Statistica Sinica 3(1993), 367-390

A BAYESIAN LOOK AT DIAGNOSTICS IN THE UNIVARIATE LINEAR MODEL

Irwin Guttman and Daniel Peña

University of Toronto and University Carlos III de Madrid

Abstract: This paper develops diagnostics for data thought to be generated in accordance with the general univariate linear model. A first set of diagnostics is developed by considering posterior probabilities of models that dictate which of k observations from a sample of n observations ($k < n/2$) are *spuriously* generated, giving rise to the possible *outlyingness* of the k observations considered. This is turn gives rise to diagnostics to help assess (estimate) the value of k. A second set of diagnostics is found by using the Kullback-Leibler symmetric divergence, which is found to generate measures of *outlyingness* and *influence*. Both sets of diagnostics are compared and related to each other and to other diagnostic statistics suggested in the literature. An example to illustrate to the use of these diagnostic procedures is included.

Key words and phrases: Spurious and outlying observations, posteriors of models, leverage, Kullback-Leibler measures, outlying and influential observations.

1. Introduction

According to Webster's dictionary, diagnosis is the art of inferring from symptoms or manifestations the nature of an illness or the cause of a situation. One of the most serious "illnesses" that can occur in linear statistical model situations is the presence of outliers, and this fact has motivated the creation of the whole area of robust estimation and outlier testing. From the Bayesian point of view the study of outliers in linear models has already induced a long tradition. In a seminal paper, Box and Tiao (1968) showed that, assuming a normal contaminated distribution for the generation of observations of a linear model, the estimation of the parameters involve a weighted average of estimators from 2^n distributions, where n is the number of observations in the contaminating distribution. Although they were more concerned with estimation than with outlier identification, their approach leads to diagnostics for model heterogeneity, further investigated by Pēna and Tiao (1992). Abraham and Box (1978) introduced heterogenety in the mean instead of in the variance. This mean-shift model was also suggested by Guttman, Dutter and Freeman (1978). These models have been compared by Eddy (1980), Freeman (1980), and Pettit and Smith (1985). In

Section 3 of this paper it is shown that one of the diagnostic measures we suggest can be justified if sampling is from either one of the aforementioned models.

Zellner (1975), Zellner and Moulton (1985) and Chaloner and Brant (1988) define outliers as extreme observations arising from the model under consideration and do not view these as being generated from a mean-shift or variance-shift model. Outliers are then detected by examining the posterior distribution of the random errors.

Since the work of Cook (1977), Cook and Weisberg (1982) and Belsley, Kuh and Welsch (1980), the study of influential observations in a linear model has been an area of very active research. Johnson and Geisser (1983, 1985) built measures of influence in univariate and multivariate linear models by using the Kullback-Leibler divergence between certain predictive or posterior distributions. Related work is found in Pettit and Smith (1985). Guttman and Peña (1988) showed, using the same approach, that a global influence measure built from a certain joint posterior distribution can be decomposed into a measure of outlyingness and a measure of influence, and that this Bayesian diagnostic emcompasses the frequentist diagnostics for outliers and influence. Related work can be found in Ali (1990). Kempthorne (1986) used a formal decision-theoretic set up to justify influence measures in a Bayesian framework. In a similar spirit, Carlin and Polson (1991) have justified taking the Kullback-Leibler divergence as the utility function, and have shown how to compute diagnostics using the Gibbs sampling method.

The objective of this paper is (i) to present diagnostics for heterogeneity based on mean shift or variance-shift models, and (ii) to present diagnostics based on measures of influence derived from Kullback-Leibler divergences. Doing this requires different approaches and assumptions, so that a further objective is to show the relationship of the diagnostics found from (i) and (ii).

In Section 2, we describe two variants of the usual linear model which allow for the generation of spurious observations, namely the so-called 'mean-shift' and 'variance-inflation' models. In Section 3, we derive our first diagnostic c_I, the conditional posterior probability that for given k, a certain set of k out of n observations are generated by the mean-shift model, and show the connection of c_I with the leverage of these k observations. We also demonstrate that c_I is, approximately for large n, the conditional (on k) posterior probability that the k observations have been generated according to the variance-inflation model. Section 4 allows for diagnostics concerning the determination of k. The Kullback-Leibler divergence is used in Section 5 to measure the disparity between various posteriors based on the full sample with those based on a set of $n-k$ observations. With the measures obtained in Section 5, we turn to comparing the behaviour of c_I and the Kullback-Leibler induced diagnostics in Section 6. We then indi-

cate in Section 7, how a procedure using all these diagnostics would proceed, by illustrating with a real set of data.

2. The General Setting

We focus on the analysis of data thought to be generated in accordance with the general univariate linear model, universally denoted as

$$y = X\beta + \epsilon \tag{2.1}$$

where

$$\begin{aligned} X &\quad \text{is} \quad (n \times p),\ r(X) = p < n, \\ \beta &\quad \text{is} \quad (p \times 1), \\ \epsilon &\quad \text{is} \quad N(O, \sigma^2 I_n). \end{aligned} \tag{2.1a}$$

We envisage that, although (2.1) is the intended situation, the experimenter fears (because of experience) that some observations, say y_{i_t}, $t = 1, \ldots, k$, with k fixed and such that $k << n/2$, are *spuriously* generated, with mean-shift spuriosity parameter a_t, that is

$$\begin{aligned} E(y_{i_t}) &= \mathbf{x}'_{i_t}\beta + a_t, \\ V(y_{i_t}) &= \sigma^2, \end{aligned} \qquad t = 1, \ldots, k. \tag{2.2}$$

Denote the set $\{i_1, \ldots, i_k\}$ by I, that is, I is the set of k distinct integers chosen from the set $\{1, \ldots, n\}$. The use of the term "Spurious" above implies that the observations indexed by the set I were generated *not in the manner intended* (as described by (2.1)), but specifically by the generation process (2.2), called the *mean-shift spuriosity model.* If for a given set of observations indexed by $\{i_1, \ldots, i_k\} = I$, (2.2) holds, then, after permutation, write

$$\begin{pmatrix} y_{(I)} \\ \cdots \\ y_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} O \\ \cdots \\ a \end{pmatrix} + \epsilon, \tag{2.3a}$$

where notationally, we mean:

$$(I) = \text{exclude or omit objects connected with the elements of } I = \{i_1, \ldots, i_k\} \tag{2.3b}$$

so that, for example,

$$y_{(I)} = (y_{j_1}, \ldots, y_{j_{n-k}})' \tag{2.3c}$$

where the complement of I is $\{j_1, \ldots, j_{n-k}\} \subset (1, \ldots, n)$. Further, $X_{(I)}$ is the $[(n-k) \times p]$ matrix formed by omitting rows $(i_1, \ldots, i_k)$ from the matrix X of (2.1); we use the notation I to denote:

$$I = \text{use the data indexed by } I \text{ only.} \tag{2.3d}$$

Denote the model described by (2.3a) by $M_I = M_{i_1,\ldots,i_k}$, and note that it says that k observations $y_I = (y_{i_1}, \ldots, y_{i_k})'$, are generated spuriously, while the rest, that is, $n-k$ observations $(y_{i_1}, \ldots, y_{j_{n-k}})' = y_{(I)}$ have been generated as intended. We make one additional assumption, which is:

$$r\left(X_{(I)}\right) = p < n - k. \tag{2.4}$$

Note that in the ensuing sections, the special case $k = 1$ will be delineated and discussed and for this situation we will use the notation $I = i$, etc.

Also note that if we knew one of the $\binom{n}{k}$ models M_I holds, and if we knew exactly which one of these holds, say M_I, then it would be natural to regress $y_{(I)}$ on $X_{(I)}$, forming

$$\hat{\beta}_{(I)} = \left(X'_{(I)}X_{(I)}\right)^{-1} X'_{(I)}\mathbf{y}_{(I)} \tag{2.5}$$

$$S_{(I)} = \mathbf{y}'_{(I)}\left[I_{n-k} - X_{(I)}(X'_{(I)}X_{(I)})^{-1}X'_{(I)}\right]\mathbf{y}_{(I)} \tag{2.5a}$$

etc. $S_{(I)}$ is the sum of squares of residuals based on what is thought to be the "good" data, $(\mathbf{y}_{(I)}, X_{(I)})$, so that $S_{(I)}$ is a measure of scatter.

