Biometrika (1988), **75**, 4, *pp.* 651-9
Printed in Great Britain

A Bayesian approach to outlier detection and residual analysis

BY KATHRYN CHALONER

Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota 55108, U.S.A.

AND ROLLIN BRANT

Department of Preventive Medicine and Biostatistics, University of Toronto, Toronto M5S 1A8, Canada

SUMMARY

An approach to detecting outliers in a linear model is developed. An outlier is defined to be an observation with a large random error, generated by the linear model under consideration. Outliers are detected by examining the posterior distribution of the random errors. An augmented residual plot is also suggested as a graphical aid in finding outliers.

Some key words: Leverage; Linear model; Posterior distribution; Residual plot.

1. INTRODUCTION

We propose a precise definition of an outlier in a linear model which appears to lead to simple ways of exploring data for the possibility of outliers. The definition is such that, if the parameters of the model are known, then it is also known which observations are outliers. Alternatively, if the parameters are unknown, the posterior distribution can be used to calculate the posterior probability that any observation is an outlier. In a linear model with normally distributed random errors, ε_i, with mean zero and variance σ^2, we declare the ith observation to be an outlier if $|\varepsilon_i| > k\sigma$ for some choice of k. The value of k can be chosen so that the prior probability of an outlier is small and thus outliers are observations which are more extreme than is usually expected. Realizations of normally distributed errors of more than about three standard deviations from the mean are certainly surprising, and worth further investigation. Such outlying observations can occur under the assumed model, however, and this should be taken into account when deciding what to do with outliers and in choosing k. Note that ε_i is the actual realization of the random error, not the usual estimated residual $\hat{\varepsilon}_i$.

The problem of outliers is studied and thoroughly reviewed by Barnett & Lewis (1984), Hawkins (1980), Beckman & Cook (1983) and Pettit & Smith (1985). The usual Bayesian approach to outlier detection uses the definition given by Freeman (1980). Freeman defines an outlier to be 'any observation that has not been generated by the mechanism that generated the majority of observations in the data set'. Freeman's definition therefore requires that a model for the generation of outliers be specified and is implemented by, for example, Box & Tiao (1968), Guttman, Dutter & Freeman (1978) and Abraham & Box (1978). Our method differs in that we define outliers as arising from the model under consideration rather than arising from a separate, expanded, model. Our approach is similar to that described by Zellner & Moulton (1985) and is an extension of the philosophy of Zellner (1975). Geisser (1980, 1987, 1988a, b) develops a different set of diagnostic tools, called predictive discordancy diagnostics, which also do not require a model for the generation of outliers.

2. A METHOD OF OUTLIER DETECTION

Assume that the model under consideration is the usual linear model with parameters $\theta^{\mathrm{T}}=(\theta_1,\ldots,\theta_p)$ and normally distributed, $N(0,\sigma^2)$, independent random errors $\varepsilon^{\mathrm{T}}=(\varepsilon_1,\ldots,\varepsilon_n)$. The $n\times p$ design matrix is X with observations taken at $x_1,\ldots,x_n$ and $y=X\theta+\varepsilon$. To compute the posterior probability that $|\varepsilon_i|$ is greater than $k\sigma$ we need the posterior distribution of ε, $p(\varepsilon|y)$. This posterior distribution is derived assuming an improper prior distribution of Zellner (1975) and can be derived in a similar way assuming a normal-gamma prior distribution.

Define $\tau=\sigma^{-2}$ and let R be a specified positive-definite matrix. The normal-gamma conjugate prior distribution takes θ to have a normal distribution, conditional on τ, with mean m_0 and variance $\tau^{-1}R^{-1}$. The prior distribution of τ is a gamma distribution with parameters a and b; that is $p(\tau)\propto\tau^{a-1}e^{-b\tau}$ and the prior mean is ab^{-1}. The posterior distribution of θ, conditional on τ, is a normal distribution with mean $m_1=(R+X^{\mathrm{T}}X)^{-1}(Rm_0+X^{\mathrm{T}}y)$ and covariance matrix $\tau^{-1}(R+X^{\mathrm{T}}X)^{-1}$. The posterior distribution of τ is a gamma distribution with parameters a_1 and b_1, where $a_1=a+\frac{1}{2}n$ and

