A Bayesian Approach to Residual Analysis and Outlier Detection.

Kathryn Chaloner and Rollin Brant
Department of Applied Statistics
University of Minnesota
St. Paul, MN 55108

Technical Report No. 480
October, 1986

SUMMARY

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

Keywords: Leverage; Linear model; Posterior distribution; Residual plot.

1. Introduction

Methods for detecting outliers have been plagued by the lack of a satisfactory definition for an outlier. We propose a very precise definition for an outlier which simplifies the concepts behind outlier detection and allows a simple Bayesian analysis. In a linear model with normally distributed random errors, e_i, with mean zero and variance σ^2, we declare the ith observation to be an outlier if $|e_i| > k\sigma$. The choice of k can be left to the data analyst and we will use values of k=2 and k=3 in our illustrative example. Realizations of normally distributed random variables of more than two or three standard deviations from the mean are certainly surprising.

The method will use the posterior distribution of the random errors and it will be seen that the leverage of an observation is accounted for in a very natural way through the posterior variance. A residual plot augmented with interval estimates of the errors is suggested.

The problem of outliers is studied and thoroughly reviewed in Barnett and Lewis (1984), Hawkins (1980) and Beckman and Cook (1983). Bayesian methods are reviewed in Chapter 12 of Barnett and Lewis and further developments given in Pettit and Smith (1984). Our method differs from the usual approach in that we define the outliers as arising from the model under consideration rather than arising from a separate model. If the situation is such that there is clearly an appropriate model for contamination such as a shift in the mean or an inflated variance, then such a model should be used. Our approach therefore is a general approach where no obvious contamination model is available. Bayesian methods for contamination models are given in, for

example, Box and Tiao (1968), Guttman, Dutter and Freeman (1978), Abraham and Box (1978) and Freeman (1980).

2. A Method for Outlier Detection

We will assume that the model under consideration for the data $y^T = (y_1,,..y_n)$ is the usual linear model with parameters $\theta^T = (\theta_1,,..\theta_p)$ and normally distributed, $N(0,\sigma^2)$, independent random errors $e^T = (e_1,...e_n)$. The $n \times p$ design matrix is X and $y = X\theta + e$. To compute the posterior probability that $|e_i|$ is greater than $k\sigma$ we need the posterior distribution of e, $p(e|y)$. We first derive this distribution using a normal gamma prior distribution and then take a limit to give the posterior distribution for an improper prior distribution.

The prior distribution on θ is taken to be such that conditional on the precision τ, $\tau = \sigma^{-2}$, θ is normally distributed with mean θ_0 and variance $\tau^{-1}R^{-1}$. The prior distribution on τ is a Gamma distribution with parameters α and β, the prior mean is $\alpha\beta^{-1}$. The posterior distribution is of the same form and, following DeGroot (1970) page 252, conditional on τ the posterior distribution of θ is a normal distribution with mean $\theta_1=(R+X^TX)^{-1}(R\theta_0+X^Ty)$ and variance $\tau^{-1}(R+X^TX)^{-1}$. The posterior distribution of τ is a Gamma distribution with parameters α_1 and β_1 where $\alpha_1 = \alpha + n/2$ and

$$\beta_1 = \beta + \frac{1}{2}\left[(y - X\theta_1)^T y + (\theta_0 - \theta_1)^T R\, \theta_0\right].$$

The posterior distribution of e is easily derived by writing $e=y-X\theta$. The distribution is singular in that mass is on a p-dimensional space only.

Denote $H = X(R+X^TX)^{-1}X$ then conditional on τ the posterior on e is a singular multivariate normal distribution with mean $\hat{e} = y - X\theta_1$, and variance $\tau^{-1}H$. Denote the elements of H as h_{ij} then each e_i, i=1,..n, has a t-distribution with location $\hat{e}_i$, precision $(\alpha_1/\beta_1)h_{ii}^{-1}$, and $2\alpha_1$ degrees of freedom (see DeGroot (1970) page 42). The variance covariance matrix of e is proportional to H.

To compute the posterior distribution corresponding to the improper prior distribution $p(\theta,\tau) = \tau^{-1}$ for $\tau>0$, $\theta \in \mathbb{R}^p$, let $R \to 0$, $\alpha \to p/2$ and $\beta \to 0$. Denote $\hat{\theta} = (X^TX)^{-1}X^Ty$ and $s^2 = (y - X\hat{\theta})^T(y - X\hat{\theta})/(n-p)$ then, as in DeGroot (1970) page 252, the posterior distribution of θ is multivariate-t with location vector $\hat{\theta}$, precision matrix $s^{-2}(X^TX)$ and n-p degrees of freedom. The posterior distribution of e is therefore as it is for the informative prior distribution, but with $\theta_1 = \hat{\theta}$, $H = X(X^TX)^{-1}X^T$, $\alpha_1 = (n-p)/2$ and $\beta_1=(n-p)s^2/2$.