There are of course other models than (2.3a) for describing the generation of spurious observations; for example, we might have

$$\begin{pmatrix} \mathbf{y}_{(I)} \\ \cdots \\ \mathbf{y}_I \end{pmatrix} = \begin{pmatrix} X_{(I)} \\ \cdots \\ X_I \end{pmatrix} \beta + \begin{pmatrix} \epsilon_{(I)} \\ \cdots \\ \epsilon_I \end{pmatrix} \tag{2.6}$$

with $\epsilon_I \sim N(O, \delta^2\sigma^2 I_k)$, but, as usual, $\epsilon_{(I)} \sim N(O, \sigma^2 I_{n-k})$ and where $\delta^2 > 1$. The model (2.6) is referred to as the variance-inflation model in the literature (see for example, Box and Tiao (1968)).

We turn now to our first diagnostic, and its use in a first part of a diagnosis of a set of data, namely, a diagnostic to detect spurious observations.

3. Diagnosis – Part 1

Faced with the possibility that one of the mean-shift models $\{M_I\}$ as specified by (2.3a) holds, where I ranges over the $\binom{n}{k}$ sets of form $I = \{i_1, \ldots, i_k\}$, a

Bayesian might want to calculate the posterior probability, say c_I, that M_I holds, that is, that $(y_{i_1}, \ldots, y_{i_k}) = \mathbf{y}'_I$ is spuriously generated, and use the $\binom{n}{k}$ c_I's as a set of diagnostics. It turns out that this probability, as derived by Guttman, Dutter and Freeman (1978), is given by

$$c_I = K S_{(I)}^{-(n-k-p)/2} \left| X'_{(I)} X_{(I)} \right|^{-1/2}$$

with (3.1)

$$K^{-1} = \sum{}' S_{(I)}^{-(n-k-p)/2} \left| X'_{(I)} X_{(I)} \right|^{-1/2},$$

where $\sum'$ denotes sum over all the $\binom{n}{k}$ possible sets I. (An alternative approach that leads to (3.1) is given in Draper and Guttman (1987).)

To help interpret the role of c_I's as diagnostics, suppose we consider first the simplest case, where $p = 1$, and it is thought that the generating process of the y's is such that

$$E(y) = \mu \tag{3.2}$$

but it is feared that in a sample of n, model M_I holds, which is to say,

$$E(y_{i_t}) = \mu + a_t, \quad t = 1, \ldots, k$$

while (3.3)

$$E(y_{j_u}) = \mu, \quad u = 1, \ldots, n-k.$$

Suppose indeed that the experimenter fears M_I may hold for $k = 2$, and that a sample of n yielded data which when plotted exhibits an extreme case such as depicted in Figure 3.1.

Figure 3.1. A sample of $n = 10$ observations

Now for this problem $X = \mathbf{1}_{10}$, a (10×1) vector of ones, so that $X_{(I)} = 1_8$ as i ranges over the 45 different sets of 2 integers chosen from $\{1, \ldots, 10\}$. Hence $X'_{(I)} X_{(I)} = n - k = 8$ for all I. Further, for this example, $I = (i_1, i_2) \subset (1, \ldots, 10)$, and

$$\begin{aligned} S_{(I)} &= \mathbf{y}'_{(I)} \left[I_8 - \frac{1}{8} \mathbf{1}_8 \mathbf{1}'_8 \right] y_{(I)} \\ &= \sum_{j \neq i_1, i_2} (y_j - \bar{y}_{(I)})^2 \end{aligned} \tag{3.4}$$

with $\bar{y}_{(I)} = \sum_{j \neq i_1, i_2} y_j/(n-k) = \sum_{j \neq i_1, i_2} y_j/8.$

Hence

$$c_I = \tilde{K}\left(S_{(I)}\right)^{-(n-k-p)/2} = \tilde{K}(S_{(I)})^{-7/2} \tag{3.5}$$

with

$$\tilde{K}^{-1} = \sum{}'(S_{(I)})^{-7/2},$$

since $|X'_{(I)}X_{(I)}| = 8$ for all I. For this example, $\sum'$ denotes the sum over all 45 sets $I = (i_1, i_2)$ of 2 integers chosen from $(1, \ldots, 10)$. Now, as we cycle through the 45 different sets $I = \{i_1, i_2\}$, we will eventually come to the set that excludes the minimum and maximum of the observations shown in Figure 3.1, so that the $S_{(I)}$ that we will then be concerned with, will be minimum amongst all the $S_{(I)}$; and since c_I is proportional to $S_{(I)}^{-(n-k-p)/2} = S_{(I)}^{-7/2}$, the c_I for the case we are discussing will be largest, and in this extreme case, near 1.

We remark that c_I as defined in (3.1) can be expressed as a function of *leverage*. We first note that since $X = (X'_{(I)} : X'_I)'$,

$$X'X = X'_{(I)}X_{(I)} + X'_I X_I \tag{3.6}$$

so that

$$\begin{aligned} \left|X'_{(I)}X_{(I)}\right| &= |X'X| \cdot \left|I_p - (X'X)^{-1}X'_I X_I\right| \\ &= |X'X| \cdot \left|I_k - X_I(X'X)^{-1}X'_I\right|. \end{aligned} \tag{3.7}$$

Absorbing $|X'X|$ into the constant of proportionality K of (3.1), we thus have

$$c_I = K S_{(I)}^{-(n-k-p)/2} \cdot |I_k - H_I|^{-1/2} \tag{3.8}$$

with K defined in the obvious way (see (3.1)), and where

$$H_I = X_I(X'X)^{-1}X'_I \tag{3.9}$$

is that block of the so called "hat matrix" H,

$$H = X(X'X)^{-1}X' \tag{3.10}$$

that is found by using columns and rows of H indexed by $I = (i_1, \ldots, i_k)$. We note that for $k = 1$ we have

$$c_i = K S_{(i)}^{-(n-p-1)/2} \cdot (1 - h_i)^{-1/2} \tag{3.11}$$

where h_i is the ith diagonal element of H. Now, the element h_i is said to be the "leverage" of the observation y_i, and we note that if this is large (i.e., close to 1),

then c_i of (3.11), which takes the leverage of y_i into account, tends to be large, since c_i is increasing in h_i. For general k, $|I - H_I|^{-1}$ is a general function of the leverages of $(y_{i_1}, \ldots, y_{i_k}) = \mathbf{y}_I'$ etc.

There may be a concern that the diagnostics c_I are only useful for the mean-shift model (2.3a), and not at all useful for diagnostics concerning the variance-inflation model (2.6). Pĕna and Tiao (1992) address the question of diagnostics for the variance-inflation model (2.6), and it turns out that their diagnostics have an important and surprising connection with c_I. In fact, conditional on $\mathbf{y}$ containing k spuriously generated observations, the posterior probability, say $\wp_k(I)$, that the set $\mathbf{y}_I$ is spuriously generated according to the variance-inflation model (2.6) is given by (see Pĕna and Tiao (1992))

$$\wp_k(I) = K_0 \left\{ \frac{|X'X|}{|X'X - \phi X_I' X_I|} \right\}^{1/2} \left\{ \frac{s^2}{\hat{s}^2_{(I)}} \right\}^{\frac{n-p}{2}}, \tag{3.12}$$

with

$$\phi = 1 - \delta^{-2}, \quad (n-p)s^2 = S = \mathbf{y}'(I - H)\mathbf{y}. \tag{3.12a}$$

As to $\hat{s}^2_{(I)}$, a precise definition is given in Peña and Tiao (1992), and it turns out that

$$\lim_{\delta^2 \to \infty} \hat{s}^2_{(I)} = s^2_{(I)} \tag{3.12b}$$

where

$$(n-k-p)s^2_{(I)} = S_{(I)} = \mathbf{y}' \left[I_{n-k} - X_{(I)}(X_I' X_{(I)})^{-1} X_{(I)}' \right] \mathbf{y}_{(I)}.$$

Hence, if δ^2 is large, so that $\phi \simeq 1$, we have, on referring to (3.6), that

$$\wp_K(I) = K_1 S_{(I)}^{-\frac{n-p}{2}} \cdot \left| X_{(I)}' X_{(I)} \right|^{-1/2} \tag{3.13}$$

which, for moderate or large n, is essentially c_I.

A word here about k, the "order" of the model M_I is appropriate. In practice, this is not known, but a realistic range of values for k may often be stated by the experimenter, based on his/her experience in the subject field, say $0 \leq k \leq k_0$. (Interesting comments on the "choice" of k_0 have been made by Daniel (1959) and Box and Tiao (1968). For α = Probability that an observation is spuriously generated, these authors choose $k_0 = \alpha n$, with $\alpha = .10$, with supporting arguments.) Hence, a second part of the diagnosis involves "estimating" k. This generates other diagnostic procedures, explained in Section 4 and illustrated in Section 6.

