Working Paper 92-32
July 1992

División de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

COMPARING PROBABILISTIC METHODS FOR OUTLIER DETECTION

Daniel Peña and Irwin Guttman*

Abstract

This paper compares the use of two posterior probability methods to deal with outliers in linear models. We show that putting together diagnostics that come from the mean-shift and variance-shift models yields a procedure that seems to be more effective than the use of probabilities computed from the posterior distributions of actual realized residuals. The relation of the suggested procedure to the use of a certain predictive distribution for diagnostics is derived.

Key words:
Diagnostic, Posterior and Predictive distributions, leverage, linear models.

*Peña, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Getafe, 28903 Spain; Guttman, Department of Statistics, University of Toronto, Toronto, Ontario M5S1A1.

1. INTRODUCTION.

The analysis of outliers from the Bayesian point of view has become increasingly interesting to the Statistical Profession because of the possibility of carrying out the difficult computations involved using new algorithms such as the Gibbs sampling method (see Gelfand and Smith (1990), or Casella and George (1992)). Also, specific algorithms have been developed to deal with the Bayesian treatment of outliers (Peña and Tiao (1992)).

Not only has the work on analysis of outliers proved useful in its own right, but it turns out that many other statistical problems can also be usefully analyzed by approaching these problems from the outlier point of view. As examples, we would cite: analysis of unreplicated fractional factional factorial designs to detect significant effects, see Box and Meyer (1986), Juan and Peña (1992); detection of interaction in unreplicated ANOVA designs, see Tussell (1990); estimation of missing observations in time series models, see Ljung (1989) and Peña and Maravall (1991).

There have been three main approaches to the problem of outliers in the literature. Succinctly, these may be classified as (i) the diagnostic approach (ii) the Bayesian approach and (iii) 'robust' approach to estimation and tests of hypothesis in the presence of outliers. The first approach is clearly identified with the work of Cook and Weisberg (1982), Belsley, Kuh and Welsch (1980), and Atkinson (1985), and the aim of workers in this area is mostly that of identification of observations that may be deemed outlying and/or influential. The approach listed as (iii) above has been motivated by the work of Huber (1981), and Hampel et al (1986), the aim here being to build estimatiors that are not affected by the fraction of the sample that is outlying. Truly in the middle and listed as such above, is the Bayesian approach, which seeks to combine identification with estimation: see for example Box and Tiao (1968); Guttman et al (1978), etc. Here, the identification is carried out using the posterior probabilities for an observation or a set of observations being outlying, and these are used as weights in estimation procedures.

conditional on σ^2, we have

$$\boldsymbol{\beta} \sim N\left(\hat{\beta}; \sigma^2 (X'X)^{-1}\right) \tag{2.3}$$

where $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$, and $p(\sigma^2|\mathbf{y}, x)$ is such that

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}\,, \tag{2.4}$$

where

$$(n-p)s^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'[I-H]\mathbf{y}\,, \tag{2.5}$$

with H denoting the so called hat matrix

$$H = X(X'X)^{-1}X'\,. \tag{2.6}$$

We denote a set of k distinct integers chosen from the set $(1,\ldots,n)$ by I. Then, the vector $\mathbf{y}$ can be decomposed as

$$\mathbf{y}' = (\mathbf{y}'_I\ \mathbf{y}'_{(I)}) \tag{2.7}$$

where (I) means "delete set I". Similarly, the X matrix can be partitioned as

$$X' = [X'_I\ X'_{(I)}]\,. \tag{2.8}$$

(The use of the symbol I without brackets means restrict information to the set I.) Consistent with the above notation, we will use in the rest of the paper the designations

$$\hat{\boldsymbol{\beta}}_{(I)} = \left(X'_{(I)}X_{(I)}\right)^{-1}X'_{(I)}\mathbf{y}_{(I)} \tag{2.9}$$

$$s^2_{(I)} = \left(y_{(I)} - X_{(I)}\hat{\boldsymbol{\beta}}_{(I)}\right)'\left(y_{(I)} - X_{(I)}\hat{\boldsymbol{\beta}}_{(I)}\right)/(n-p-k) \tag{2.10}$$

that is, $\hat{\boldsymbol{\beta}}_{(I)}$ and $s^2_{(I)}$ are estimators of $\boldsymbol{\beta}$ and σ^2 based on $(X_{(I)}, \mathbf{y}_{(I)})$, etc.

In contrast with the null model (2.1) we will be concerned in this paper with two alternative models. The first is the mean-shift model and takes the form for the generation of the observations $\mathbf{y} = (\mathbf{y}'_I, \mathbf{y}'_{(I)})'$,

$$\begin{aligned} \mathbf{y}_I &= X_I\boldsymbol{\beta} + \mathbf{a} + \boldsymbol{\epsilon}_I \\ \mathbf{y}_{(I)} &= X_{(I)}\boldsymbol{\beta} \quad\ + \boldsymbol{\epsilon}_{(I)} \end{aligned} \tag{2.11}$$

Here H_I is the $k \times k$ block of the hat matrix H of (2.6) formed by using the k rows and columns of H indexed by I, or

$$H_I = X_I(X'X)^{-1}X_I' \,. \tag{3.4a}$$

Indeed, H_I is refereed to as the "leverage of observations $\mathbf{y}_I$".

Expression (3.2) could be written in another form that will be useful when comparing it to the other approaches to be discussed in this paper. In order to do so, we use the identity (see Cook and Weisberg, 1982, pg. 191)

$$(n-p-k)s_{(I)}^2 = (n-p)s^2 - \mathbf{e}_I'(I-H_I)^{-1}\mathbf{e}_I \,, \tag{3.5}$$

for then

$$\frac{s^2}{s_{(I)}^2} = \frac{n-p-k}{n-p}\left(1 + \frac{\mathbf{e}_I'(I-H_I)^{-1}\mathbf{e}_I}{(n-p-k)s_{(I)}^2}\right) \,. \tag{3.6}$$

Here, we have partitioned the residual vector $\mathbf{e}$,

$$\mathbf{e} = (I-H)\mathbf{y} \tag{3.6a}$$

using

$$\mathbf{e} = (\mathbf{e}_I', \mathbf{e}_{(I)}')' = (I-H)(\mathbf{y}_I', \mathbf{y}_{(I)}')' \tag{3.6b}$$

with $X = (X_I' \vdots X_{(I)}')'$ used in constructing $H = X(X'X)^{-1}X'$. Now, it can be shown that

$$\mathbf{y}_I - x_I\hat{\boldsymbol{\beta}}_{(I)} = (I-H_I)^{-1}\mathbf{e}_I \tag{3.7}$$

and, therefore, (3.4) itself can be written as

$$Q_I = \frac{\mathbf{e}_I'(I-H_I)^{-1}\mathbf{e}_I}{(n-p-k)s_{(I)}^2} \,. \tag{3.8}$$

Hence, we may rewrite (3.2) as follows

$$p(\mathbf{y}_I|\mathbf{y}_{(I)}) = K_1 \frac{|I-H_I|^{1/2}}{\left(\frac{n-p}{n-p-k}\,\frac{s^2}{s_{(I)}^2}\right)^{-k/2}}\,[1+Q_I]^{-(n-p)/2} \tag{3.9}$$

Following the results outlined in Section 2, we have that the posterior distribution of the ε_i, as given by (3.13) conditional on σ^2, is easily seen to be on using (2.3), such that

$$\varepsilon_i \sim N(e_i, \sigma^2 h_i) , \tag{3.14}$$

where $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is the residual of the observation y_i, and h_i is the i-th diagonal element of H given in (2.6), which is to say, h_i is the leverage of y_i. Chaloner and Brant (1988) have shown that P_i of (3.12) can be written in the form

$$P_i = 1 - \int_0^\infty [\Phi(z_1) - \Phi(z_2)] p(\tau|\mathbf{y}, \mathbf{x}) d(\tau) . \tag{3.15}$$

where $\tau = \sigma^{-2}$, so that $p(\tau|\mathbf{y}, \mathbf{x})$ is the density of a $\dfrac{\chi^2_{n-p}}{(n-p)s^2}$ variable, with

$$z_1 = \frac{(q - e_i\sqrt{\tau})}{\sqrt{h_i}} , \; z_2 = \frac{(q + e_i\sqrt{\tau})}{\sqrt{h_i}} . \tag{3.16}$$

Chaloner and Brant (1988) discuss appropriate choices of q and then declare an observation y_i to be an "outlier" if its P_i is large.