The usual residual, $\hat{e}_i$, can therefore be viewed as the posterior mean of the actually occuring random error, e_i. A $(1-\alpha)$ highest posterior density (hpd) interval for e_i is given by

$$\hat{e}_i \pm t(\alpha/2,\ n-p)\ s\sqrt{h_{ii}}$$

where $t(\alpha/2,\ n-p)$ is the upper $\alpha/2$ percentage point for a t-distribution with n-p degrees of freedom.

The probability $P(|e_i| > k\tau^{-1/2}|y)$ is the probability that the ith observation is what we have defined to be an outlier. These probabilities can be easily computed using numerical integration. Denote $\Phi(z)$ to be the

probability that a standard normal, N(0,1), random variable is less than z. Further denote

$$z_1 = (k - \tau^{1/2}\,\hat{e}_i)\,h_{ii}^{-1/2} \quad \text{and} \quad z_2 = (-k - \tau^{1/2}\,\hat{e}_i)\,h_{ii}^{-1/2}$$

then we have

$$\begin{aligned} & P(|e_i| > k\,\tau^{-1/2} \mid y) \\ &= \int P(|e_i| > k\,\tau^{-1/2} \mid y, \tau)\; p(\tau|y)\; d\tau \\ &= \int (1 - \Phi(z_1) + \Phi(z_2))\; p(\tau|y)\; d\tau. \end{aligned}$$

As these diagnostic measures are probabilities they are easy to interpret.

Points with high probabilities of being an outlier will have a large $|\hat{e}_i|$ a large h_{ii}, or both. When $|\hat{e}_i|$ is large this suggests that $|e_i|$ is large. When h_{ii} is large there is uncertainty about e_i as reflected in the posterior variance. The quantity h_{ii} is often referred to as leverage, as points with large h_{ii} are potentially highly influential (see Cook and Weisberg (1982) page 15).

Using prior information can clearly help in detecting outliers. In order to compare this method to others which do not use prior information, however, we use the posterior distribution from the improper prior distribution in our example.

The posterior distribution of the e_i's is in marked contrast to the sampling distribution of the $\hat{e}_i$'s upon which outlier tests are sometimes based. The variance of $\hat{e}$ is proportional to (I-H) and the distribution of $\hat{e}$ is over (n-p) dimensions. Specifically the sampling distribution of $\hat{e}_i$ is

normal with mean 0 and variance $\tau^{-1}(1-h_{ii})$. Traditionally residuals are standardized by dividing $\hat{e}_i$ by $s(1 - h_{ii})^{-1/2}$. In the extreme if $\hat{e}_i = 0$ then no amount of standardization will yield anything but a value of 0. If the posterior variance of e_i is large however even if $\hat{e}_i = 0$ there may be some probability that $|e_i|$ is large. If we are interested in the $\hat{e}_i$'s as estimates of the e_i's putting error bars on $\hat{e}_i$ is sensible.

The calculations in the example were done by implementing the procedure as an S function (see Becker and Chambers (1985)). The pnorm function from S was used for $\Phi(z)$. The double precision numerical integration routine dqagi and the gamma function gamln from the core math library from the National Bureau of Standards (cmlib) were used. The calculations are easily incorporated into an interactive computing environment.

3. Illustration

The Gessel adaptive score data from Mickey, Dunn and Clark (1967) has been used extensively to illustrate outlier detection. The values of x are the age of a child in months at first word and y is the Gessel adaptive score. The data are given in Table 1 together with the posterior probabilities $P(|e_i| > k\tau^{-1/2}|y)$ for k=2 and k=3. For comparison, the prior probabilities of these events are .0455 and .0027 respectively. Figure 1 is a plot of $\hat{e}_i$ against the case number, i, with 95% hpd regions drawn in. The plot provides a simple summary of the information in the data about the e_i's.

The value of e_{19} is almost certainly large. The posterior probability that e_{19} is more than two standard deviations away from zero is approximately .93. The posterior probability of it being more than three standard deviations from the mean is .28 which is certainly surprising compared to the

prior probability of much less than .01. Observation 18 is associated with the largest leverage h_{18} and therefore there is uncertainty about e_{18}. Observations 2,3,11,13 and 14 might be worth a cursory inspection but the posterior probability of being an outlier at k=2 is less than the prior probability. We see that the ordering of the probabilities changes as k changes, observation 18 is not particularly apparent at k=2 but is noticed at k=3. Observations 18 and 19 are clearly noticeable on the plot.

Without further information on this data set it is impossible to give a definitive analysis. Observation 19 should be examined for a possible typographical error. Alternatively, the child associated with that score may be different from the other 20 children and another explanatory variable might be suggested. Observation 19 might be considered to be an outlier that does not greatly influence the estimates $\hat{\theta}$ but it is suggestive of an inappropriate model. Similarly the child of observation 18 was relatively old at the age of first word and again that child may suggest an additional explantory variable.

Whether the data refer to score on a test or reading of an instrument any large values of $|e_i|$ should be noted. What to do with a large value once it is detected depends on what is being measured and modelled.

4. Conclusion

This method of computing posterior probabilities of an observation being an outlier is very simple and our definition of an outlier is very simple. The probability that observation i is an outlier is computed under the model of interest and some of the computational difficulties of multiple outliers are avoided. The conceptual difficulties of outliers are also avoided by using a precise definition.