$$b_1=b+\tfrac{1}{2}\{(y-xm_1)^{\mathrm{T}}y+(m_0-m_1)^{\mathrm{T}}Rm_0\}.$$

The posterior distribution of ε is easily derived by writing $\varepsilon=y-Xe$, and noting that ε is a linear function of θ. The distribution of ε is singular as its mass is on a p-dimensional space only. Let $H=X(R+X^{\mathrm{T}}X)^{-1}X$. Then, conditional on τ, the posterior distribution of ε is a singular multivariate normal distribution with mean $\hat\varepsilon=y-Xm_1$ and covariance matrix $\tau^{-1}H$. Denote the elements of H by h_{ij}. Each ε_i, for $i=1,\ldots,n$, has a t distribution with location $\hat\varepsilon_i$, precision $a_1/(b_1h_{ii})$, and $2a_1$ degrees of freedom (DeGroot, 1970, p. 42). The covariance matrix of ε is proportional to H.

To compute the posterior distribution corresponding to the improper prior distribution $p(\theta,\tau)=\tau^{-1}$, let $R\to 0$, $a\to-\frac{1}{2}p$ and $b\to 0$. Let

$$\hat\theta=(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y,\quad s^2=(y-X\hat\theta)^{\mathrm{T}}(y-X\hat\theta)/(n-p);$$

then, as DeGroot (1970, p. 252) shows, the posterior distribution of ε is again a multivariate t distribution on a p-dimensional space, with

$$\hat\varepsilon=y-X\hat\theta,\quad H=X(X^{\mathrm{T}}X)^{-1}X,\quad a_1=\tfrac{1}{2}(n-p),\quad b_1=\tfrac{1}{2}(n-p)s^2.$$

To detect which observations are outliers, define the probability p_i to be $\mathrm{pr}(|\varepsilon_i|>k\sigma|y)$, the posterior probability that the ith observation is what we have defined to be an outlier. Let $\Phi(z)$ be the standard normal distribution function. Further let

$$z_1=(k-\hat\varepsilon_1\surd\tau)/(\surd h_{ii}),\quad z_2=(-k-\hat\varepsilon_i\surd\tau)/(\surd h_{ii}); \qquad (1)$$

then we have

$$p_i=\mathrm{pr}(|\varepsilon_i|>k\sigma|y)=\int\{1-\Phi(z_1)+\Phi(z_2)\}p(\tau|y)\,d\tau. \qquad (2)$$

The p_i's can be compared to the prior probability $2\Phi(-k)$. Points with a high posterior probability p_i of being an outlier will have a large $|\hat\varepsilon_i|$, or a large h_{ii}, or both. When $|\hat\varepsilon_i|$ is large this suggests $|\varepsilon_i|$ is large. When h_{ii} is large there is uncertainty about

ε_i. The quantity h_{ii} is often referred to as leverage, for example by Cook & Weisberg (1982, p. 15), as points with large h_{ii} are potentially highly influential.

In addition to examining the p_i's we can also examine the joint posterior probability that two observations are outliers. Define z_{1i} and z_{2i} to be the values of z_1 and z_2 defined in (1) for the ith observation and z_{1j} and z_{2j} the corresponding values for the jth observation. In addition, define ρ_{ij} to be $h_{ij}/\surd(h_{ii}h_{jj})$, the correlation between ε_i and ε_j. Let $B(a, b, \rho)$ be the probability, for a standard bivariate normal random variable with correlation ρ, that the first coordinate is larger than a and the second is larger than b. Then the posterior probability that ε_i and ε_j are both outliers is

$$\begin{aligned} p_{ij} &= \text{pr}(|\varepsilon_i| > k\sigma \text{ and } |\varepsilon_j| > k\sigma | y) \\ &= \int \{B(z_{1i}, z_{1j}, \rho_{ij}) + B(-z_{2i}, -z_{2j}, \rho_{ij}) + B(z_{1i}, -z_{2j}, -\rho_{ij}) \\ &\qquad + B(-z_{2i}, z_{1j}, -\rho_{ij})\} p(\tau|y)\, d\tau. \qquad (3) \end{aligned}$$