4. Towards Completing the Diagnosis – Part 2

Diagnostics for k are readily available, but to describe this aspect of the diagnostic procedure we now present some results, interesting in themselves, which turn out to be useful in making diagnoses about k.

First assume that we are interested in making inference about $\boldsymbol{\beta}$, the regression coefficients involved in our linear model. It is well known (see for example, Box and Tiao (1973)), that if all the data were generated in the manner intended, the posterior of $\boldsymbol{\beta}$ when non-informative priors are appropriate is such that

$$\wp(\boldsymbol{\beta}|\mathbf{y}; X) = h_p\left(\boldsymbol{\beta}\middle|\hat{\boldsymbol{\beta}}; \frac{\nu}{S}(X'X); \nu\right) \tag{4.1}$$

where

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}, \quad S = (\mathbf{y} - \hat{\mathbf{y}})'(\mathbf{y} - \hat{\mathbf{y}}) = \mathbf{y}'[I - H]\mathbf{y} \tag{4.1a}$$

and

$$\nu = n - p. \tag{4.1b}$$

In (4.1), the density function h is that of a general multivariate-t density, of order p, where, in general, letting $\mathbf{x}$ and $\mathbf{x}_0$ be p order vectors and B a $(p \times p)$ positive definite matrix, h is defined as

$$h_p(\mathbf{x}|\mathbf{x}_0; B; \nu) = \frac{\Gamma\left(\frac{p+\nu}{2}\right)|B|^{1/2}}{(\pi\nu)^{p/2}\Gamma\left(\frac{\nu}{2}\right)}\left[1 + \frac{(\mathbf{x} - \mathbf{x}_0)'B(\mathbf{x} - \mathbf{x}_0)}{\nu}\right]^{-(p+\nu)/2} \tag{4.2}$$

with $\nu > 0$. It is well known that the following properties of (4.2) hold, viz

$$E(\mathbf{x}) = \mathbf{x}_0; \quad V(\mathbf{x}) = B^{-1}\frac{\nu}{\nu - 2}, \quad \nu > 2, \tag{4.3}$$

and further, that

$$(\mathbf{x} - \mathbf{x}_0)'B(\mathbf{x} - \mathbf{x}_0) \sim pF_{p,\nu}. \tag{4.4}$$

Using the properties (4.3)–(4.4), it follows that (4.1) implies

$$E(\boldsymbol{\beta}|\mathbf{y}; X) = \hat{\boldsymbol{\beta}}; \quad \mathrm{Var}(\boldsymbol{\beta}|\mathbf{y}; X) = \frac{S}{n - p - 2}(X'X)^{-1} \tag{4.5}$$

and that a $(1 - \alpha)$ posterior region for $\boldsymbol{\beta}$ is

$$C = \left\{\boldsymbol{\beta}\middle|(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})'(X'X)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) \leq p\frac{S}{n - p}F_{p,n-p;\alpha}\right\} \tag{4.6}$$

where $F_{p,n-p;\alpha}$ is the point exceeded with probability α when using the $F_{p,n-p}$ distribution. It turns out that (4.6) is the H.P.D. posterior region for $\boldsymbol{\beta}$ (see Box and Tiao (1973)).

We now explore the situation when one of the models M_I holds for a given value of k. Then, as derived in Guttman, Dutter and Freeman (1978), it turns out that the posterior of β takes the form

$$\wp(\beta|\text{data}; k) = \sum{}' c_I h_p\left(\beta\middle|\hat{\beta}_{(I)}; \frac{n-k-p}{S_{(I)}}\left(X'_{(I)}X_{(I)}\right); n-k-p\right) \tag{4.7}$$

where $\hat{\beta}_{(I)}$ and $S_{(I)}$ have been defined in (2.5) and (2.5a) respectively. That is, the posterior of β is now a weighted combination of p-order multivariate t-densities, and a typical term says: omit k y's indexed by the set I and compute the posterior based on the remaining data whose effective sample size is $n-k$ (see (4.1) and (4.7)), and weight that density with c_I, the posterior probability that the k observations now ignored, are spuriously generated or, put another way, the density based on the $(n-k)$ observations indexed by the complement of the set I is weighted with the probability that the k observations indexed by the set I itself should indeed be dropped.

Now, using properties of the p-order t-distribution given in (4.2), we find that

$$E(\beta|\text{data}; k) = \sum{}' c_I \hat{\beta}_{(I)} = \boldsymbol{b}_k, \quad \text{say} \tag{4.8}$$

and

$$\begin{aligned} E(\beta\beta'|\text{data}; k) &= \sum{}' c_I \left[\frac{S_{(I)}}{n-k-p-2}(X'_I X_{(I)})^{-1} + \hat{\beta}_{(I)}\hat{\beta}'_{(I)}\right] \\ &= D_k, \quad \text{say} \end{aligned} \tag{4.9}$$

so that

$$V(\beta|\text{data}; k) = D_k - \boldsymbol{b}_k \boldsymbol{b}'_k. \tag{4.10}$$

With these results in mind, we may now turn to the question of diagnostics for the likely value of k. To begin with, suppose the values x_{ju}, $u = 1, \ldots, n$ of the jth independent variable ($j = 1, \ldots, p$) used to generate the y_u are coded i.e., dimensionaless variables. This would mean that each of the diagonal elements V_{tt} of the variance-covariance matrix given in (4.10) are in the same units, namely, "y^2" units. Now, a measure of dispersion of the densities (4.1) and (4.7) is the trace (tr) of their variance-covariance matrices, and from (4.5) and (4.10) these are given by

$$trV(\beta|\text{data}; k) = \begin{cases} \dfrac{S}{n-p-2} tr(X'X)^{-1} & \text{if } k = 0, \\ trD_k - \boldsymbol{b}'_k \boldsymbol{b}_k & \text{if } k > 0. \end{cases} \tag{4.11}$$

Of course, (4.11) is in units of "y^2". We now compare these traces for values $k = 0, 1, \ldots, k_0$. If a data set contains spurious observations which give rise to k (extreme) outlying observations, then (4.11) tends to have a minimum as a function of k, about some value, say $\hat{k} > 0$, and we would use $\hat{k}$ as our estimator of k. This in turn means that we would use $p(\boldsymbol{\beta}|\text{data}; \hat{k})$ — see (4.7) — to make inferences about $\boldsymbol{\beta}$. Of course, the x_{ju}'s could be in original units — for example, pressure in units of lbs./sq.in., time allowed for the process to run in minutes, temperature in $^\circ C$, etc. Hence V_{tt} is in units of "y^2/x_j^2", so that values of $trV(\boldsymbol{\beta}|\text{data}; k)$ cannot be used.

But in this case, we can easily examine separately the diagonal elements $V_{tt}(k)$, and do this for each $t, t = 1, \ldots, p$. Their minima will usually be attained for each t at the same value of k. (Of course, we can also do this for the previous case where x_{ju}'s are in coded units, $u = 1, \ldots, n$; $j = 1, \ldots, p$.)

Another source of a possible diagnosis for k is the c_I's themselves. For each k we may compute the $\binom{n}{k}$ c_I's and note the maximum, say c_I^*, that is,

$$c_I^* = \max_I c_I. \tag{4.12}$$

We now do this for each $k = 1, \ldots, k_0$ and find

$$c^{**} = \max_k c_I^*. \tag{4.13}$$

The pattern of the individual c_I's for given k and the value of k for which (4.13) is attained, together with the analysis of the variance-covariance matrices as described above, gives much information about the likely value of k. This is illustrated in the example of Section 7.

Before turning to an example, we discuss the use of the Kullback-Leibler information to generate other diagnostics for spuriosity, and, it turns out, of influence.

5. Diagnosis: Part 3; The Use of Kullback-Leibler Divergence

The motivation for the approach of this section is as follows: Suppose (2.1) holds, so that, in particular, all observations, have been generated as intended. Now consider the posterior $\wp$ of any or all the parameters of model (2.1), based on all observations, and contrast this with the posterior $\wp_{(I)}$, the posterior based on the $n - k$ observations, $\mathbf{y}_{(I)} = (y_{j_1}, \ldots, y_{j_{n-k}})'$ with $k << n/2$. The pair $(\wp, \wp_{(I)})$ should not differ too markedly, reflecting, basically, the same information about the parameters, except for the fact that $\wp_{(I)}$ is based on fewer observations than $\wp$. So as we let I range over the possible $\binom{n}{k}$ available sets $I = (i, \ldots, i_k) \subset (1, \ldots, n)$, the pairs $(\wp, \wp_{(I)})$ should differ in much the same fashion as each other.