To help us interpret the properties of the Chaloner-Brant procedure, we will obtain an explicit formula for (3.15) as a function of the standard diagnostic measures. We may, of course, write $p(\varepsilon_i, \sigma^2|\mathbf{y}, X)$ as

$$p(\varepsilon_i, \sigma^2|\mathbf{y}, X) = p(\varepsilon_i|\sigma^2; \mathbf{y}, X) p(\sigma^2|\mathbf{y}; X) \tag{3.17}$$

and from (3.14) and results of Section 2, we have that the right hand side of (3.17) is

$$\frac{1}{\sqrt{2\pi\sigma^2 h_i}} \exp\left\{-\frac{1}{2\sigma^2 h_i}(\varepsilon_i - e_i)^2\right\} \cdot K(\sigma^2)^{-[(n-p)/2+1]} \exp -\left\{\frac{1}{2\sigma^2}(n-p)s^2\right\} . \tag{3.18}$$

Now P_i as given in (3.15) may be written as $P_i = P_{i1} + P_{i2}$, where

$$P_{i1} = P(\varepsilon_i > q\sigma|\mathbf{y}; X) , \; P_{i2} = P(\varepsilon_i < -q\sigma|\mathbf{y}; X) \tag{3.19}$$

Now

$$P_{1i} = \int_0^\infty \int_{q\sigma}^\infty p(\varepsilon_i|\sigma^2; \mathbf{y}, X) p(\sigma^2|\mathbf{y}, X) d\varepsilon_i d\sigma^2 \tag{3.20}$$

$$= 1 - P(T \leq \frac{e_i}{\sqrt{h_i s^2}} | \Delta = -\frac{q}{\sqrt{h_i}}). \tag{3.26a}$$

Hence,

$$P_{2i} \simeq 1 - \Phi\left[\frac{\left(\frac{e_i}{\sqrt{h_i s^2}} + \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}}}\right] . \tag{3.27}$$

To summarize we have

$$P_i \simeq 1 - \Phi(u_1) + \Phi(u_2) \tag{3.28}$$

with

$$u_1 = \frac{\left(\frac{e_i}{\sqrt{h_i s^2}} + \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}}} \tag{3.28a}$$

and

$$u_2 = \frac{\left(\frac{e_i}{\sqrt{h_i s^2}} - \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}}} \tag{3.28b}$$

It can be shown that u_1 and u_2 can be written as:

$$u_1 = \frac{\left(\frac{r_i}{\sqrt{l_i}} + \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{1}{2(n-p)}\frac{r_i^2}{l_i}}} \tag{3.29}$$

$$u_2 = \frac{\left(\frac{r_i}{\sqrt{l_i}} - \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{1}{2(n-p)}\frac{r_i^2}{l_i}}} \tag{3.30}$$

where r_i is the studentized residual $\frac{e_i}{\sqrt{s^2(1-h_i)}}$; and l_i is the measure of leverage given by

$$l_i = x_i'(X_{(i)}'X_{(i)})^{-1}x_i = \frac{h_i}{(1-h_i)} . \tag{3.31}$$

Now, suppose that r_i is positive and fixed. If we now let $h_i \to 1$, that is, the leverage of the observation is very high, then $l_i \to \infty$ and $u_1 \to q$, $u_2 \to -q$ and from (3.28), we see that P_i goes to $2\Phi(-q)$, which is the conditional probability, given $\boldsymbol{\beta}, \sigma^2$ that y_i is outlying in the Chaloner-Brant sense, that is,

$$2\Phi(-q) = P[|y_i - \mathbf{x}_i'\boldsymbol{\beta}| > q\sigma \Big| \boldsymbol{\beta}, \sigma^2] \tag{3.32}$$

by I to be spuriously generated is

$$c_I = K(s^2_{(I)})^{-(n-p-k)/2}|I_k - H_I|^{-1/2} \tag{4.1}$$

where

$$K^{-1} = \sum (s^2_{(I)})^{-(n-p-k)/2}|I_k - H_I|^{-1/2} \tag{4.2}$$

and where the sum is taken over all sets I of size k of distinct integers from $(1,\ldots,n)$. Now, on the other hand, it is interesting to note that for the variance inflation model (2.12), the probability that k observations indexed by I are generated with 'noise' ϵ given by $N(O,\delta^2\sigma^2)$, $\delta^2 > 1$, and $(n-k)$ generated by $N(O,\sigma^2)$ takes the form, as proved in Box and Tiao (1968),

$$w_I = C\left(\frac{\alpha}{1-\alpha}\right)^k \delta^{-k}\left(\frac{|X'X|}{|X'X - \phi X_I' X_I|}\right)^{1/2}\left(\frac{s^2}{\hat{s}^2_{(I)}}\right)^{\frac{n-p}{2}}, \tag{4.3}$$

where C is a normalizing constant that can be shown to be the probability of no outliers, and $\phi = 1 - \delta^{-2}$. (For a precise definition of $\hat{s}^2_{(I)}$ see Box and Tiao (1968).) When δ is large, it can be shown (Peña and Tiao (1992)) that w_I is approximately

$$w_I = C\left(\frac{\alpha}{1-\alpha}\right)^k \delta^{-K}\left(\frac{|X'X|}{|X'_{(I)}X_{(I)}|}\right)^{1/2}\left(\frac{s^2}{s^2_{(I)}}\right)^{\left(\frac{n-p}{2}\right)}. \tag{4.3a}$$

Adding up the values w_I for all sets of size k we obtain the probability of exactly k outliers in the sample, and, in turn, by adding all the w_I's up, the constant C could be obtained.

For fixed k and δ large, the conditional probability that a particular set of k observations indexed by I are spuriously generated with noise variances $\delta^2\sigma^2$ is (Peña and Tiao (1992)):

$$p_I = C'\left(\frac{|X'X|}{|X'_{(I)}X_{(I)}|}\right)^{1/2}\left(\frac{s^2}{s^2_{(I)}}\right)^{\frac{n-p}{2}}, \tag{4.4}$$

which in turn can be written as

$$p_I = C'' \cdot |I - H_I|^{-1/2}(s^2_{(I)})^{-(n-p)/2}, \tag{4.5}$$

P_i given by (3.15) with $q = 2$, and the c_i as given by (4.1), which is, as demonstrated in Section 4, inversely proportional to the predictive ordinate. As anticipated, the values for P_i are all very small, except for a large value of .9552 at $i = 11$, which is 2.2 times greater than the next largest value that occurs at $i = 14$. As far as the c_i, the largest occurs at $i = 11$ with a value of .9708 which is 511 times greater than the next largest value, that occurs at $i = 20$. We note that the outlier will be idenitified by both procedures, although in a more powerful way by c_i then by P_i.

The next experiment we carried out was to introduce to the original set of data of Table 5.1 an outlier at the high leverage point $x_{20} = 40$, by again adding 4 to the original

Table 5.1.

Data for the simulated example

	x	y	h
1	1	1.955	.13
2	2	2.201	.11
3	3	3.235	.10
4	4	5.862	.09
5	5	5.944	.08
6	6	7.514	.07
7	7	8.397	.06
8	8	9.756	.06
9	9	10.401	.05
10	10	9.659	.05
11	11	12.375	.05
12	12	14.125	.05
13	13	14.729	.05
14	14	12.622	.05
15	15	15.726	.06
16	16	16.677	.06
17	17	18.318	.07
18	18	18.489	.08
19	19	19.998	.09
20	40	42.607	.62

observed data point. Here, however, the largest P_i value is .6758 and occurs at $i = 14$, instead of the expected $i = 20$. Indeed the next largest of the P_i's is P_{20}, with the value

country involved in this set of data. We should note that the data is from 1959 and 1960 UN data.