The p_{ij}'s can be compared to the prior probability $\{2\Phi(-k)\}^2$. In the presence of multiple outliers masking is said to occur when a test for a single outlier does not detect one outlier in the presence of another outlier. In the approach described here masking could be said to occur when one of the individual probabilities p_i, say, does not indicate an outlier, but for some j the probability p_{ij} is larger than the prior probability of two outliers. This will necessarily entail some correlation between ε_i and ε_j. Such masking occurs in an example in § 5.

The usual residuals $\hat{\varepsilon}_i$ are the posterior means of the ε_i and can be thought of as point estimates of the ε_i. Interval estimates of ε_i are easily constructed. Let $t(\frac{1}{2}\alpha, n-p)$ be the upper $\frac{1}{2}\alpha$ point for a t distribution with $n-p$ degrees of freedom; then a $(1-\alpha)$ highest posterior density interval for ε_i is $\hat{\varepsilon}_i \pm t(\frac{1}{2}\alpha, n-p)s\surd h_{ii}$. These intervals can be added to a residual plot; for examples, see § 5. The plots can be thought of as describing the marginal posterior distributions of the ε_i, the realized errors. This is a different interpretation from the usual approach of thinking of a residual plot as representing the sampling distribution of the $\hat{\varepsilon}_i$.

The posterior distribution of the realized errors is quite different from the sampling distribution of the $\hat{\varepsilon}_i$. Note that the posterior distribution is over p dimensions whereas the sampling distribution of the $\hat{\varepsilon}_i$ is over $(n-p)$ dimensions. The posterior distribution of the ε_i treats the $\hat{\varepsilon}_i$ and $\hat{\theta}$ as fixed and the uncertainty is in $X\theta$ over the p-dimensional space which is orthogonal to the space spanned by the $\hat{\varepsilon}_i$. In contrast the sampling distribution of the $\hat{\varepsilon}_i$ has mean zero and a covariance matrix proportional to $(I-H)$. The residual plot, as a representation of the posterior distribution of the ε_i, represents p correlated quantities, whereas the frequentist interpretation of the residual plot represents $n-p$ correlated quantities.

3. Choice of k

The value of k can be chosen so that the prior probability of no outliers is large, say 0·95. This gives $k = \Phi^{-1}\{0{\cdot}5 + \frac{1}{2}(0{\cdot}95^{1/n})\}$, which for $n = 20$ is $k = 3{\cdot}0$, for $n = 50$ is $k = 3{\cdot}3$, for $n = 100$ is $k = 3{\cdot}5$ and for $n = 1000$ is $k = 4{\cdot}0$. Any observation with posterior probability of being an outlier larger than the prior probability $2\Phi(-k)$ would be suspect.

Alternatively, if the model under consideration is required to describe the data, rather than being considered as a stochastic model, then $k = 2$ might be used to find observations which are not well described by the data no matter what the sample size. If more than 5% have high posterior probability of being greater than $k=2$ standard deviations then this is cause for concern in that the model does not describe the complete data set well.

4. OTHER METHODS

Standardized residuals are often used to detect outliers. These standardized residuals, $r_i=\hat{\varepsilon}_i/\{s\surd(1-h_{ii})\}$, have a constant sampling variance and are often plotted in residual plots instead of the $\hat{\varepsilon}_i$. A monotone transformation of r_i, called an externally standardized residual, forms the basis of an outlier test. Let $y_{(i)}$ be the observations excluding y_i; then the externally standardized residuals are $t_i=\hat{\varepsilon}_i/\{s_{(i)}\surd(1-h_{ii})\}$, where $s_{(i)}$ is based on the regression of $y_{(i)}$ only. The distribution of t_i is a t distribution with $(n-p-1)$ degrees of freedom under the null hypothesis of no shift in the mean of the ith observation. The Bonferroni inequality is commonly used to provide critical values (Weisberg, 1985, Ch. 5).