A residual plot is routinely drawn in analysing data from a linear model. We believe that thinking of the residuals as estimates of the random errors and including interval estimates on the plot will help in their interpretation.

References

Abraham, B. and Box, G.E.P. (1978) Linear models and spurious observations. Appl. Statist., 27, 120-130.

Barnett, V. and Lewis, T. (1984) Outliers in Statistical Data. Chichester: Wiley.

Becker, R.A. and Chambers, J.W. (1985) Extending the S System. Monterey: Wadsworth.

Beckman, R.J. and Cook, R.D. (1983) Outlier....s (with Discussion). Technometrics. 25, 119-163.

Box, G.E.P. and Tiao, G.C. (1968) A Bayesian approach to some outlier problems. Biometrika, 49, 419-432.

Cook, R.D. and Weisberg, S. (1982) Residuals and Influence in Regression. New York and 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, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., Valencia: University Press.

Guttman, I., Dutter, R. and Freeman, P.R. (1978) Care and handling of univariate outliers in the general linear model to detect spuriosity- A Bayesian approach. Technometrics, 20, 187-193.

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

Mickey, M.R., Dunn, O.J., and Clark, V. (1967) Note on the use of stepwise regression in detecting outliers. Computers and Biomed. Res., 1, 105-111.

Pettit, L.I. and Smith, A.F.M. (1984) Outliers and influential observations in linear models, In Bayesian Statistics 2 J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, eds., Amsterdam: North Holland.

i	y_i	x_i	h_{ii}	$P(\lvert e_i \rvert > k\tau^{-1/2} \mid y)$ k=2	k=3
1	95	15	.05		
2	71	26	.15	.0031	
3	83	10	.06	.0391	
4	91	9	.07		
5	102	15	.05		
6	87	20	.07		
7	93	18	.06		
8	100	11	.06		
9	104	8	.08		
10	94	20	.07		
11	113	7	.09	.0016	
12	96	9	.07		
13	83	10	.06	.0391	
14	84	11	.06	.0057	
15	102	11	.06		
16	100	10	.06		
17	105	12	.05		
18	57	42	.65	.0329	.0010
19	121	17	.05	.9261	.2778
20	86	11	.06	.0005	
21	100	10	.06		

TABLE 1: Gessel adaptive score data with posterior probabilities of being an outlier.

(Probabilities less than 10^{-4} are omitted)

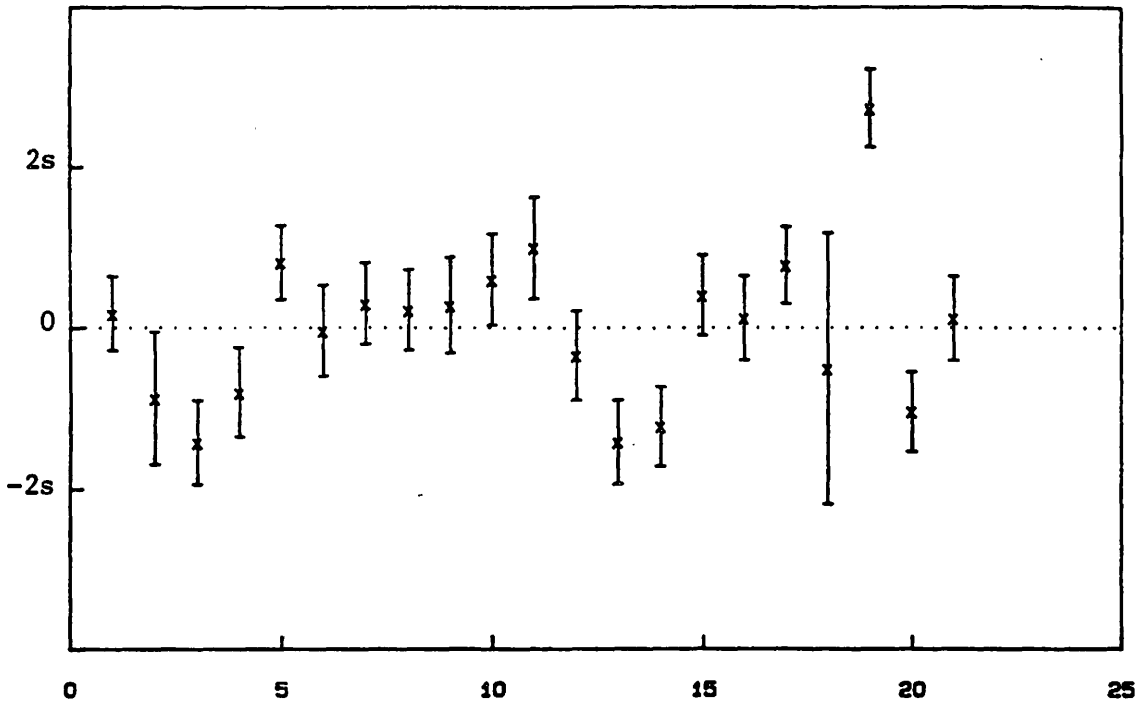

Figure 1: Residual plot of residuals against case number with 95% hpd regions.