Table 5.2.

Probability of k outliers for the logistic model with Zellner-Moulton data.

k	0	1	2	3	4
$p(k)$	.3566	.3732	.1961	.0611	.0130

ACKNOWLEDGEMENTS

The authors would like to acknowledge support, with thanks, for this research. Daniel Peña was supported by DGICYT (Spain), under Grant No. PB90-0266. Irwin Guttman was supported by NSERC (Canada) through Grant No. A8743.

APPENDIX I. The non-central T-distribution.

We remind the reader that the definition of the classical non-central T-variable comes about as follows. Suppose Z and W are independent random variables with distributions given by

$$Z \sim N(O,1) \quad \text{and} \quad W \sim \chi^2_v \tag{AI.1}$$

Then the non-central T-variable with non-central parameter Δ, degrees of freedom v, is defined as

$$T = \frac{(Z+\Delta)}{\sqrt{W/v}} \tag{AI.2}$$

It is very easy to see that the density of the random variable T as defined in (AI.2) may be written as

$$\mathfrak{p}(t) = \int_0^\infty \sqrt{\frac{w}{v}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\sqrt{\frac{w}{v}}t - \Delta\right]^2 h_v(w)dw \right. \tag{AI.3}$$

or

$$P(T \leq t) \doteq P\left(Z \leq \frac{t-\Delta}{\sqrt{1+\frac{t^2}{2v}}}\right) = \Phi(u) \qquad (AI.10)$$

where $u = \dfrac{(t-\Delta)}{\sqrt{1+t^2/2v}}$, and, $Z \sim N(O,1)$. Hence, the density function of T is , differentiating, such that

$$\mathfrak{p}(t) \dot{\sim} \phi(u) \left|\frac{du}{dt}\right| , \qquad (AI.11)$$

where of course $\phi(u)$ is the density of a $N(O,1)$ random variable, and, to repeat.

$$u = \frac{(t-\Delta)}{\sqrt{1+\frac{t^2}{2v}}} . \qquad (AI.12)$$

We have that the Jacobian of the transformation from t to u given by (AI.12) has absolute value

$$|J| = \frac{1}{\left|\frac{du}{dt}\right|}$$

so that the density of U is

$$g(u) \doteq \phi(u) \left|\frac{du}{dt}\right| \times \frac{1}{\left|\frac{du}{dt}\right|} \qquad (AI.13)$$

that is, the density of U is

$$g(u) = \phi(u) ,$$

and we have that, approximately, U has the density of a standard normal $N(O.1)$ variable. for large v , and the theorem is proved.

REFERENCES

Atkinson. A.C. (1975). *Plots, Transformations and Regression.* Oxford: Clarendon Press.

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

Box, G.E.P. (1980). "Sampling and Bayes Inference in Scientific Modelling and Robustness." *Journal of the Royal Statistical Society*, Series A 143, 383-430 (with discussion).

Box, G.E.P., and Meyer, R. (1986). "An Analysis for Unreplicated Fractional factorials." *Technometrics* 28, 11-18.

Box, G.E.P., and Tiao, G.C. (1968). "A Bayesian Approach to Some Outlier Problems." *Biometrika* 55, 119-129.

Casella, G. and George, E.I. (1992). "An Introduction to Gibbs Sampling." *American Statistician.* (To Appear)

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

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

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

Freeman, P.R. (1980). "On the number of outliers in data from a linear model." *Bayesian Statistics*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Editors. Valencia University Press, 349-365.

Geisser, S. (1980). "Discussion of a paper by G.E.P. Box", *Journal of the Royal Statistical Society.* Series A 143, 416-7.

Geisser, S. (1987). "Influential observations, diagnostics and discordancy tests". *Journal of Applied Statistics* 14, 133-42.

Geisser, S. (1988). "Predictive approaches to discordancy testing". *Bayesian and Likelihood Influence: Essays in Honor of George Barnard*, S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, Editors. Amsterdam: North Holland.

Kendall, M., and Stuart, A. (1977). *The Advanced Theory of Statistics; Volume 1 - Distribution Theory* (Fourth Edition): Macmillan Publishing Company, New York.

Ljung, G.M. (1989). "A note on estimation of Missing Values in Time Series". *Communications in Statistics, Simulation and Computation* 18, 459-465.

Peña, D. and Maravall, A. (1991). "Interpolation, Outliers and Inverse Autocorrelations." *Communications in Statistics, Theory and Methods* 20, 3175-3186.

Peña, D. and Tiao, G.C. (1992). "Bayesian Robustness Functions for Linear Models." *Bayesian Statistics* 4., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds). Oxford University Press, 365-388.

Pettit, L.I. and Smith, A.F.M. (1985). "Outliers and Influential Observations in Linear Models." *Bayesian Statistics* 2, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (eds). Elsevier Science Publishers, 473-494.

Tussell, F. (1990). "Testing for Interaction in Two-Way ANOVA tables with no replication." *Computational Statistics and Data Analysis* 10, 29-45.

WORKING PAPERS 1992

92-01 Carlos Ocaña Pérez de Tudela and Joan Pasqual i Rocabert
"Environmental Costs of Residuals: A Characterization of Efficient Tax Policies"

92-02 Nélida E. Ferreti, Diana M. Kelmansky and Victor J. Yohai
"A Goodness-of-Fit Test Based on Ranks for Arma Models"

92-03 Jesús Juan and Daniel Peña
"A Simple Method to Identify significant Effects in Unreplicated Two-Level Factorial Designs"

92-04 Jacek Osiewalski and Mark. F.J. Steel
"Bayesian Marginal Equivalence of Elliptical Regression Models"

92-05 Jacek Osiewalski and Mark. F.J. Steel
"Posterior Moments of Scale Parameters in Elliptical Regression Models"

92-06 Omar Licandro
"A Non-Walrasian General Equilibrium Model With Monopolistic Competition and Bargaining"

92-07 David de la Croix and Omar Licandro
"The q Theory of Investment Under Unit Root Tests"

92-08 Santiago Velilla
"A Note on the Multivariate Box-Cox Transformation to Normality"

92-09 T. Cipra, R. Romera and A. Rubio
"Square Root Kalman Filter with Contaminated Observations"

92-10 Gary Koop, Jacek Osiewalski and Mark F.J. Steel
"Baysian Long-Run Prediction in Time Series Models"

92-11 Eduardo Ley
"Switching Regressions and Activity Analysis

92-12 Julie van den Broeck, Gary Koop, Jacek Osiewalski and Mark F.J. Steel
"Stochstic Frontier Models: A Bayesian Perspective"

92-13 Carlos San Juan Mesonada
"Costs and Benefits of the Cap Reform"

92-14 Eduardo Ley
"A Note on Ramsey and Corlett-Hague Rules"

92-15 Miguel A. Delgado
"Testing the Equality of Nonparametric Regression Curves"

92-16 Miguel A. Delgado and Peter M. Robinson
"Nonparametric and Semiparametric Methods for Economic Research"

92-17 Miguel A. Delgado
"Computing Nonparametric Functional Stimates in Semiparametric Problems"

92-18 Pedro Fraile Balbin
"The Difussion of Modern Iron and Steed Technology in France, Spain and Italy"

DOCUMENTOS DE TRABAJO 1992

92-01 Daniel Peña
"Reflexiones sobre la enseñanza experimental de la Estadística"

92-02 Carlos Newland y Mª Jesús San Segundo
"Una contrastación de la Teoría del Capital Humano: ingresos y educación en Buenos Aires a mediados del siglo XIX"

92-03 Luis Rodríguez Romero
"Actividad económica y actividad tecnológica: un análisis simultáneo de datos de panel"

92-04 Alfonso Alba Ramírez
"El empleo asalariado en España desde 1987 hasta 1991. Especial referencia al tipo de contrato"

92-05 Juan Ignacio Peña
"Contratación Asíncrona, Riesgo Sistemático y Contrastes de Eficiencia"

92-06 Pedro F. Delicado Useros
"NOPARAM: Estimación Funcional No-Paramétrica. Un acercamiento del Software CURVDAT al usuario"