Another approach is Geisser's (1980, 1987, 1988a, b) predictive method, based on the predictive distribution of the ith observation given $y_{(i)}$, also discussed by Pettit & Smith (1985). The conditional predictive ordinate, c_i, is the predictive density of the ith observation given $y_{(i)}$, evaluated at the observed value y_i; that is $c_i=p(y_i\,|\,y_{(i)})$. The values c_i give a ranking of the observations, with the most discordant having the smallest value of c_i. The c_i work best for situations with y_i independent and identically distributed where $p(y_i\,|\,y_{(i)})$ has the same scale factor for all i. Geisser (1987) suggests a further diagnostic measure which he recommends for regression problems where $p(y_i\,|\,y_{(i)})$ have different scale factors. This measure is the tail area of $p(y_i\,|\,y_{(i)})$, and is interpreted as a p-value for a 'predictive discordancy test'. This is the probability, under $p(y_i\,|\,y_{(i)})$, of an observation with a smaller c_i than that of the observed y_i, and will be denoted pd_i. In the case of the linear model with a noninformative prior distribution it can be shown that algebraically the predictive discordancy diagnostics are closely related to standardized residuals. The value of pd_i, for example, is the p-value from the outlier test based on comparing t_i, the externally standardized residual, to its t-distribution, not using the Bonferroni inequality. Similarly, the value of c_i is exactly the density of a t distribution with $(n-p-1)$ degrees of freedom evaluated at t_i, multiplied by a scale factor $(1-h_{ii})^{\frac{1}{2}}/s_{(i)}$. Other related conditional predictive discordancy diagnostics are suggested and discussed by Geisser. These diagnostics are very general in that they can be applied to many different models.

Our posterior probabilities are based on all the data y rather than $y_{(i)}$. Conditioning on all the data including the ith case is very natural when interest is in $\varepsilon_i=y_i-x_i^{\mathrm{T}}\theta$, a function of y_i. The observation y_i provides almost the only information available about the location of the realized error ε_i. Indeed, if the precision, τ, is known then the posterior distribution $p(\varepsilon_i\,|\,y_{(i)})$ is exactly the prior distribution of ε_i. If τ is unknown the posterior distribution $p(\varepsilon_i\,|\,y_{(i)})$ has a mean of zero and a variance which does not approach zero as the sample size increases. Our posterior probability is computed under the assumption that the model is correct and using all the data is not unreasonable under that assumption.

5. EXAMPLES

The use of the residual plot and the diagnostic posterior probabilities will now be illustrated using two examples. The residuals and highest posterior density intervals are plotted against case number for both examples, but they could also have been plotted against the fitted values or values of an explanatory variable. The posterior distribution from the noninformative prior distribution is used in both examples. The first example uses the Gesell adaptive score data (Mickey, Dunn & Clark, 1967) given in Table 1.

Setting the prior probability of no outliers to be 0·95 gives $k=3{\cdot}0$ to define an outlier. Table 2 gives the posterior probabilities $\mathrm{pr}\,(|\varepsilon_i|>3\sigma|y)$ and, for illustration, $\mathrm{pr}\,(|\varepsilon_i|>2\sigma|y)$; these are denoted as $p(3)_i$ and $p(2)_i$ respectively. Other diagnostic measures are also given in Table 2 for comparison. Figure 1 is the augmented residual plot of $\hat{\varepsilon}_i$ against the case number, i, with 95% highest posterior density intervals drawn in. The plot provides a simple summary of the information in the data about the ε_i's. Observation 19 is noticeable as a point with large $\hat{\varepsilon}_i$ and small posterior variance. Observation 18 has a residual with large posterior variance.

