative* deviance residual of -0.37. In the upper part of Fig. 9 there are three units (30, 27 and 29) which show deviance residuals that are always above 3 in the central part of the forward search. However, in the last step of the forward search unit 29 has a deviance residual which is equal to only 0.22. We may thus expect some problems for backward methods due to masking. What these are will be shown in the next section. Finally, in Fig. 9 we can see two units (7 and 21) whose negative residuals (less than -2) remain virtually constant in all steps of the forward search.

To interpret these results we give in Table 4 the order in which the forward search causes the observations to enter the fit. Also given is the estimate of the dispersion parameter $\hat{\phi}_{S_*^m} =$

Table 4. Toxoplasmosis data: last steps of the forward search — subset size, observation introduced and estimate of dispersion parameter ϕ

Subset size m	*Observation*	*Estimate* $\hat{\phi}_{S_*^m}$
26	7	0.76
27	29	0.92
28	27	1.05
29	21	1.33
30	30	1.64
31	23	1.72
32	19	1.76
33	34	1.73
34	14	1.94

$X^2_{S_*^{m}}/(m-p)$ where $X^2_{S_*^{m}}$ is the observed value of the Pearson χ^2 goodness-of-fit statistic. Of course, for binomial data, we hope for a value around 1 for $\hat{\phi}$.

We notice that the observations entering at the end of the search are those that are identified as different by the forward plot of residuals. To see whether these observations affect inferences drawn from the model, we studied forward plots of a number of statistics. The plot of the goodness-of-link test shows that the exclusion of observations 14, 34 and 19 causes the statistic to change from 1.58 to −1.92. Monitoring the forward plot of the Cook statistic indicates that the addition of observation 23 causes a significant change in the parameter values. These effects on parameter estimates are most easily seen by looking at the plot of the individual *t*-statistics in Fig. 10. The statistics for the linear and cubic terms (b_1 and b_3) remain sensibly unchanged for most of the search. But those for the intercept (b_0) and the quadratic term (b_2) change sign and become less significant in the last five steps. These results have a straightforward interpretation if we go back to the data as plotted in Fig. 11, in which the diameter of the dots is proportional to the binomial denominator raised to the power 0.8. As a referee has commented, the observations from cities where the denominator is large do not suggest a strong relationship between toxoplasmosis and rainfall. However, the deviance explained by fitting the cubic model is 11.6, rising to 13.8 if observations 23, 19, 14 and 34 are removed, values which are significant at the 1% level. The full curve shows the cubic fit using all data, and the curve with short dashes shows a cubic fit without observations 23, 19, 14 and 34 (the last four in the forward search). These four observations form a group with the highest rainfall and are clearly all influencing the shape of the cubic curve in the same way, lessening the curvature at the second point of inflection. The first of the four to be included is 23. Once it has been included the other points do not greatly change the shape of the curve. When all are included observation 34 is virtually on the fitted curve. But when $m = n - 4$ this observation has a deviance residual of −12.9. This dramatic change can be seen in Fig. 9. The last observation to be considered is 30, which enters immediately before this group of four. The effect resulting from its additional deletion, shown by the curve with long dashes in Fig. 11, is to

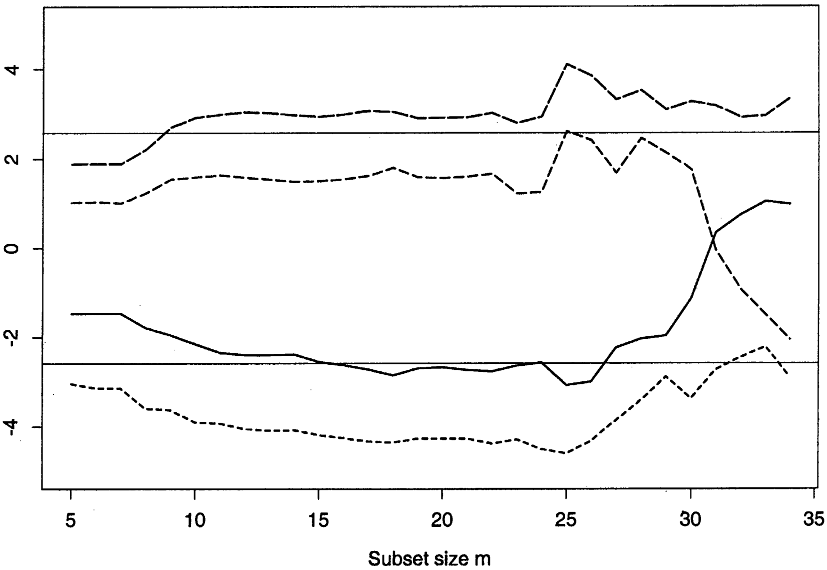

Fig. 10. Toxoplasmosis data—forward plot of *t*-statistics for explanatory variables: ———, constant; --------, *z*; — — — —, z^2; — —, z^3