92-07 Carmen Higuera y Javier Ruiz-Castillo
"Indices de Precios Individuales para la Economía Española con base de 1976 y 1983"

92-08 F. Javier Suárez Bernaldo de Quirós
"Seguro de depósitos y comportamiento bancario: un análisis basado en la teoría de opciones"

92-09 Juan Urrutia
"La moda en Economía. (El caso del ajuste liberal a la crisis)"

92-10 Antoni Espasa
"El Análisis de la Coyuntura Económica: un Ejercicio basado en Modelos"

92-11 J. Ignacio Peña
"Nuevos Modelos Estadísticos para el Análisis de Mercados Financieros"

92-12 Miguel Angel López López
"Sobre la Representación Continua de Relaciones de Preferencia"

REPRINT 1992

- Antoni Espasa
"Un análisis de la inflación en la economía española a través de los índices de precios al consumo"

- Eduardo Morales, Antoni Espasa and Mª Luisa Rojo
"Univariate Methods for the Analysis of the Industrial Sector in Spain"

Working Paper 92-32
July 1992

División de Economía
Universidad Carlos III de Madrid
Calle Madrid, 126
28903 Getafe (Spain)
Fax (341) 624-9849

COMPARING PROBABILISTIC METHODS FOR OUTLIER DETECTION

Daniel Peña and Irwin Guttman*

Abstract

This paper compares the use of two posterior probability methods to deal with outliers in linear models. We show that putting together diagnostics that come from the mean-shift and variance-shift models yields a procedure that seems to be more effective than the use of probabilities computed from the posterior distributions of actual realized residuals. The relation of the suggested procedure to the use of a certain predictive distribution for diagnostics is derived.

Key words:
Diagnostic, Posterior and Predictive distributions, leverage, linear models.

*Peña, Departamento de Estadística y Econometría, Universidad Carlos III de Madrid, Getafe, 28903 Spain; Guttman, Department of Statistics, University of Toronto, Toronto, Ontario M5S1A1.

1. INTRODUCTION.

The analysis of outliers from the Bayesian point of view has become increasingly interesting to the Statistical Profession because of the possibility of carrying out the difficult computations involved using new algorithms such as the Gibbs sampling method (see Gelfand and Smith (1990), or Casella and George (1992)). Also, specific algorithms have been developed to deal with the Bayesian treatment of outliers (Peña and Tiao (1992)).

Not only has the work on analysis of outliers proved useful in its own right, but it turns out that many other statistical problems can also be usefully analyzed by approaching these problems from the outlier point of view. As examples, we would cite: analysis of unreplicated fractional factional factorial designs to detect significant effects, see Box and Meyer (1986), Juan and Peña (1992); detection of interaction in unreplicated ANOVA designs, see Tussell (1990); estimation of missing observations in time series models, see Ljung (1989) and Peña and Maravall (1991).

There have been three main approaches to the problem of outliers in the literature. Succinctly, these may be classified as (i) the diagnostic approach (ii) the Bayesian approach and (iii) 'robust' approach to estimation and tests of hypothesis in the presence of outliers. The first approach is clearly identified with the work of Cook and Weisberg (1982), Belsley, Kuh and Welsch (1980), and Atkinson (1985), and the aim of workers in this area is mostly that of identification of observations that may be deemed outlying and/or influential. The approach listed as (iii) above has been motivated by the work of Huber (1981), and Hampel et al (1986), the aim here being to build estimatiors that are not affected by the fraction of the sample that is outlying. Truly in the middle and listed as such above, is the Bayesian approach, which seeks to combine identification with estimation: see for example Box and Tiao (1968); Guttman et al (1978), etc. Here, the identification is carried out using the posterior probabilities for an observation or a set of observations being outlying, and these are used as weights in estimation procedures.

In the Bayesian approach, two main categories have emerged. The first of these confines itself to postulating a (null) model for the generation of the data and then seeks identification methods for outliers with no alternative model to the null entertained. Examples of work in the category are (i) use of the predictive distribution for detection (ii) using the posterior probabilities of various unobserved pertubations and (iii) looking at the change in a posterior of interest when some observations are deleted. These methods will be discussed in Section 3 of this paper.

The second category that has emerged takes into account an alternative model for the generation of a subset of the sample. As examples, various authors have proposed and utilized the mean-shift model and the variance-inflation model. These methods are discussed in Section 4 of this paper.

All of the above will be compared and discussed in Section 5, where an illustrative example based on real data will be given.

2. THE BASIC MODEL.

In this paper we will be concerned with the univariate linear model

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

where $\mathbf{y}$ is a $(n \times 1)$ vector of normal random variables, X is a $(n \times p)$ matrix of full column rank $p < n$, $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of parameters and $\boldsymbol{\epsilon}$ is a $(n \times 1)$ vector of normal variables, mean vector $\mathbf{O}$ and with covariance matrix $\sigma^2 I_n$. This will be called the null model in the rest of the paper. The estimation of the parameters $(\boldsymbol{\beta}, \sigma^2)$ will be done assuming a non-informative prior

$$p(\boldsymbol{\beta}, \sigma^2) \propto (\sigma^2)^{-1} . \tag{2.2}$$

It is well known that in this case the posterior distribution for the parameters is such that

conditional on σ^2, we have

$$\boldsymbol{\beta} \sim N\left(\hat{\beta}; \sigma^2 (X'X)^{-1}\right) \tag{2.3}$$

where $\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$, and $p(\sigma^2|\mathbf{y}, x)$ is such that

$$\frac{(n-p)s^2}{\sigma^2} \sim \chi^2_{n-p}\,, \tag{2.4}$$

where

$$(n-p)s^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'[I-H]\mathbf{y}\,, \tag{2.5}$$

with H denoting the so called hat matrix

$$H = X(X'X)^{-1}X'\,. \tag{2.6}$$

We denote a set of k distinct integers chosen from the set $(1,\ldots,n)$ by I. Then, the vector $\mathbf{y}$ can be decomposed as

$$\mathbf{y}' = (\mathbf{y}'_I\ \mathbf{y}'_{(I)}) \tag{2.7}$$

where (I) means "delete set I". Similarly, the X matrix can be partitioned as

$$X' = [X'_I\ X'_{(I)}]\,. \tag{2.8}$$

(The use of the symbol I without brackets means restrict information to the set I.) Consistent with the above notation, we will use in the rest of the paper the designations

$$\hat{\boldsymbol{\beta}}_{(I)} = \left(X'_{(I)}X_{(I)}\right)^{-1}X'_{(I)}\mathbf{y}_{(I)} \tag{2.9}$$

$$s^2_{(I)} = \left(y_{(I)} - X_{(I)}\hat{\boldsymbol{\beta}}_{(I)}\right)'\left(y_{(I)} - X_{(I)}\hat{\boldsymbol{\beta}}_{(I)}\right)/(n-p-k) \tag{2.10}$$

that is, $\hat{\boldsymbol{\beta}}_{(I)}$ and $s^2_{(I)}$ are estimators of $\boldsymbol{\beta}$ and σ^2 based on $(X_{(I)}, \mathbf{y}_{(I)})$, etc.

In contrast with the null model (2.1) we will be concerned in this paper with two alternative models. The first is the mean-shift model and takes the form for the generation of the observations $\mathbf{y} = (\mathbf{y}'_I, \mathbf{y}'_{(I)})'$,

$$\begin{aligned} \mathbf{y}_I &= X_I\boldsymbol{\beta} + \mathbf{a} + \boldsymbol{\epsilon}_I \\ \mathbf{y}_{(I)} &= X_{(I)}\boldsymbol{\beta} \quad\ + \boldsymbol{\epsilon}_{(I)} \end{aligned} \tag{2.11}$$

where $\mathbf{a}$ is a $(k \times 1)$ vector of mean-shift constants, and $\boldsymbol{\epsilon}_I \sim N(\mathbf{O}, \sigma^2 I_k)$ independent of $\boldsymbol{\epsilon}_{(I)} \sim N(\mathbf{O}, \sigma^2 I_{n-k})$. This model was used by Guttman (1973) and further utilized by Guttman et al (1978).