Table 1. *Gesell adaptive score data*: y_i *is adaptive score and* x_i *is age, in months, at first word*

i	y_i	x_i	i	y_i	x_i	i	y_i	x_i
1	95	15	8	100	11	15	102	11
2	71	26	9	104	8	16	100	10
3	83	10	10	94	20	17	105	12
4	91	9	11	113	7	18	57	42
5	102	15	12	96	9	19	121	17
6	87	20	13	83	10	20	86	11
7	93	18	14	84	11	21	100	10

Table 2. *Table of residuals* $\hat{\varepsilon}_i$, *leverages* h_{ii}, *posterior probabilities* $\mathrm{pr}\,(|\varepsilon_i|>2\sigma|y)=p(2)_i$ *and* $\mathrm{pr}\,(|\varepsilon_i|>3\sigma|y)=p(3)_i$, *standardized residuals* r_i, *externally standardized residuals* t_i, *conditional predictive ordinates* c_i, *and predictive discordancy p-values* pd_i, *for Gesell adaptive score data*

i	$\hat{\varepsilon}_i$	h_{ii}	$p(2)_i$	$p(3)_i$	r_i	t_i	c_i	pd_i
1	2·03	0·05			0·19	0·18	0·0333	0·8561
2	−9·57	0·15	0·0031		−0·94	−0·94	0·0207	0·3589
3	−15·60	0·06	0·0391		−1·46	−1·51	0·0114	0·1482
4	−8·73	0·07			−0·82	−0·81	0·0242	0·4261
5	9·03	0·05			0·84	0·83	0·0241	0·4158
6	−0·33	0·07			−0·03	−0·03	0·0334	0·9759
7	3·41	0·06			0·32	0·31	0·0321	0·7592
8	2·52	0·06			0·24	0·23	0·0329	0·8209
9	3·14	0·08			0·30	0·29	0·0320	0·7752
10	6·67	0·07			0·63	0·62	0·0277	0·5445
11	11·02	0·09	0·0016		1·05	1·05	0·0194	0·3072
12	−3·73	0·07			−0·35	−0·34	0·0316	0·7357
13	−15·60	0·06	0·0391		−1·46	−1·51	0·0148	0·1482
14	−13·48	0·06	0·0057		−1·26	−1·28	0·0154	0·2169
15	4·52	0·06			0·42	0·41	0·0310	0·6844
16	1·40	0·06			0·13	0·13	0·0334	0·9000
17	8·65	0·05			0·81	0·80	0·0247	0·4351
18	−5·54	0·65	0·0329	0·0010	−0·85	−0·85	0·0144	0·4091
19	30·28	0·05	0·9261	0·2776	2·82	3·61	0·0002	0·0020
20	−11·48	0·06			−1·07	−1·08	0·0192	0·2959
21	1·40	0·06			0·13	0·13	0·0334	0·9000

Posterior probabilities less than 10^{-4} have been omitted.

The prior probability of $|\varepsilon_i| > 3\sigma$ is 0·0027 and observation 19 is the only observation with posterior probability larger than this of being an outlier, with a posterior probability of 0·28. Observation 18 is associated with the next largest probability of 0·0010 but this is, however, smaller than the prior probability. Observation 18 has largest leverage, h_{18}, and although there is uncertainty about ε_{18}, the uncertainty does not indicate that observation 18 is an outlier. All other observations have posterior probability of being an outlier less than 10^{-4}. The ordering of the posterior probabilities depends on k. For $k=2$ the ordering is different from that for $k=3$.

The standardized residuals and predictive discordancy diagnostics are shown for comparison in Table 2. The outlier test using t_{19} and the Bonferroni inequality is just significant at a 0·05 level. The predictive discordancy diagnostics identify observation 19 as the most discordant. It is not clear how small either c_i or pd_i should be to indicate an observation is discordant or unusual but the values associated with observation 19 are much smaller than the values for the other observations. All methods therefore point to observation 19 as an outlier.

Our second example, a more complicated one, is the stack loss data from Brownlee (1965, p. 454), with three explanatory variables, discussed by Atkinson (1985), and used by Atkinson (1986) and Hawkins, Bradu & Kass (1984) to demonstrate masking and the detection of multiple outliers.