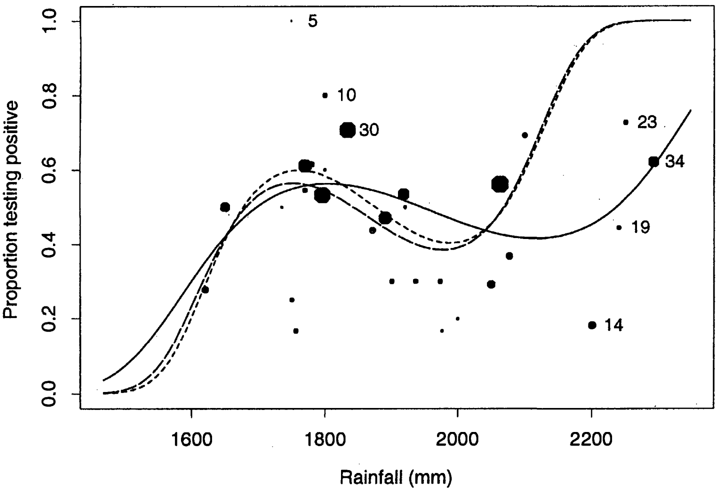

Fig. 11. Toxoplasmosis data—proportion testing positively *versus* rainfall for 34 cities in El Salvador (observed proportions are plotted with diameter proportional to the binomial denominator to the power 0.8): ———, fitted cubic function using all the observations ($m = 34$); ---------, fitted cubic function when $m = 30$; – – – –, fitted cubic function when $m = 29$

reduce the curvature of the fitted cubic model. Observations 5 and 10 do not have a similar effect, but they are for two and 10 subjects, whereas observation 30 is from 75.

The deletion of these five observations has other beneficial effects. The residual deviance is 36.42 on 25 degrees of freedom, which is still perhaps some evidence of a lack of fit, if asymptotic theory is a good guide, but a decided improvement on the previous value. The deletion of one further observation gives a value of 1.05 for $\hat{\phi}_{S_*^{28}}$, removing any evidence of that overdispersion which caused Firth (1991) to wonder whether the model was appropriate. Of course, to remove observations solely to achieve a small deviance is not likely to lead to a correct model for the data. But our results show how many aspects of model building and criticism come together once the observations have been ordered by the forward search. As one further example, the t-statistics for the parameters in Fig. 10 are reasonably stable up to $m = 29$. Our analysis was sensibly unchanged if we used the complementary log–log-link.

4. Comparison with backward methods

The structure that we have discussed was found in the data by a forward search. An important part of our argument in favour of our method is that the conventional backward methods using deletion diagnostics can fail in the presence of the masking that is visible in Fig. 9. We accordingly compare our results with those of Lee and Fung (1997) based on deletion diagnostics and show that their method fails to detect the group of four influential observations. A typical starting-point for such a deletion approach is the normal plot of deletion residuals shown in Fig. 12, which are the signed square roots of the changes in deviance as each observation is deleted in turn. In order, the five observations with largest absolute residuals are 27, 30, 14, 21 and 28. These observations were investigated by Lee and Fung for outlyingness and influence. As the plot of the related deviance residuals in Fig. 9 shows, there is appreciable evidence of masking, shown by the rapid change in the plot in the final stages of the forward search. Fig. 12 corresponds to a normal plot of

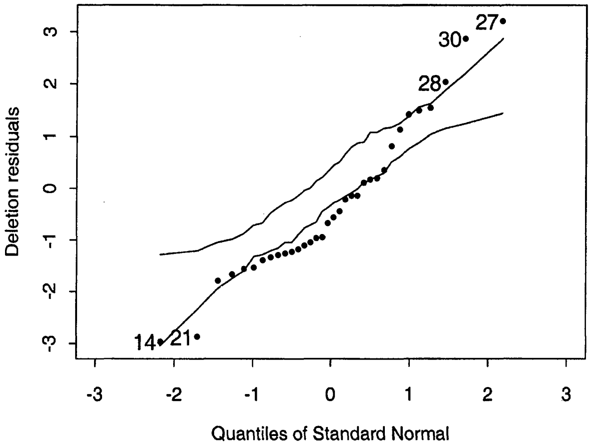

Fig. 12. Toxoplasmosis data—normal plot of deletion residuals with simulation envelope

the residuals in Fig. 9 when $m = 34$, with allowance for individual leverages. It is therefore not at all surprising that Lee and Fung failed to identify the group of four influential observations which we found for high rainfall. In contrast, the points that they investigated form no particular pattern in plots of the data. A final feature of Fig. 12 that is of interest is that the plot on its own looks curved, but arguably not too far from what might be expected. However, the superimposition of an approximate 90% envelope from 100 simulations shows that there is some systematic lack of fit, with too many negative residuals. Although the backward method alerts us to the fact that something might be wrong, it does not provide any suggestions about what needs to be improved. It seems to us that this comparison of the forward and backward analyses provides a confirmation of the power of our procedure.