The second model we will be involved with is the so-called variance-inflation model which says that the distribution of $\boldsymbol{\epsilon}$ of (2.11) is such that

$$\epsilon_i \sim (1-\alpha)N(O, \sigma^2) + \alpha N(O, \delta^2\sigma^2) \tag{2.12}$$

where α is small and $\delta^2 > 1$ is usually thought of as being large. These models have been compared in Freeman (1980), Eddy (1980) and Pettit and Smith (1985).

3. BAYESIAN IDENTIFICATION METHODS USING THE NULL MODEL.

Several authors (for example, see Box (1980) , Geisser (1980,1987,1988), Pettit and Smith (1985) and Peña and Tiao (1992)) have advocated the use of the predictive distribution to arrive at methods for detecting outlying observations. A main idea here is to compute the predictive density

$$p(\mathbf{y}_I|\mathbf{y}_{(I)}) = \int p(\mathbf{y}_I|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{(I)})d\boldsymbol{\theta} \tag{3.1}$$

where $\mathbf{y}_I$ is a vector of k observations, and where $\mathbf{y}_{(I)}$ is the sample data at hand on which the posterior for the vector of parameters $\boldsymbol{\theta}$ is based. For the linear models with normal noise as given in (2.1), it is well known that for the non-informative prior (2.2)

$$p(\mathbf{y}_I|\mathbf{y}_{(I)}) = K(s^2_{(I)})^{-k/2}|I - H_I|^{1/2}(1+Q_I)^{-(n-p)/2} \tag{3.2}$$

where

$$K = \frac{\Gamma\left(\frac{n-p}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^k \Gamma\left(\frac{n-p-k}{2}\right)(n-p-k)^{k/2}}\,, \tag{3.3}$$

and

$$Q_I = \frac{(\mathbf{y}_I - X_I\hat{\boldsymbol{\beta}}_{(I)})'(I-H_I)(\mathbf{y}_I - X_I\hat{\boldsymbol{\beta}}_{(I)})}{(n-p-k)s^2_{(I)}}\,. \tag{3.4}$$

Here H_I is the $k \times k$ block of the hat matrix H of (2.6) formed by using the k rows and columns of H indexed by I, or

$$H_I = X_I(X'X)^{-1}X_I' \,. \tag{3.4a}$$

Indeed, H_I is refereed to as the "leverage of observations $\mathbf{y}_I$".

Expression (3.2) could be written in another form that will be useful when comparing it to the other approaches to be discussed in this paper. In order to do so, we use the identity (see Cook and Weisberg, 1982, pg. 191)

$$(n-p-k)s^2_{(I)} = (n-p)s^2 - \mathbf{e}_I'(I-H_I)^{-1}\mathbf{e}_I \,, \tag{3.5}$$

for then

$$\frac{s^2}{s^2_{(I)}} = \frac{n-p-k}{n-p}\left(1 + \frac{\mathbf{e}_I'(I-H_I)^{-1}\mathbf{e}_I}{(n-p-k)s^2_{(I)}}\right) \,. \tag{3.6}$$

Here, we have partitioned the residual vector $\mathbf{e}$,

$$\mathbf{e} = (I-H)\mathbf{y} \tag{3.6a}$$

using

$$\mathbf{e} = (\mathbf{e}_I', \mathbf{e}_{(I)}')' = (I-H)(\mathbf{y}_I', \mathbf{y}_{(I)}')' \tag{3.6b}$$

with $X = (X_I' \vdots X_{(I)}')'$ used in constructing $H = X(X'X)^{-1}X'$. Now, it can be shown that

$$\mathbf{y}_I - x_I\hat{\boldsymbol{\beta}}_{(I)} = (I-H_I)^{-1}\mathbf{e}_I \tag{3.7}$$

and, therefore, (3.4) itself can be written as

$$Q_I = \frac{\mathbf{e}_I'(I-H_I)^{-1}\mathbf{e}_I}{(n-p-k)s^2_{(I)}} \,. \tag{3.8}$$

Hence, we may rewrite (3.2) as follows

$$p(\mathbf{y}_I|\mathbf{y}_{(I)}) = K_1 \frac{|I-H_I|^{1/2}}{\left(\frac{n-p}{n-p-k}\,\frac{s^2}{s^2_{(I)}}\right)^{-k/2}}\,[1+Q_I]^{-(n-p)/2} \tag{3.9}$$

and using (3.6), we find

$$p(\mathbf{y}_I|\mathbf{y}_{(I)}) = K_1|I - H_I|^{1/2}\,[1 + Q_I]^{-(n-p-k)/2}\,, \tag{3.10}$$

where

$$K_1 = \left(\frac{(n-p)s^2}{n-p-k}\right)^{-k/2} K\,. \tag{3.11}$$

Note that $p(\mathbf{y}_I|\mathbf{y}_{(I)})$ behaves in a rather expected way: the larger the studentized residual statistic (3.8), as measured by the quadratic function (3.8) that takes into account the leverage - the smaller the ordinate of the density (3.10). We also note that small ordinates could occur because of large leverage, due to the presence of the factor $|I - H_I|^{1/2}$ in (3.10). Because of all this, many authors use the predictive to rank sets $\mathbf{y}_I$, deeming these with lowest $p(y_I|y_{(I)})$ as outliers.

Another approach that utilizes the predictive distribution is the one by Johnson and Geisser (1983). They showed that when monotoring the change in the predictive distribution when some observations are dropped, measures of influence and outlyingness can be built that are related to the standard ones used in the literature. Johnson and Geisser (1985) and Guttman and Peña (1988) extended these ideas to changes in certain posterior distributions. Recently, Guttman and Peña (1992) have shown that the behaviour of these changes are related to the probability of a group of observations being spuriously generated that is, generated not according to the null model. The probability will be precisely defined and discussed in section 4.

The third approach has been suggested by Chaloner and Brant (1988) and is based on the posterior probabilities

$$P_i = P(|\varepsilon_i| > q\sigma|\mathbf{y}..X) \tag{3.12}$$

where the ε_i are the unobserved residuals, given by

$$\varepsilon_i = y_i - \mathbf{x}_i'\boldsymbol{\beta}\,. \tag{3.13}$$

Following the results outlined in Section 2, we have that the posterior distribution of the ε_i, as given by (3.13) conditional on σ^2, is easily seen to be on using (2.3), such that

$$\varepsilon_i \sim N(e_i, \sigma^2 h_i) , \tag{3.14}$$

where $e_i = y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$ is the residual of the observation y_i, and h_i is the i-th diagonal element of H given in (2.6), which is to say, h_i is the leverage of y_i. Chaloner and Brant (1988) have shown that P_i of (3.12) can be written in the form

$$P_i = 1 - \int_0^\infty [\Phi(z_1) - \Phi(z_2)] p(\tau|\mathbf{y}, \mathbf{x}) d(\tau) . \tag{3.15}$$

where $\tau = \sigma^{-2}$, so that $p(\tau|\mathbf{y}, \mathbf{x})$ is the density of a $\dfrac{\chi^2_{n-p}}{(n-p)s^2}$ variable, with

$$z_1 = \frac{(q - e_i\sqrt{\tau})}{\sqrt{h_i}} , \; z_2 = \frac{(q + e_i\sqrt{\tau})}{\sqrt{h_i}} . \tag{3.16}$$

Chaloner and Brant (1988) discuss appropriate choices of q and then declare an observation y_i to be an "outlier" if its P_i is large.