Now suppose the k observations $(y_{i_1}, \ldots, y_{i_k})$ have been generated spuriously (models (2.3) and (2.6) are examples) and we based our posterior on $(y_{j_1}, \ldots, y_{j_{n-k}})$. Then we would expect much divergence between $\wp$ and $\wp_{(I)}$, since $\wp$ is based on data that contains spuriously generated observations, while $\wp_{(I)}$ does not. Of course, we do not know which set $(y_{i_1}, \ldots, y_{i_k})$ is the spurious set, so that we would like to examine the $\binom{n}{k}$ possible cases, noting the pairs $(\wp, \wp_{(I)})$ that seem to diverge markedly, thus indicating that $(y_{i_1}, \ldots, y_{i_k})$ has been generated spuriously.

The question, then, at this point is how to measure divergence between two densities. In this paper, we utilize the Kullback-Leilber symmetric divergence, defined as follows.

Definition 5.1. If f_1 and f_2 are densities that are absolutely continuous with respect to measures μ_1 and μ_2, respectively, then the Kullback-Leibler information measure, often called the symmetric *divergence*, is

$$J(f_1, f_2) = I(f_1, f_2) + I(f_2, f_1) \tag{5.1}$$

where, for example, the *directed* divergence of f_2 from f_1, $I(f_1, f_2)$ is defined as

$$\begin{aligned} I(f_1, f_2) &= E_{f_1}[\log(f_1/f_2)] \\ &= \int \log(f_1/f_2) f_1(x) d\mu_1(x). \end{aligned} \tag{5.2}$$

For an interpretation of the measure of divergence J and its properties, see Kullback (1959) and Kullback and Leibler (1951).

We use (5.1)–(5.2) as follows. First assume that all observations are generated as intended, and set $f_1 = \wp$ and $f_2 = \wp_{(I)}$, letting I range over the $\binom{n}{k}$ sets $i = (i_1, \ldots, i_k) \subset (1, \ldots, n)$. Here, $\wp$ and $\wp_{(I)}$ would then be reflecting essentially the same information about the parameters, expect for the fact that $\wp$ is based on n observations, while $\wp_{(I)}$ is based on only $n - k$ observations, so that we would expect the set of values

$$J(f_1, f_2) = J(\wp, \wp_{(I)}) = J_I, \quad \text{say}, \tag{5.3}$$

to have, more or less, the same (low) values as I ranges over the $\binom{n}{k}$ sets $\{i_1, \ldots, i_k\}$. This, of course, is in contrast to the case where there is a set of observations, say $y_I = (y_{i_1}, \ldots, y_{i_k})'$, which have been generated spuriously. This approach has been used by Johnston and Geisser (1985) and Guttman and Peña (1988). Both papers derive results for (5.3) when $f_1 = \wp$ and $f_2 = \wp_{(I)}$ are (i) the joint posterior distributions of β and σ^2, (ii) the marginal distributions for β, and (iii) the marginal distributions for σ^2. The latter authors showed that, whereas all

interesting observation points are picked up by the change in the posterior for β and σ^2, the change in the marginal for β is linked to influential observations, whereas the change in the marginal for σ^2 is linked to outliers. These measures have the forms given in the next three theorems of this section. We first state the following lemma:

Lemma 5.1. *Suppose* $\mathbf{x}$ *is a* $(p \times 1)$ *random vector variable whose density is one of*

$$f_j = \frac{|\sum^{-1}|^{1/2}}{(2\pi)^{p/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\mu_j)' \sum_j^{-1}(\mathbf{x}-\mu_j)\right\}$$

$(j = 1, 2)$. *Then the Kullback-Leibler divergence between* f_1 *and* f_2 *is*

$$J(f_1, f_2) = \frac{1}{2}(\mu_1-\mu_2)'\left(\sum_1^{-1}+\sum_2^{-1}\right)(\mu_1-\mu_2) + \frac{1}{2}tr\left(\sum_1\sum_2^{-1}+\sum_2\sum_1^{-1}\right) - p. \tag{5.4}$$

The proof of this lemma is a straightforward application of (5.1)–(5.2) and is left to the reader. We need Lemma 5.1 for the following situation. Suppose we assume that data is generated in accordance with

$$\mathbf{y} = X\beta + \epsilon, \quad \epsilon \sim N(O, \sigma^2 I) \tag{5.5}$$

and in (2.1), and that use of non-informative priors for β and σ^2 is made, so that, in particular the posterior of β is as stated in (4.1). This, of course, means that for moderate to large n,

$$\beta|\mathbf{y} \dot{\sim} N(\hat{\beta}; s^2(X'X)^{-1}). \tag{5.6}$$

Here, $s^2 = S/(n-p)$, where S has been defined in (4.1a). (The symbol "$\dot{\sim}$" means "approximately distributed as".) Denote the density involved in (5.6) as $\wp$.

Now suppose our posterior is based on the data $(\mathbf{y}_{(I)}; X_{(I)})$, where $\mathbf{y}_{(I)}, X_{(I)}$, $\hat{\beta}_{(I)}$, and $S_{(I)}$ have been defined in Section 2. Then, for moderate to large n,

$$\beta|\mathbf{y}_{(I)} \dot{\sim} N\left(\hat{\beta}_{(I)}, s^2_{(I)}(X'_{(I)}X_{(I)})^{-1}\right) \tag{5.7}$$

with $s^2_{(I)} = S_{(I)}/(n-k-p)$. Denote the density involved in (5.7) as $\wp_{(I)}$. Setting $f_1 = \wp$ of (5.6) and $f_2 = \wp_{(I)}$ of (5.7), we may now use Lemma 5.1 to state the following theorem.

Theorem 5.1. *Using the above notation, and assuming* (5.5) *holds with* n *moderate to large, then*

$$J_\beta(\wp, \wp_{(I)}) \simeq p(D_I^2 + D_{(I)}^2)/2 + \frac{s^2}{s^2_{(I)}}(p - trH_I)/2$$

$$+ \frac{s^2_{(I)}}{s^2}\Big\{p + tr H_I[I_k - H_I]^{-1}\Big\}/2 - p, \tag{5.8}$$

and

$$\frac{s^2}{s^2_{(I)}} = \frac{(n-p-k)}{(n-p)}\frac{(1+e_I'(I-H_I)^{-1}e_I)}{(n-p-k)s^2_{(I)}} \tag{5.9}$$

$$D_I^2 = \frac{e_I'(I-H_I)^{-1}H_I(I-H_I)^{-1}e_I}{ps^2} \tag{5.9a}$$

$$D^2_{(I)} = \frac{e_I'(I-H_I)^{-1}H_I e_I}{ps^2_{(I)}}, \tag{5.9b}$$

where, with $\mathbf{y} = (\mathbf{y}'_{(I)}, \mathbf{y}_I)$ *and* $X = (X'_{(I)}, X'_I)'$, *the* $(k \times 1)$ *vector* $\boldsymbol{e}_I$ *is given by*

$$\boldsymbol{e} = (\boldsymbol{e}'_{(I)}, \boldsymbol{e}'_I)' = (I-H)\mathbf{y}. \tag{5.9c}$$

The proof of Theorem 5.1 is obtained by straightforward algebra using the results of Guttman and Peña (1988).

The quantity D_I^2 has long been advocated by Dennis Cook and fellow workers as a *measure of influence* — see for example Cook (1977, 1979), and Cook and Weisberg (1982) and the references therein. Of course, $D^2_{(I)}$, then, is also a measure of influence, albeit in a slightly different metric then D_I^2. We remark that, because of this, J_β of Theorem 5.1 is essentially a measure of influence, due to the presence of the terms pD_I^2 and $pD^2_{(I)}$.

A Corollary to Theorem 5.1 for the special case of interest when $k = 1$ is the following: We have denoted (5.8) by $J_\beta(\wp, \wp_{(I)}|k)$ in the following Corollary, and we note that when $k = 1$, then $I = \{i_1\}$ which we may denote by i, i varying over $(1, \ldots, n)$.

Corollary 5.1.1. *Setting* $J_\beta(\wp, \wp_{(i)}|k = 1) = M_i(\boldsymbol{\beta})$, *we have*

$$M_i(\boldsymbol{\beta}) = p\Big(D_i^2 + D^2_{(i)}\Big)/2 + \frac{s^2_{(i)}}{2s^2}\left(p + \frac{h_i}{1-h_1}\right) + \frac{s^2}{2s^2_{(i)}}(p-h_i) - p \tag{5.10}$$

where h_i *is the* i*th diagonal element of* H.

The proof of this Corollary is a straightforward application of Theorem 5.1 for the case $k = 1$. (Since $k = 1$, the sets I are singletons i_1, etc.) We shall use $M_i(\boldsymbol{\beta})$ for all n sets $i = \{i_u\}$, $u = 1, \ldots, n$ as diagnostics in our example of Section 7.