5. Comments and conclusions

During the last few years there have been attempts to develop robust algorithms for generalized linear models. Stefanski *et al.* (1986) proposed bounded influence estimators which minimize certain functionals of the asymptotic covariance matrix. Bedrick and Hill (1990) developed tests for single and multiple outliers assuming a logistic slippage model. Morgenthaler (1992) explored the consequences of replacing the L_2-norm by the L_1-norm in the derivation of quasi-likelihoods. Christmann (1994) suggested transforming the data for large strata, followed by the application of the least median of squares algorithm to the transformed data.

The forward search algorithm described here is a powerful and practical high breakdown method for generalized linear models. We stress that the forward analysis is not only an alternative way of looking at the data but also leads naturally to the definition and calculation of a robust and fully efficient forward search estimator. The presentation of the results through plots provides a powerful method for revealing the structure of the data.

For large samples (e.g. $n > 1000$), slight variations of the method must be considered. For example, after choosing the best subset of dimension p we can obtain the k units (e.g. $k = n/2$) with the smallest deviance residuals. The forward search estimator and the monitoring of the statistics can start at step k. Another alternative is initially to divide the data into smaller subgroups and then to amalgamate the best initial subset from each subgroup. This technique is related to that of Woodruff and Rocke (1994).

In this paper we have considered the analysis of binomial data. Applications of our forward search to regression and the analysis of transformations are described in Atkinson and Riani (2000). Programs for our method are available at `http://stat.econ.unipr.it/riani/ar`. The data for the examples in this paper are on `http://stat.econ.unipr.it/riani/statist01`.

Acknowledgement

We are grateful to the referees for their comments which helped to improve the presentation and for, in one case, a heartening enthusiasm for our forward procedure.

References

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

———(1994) Fast very robust methods for the detection of multiple outliers. *J. Am. Statist. Ass.*, **89**, 1329–1339.

Atkinson, A. C. and Riani, M. (1997) Bivariate boxplots, multiple outliers, multivariate transformations and discriminant analysis: the 1997 Hunter lecture. *Environmetrics*, **8**, 583–602.

———(2000) *Robust Diagnostic Regression Analysis*. New York: Springer.

Barrett, B. E. and Gray, J. B. (1997) Leverage, residual, and interaction diagnostics for subsets of cases in least squares regression. *Comput. Statist. Data Anal.*, **26**, 39–52.

Bedrick, E. J. and Hill, J. R. (1990) Outlier tests for logistic regression: a conditional approach. *Biometrika*, **77**, 815–827.

Bliss, C. I. (1935) The calculation of the dosage-mortality curve. *Ann. Appl. Biol.*, **22**, 134–167.

Christmann, A. (1994) Least median of weighted squares in logistic regression with large strata. *Biometrika*, **81**, 413–417.

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

Efron, B. (1978) Regression and ANOVA with zero–one data: measures of residual variation. *J. Am. Statist. Ass.*, **73**, 113–121.

Firth, D. (1991) Generalized linear models. In *Statistical Theory and Modelling* (eds D. V. Hinkley, N. Reid and E. J. Snell), pp. 55–82. London: Chapman and Hall.

Flury, B. (1997) *A First Course in Multivariate Statistics*. New York: Springer.

Haslett, J. (1999) A simple derivation of deletion diagnostic results for the general linear model with correlated errors. *J. R. Statist. Soc.* B, **61**, 603–609.

Hawkins, D. M. (1993) The accuracy of elemental set approximations for regression. *J. Am. Statist. Ass.*, **88**, 580–589.

Lee, A. H. and Fung, W. K. (1997) Confirmation of multiple outliers in generalized linear and nonlinear regressions. *Comput. Statist. Data Anal.*, **25**, 55–65.

McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. London: Chapman and Hall.

Morgenthaler, S. (1992) Least-absolute-deviations fits for generalized linear models. *Biometrika*, **79**, 747–754.

Riani, M. and Atkinson, A. C. (2000) Robust diagnostic data analysis: transformations in regression (with discussion). *Technometrics*, **42**, 384–398.

Rousseeuw, P. J. (1984) Least median of squares regression. *J. Am. Statist. Ass.*, **79**, 871–880.

Stefanski, L. A., Carroll, R. J. and Ruppert, D. (1986) Optimally bounded score functions for generalized linear models with applications to logistic regression. *Biometrika*, **73**, 413–424.

Woodruff, D. and Rocke, D. M. (1994) Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *J. Am. Statist. Ass.*, **89**, 888–896.

Zelterman, D. (1999) *Models for Discrete Data*. Oxford: Oxford University Press.