To help us interpret the properties of the Chaloner-Brant procedure, we will obtain an explicit formula for (3.15) as a function of the standard diagnostic measures. We may, of course, write $p(\varepsilon_i, \sigma^2|\mathbf{y}, X)$ as

$$p(\varepsilon_i, \sigma^2|\mathbf{y}, X) = p(\varepsilon_i|\sigma^2; \mathbf{y}, X) p(\sigma^2|\mathbf{y}; X) \tag{3.17}$$

and from (3.14) and results of Section 2, we have that the right hand side of (3.17) is

$$\frac{1}{\sqrt{2\pi\sigma^2 h_i}} \exp\left\{-\frac{1}{2\sigma^2 h_i}(\varepsilon_i - e_i)^2\right\} \cdot K(\sigma^2)^{-[(n-p)/2+1]} \exp -\left\{\frac{1}{2\sigma^2}(n-p)s^2\right\} . \tag{3.18}$$

Now P_i as given in (3.15) may be written as $P_i = P_{i1} + P_{i2}$, where

$$P_{i1} = P(\varepsilon_i > q\sigma|\mathbf{y}; X) , \; P_{i2} = P(\varepsilon_i < -q\sigma|\mathbf{y}; X) \tag{3.19}$$

Now

$$P_{1i} = \int_0^\infty \int_{q\sigma}^\infty p(\varepsilon_i|\sigma^2; \mathbf{y}, X) p(\sigma^2|\mathbf{y}, X) d\varepsilon_i d\sigma^2 \tag{3.20}$$

and in view of (3.14), we have

$$P_{1i} = \int_0^\infty [1 - \Phi\left(\frac{q\sigma - e_i}{\sqrt{h_i\sigma^2}}\right)]p(\sigma^2|\mathbf{y}; X)d\sigma^2 \ , \tag{3.21}$$

and letting $w = (n-p)s^2/\sigma^2$, we find that

$$P_{1i} = \int_0^\infty \Phi\left(\frac{e_i}{\sqrt{h_i s^2}}\sqrt{\frac{w}{n-p}} - \frac{q}{\sqrt{h_i}}\right) h_{n-p}(w)dw \ , \tag{3.22}$$

where, in general,

$$h_\nu(w) = [2^{\frac{\nu}{2}}\Gamma\left(\frac{\nu}{2}\right)]^{-1} w^{\nu/2-1}\exp(-w/2) \tag{3.23}$$

is the density of a χ^2_ν-variable. Now it is proved in Appendix I of this paper that the right hand side of (3.22) is in the form of the probability that a non-central T-variable with $(n-p)$ degrees of freedom, non-central parameter $\Delta = q/\sqrt{h_i}$ has value less than or equal to $t = \frac{e_i}{\sqrt{h_i s^2}}$, that is,

$$P_{1i} = P(T \leq \frac{e_i}{\sqrt{h_i s^2}}|\Delta = \frac{q}{\sqrt{h_i}}; \nu = n-p) \tag{3.24}$$

and, as is well known, (details given in Appendix I), (3.24) has as an approximation, for moderate to large n , given by

$$P_{1i} \simeq P(Z \leq \frac{t-\Delta}{\sqrt{1+\frac{t^2}{2(n-p)}}}) \tag{3.25}$$

where $Z \sim N(O,1)$, that is, we have

$$P_{1i} \simeq \Phi\left(\frac{\frac{e_i}{\sqrt{h_i s^2}} - \frac{q}{\sqrt{h_i}}}{\sqrt{1+\frac{e_i^2}{2h_i s^2(n-p)}}}\right) \ . \tag{3.25a}$$

Similarily, we may easily find that

$$\begin{aligned} P_{2i} &= P(\varepsilon_i < -q\sigma|\mathbf{y}; X) \\ &= 1 - \int_0^\infty \Phi\left[\frac{e_i}{\sqrt{h_i s^2}}\sqrt{\frac{w}{n-p}} + \left(\frac{q}{\sqrt{h_i}}\right)\right] h_{n-p}(w)dw \end{aligned} \tag{3.26}$$

$$= 1 - P(T \leq \frac{e_i}{\sqrt{h_i s^2}} | \Delta = -\frac{q}{\sqrt{h_i}}). \tag{3.26a}$$

Hence,

$$P_{2i} \simeq 1 - \Phi\left[\frac{\left(\frac{e_i}{\sqrt{h_i s^2}} + \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}}}\right] . \tag{3.27}$$

To summarize we have

$$P_i \simeq 1 - \Phi(u_1) + \Phi(u_2) \tag{3.28}$$

with

$$u_1 = \frac{\left(\frac{e_i}{\sqrt{h_i s^2}} + \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}}} \tag{3.28a}$$

and

$$u_2 = \frac{\left(\frac{e_i}{\sqrt{h_i s^2}} - \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}}} \tag{3.28b}$$

It can be shown that u_1 and u_2 can be written as:

$$u_1 = \frac{\left(\frac{r_i}{\sqrt{l_i}} + \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{1}{2(n-p)}\frac{r_i^2}{l_i}}} \tag{3.29}$$

$$u_2 = \frac{\left(\frac{r_i}{\sqrt{l_i}} - \frac{q}{\sqrt{h_i}}\right)}{\sqrt{1 + \frac{1}{2(n-p)}\frac{r_i^2}{l_i}}} \tag{3.30}$$

where r_i is the studentized residual $\frac{e_i}{\sqrt{s^2(1-h_i)}}$; and l_i is the measure of leverage given by

$$l_i = x_i'(X_{(i)}'X_{(i)})^{-1}x_i = \frac{h_i}{(1-h_i)} . \tag{3.31}$$

Now, suppose that r_i is positive and fixed. If we now let $h_i \to 1$, that is, the leverage of the observation is very high, then $l_i \to \infty$ and $u_1 \to q$, $u_2 \to -q$ and from (3.28), we see that P_i goes to $2\Phi(-q)$, which is the conditional probability, given $\boldsymbol{\beta}, \sigma^2$ that y_i is outlying in the Chaloner-Brant sense, that is,

$$2\Phi(-q) = P[|y_i - \mathbf{x}_i'\boldsymbol{\beta}| > q\sigma \Big| \boldsymbol{\beta}, \sigma^2] \tag{3.32}$$

Therefore, the *posterior* probability of a high leverage observation $y_i (h_i \simeq 1)$ being an outlier in the Chaloner-Brant sense, is, in moderate to large samples, very nearly

$$\lim_{h_i \to 1} P_i = 2\Phi(-q) \tag{3.33}$$

regardless of the data, so that leverage is not always being treated by P_i of (3.15) in the way we would wish in moderate to large numbers.

It is interesting to note the similarities between the Chaloner and Brant (1988) result for P_i, as stated in (3.15), and the approximation derived in this paper stated in (3.28). In fact, it is easy to see that after setting τ in (3.15) equal to

$$\tau = \frac{w}{(n-p)s^2} \ , \ w \sim \chi^2_{n-p} \tag{3.34}$$

that P_i of (3.15) can be written as

$$\begin{aligned} &\int_0^\infty \Phi\left(-\frac{e_i}{\sqrt{h_i s^2}}\sqrt{\frac{w}{n-p}} - \frac{q}{\sqrt{h_i}}\right) h_{n-p}(w)dw \\ &+ \int_0^\infty \Phi\left(\frac{e_i}{\sqrt{h_i s^2}}\sqrt{\frac{w}{n-p}} - \frac{q}{\sqrt{h_i}}\right) h_{n-p}(w)dw \end{aligned} \tag{3.35}$$

while the approximation (3.28) is of course,

$$\Phi\left(\frac{\left[-\frac{e_i}{\sqrt{h_i s^2}} - \frac{q}{\sqrt{h_i}}\right]}{a}\right) + \Phi\left(\frac{\left[\frac{e_i}{\sqrt{h_i s^2}} - \frac{q}{\sqrt{h_i}}\right]}{a}\right) \tag{3.36}$$

where

$$a = \sqrt{1 + \frac{e_i^2}{2h_i(n-p)s^2}} \ . \tag{3.36a}$$

We note that in (3.35) and (3.36), the signs attached to $\frac{e_i}{\sqrt{h_i s^2}}$ and $\frac{q}{\sqrt{h_i}}$ that appear in Φ functions are the same, but that the (approximate) effect of integration with respect to $w \sim \chi^2_{n-p}$ is to remove $\sqrt{\frac{w}{n-p}}$ and replace it with $\frac{1}{a}$, while changing $\frac{q}{\sqrt{h_i}}$ to $\frac{q}{a\sqrt{h_i}}$.