We may also want to inquire about the divergence between posteriors of σ^2, as we withdraw observations $(y_{i_1}, \ldots, y_{i_k})$. As is well known (for the case of non-informative priors), we have

$$\sigma^2|\mathbf{y} \sim (n-p)s^2/\chi^2_{n-k-p}, \tag{5.11}$$

and

$$\sigma^2|\mathbf{y}_{(I)} \sim (n-k-p)s^2_{(I)}/\chi^2_{n-k-p}. \tag{5.12}$$

Identifying the posterior density of σ^2 in (5.11) as $\wp$ and that of (5.12) as $\wp_{(I)}$, we have the following Theorem.

Theorem 5.2. *Suppose* (5.5) *holds and non-informative priors are used, so that* (5.11) *and* (5.12) *apply. Then the posteriors of* σ^2 *of* (5.11) *and* (5.12) *have Kullback-Leibler symmetric divergence* $J_{\sigma^2}(\wp, \wp_{(I)}|k)$, *which, to order of* n^{-2} *is*

$$J_{\sigma^2}(\wp, \wp_{(I)}|k) = \frac{k}{2}\ln\frac{s^2_{(I)}}{s^2} + \frac{1}{2}\left[e_I'(I_k - H_I)^{-1}e_I\right]\left[\frac{1}{s^2_{(I)}} - \frac{1}{s^2}\right]. \tag{5.13}$$

The proof of this theorem is given in Guttman and Peña (1988).

The case $k = 1$ will be of special interest, and we have

Corollary 5.2.1. *If* $J_{\sigma^2}(\wp, \wp_{(I)}|k=1) = M_i(\sigma^2)$, *then to terms of order* n^{-2},

$$M_i(\sigma^2) = \frac{1}{2}\ln\frac{s^2_{(i)}}{s^2} + \frac{1}{2}(t_i^2 - r_i^2) \tag{5.14}$$

where

$$t_i^2 = \frac{e_i^2}{s^2_{(i)}(1-h_i)}, \quad r_i^2 = \frac{e_i^2}{s^2(1-h_i)} \tag{5.15}$$

with $e_I = e_i$, *where we have set* $i_1 = i$.

The statistic r_i defined by (5.15) has been extensively used in the literature as a test for spuriosity, and, of course, t_i is a similar statistic, using a slightly different estimator of $\mathrm{Var}(e_i) = \sigma^2(1-h_i)$ in its denominator. It can be shown that $M_i(\sigma^2)$ is an increasing function of t_i^2, and, hence is essentially a measure of outlyingness of y_i (we have set $i_1 = i$, since $k = 1$ in the above).

Finally, we may ask about the divergences of the posterior of $(\boldsymbol{\beta}, \sigma^2)$, based on $\mathbf{y}$ and $\mathbf{y}_{(i)}$ respectively. We have

Theorem 5.3. *Suppose* (5.5) *holds, and non-informative priors are used. Then the Kullback-Leibler divergence between* $\wp_{(I)} = p(\boldsymbol{\beta}, \sigma^2|\mathbf{y}_{(I)})$ *and* $\wp = p(\boldsymbol{\beta}, \sigma^2|\mathbf{y})$

is, to terms of order n^{-2},

$$J_{\beta,\sigma^2}(\wp, \wp_{(I)}|k) = \left[\frac{e'_{(I)}(I_k - H_I)^{-1}e_I}{2}\right]\left[\frac{1}{s^2_{(I)}} - \frac{1}{s^2}\right]$$

$$+\frac{p}{2}\left[\frac{s^2_{(I)}}{s^2}D^2_{(I)} + \frac{s^2}{s^2_{(I)}}D^2_I\right] + \frac{1}{2}tr\left\{H_I[I_k - H_I]^{-1}H_I\right\} + \frac{k}{2}\ln\frac{s^2_{(I)}}{s^2}. \quad (5.16)$$

The proof of this Theorem is given in Guttman and Peña (1988). This proof uses a key relation about conditional-unconditional divergences used for a more general model by Johnston and Geisser (1985). For the special case $k = 1$, we have:

Corollary 5.3.1. *Letting* $J_{\beta,\sigma^2}(\wp, \wp_{(I)}|k = 1) = M_i(\beta, \sigma^2)$, *then*

$$M_i(\beta, \sigma^2) = \frac{1}{2}(t_i^2 - r_i^2) + \frac{p}{2}\left[\frac{s^2_{(i)}}{s^2}D^2_{(i)} + \frac{s^2}{s^2_{(i)}}D^2_i\right] + \frac{1}{2}\frac{h_i^2}{1-h_i} + \frac{1}{2}\ln\frac{s^2_{(I)}}{s^2}. \quad (5.17)$$

With the above Theorems and Corollaries in mind, we now turn, in the next Section, to a description of their behaviour, which will help map a strategy on how to use these results in a diagnostic procedure.

6. Comparison of the Various Diagnostic Measures

We have presented various statistics to identify spurious observations. These are the probability c_I, and the distance $J_I(\beta, \sigma^2)$. We have also shown that this latter portmanteau measure is related to the specific measures $J_I(\beta)$ and $J_I(\sigma^2)$, which, of course, can be used to identify influential and outlying observations.

To illustrate the relationship between c_I and $J_I(\beta, \sigma^2)$, let us consider the case $k = 1$. Then, (5.17) can be written, after some algebra, as

$$M_i(\beta, \sigma^2) = \frac{r_i^2}{2}\left[\frac{s^2}{s^2_{(i)}}\frac{1}{(1-h_i)} - (1-h_i)\right] + \frac{1}{2}\frac{h_i^2}{(1-h_i)} - \frac{1}{2}\ln\frac{s^2}{s^2_{(i)}} \quad (6.1)$$

and using the fact that, for n large,

$$\ln\frac{s^2}{s^2_{(i)}} \doteq \ln\left(1 + \frac{t_i^2}{n}\right) \doteq \frac{t_i^2}{n} \quad (6.2)$$

and since, when n is large, $t_i \simeq r_i$, we have, asymptotically,

$$M_i(\beta, \sigma^2) \doteq \frac{1}{2}\frac{t_i^2}{(1-h_i)} - (1-h_i)\frac{t_i^2}{2} + \frac{1}{2}\frac{h_i^2}{(1-h_i)} - \frac{1}{2}\frac{t_i^2}{n}, \quad (6.3)$$

so that, for n large,

$$M_i(\beta, \sigma^2) \doteq \frac{1}{2}\frac{h_i}{(1-h_i)}\left[(2-h_i)t_i^2 + h_i\right]. \tag{6.4}$$

The above shows that $M_i(\beta, \sigma^2)$ is a linear increasing function of t_i^2. The slope depends on h_i, and the scale factor is a standard measure of leverage.

In order to discuss the relationship between $M_i(\beta, \sigma^2)$ and c_i, put both in the same scale by comparing $M_i(\beta, \sigma^2)$ with $\log c_i$. Then

$$\log c_i = K - \frac{n}{2}\log\frac{s_{(i)}^2}{s^2} - \frac{1}{2}\log(1-h_i), \tag{6.5}$$

and using (6.2),

$$\log c_i = K + \frac{t_i^2}{2} - \frac{1}{2}\log(1-h_i), \tag{6.6}$$

which shows that $\log c_i$ is also a linear increasing function of t_i^2. The main difference between (6.4) and (6.6) is the way each of them deals with the leverage. $M_i(\beta, \sigma^2)$ is concerned with both outliers and influential points and the leverage factor $h_i/(1-h_i)$ is the one that appears naturally in the standard influence measures such as Cook's statistics. On the other hand, $\log c_i$ is a measure of spuriosity and does not include a product term between the outlier measure t_i^2 and the leverage measure $(1-h_i)$.

It is interesting to relate these measures to other statistics suggested in the literature to achive the same objective. Andrews and Pregibon (1978) proposed the ratio

$$R_i = \left(\frac{n-p-1}{n-p}\right)\frac{s_{(i)}^2}{s^2}(1-h_i) \tag{6.7}$$

and they identify outliers with the association of small values of this statistic. Belsley, Kuh and Welsch (1980) suggested a similar statistic based on the volume of confidence ellipsoids. See Cook and Weisberg (1982) and Chatterjee and Hadi (1986) for a comparison of these measures. Now to compare (6.7) with the previous statistics in the same scale we take minus the logarithm of R_i to obtain, for large n,

$$-\log R_i = -\log\frac{s_{(i)}^2}{s^2} - \log(1-h_i) \tag{6.8}$$

and if we compare the above with (6.5) it is obvious that c_i is taking into account the sample size in the evaluation of the observation point whereas, the Andrews and Pregibon statistic does not, for large n.