The residual plot for the stack loss data augmented with highest posterior density intervals is given as Fig. 2. Taking the prior probability of no outliers to be 0·95 again leads to $k=3{\cdot}0$. Table 3 gives the posterior probabilities for being an outlier for all observations together with the residuals $\hat{\varepsilon}_i$, standardized residuals r_i and t_i, and predictive discordancy diagnostics c_i and pd_i.

Fig. 1

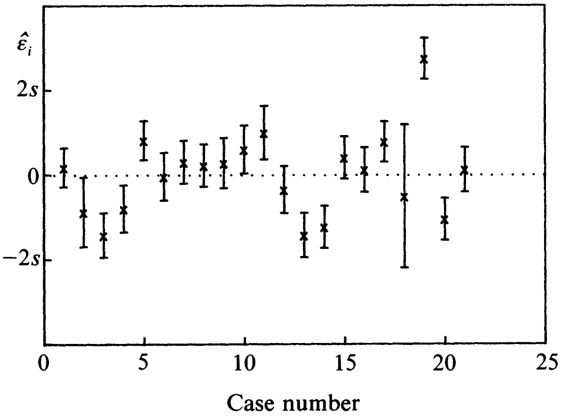

Fig. 2

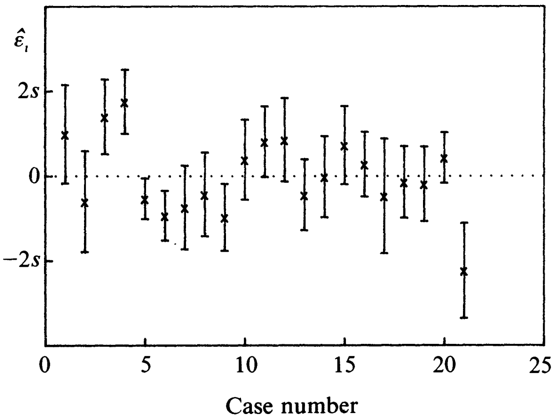

Fig. 1. Plot of residuals against case numbers with 95% highest posterior density regions: Gessel adaptive score data.

Fig. 2. Plot of residuals against case number with 95% highest posterior density regions: stack loss data.

Figure 2 indicates that observation 21 has a large $\hat{\varepsilon}_i$ and also a relatively large posterior variance. The posterior probability that observation 21 is an outlier is 0·11 and is large compared to the prior probability of 0·0027. Observation 4 has the next largest posterior probability of 0·0038, also larger than the prior probability.

To examine further observations 4 and 21 and to examine the possibility of additional outliers, consider the posterior probabilities p_{ij} as given in (3) for all pairs i and j, for

Table 3. *Table of residuals $\hat{\varepsilon}_i$, leverages h_{ii}, probabilities* $\mathrm{pr}\,(|\varepsilon_i| > 2\sigma|y) = p(2)_i$ *and* $\mathrm{pr}\,(|\varepsilon_i| > 3\sigma|y) = p(3)_i$, *standardized residuals r_i, externally standardized residuals t_i, conditional predictive ordinates c_i, and predictive discordancy p-values pd_i, for the stack loss data*

i	$\hat{\varepsilon}_i$	h_{ii}	$p(2)_i$	$p(3)_i$	r_i	t_i	c_i	pd_i
1	3·23	0·30	0·0385	0·0002	1·19	1·21	0·0488	0·2440
2	−1·92	0·32	0·0067		−0·72	−0·71	0·0760	0·4909
3	4·56	0·17	0·1008	0·0004	1·55	1·62	0·0317	0·1252
4	5·70	0·13	0·2806	0·0038	1·88	2·05	0·0169	0·0569
5	−1·71	0·05			−0·54	−0·53	0·0995	0·6030
6	−3·01	0·08	0·0004		−0·97	−0·96	0·0719	0·3500
7	−2·39	0·22	0·0043		−0·83	−0·83	0·0743	0·4210
8	−1·39	0·22	0·0004		−0·48	−0·47	0·0929	0·6421
9	−3·14	0·14	0·0055		−1·05	−1·05	0·0640	0·3100
10	1·27	0·20	0·0002		0·44	0·43	0·0960	0·6756
11	2·64	0·16	0·0021		0·88	0·88	0·0741	0·3927
12	2·78	0·22	0·0091		0·97	0·97	0·0660	0·3481
13	−1·43	0·16			−0·48	−0·47	0·0967	0·6455
14	−0·05	0·20			−0·02	−0·01	0·1047	0·9867
15	2·36	0·19	0·0024		0·81	0·80	0·0772	0·4351
16	0·91	0·13			0·30	0·30	0·1050	0·7747
17	−1·52	0·41	0·0088		−0·61	−0·60	0·0754	0·5572
18	−0·46	0·16			−0·15	−0·15	0·1065	0·8837
19	−0·60	0·17			−0·20	−0·20	0·1047	0·8462
20	1·41	0·08			0·45	0·44	0·1022	0·6636
21	−7·24	0·28	0·6174	0·1117	−2·64	−3·33	0·0015	0·0042