4. BAYESIAN IDENTIFICATION METHODS USING ALTERNATIVE MODELS.

Starting with the mean-shift model (2.11), it can be shown (see Guttman and Peña (1992)) that, conditional on k, the probability for a given set of k observations indexed

by I to be spuriously generated is

$$c_I = K(s^2_{(I)})^{-(n-p-k)/2}|I_k - H_I|^{-1/2} \tag{4.1}$$

where

$$K^{-1} = \sum (s^2_{(I)})^{-(n-p-k)/2}|I_k - H_I|^{-1/2} \tag{4.2}$$

and where the sum is taken over all sets I of size k of distinct integers from $(1,\ldots,n)$. Now, on the other hand, it is interesting to note that for the variance inflation model (2.12), the probability that k observations indexed by I are generated with 'noise' ϵ given by $N(O,\delta^2\sigma^2)$, $\delta^2 > 1$, and $(n-k)$ generated by $N(O,\sigma^2)$ takes the form, as proved in Box and Tiao (1968),

$$w_I = C\left(\frac{\alpha}{1-\alpha}\right)^k \delta^{-k}\left(\frac{|X'X|}{|X'X - \phi X_I'X_I|}\right)^{1/2}\left(\frac{s^2}{\hat{s}^2_{(I)}}\right)^{\frac{n-p}{2}}, \tag{4.3}$$

where C is a normalizing constant that can be shown to be the probability of no outliers, and $\phi = 1 - \delta^{-2}$. (For a precise definition of $\hat{s}^2_{(I)}$ see Box and Tiao (1968).) When δ is large, it can be shown (Peña and Tiao (1992)) that w_I is approximately

$$w_I = C\left(\frac{\alpha}{1-\alpha}\right)^k \delta^{-K}\left(\frac{|X'X|}{|X_{(I)}'X_{(I)}|}\right)^{1/2}\left(\frac{s^2}{s^2_{(I)}}\right)^{\left(\frac{n-p}{2}\right)}. \tag{4.3a}$$

Adding up the values w_I for all sets of size k we obtain the probability of exactly k outliers in the sample, and, in turn, by adding all the w_I's up, the constant C could be obtained.

For fixed k and δ large, the conditional probability that a particular set of k observations indexed by I are spuriously generated with noise variances $\delta^2\sigma^2$ is (Peña and Tiao (1992)):

$$p_I = C'\left(\frac{|X'X|}{|X_{(I)}'X_{(I)}|}\right)^{1/2}\left(\frac{s^2}{s^2_{(I)}}\right)^{\frac{n-p}{2}}, \tag{4.4}$$

which in turn can be written as

$$p_I = C'' \cdot |I - H_I|^{-1/2}(s^2_{(I)})^{-(n-p)/2}, \tag{4.5}$$

and hence, for large n and small k, both probabilities c_I and p_I are essentially the same.

As indicated in section 3 the predictive density has been advocated as a diagnostic tool. Interestingly, there is a strong connection between both c_I (or p_I) with the predictive density (3.2). Writing c_I from (4.1) as

$$c_I = K'|I - H_I|^{-1/2} \left\{ \frac{s^2}{s^2_{(I)}} \right\}^{(n-p-k)/2} \tag{4.6}$$

and using (3.6) and (3.8) we then have

$$c_I = K''|I - H_I|^{-1/2}[1 + Q_I]^{(n-p-k)/2} \ . \tag{4.7}$$

Hence, from (3.10), we see that

$$c_I = K''' \cdot [p(y_I/y_{(I)})]^{-1} \ , \tag{4.8}$$

that is to say that the posterior probability c_I is inversely proportional to the ordinate of a predictive density function with is related to the general k-variate student-t distribution with $n - k - p$ degrees of freedom. The smaller the predictive ordinate, which occurs for large residuals in absolute value, the large the probability c_I that the corresponding observations are spuriously generated.

5. TWO ILLUSTRATIVE EXAMPLES.

We first illustrate the behaviour of the predictive and posterior probabilities for the residuals with a simulated example. We have generated 20 observations using the model $y = 1 + x + \varepsilon$, where ε is $N(O, 1)$. The values for y, x are given in Table 5.1. We have included a potential influential point by locating x_{20} at 40. Then, we introduced an outlier by adding 4 to the original y_{11}, and proceeding by then computing for this new set of data the probabilities for each observation to be an outlier using the posterior probabilities

P_i given by (3.15) with $q = 2$, and the c_i as given by (4.1), which is, as demonstrated in Section 4, inversely proportional to the predictive ordinate. As anticipated, the values for P_i are all very small, except for a large value of .9552 at $i = 11$, which is 2.2 times greater than the next largest value that occurs at $i = 14$. As far as the c_i, the largest occurs at $i = 11$ with a value of .9708 which is 511 times greater than the next largest value, that occurs at $i = 20$. We note that the outlier will be idenitified by both procedures, although in a more powerful way by c_i then by P_i.

The next experiment we carried out was to introduce to the original set of data of Table 5.1 an outlier at the high leverage point $x_{20} = 40$, by again adding 4 to the original

Table 5.1.

Data for the simulated example

	x	y	h
1	1	1.955	.13
2	2	2.201	.11
3	3	3.235	.10
4	4	5.862	.09
5	5	5.944	.08
6	6	7.514	.07
7	7	8.397	.06
8	8	9.756	.06
9	9	10.401	.05
10	10	9.659	.05
11	11	12.375	.05
12	12	14.125	.05
13	13	14.729	.05
14	14	12.622	.05
15	15	15.726	.06
16	16	16.677	.06
17	17	18.318	.07
18	18	18.489	.08
19	19	19.998	.09
20	40	42.607	.62

observed data point. Here, however, the largest P_i value is .6758 and occurs at $i = 14$, instead of the expected $i = 20$. Indeed the next largest of the P_i's is P_{20}, with the value

.3761. This is in contrast to the behaviour of the c_i: the largest value is $c_{20} = .9393$ with the next largest $c_{14} = .0326$, that is $c_{20} = 28.8c_{14}$.

These results are in agreement with the theoretical results of Section 3, in which we have shown that the behaviour of P_i could be unsatisfactory for high leverage points.

As a second example we have chosen a set of data in which there are not pronounced differences among the leverages. For this purpose, we will use the consumption/income data, reported by Zellner and Moulton (1985). They compared the linear, log and logistic transformation for data obtained from 26 countries. The model we will use as our example is described as follows. Let, for the ith country, z_i be the permanent consumption expenditures and u_i be the permanent disposable income, both in a per capita basis and define

$$x_i = \log u_i \quad \text{and} \quad y_i = \log \frac{z_i/u_i}{1 - z_i/u_i} \tag{5.1}$$

The logit model is, in terms of x_i and y_i, is estimated by

$$\hat{y}_i = 3.828 - .215x_i \; ; \; s^2 = .2236 \; . \tag{5.2}$$

Let us compare the behaviour of P_i and c_i for the logit model applied to this set of data. In order to have a prior probability that an observation is not an outlier equal to .997, the value of q should be 3. Now for $q = 3$, the maximum P_i is $P_{14} = .1634$, followed by $P_{18} = .0020$. We conclude that the use of the P_i's does not focus attention on any one observation.

Table 5.2 provides the probabilities of having exactly k outliers in the sample using the scale contaminated linear model (4.3a), for $\delta = 5$ and the prior probability of no outliers equalling .95, as before. It can be seen that the most likely event is the presence of one outlier. To identify which point has been spuriously generated, we turn to the conditional probabilities c_i as given by (4.1). The largest value is attained at $i = 14$, and is $c_{14} = .5665$, followed in magnitude by $c_{18} = .0813$, a factor of 7, approximately. The eighteenth observation corresponds to Japan, and, as it happens, it was the only Asian

country involved in this set of data. We should note that the data is from 1959 and 1960 UN data.

Table 5.2.

Probability of k outliers for the logistic model with Zellner-Moulton data.

k	0	1	2	3	4
$p(k)$	.3566	.3732	.1961	.0611	.0130

ACKNOWLEDGEMENTS

The authors would like to acknowledge support, with thanks, for this research. Daniel Peña was supported by DGICYT (Spain), under Grant No. PB90-0266. Irwin Guttman was supported by NSERC (Canada) through Grant No. A8743.