In summary, $M_i(\beta, \sigma^2)$ and c_i provide us with complementary information about interesting points in the data set. The points identified as interesting by

all the above measures could be further analyzed using $M_i(\beta)$ and $M_i(\sigma^2)$ to differentiate between influential observations and outliers.

7. An Illustrative Example – The Mickey Dunn Clark Data

For this example, we refer to the famous "MDC data set" due to Mickey, Dunn and Clark (1967), and reported on in Cook and Weisberg (1982), Draper and Smith (1981), amongst others. We list the data in Table 7.1, and a plot is given in Figure 7.1.

This data gives (X, Y) values for $n = 21$ students, where X = age at first word (months) and Y = score of Gessell aptitude test. It is assumed that the linear relation $E(Y|x) = \beta_0 + \beta_1 x$ is appropriate, so that, in our notation, $p = 2$.

Table 1. The Mickey Dunn Clark data ($n = 21$)

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

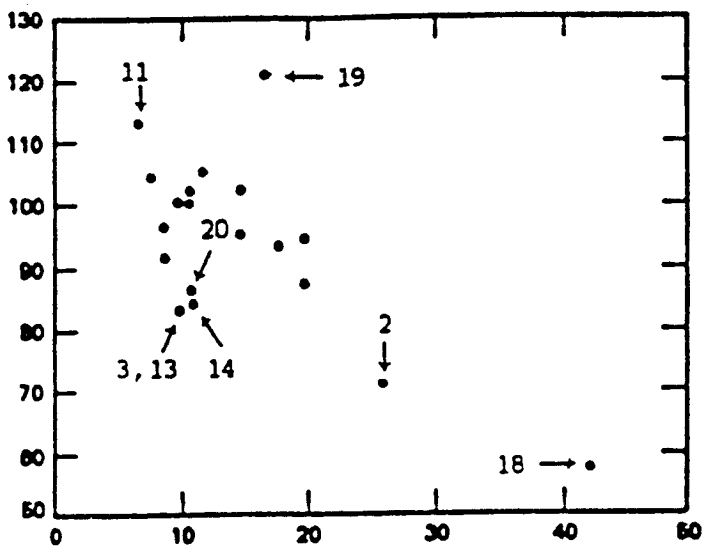

Figure 7.1. A plot of the MDC data ($n = 21$)

In the language of Sections 3 and 4, we first set k and compute the resulting $\binom{n}{k}$ c_I's, given by (3.1). For the MDC data, we let $k = 1, 2, 3$. (That is, bearing

in mind that $.1n = 2.1$ for this set of data, we have set $k_0 = 3$.) The largest 6 c_I's are given in Table 7.2.

Table 7.2. The 6 largest c_I's for the MDC data*

$k=1$	$k=2$	$k=3$
.8153(19)	.1096(13,19)	.0459(3,13,19)
.0180(13)	.1096(3,19)	.0246(13,14,19)
.0180(3)	.0709(11,19)	.0246(3,14,19)
.0143(18)	.0685(14,19)	.0163(11,13,19)
.0134(14)	.0517(5,19)	.0163(3,11,19)
.0107(20)	.0490(19,20)	.0160(3,19,20)

*The numbers in brackets are the $(i_1, i_2, \ldots, i_k)$ that correspond to the accompanying c_I value.

From Table 7.2 it is evident that the maximum of the maximum c_I's occurs at $k = 1$ with $c_{19} = \text{Prob}(y_{19} \text{ is spurious } | k = 1) = .8153$. Note too, that for $k = 1$, the second largest c is c_3 or c_{13} with value .0180, or put more dramatically, $c_{19}/c_{13} = c_{19}/c_3 = 45.3$. Also note the consistency with which observations y_j, for $j = 19, 3, 13$, get into the act – for $k = 2$, $\max c_{i_1,i_2} = c_{13,19} = c_{3,19} = .1096$ and for $k = 3$, $\max c_{i_1,i_2,i_3} = c_{3,13,19} = .0459$. Note too, that for $k = 2$, $c_{13,19}/c_{11,19} = c_{3,19}/c_{11,19} = 1.55$, and for $k = 3$, $c_{19,3,13}/c_{19,13,14} = c_{19,3,13}/c_{19,3,14} = 1.87$, and these ratios are pedestrian when compared with $c_{19}/c_{13} = 45.3 = c_{19}/c_3(k = 1)$. Thus, even at this stage of the diagnosis, evidence is building that $k = 1$, and indeed that y_{19} is the spurious observation.

Using (4.5) and (4.11), we obtain the numerical results of Table 7.3. (Complete listings of values of c_I's β_k's and D_k's are available from the authors.)

Table 7.3. The diagonal elements, V_{tt}, of the matrices $V(\beta|\text{data}; k)$

	$k=0$	$k=1$	$k=2$	$k=3$
$V_{11}(\beta\|k)$	28.70410	20.90559	22.67359	26.20836
$V_{22}(\beta\|k)$	0.10753	0.08073	0.08954	0.10699

Note that the diagonal elements V_{11} and V_{22} attains their minimums in both cases at $k = 1$, providing yet more evidence that there seems to be one spurious observation, the observation y_{19}, in this data set. Tentatively, then, we consider the use of $\wp(\beta|\text{data}; k = 1)$ to do inference re β (and/or $\wp(\sigma^2|\text{data}; k = 1)$ and/or $\wp(\beta, \sigma^2|\text{data}; k = 1)$, depending on objectives). Indeed, for $\wp(\beta|\text{data}; k = 1)$ it

turns out that

$$E(\boldsymbol{\beta}|\text{data}; k=1) = \begin{pmatrix} 109.40284 \\ -1.17759 \end{pmatrix};$$
$$V(\boldsymbol{\beta}|\text{data}, k=1) = \begin{pmatrix} 20.90559 & -1.12645 \\ -1.12645 & 0.08073 \end{pmatrix}. \tag{7.1}$$

Now, we have calculated $M_i(\boldsymbol{\beta}, \sigma^2)$ of (5.17), and have tabulated the results in Table 7.4. Examination of the values of $M_i(\boldsymbol{\beta}, \sigma^2)$ yields the fact that for this measure, nineteen of the $n = 21$ have value less than or equal to .1853, but $M_{18}(\boldsymbol{\beta}, \sigma^2)$ and $M_{19}(\boldsymbol{\beta}, \sigma^2)$ have values of 1.5157 and 2.8919, respectively, which are 8.18 and 15.61 times larger, respectively than .1853.

Table 7.4. Values of the diagnostics c_i, $M_i(\boldsymbol{\beta}, \sigma^2)$, $M_i(\sigma^2)$, $M_i(\boldsymbol{\beta})$, h_i and t_i^2 for the Mickey Dunn Clark data

Observation number	c_i	$M_i(\boldsymbol{\beta}, \sigma^2)$	$M_i(\sigma^2)$	$M_i(\boldsymbol{\beta})$	h_i	t_i^2
1	0.0062	0.0281	0.0252	0.0082	0.0479	0.0338
2	0.0099	0.1644	0.0003	0.1652	0.1545	0.8866
3	0.0180	0.1853	0.0395	0.1455	0.0628	2.2826
4	0.0085	0.0546	0.0030	0.0533	0.0705	0.6630
5	0.0085	0.0380	0.0024	0.0366	0.0479	0.6937
6	0.0062	0.0299	0.0270	0.0099	0.0726	0.0009
7	0.0064	0.0296	0.0219	0.0130	0.0580	0.0969
8	0.0063	0.0290	0.0242	0.0105	0.0567	0.0528
9	0.0065	0.0332	0.0226	0.0172	0.0799	0.0840
10	0.0074	0.0422	0.0101	0.0357	0.0726	0.3815
11	0.0106	0.1098	0.0003	0.1090	0.0908	1.1043
12	0.0065	0.0324	0.0209	0.0172	0.0705	0.1175
13	0.0180	0.1853	0.0395	0.1455	0.0628	2.2826
14	0.0134	0.1059	0.0101	0.0949	0.0567	1.6378
15	0.0067	0.0303	0.0184	0.0165	0.0567	0.1707
16	0.0062	0.0293	0.0261	0.0095	0.0628	0.0162
17	0.0083	0.0393	0.0034	0.0373	0.0521	0.6373
18	0.0143	1.5157	0.0021	1.5396	0.6516	0.7142
19	0.8153	2.8519	2.2745	0.7871	0.0531	13.0103
20	0.0107	0.0697	0.0006	0.0686	0.0567	1.1588
21	0.0062	0.0293	0.0261	0.0095	0.0628	0.0162

We have plotted $M_i(\beta, \sigma^2)$ versus $\log c_i$ in Figure 7.2. The graph shows that the $M_i(\beta, \sigma^2)$'s have the same behaviour in all points except for observation 18. The probability c_i says that this observation is not likely to be spurious, whereas $M_i(\beta, \sigma^2)$ says that the 18th point is either outlying, influential or both.