Posterior probabilities less than 10^{-4} have been omitted.

$i \neq j$. Table 4 gives the seven values which are larger than the prior probability of $7{\cdot}3 \times 10^{-6}$. The ratio of the posterior probabilities to the prior probability and the posterior correlations ρ_{ij} are also given. Examining these joint probabilities, p_{ij}, has led to the indication that, in addition to observations 4 and 21 being outliers, 1 and 3 are also outliers. The two observations 1 and 3 are difficult to detect and masking has occurred as both p_1 and p_3 are less than the prior probability of 0·0027. The value of $\rho_{1,3}$, the posterior correlation, is 0·96. The high posterior correlation leads to a large posterior probability that both ε_1 and ε_3 are large. This dependence must be present for such masking to occur.

Table 4. *Table of posterior probabilities* $p_{ij} = \mathrm{pr}\,(|\varepsilon_i| > 3\sigma$ *and* $|\varepsilon_j| > 3\sigma|y)$ *for the stack loss data, for pairs with p_{ij} larger than the prior probability of* $(0{\cdot}0027)^2$. *Also given is the odds ratio of posterior probability to prior probability and ρ_{ij} the posterior correlation between ε_i and ε_j*

i, j	p_{ij}	$p_{ij}/(0{\cdot}0027)^2$	ρ_{ij}
1, 3	$1{\cdot}6 \times 10^{-4}$	22·1	0·96
1, 4	$4{\cdot}5 \times 10^{-5}$	6·2	0·43
1, 21	$1{\cdot}2 \times 10^{-5}$	1·7	0·40
2, 21	$8{\cdot}9 \times 10^{-6}$	1·2	0·37
3, 4	$1{\cdot}2 \times 10^{-4}$	16·8	0·40
3, 21	$3{\cdot}4 \times 10^{-5}$	4·7	0·58
4, 21	$2{\cdot}8 \times 10^{-3}$	384·7	−0·31

The conditional probability that observation 4 is an outlier given that observation 21 is an outlier can also be obtained by dividing the joint probability $p_{4,21}$ by the marginal probability p_{21}. This conditional probability is 0·0251, which is larger than the prior probability of 0·0027 and also larger than the unconditional probability of 0·0038. Observations 1, 3, 4 and 21 were identified as outliers by Atkinson (1986) who used this data set to illustrate a method for detecting outliers based on a robust regression. Hawkins et al. (1984), using a different method, find the same 4 outliers but give some evidence that there may, perhaps, be more outliers and observation 2 may be suspect.

Comparing these posterior probability calculations to other methods, the largest standardized residual is $t_{21} = -3{\cdot}33$. It is not clear what significance level to use for comparison, but as k was chosen to give the prior probability of no outliers to be 0·95, we use an overall level of 0·05. The outlier test for t_{21} is then not quite significant using the Bonferroni inequality and therefore this test has not identified any outliers.

The conditional predictive ordinate and the predictive discordancy p-value both identify observation 21 as the most discordant, with c_{21} and pd_{21} much smaller than the other values. These diagnostics do not, however, clearly identify multiple outliers. Our method has diagnosed a problem, if a regression model is used, with multiple outliers and masking. The other methods, for detecting single outliers, do not clearly indicate the multiple outliers.