APPENDIX I. The non-central T-distribution.

We remind the reader that the definition of the classical non-central T-variable comes about as follows. Suppose Z and W are independent random variables with distributions given by

$$Z \sim N(O,1) \quad \text{and} \quad W \sim \chi^2_v \tag{AI.1}$$

Then the non-central T-variable with non-central parameter Δ, degrees of freedom v, is defined as

$$T = \frac{(Z+\Delta)}{\sqrt{W/v}} \tag{AI.2}$$

It is very easy to see that the density of the random variable T as defined in (AI.2) may be written as

$$\mathfrak{p}(t) = \int_0^\infty \sqrt{\frac{w}{v}} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\sqrt{\frac{w}{v}}t - \Delta\right]^2 h_v(w)dw \right. \tag{AI.3}$$

where $h_v(w)$ is as given in (3.23). Now consider $P(T \leq t_0)$ - we have

$$P(T \leq t_0) = \int_{-\infty}^{t_0} \mathfrak{p}(t)dt \qquad (AI.4)$$

so that, inverting the order of integration, we easily find (dropping the subscript zero on the particular value of T given by t_0 that we were concerned with) that

$$P(T \leq t) = \int_0^{\infty} \Phi\left[\sqrt{\frac{w}{v}}t - \Delta\right] h_v(w)dw \; . \qquad (AI.5)$$

Hence, the right hand side of (3.22) and the integral of the second line of (3.26) are easily related to (AI.5), as indicated in (3.24) and (3.26a) respectively.

Now in general, $P(T \leq t)$, where T is non-central with ν, degrees of freedom and non-central parameter Δ, has, for moderate to large ν, a well known approximation, which we used in Section 2. This approximation, as we will see, rests on the well known result that for moderate or large ν, that $\sqrt{W} = \sqrt{\chi^2_v}$ has the approximate distribution given by ($\dot{\sim}$ denotes "approximately distributed as")

$$\sqrt{W} \dot{\sim} N\left(\sqrt{v}, \frac{1}{2}\right) \qquad (AI.6)$$

This result is easily obtained from the results of Fisher as quoted in Kendall and Stuart (1977, pg 400). Now we have that

$$P\Big(T \leq t) = P(Z + \Delta \leq \frac{t}{\sqrt{v}}\sqrt{W}\Big) \qquad (AI.7)$$

so that

$$P\Big(T \leq t) = P(Z - \frac{t}{\sqrt{v}}\sqrt{W} \leq -\Delta\Big) \; . \qquad (AI.8)$$

But for large v, using (AI.6), we have that

$$U = Z - \frac{t}{\sqrt{v}}\sqrt{W} \sim N\big(-t, 1 + \frac{t^2}{2v}\big) \; . \qquad (AI.9)$$

Hence we have that

$$P(T \leq t) \doteq P(U \leq -\Delta)$$

or

$$P(T \leq t) \doteq P\left(Z \leq \frac{t-\Delta}{\sqrt{1+\frac{t^2}{2v}}}\right) = \Phi(u) \qquad (AI.10)$$

where $u = \dfrac{(t-\Delta)}{\sqrt{1+t^2/2v}}$, and, $Z \sim N(O,1)$. Hence, the density function of T is , differentiating, such that

$$\mathfrak{p}(t) \dot{\sim} \phi(u) \left|\frac{du}{dt}\right| , \qquad (AI.11)$$

where of course $\phi(u)$ is the density of a $N(O,1)$ random variable, and, to repeat.

$$u = \frac{(t-\Delta)}{\sqrt{1+\frac{t^2}{2v}}} . \qquad (AI.12)$$

We have that the Jacobian of the transformation from t to u given by (AI.12) has absolute value

$$|J| = \frac{1}{\left|\frac{du}{dt}\right|}$$

so that the density of U is

$$g(u) \doteq \phi(u) \left|\frac{du}{dt}\right| \times \frac{1}{\left|\frac{du}{dt}\right|} \qquad (AI.13)$$

that is, the density of U is

$$g(u) = \phi(u) ,$$

and we have that, approximately, U has the density of a standard normal $N(O.1)$ variable. for large v , and the theorem is proved.

REFERENCES

Atkinson, A.C. (1975). *Plots, Transformations and Regression.* Oxford: Clarendon Press.

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

Box, G.E.P. (1980). "Sampling and Bayes Inference in Scientific Modelling and Robustness." *Journal of the Royal Statistical Society*, Series A 143, 383-430 (with discussion).

Box, G.E.P., and Meyer, R. (1986). "An Analysis for Unreplicated Fractional factorials." *Technometrics* 28, 11-18.

Box, G.E.P., and Tiao, G.C. (1968). "A Bayesian Approach to Some Outlier Problems." *Biometrika* 55, 119-129.

Casella, G. and George, E.I. (1992). "An Introduction to Gibbs Sampling." *American Statistician.* (To Appear)

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

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

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

Freeman, P.R. (1980). "On the number of outliers in data from a linear model." *Bayesian Statistics*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Editors. Valencia University Press, 349-365.

Geisser, S. (1980). "Discussion of a paper by G.E.P. Box", *Journal of the Royal Statistical Society.* Series A 143, 416-7.

Geisser, S. (1987). "Influential observations, diagnostics and discordancy tests". *Journal of Applied Statistics* 14, 133-42.

Geisser, S. (1988). "Predictive approaches to discordancy testing". *Bayesian and Likelihood Influence: Essays in Honor of George Barnard*, S. Geisser, J.S. Hodges, S.J. Press and A. Zellner, Editors. Amsterdam: North Holland.

Gelfand, A.E., and Smith, A.F.M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association* 85, 398-409.

Guttman, I. (1973). "Care and Handling of of Univariate or Multivariate Outliers in Detecting Spuriosity - a Bayesian Approach". *Technometrics* 15, 723-738.

Guttman, I., Dutter, R., and Freeman, P.R. (1978). "Care and Handling of Univariate Outliers in the General Linear Model to Detect Spuriousity - 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*, J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith, Editors. Oxford University Press, 631-640.

Guttman, I. and Peña, D. (1992). *A Bayesian Look at Diagnostics in the Univariate Linear Model.* Technical Report, Department of Statistics and Econometrics, Universidad Carlos III de Madrid, Getafe, Spain.

Hampel, F.R. et al. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley, New York.

Huber, P. (1981). *Robust Statistics.* John Wiley, New York.

Johnston, W., and Geisser, S. (1983). "A Predictive View of the Detection and Characterization of Influential Observations in Regression Analysis." *Journal of the American Statistical Association* 78, 137-144.

Johnston, W., and Geisser, S. (1985). "Estimative Influence Measures of the Multivariate General Linear Model." *Journal of Statistical Planning and Inference* 11, 33-56.

Juan, J. and Peña, D. (1992). "A Simple Method to Identify Significant Effects in Unreplicated 2-level Factorial Designs". *Communications in Statistics, Theory and Methods* 21, 1383-1403.

Kendall, M., and Stuart, A. (1977). *The Advanced Theory of Statistics; Volume 1 - Distribution Theory* (Fourth Edition): Macmillan Publishing Company, New York.

Ljung, G.M. (1989). "A note on estimation of Missing Values in Time Series". *Communications in Statistics, Simulation and Computation* 18, 459-465.

Peña, D. and Maravall, A. (1991). "Interpolation, Outliers and Inverse Autocorrelations." *Communications in Statistics, Theory and Methods* 20, 3175-3186.

Peña, D. and Tiao, G.C. (1992). "Bayesian Robustness Functions for Linear Models." *Bayesian Statistics* 4., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith (Eds). Oxford University Press, 365-388.

Pettit, L.I. and Smith, A.F.M. (1985). "Outliers and Influential Observations in Linear Models." *Bayesian Statistics* 2, J.M. Bernardo, M.H. DeGroot, D.V. Lindley, A.F.M. Smith (eds). Elsevier Science Publishers, 473-494.

Tussell, F. (1990). "Testing for Interaction in Two-Way ANOVA tables with no replication." *Computational Statistics and Data Analysis* 10, 29-45.