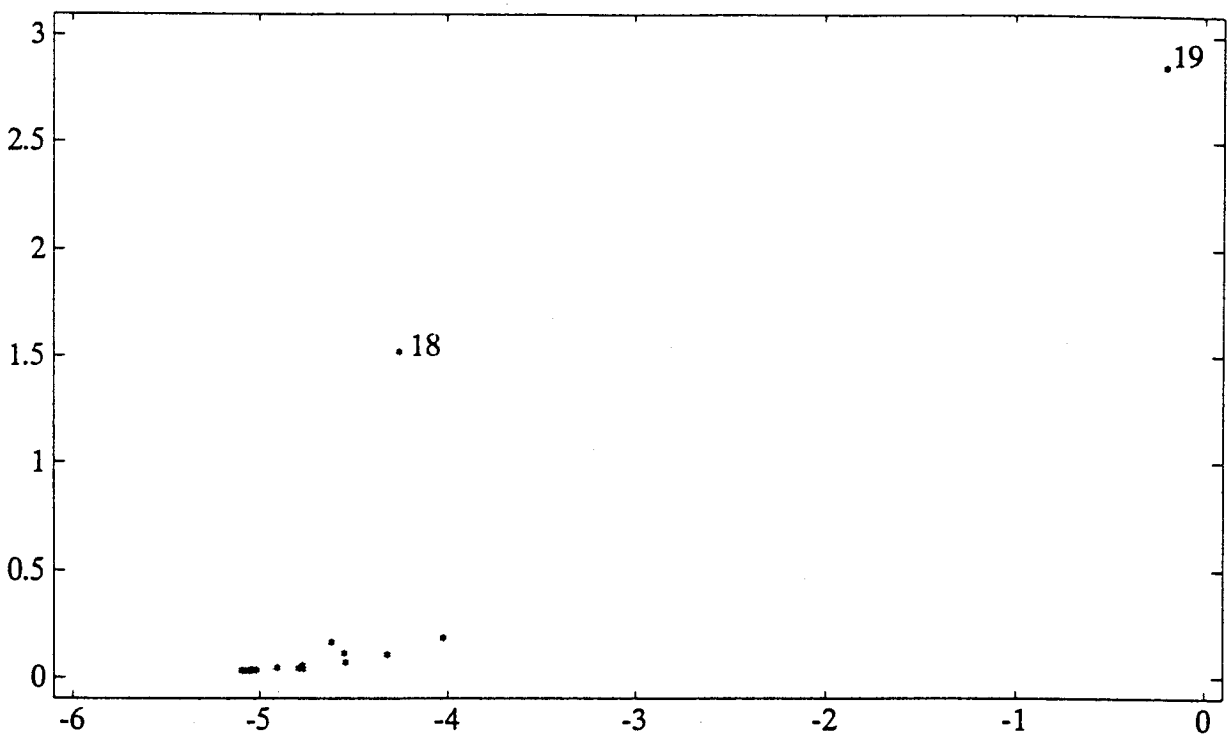

Figure 7.2. A plot of $M_i(\beta, \sigma^2)$ against $\log c_i$ (Mickey Dunn Clark data, $n = 21$)

To help differenciate between outlying and influential points, we look at the statistics $M_i(\sigma^2)$ and $M_i(\beta)$. These values are also shown in Table 7.4. $M_i(\sigma^2)$ shows clearly that the only outlying point is observation 19, with value 2.2745 which is 57.58 times greater than the next largest value, 0.0395 attained for observations 3 and 13. Going to $M_i(\beta)$, we see that the most influential point is observation 18, with a value of 1.5396 that is twice as large as the one for the spuriously generated observation 19, and 9.32 times the next largest.

Table 7.4 also shows values of h_i and t_i^2 for the MDC data. It can be seen that all observations have approximately the same leverage (between .05 and .15) except for observation eighteen that has a leverage of .65. Then, from the results of Section 6, we would expect a linear relationship between t_i^2 and $\log c_i$, except for observation 18. Figure 7.3 shows this graphically. The values of h_i and t_i^2 are given for completeness in Table 7.4.

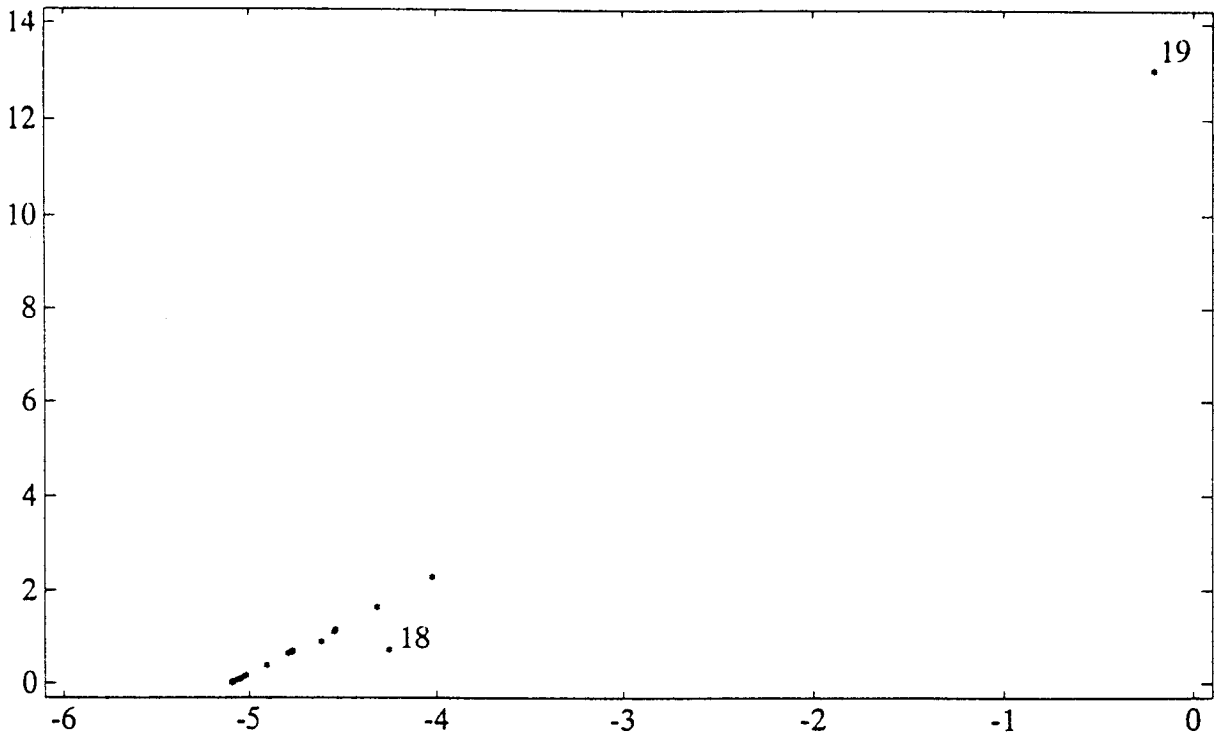

Figure 7.3. A plot of t_i^2 against $\log c_i$ (Mickey Dunn Clark data, $n = 21$)

Now from the joint distribution of $(\beta_0, \beta_1) = \boldsymbol{\beta}'$, given in (4.7) with $k = \hat{k} = 1$, $n = 21$, $p = 2$, we may find the posterior marginals of either β_0 or β_1 using properties of the bivariate t-distribution. We now illustrate the case where interest is in β_1. We need additional notation – let the (2×2) matrix

$$G_{(i)} = \left[\frac{n-k-p}{S_{(i)}}\left(X'_{(i)}X_{(i)}\right)\right]^{-1}, \tag{7.2}$$

and denote the 2 – 2 element of $G_{(i)}$ by $g_{22}^{(i)}$, and set

$$w_{22}^{(i)} = \left(g_{22}^{(i)}\right)^{-1}. \tag{7.3}$$

Then, from properties of the multivariate t-distribution, and consulting (4.7) with $p = 2$, $k = 1$, we have

$$\wp(\beta_1|\text{data}; k = 1) = \sum{}' c_i h_1\left(\beta_1|\hat{\beta}_{1(i)}, w_{22}^{(i)}; n-k-p = 18\right). \tag{7.4}$$

Here, $\sum'$ denotes the sum over all possible sets $i = \{i_1\} \subset (1, \ldots, n)$, etc. Recall from (7.1) that

$$E(\beta_1|\text{data}; k = 1) = -1.17759;\ V(\beta_1|\text{data}, k = 1) = 0.08073. \tag{7.5}$$

We have tabulated (7.4) and graphed this posterior density in Figure 7.4. The relative smooth (slightly asymmetric) curve is no doubt due to the fact the c_{19} is so much larger than all the other c_i's, so that the curve is dominated by $c_{19} \times p(\beta_1|y_{19}; k = 1)$. Using our tabulations, we have incorporated these computations into some numerical integration routines and have found posterior HPD intervals for β_1 at level $1 - \alpha = .90, .95, .99$, and tabulated these in Table 7.5.