In summary, we note that our diagnostic measure, the posterior probability, is appropriate for situations in which there is no obvious way of modelling contaminants. Our definition of an outlier is simple and requires that outliers are unusual and are outlying. Other definitions, for example, a mixture model with inflated variance, are appropriate if a model for such contamination is suggested by the information about the process generating the data.

ACKNOWLEDGEMENT

We are grateful to R. D. Cook, D. Hawkins, K. Larntz, L. Tierney, and the referees for helpful comments and also to A. Zellner and participants in the 34th NBER-NSF Seminar on Bayesian Inference in Econometrics for helpful criticism of an earlier version of this paper. This research was partly supported by a grant from the National Science Foundation.

REFERENCES

ABRAHAM, B. & BOX, G. E. P. (1978). Linear models and spurious observations. *Appl. Statist.* **27**, 120–30.

ATKINSON, A. C. (1985). *Plots, Transformations and Regression.* Oxford: Clarendon Press.

ATKINSON, A. C. (1986). Masking unmasked. *Biometrika* **73**, 533–41.

BARNETT, V. & LEWIS, T. (1984). *Outliers in Statistical Data*, 2nd ed. Chichester: Wiley.

BECKMAN, R. J. & COOK, R. D. (1983). Outlier . . . . .s (with discussion). *Technometrics* **25**, 119–63.

BOX, G. E. P. & TIAO, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–29.

BROWNLEE, K. A. (1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd ed. New York: Wiley.

COOK, R. D. & WEISBERG, S. (1982). *Residuals and Influence in Regression.* London: Chapman and Hall.

DEGROOT, M. H. (1970). *Optimal Statistical Decisions.* New York: McGraw-Hill.

FREEMAN, P. R. (1980). On the number of outliers in data from a linear model. In *Bayesian Statistics*, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 349–65. Valencia: University Press.

GEISSER, S. (1980). Discussion of a paper by G. E. P. Box. *J. R. Statist. Soc.* A **143**, 416–7.

GEISSER, S. (1987). Influential observations, diagnostics and discordancy tests. *J. Appl. Statist.* **14**, 133–42.

GEISSER, S. (1988a). Predictive approaches to discordancy testing. In *Bayesian and Likelihood Inference: Essays in Honor of George A. Barnard*, Ed. S. Geisser, J. S. Hodges, S. J. Press and A. Zellner. To appear. Amsterdam: North-Holland.

GEISSER, S. (1988b). The future of statistics in retrospect. In *Bayesian Statistics*, **3**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith. To appear. Oxford University Press.

GUTTMAN, I., DUTTER, R. & FREEMAN, P. R. (1978). Care and handling of univariate outliers in the general linear model to detect spuriosity—a Bayesian approach. *Technometrics* **20**, 187-93.

HAWKINS, D. M. (1980). *Identification of Outliers.* London: Chapman and Hall.

HAWKINS, D. M., BRADU, D. & KASS, G. V. (1984). Location of several outliers in multiple regression data using elemental sets. *Technometrics* **26**, 197-208.

MICKEY, M. R., DUNN, O. J. & CLARK, V. (1967). Note on the use of stepwise regression in detecting outliers. *Comp. Biomed. Res.* **1**, 105-11.

PETTIT, L. I. & SMITH, A. F. M. (1985). Outliers and influential observations in linear models. In *Bayesian Statistics*, **2**, Ed. J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, pp. 473-94. Amsterdam: North-Holland.

WEISBERG, S. (1985). *Applied Linear Regression*, 2nd ed. New York: Wiley.

ZELLNER, A. (1975). Bayesian analysis of regression error terms. *J. Am. Statist. Assoc.* **70**, 138-44.

ZELLNER, A. & MOULTON, B. R. (1985). Bayesian regression diagnostics with applications to international consumption and income data. *J. Econometrics* **29**, 187-211.

[*Received June* 1987. *Revised April* 1988]