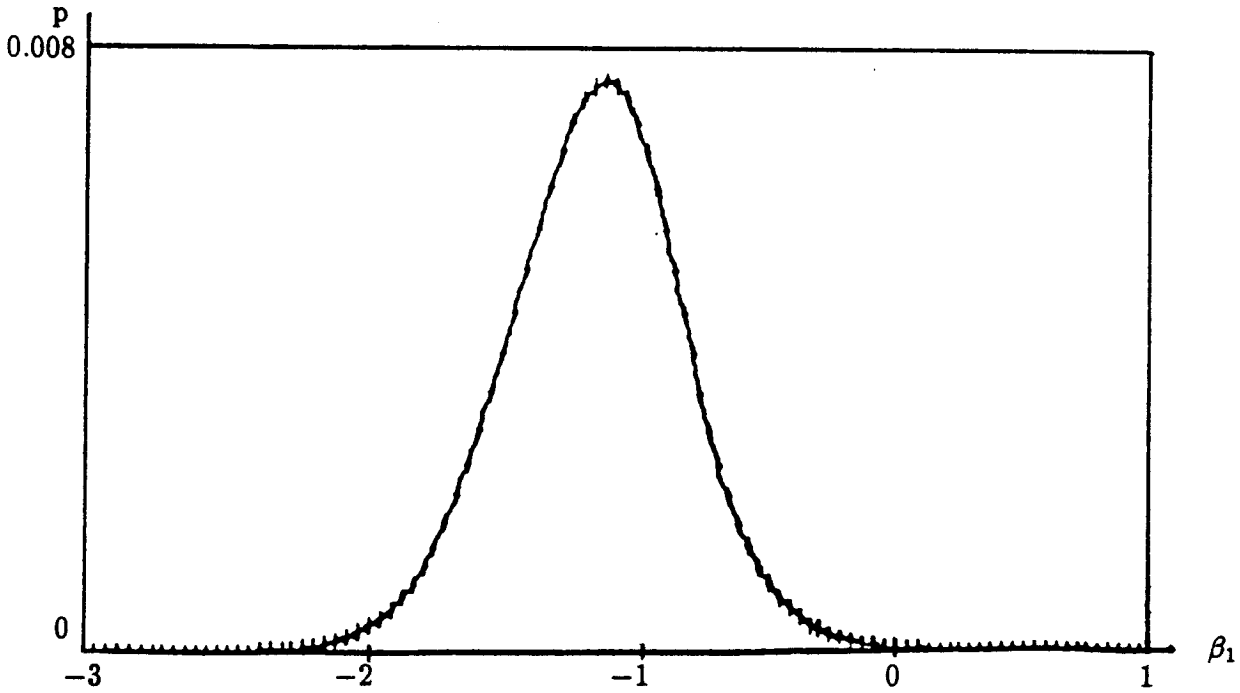

Figure 7.4. The posterior of the slope β_1 given in (7.4) based on the MDC data set.

Table 7.5. The $100(1 - \alpha)\%$ posterior HPD limits for β_1 based on (7.4)

$1 - \alpha$	lower limit	upper limit
.90	−1.637991	−0.726035
.95	−1.737931	−0.617099
.99	−1.963012	−0.347673

Acknowledgements

An extensive and thorough report from an anonymous referee proved to be very stimulating and helpful, and we wish to acknowledge and thank him/her for the report that has led to a much improved and stronger manuscript. Also discussions with Norman Draper and Muni Srivastava have improved the manuscript and are gratefully acknowledged. The authors would like to thank Boon Ping

Chew for programming assistance (all programs used are available on request). The authors' research was supported by various grants – I. Guttman by NSERC (Canada) under Grant No. A8743, while D. Peña by DGICYT (Spain), under Grant No. PB90-0266.

References

Abraham, B. and Box, G. E. P. (1978). Linear models and spurious observations. *Appl. Statist.* **27**, 131–138.

Ali, M. A. (1990). A Bayesian approach to detect informative observations in an experiment. *Comm. Statist. Theory Methods* **19**, 2567–2575.

Andrews, D. F. and Pregibon, D. (1978). Finding the outliers that matter. *J. Roy. Statist. Soc. Ser.B* **40**, 85–93.

Belsley, D. A., Kuh, E. and Welsch, R. E. (1980). *Regression Diagnostics.* John Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55**, 119–129.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Influence in Statistical Analysis.* Addison-Wesley Publishing Co., U.S.A.

Carlin, B. P. and Polson, N. G. (1991). An expected utility approach to influence diagnostics. *J. Amer. Statist. Assoc.* **86**, 1013–1021.

Chaloner, K. and Brant, R. (1988). A Bayesian approach to outlier detection and residual analysis. *Biometrika* **75**, 651–659.

Chatterjee, S. and Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statist. Sci.* **1**, 379–416.

Cook, R. D. (1977). Detection of influential observations in linear regression. *Technometrics* **19**, 15–18.

Cook, R. D. (1979). Influential observations in linear regression. *J. Amer. Statist. Assoc.* **74**, 169–174.

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

Daniel, C. (1959). Use of half-normal plots in interpreting factorial two level experiments. *Technometrics* **1**, 311–341.

Draper, N. R. and Guttman, I. (1987). A common model selection criterion. *Proceedings of the Symposium on Probability and Bayesian Statistics (in memory of Professor B. DeFinetti); Innsbruck, Austria*, 139–150, Plenum Publishing Corp.

Draper, N. R. and Smith, H. (1981). *Applied Regression Analysis*, 2nd edition. John Wiley.

Eddy, W. F. (1980). Discussion of P. Freeman's paper. *Bayesian Statistics* (Edited by J. M. Bernardo, M. H. DeGoot, D. V. Lindley and A. F. M. Smith), 370–373. Valencia University Press.

Freeman, P. R. (1980). On the number of outliers in data from a linear model. *Bayesian Statistics* (Edited by J. M. Bernardo, M. H. DeGoot, D. V. Lindley and A. F. M. Smith), 349–365. 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.

Guttman, I. and Peña, D. (1988). Outliers and influence: Evaluation of posteriors of parameters in linear model. *Bayesian Statistics* **3** (Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 631–640. Oxford University Press.

Johnston, W. and Geisser, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Assoc.* **78**, 137–144.

Johnston, W. and Geisser, S. (1985). Estimative influence measures for the multivariate general linear model. *J. Statist. Plann. Infer.* **11**, 33–56.

Kempthorne, P. J. (1986). Decision-theoretic measures of influence in regression. *J. Roy. Statist. Soc. Ser.B* **48**, 370–378.

Kullback, S. (1959). *Information Theory and Statistics*. John Wiley.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22**, 79–86.

Mickey, M. R., Dunn, O. J. and Clark, V. (1967). Note on use of stepwise regression in detecting outliers. *Computers and Biomedical Research* **1**, 105–111.

Peña, D. and Tiao, G. C. (1992). Bayesian robustness functions for linear models. *Bayesian Statistics* **4** (Edited by J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), 365–388. Oxford University Press.

Pettit, L. I. and Smith, A. F. M. (1985). Outliers and influential observations in linear models. *Bayesian Statistics* **2** (Edited by J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith), 473–494. Elsevier Science Publishers.

Zellner, A. (1975). Bayesian analysis of regression error terms. *J. Amer. Statist. Assoc.* **70**, 138–144.

Zellner, A. and Moulton, B. R. (1985). Bayesian regression diagnostics with applications to international consumption and income data. *J. Econom.* **29**, 187–211.

Department of Statistics, University of Toronto, Toronto, Ontario M5S 1A1, Canada.
Department of Statistics and Econometrics, University Carlos III de Madrid, 28903 Getafe, Madrid Spain.

(Received March 1991; accepted December 1992)