Statistics & Probability Letters 11 (1991) 321–325
North-Holland

April 1991

The detection of observations possibly influential for model selection

Philip Hans Franses

Econometric Institute and Tinbergen Institute, Erasmus University Rotterdam, P.O. Box 1738, 3000 DR Rotterdam, Netherlands

Received September 1989
Revised March 1990

Abstract: Model selection can involve several variables and selection criteria. A simple method to detect observations possibly influential for model selection is proposed. The potentials of this method are illustrated with three examples, each of which is taken from related studies.

Keywords: Model selection criteria, influential observations.

Introduction

Observations influential for regression results have been extensively studied in Cook and Weisberg (1982) and Belsley, Kuh and Welsch (1980). Typically such observations influence the sums of squared residuals, and because several model selection criteria are functions of these sums, the choice between linear regression models may depend on one or more data points.

The detection of choice influencing observations is studied in e.g. Weisberg (1981) and Chatterjee and Hadi (1988). In Weisberg (1981), a model is selected with Mallows' C_p statistic. By partitioning this statistic into individual components, one can assess the influence of observations on subset model selection. In Chatterjee and Hadi (1988, Chapter 6) the impact of simultaneously omitting one variable and one observation on a model is considered. The simple method proposed in the present paper tries to extend these approaches because model selection can involve more than one variable, and also more than one selection criterion. The latter issue is of particular interest for it gives an opportunity to compare model selection criteria (see also Jungeilges, 1989). Furthermore, the method results in a single scatterplot for inference, assuming there is an intended model selection of interest.

The outline of this paper is as follows. In Section 1, first, some notation is given with respect to model selection for the nested case and, second, a simple method is discussed to detect possibly influential observations. In Section 2, the proposed method is illustrated by means of three examples, each of which is taken from one of the above studies. The final section contains some concluding remarks.

1. The detection of influential observations

Consider the nested hypotheses

$$H_1: \ y = X_1\beta_1 + X_2\beta_2 + \varepsilon_1, \tag{1.1}$$

$$H_2: \ y = X_1\beta_1 + \varepsilon_2, \tag{1.2}$$

where y, ε_1, ε_2 are $(n \times 1)$-vectors, X_i are $(n \times k_i)$-matrices and β_i are $(k_i \times 1)$-vectors, for $i = 1, 2$. It is assumed that the disturbances are i.i.d. distributed with zero mean and variance σ^2. The choice between H_1 and H_2 can be made using

Table 1
Selection criteria and their respective q values

Selection criterion	q value
Schwarz (1978)	$n^{k_2/n}$
F-test [a]	$1+\frac{c_\alpha k_2}{n-k_1-k_2}$
Rice (1984)	$1+\frac{2k_2}{n-2k_1-2k_2}$
Craven and Wahba (1979)	$1+\frac{2k_2}{n-k_1-k_2}+(\frac{k_2}{n-k_1-k_2})^2$
Hocking's SP (1976)	$1+\frac{k_2(2n-2k_1-2k_2-1)}{(n-k_1-k_2)(n-k_1-k_2-1)}$
Amemiya's PC (1980)	$1+\frac{2nk_2}{(n-k_1-k_2)(n+k_1)}$
Akaike's IC (1974)	$e^{2k_2/n}$
Shibata (1981)	$1+\frac{2k_2}{n+2k_1}$
adj. R^2	$1+\frac{k_2}{n-k_1-k_2}$

[a] c_α denotes here the 5% critical value of the $F(k_2, n-k_1-k_2)$-distribution.

model selection criteria belonging to the class of criteria which are functions of the estimated residual sums of squares. A typical form of these criteria is to choose the alternative hypothesis H_1 if

$$\hat{\varepsilon}_2'\hat{\varepsilon}_2 > q\cdot\hat{\varepsilon}_1'\hat{\varepsilon}_1 \tag{1.3}$$

where q is a function of n, k_1 and k_2 (see Engle and Brown, 1986; Franses, 1989), and $q>1$. A summary of q values for several well-known criteria is presented in Table 1. It should be noted that the C_p statistic is equivalent to the adjusted R^2, and is therefore not mentioned in the table. From (1.3) it is easily seen that a larger q reduces the probability of choosing H_1.

With respect to the influence of a single observation two distinct situations can be recognized. The first is that H_1 is chosen with n observations and H_2 with $n-1$ observations (case 1). The second case is that after deletion of one observation H_1 is preferred, while H_2 is chosen on the whole data set (case 2). Denote $z_{(i)}$ as the scalar which is computed without using the ith row of the data set, and z_i as the ith element of a vector. Furthermore, denote h_{1ii} as the (i,i)th element of the matrix $X(X'X)^{-1}X'$, where $X=[X_1 : X_2]$, and h_{2ii} as the corresponding element of $X_1(X_1'X_1)^{-1}X_1'$. Case 1 occurs for observation i when

$$\hat{\varepsilon}_2'\hat{\varepsilon}_2 > q\cdot\hat{\varepsilon}_1'\hat{\varepsilon}_1$$
$$\text{and} \tag{1.4}$$
$$[\hat{\varepsilon}_2'\hat{\varepsilon}_2]_{(i)} < q_{(i)}\cdot[\hat{\varepsilon}_1'\hat{\varepsilon}_1]_{(i)}.$$

Since $q_{(i)}$ and q exceed 1, and $[\hat{\varepsilon}_j'\hat{\varepsilon}_j]_{(i)}=\hat{\varepsilon}_j'\hat{\varepsilon}_j-(\hat{\varepsilon}_{ji}/(1-h_{jii}))$ for $j=1,2$, (1.4) can be transformed into

$$\hat{\varepsilon}_2'\hat{\varepsilon}_2 > q\cdot\hat{\varepsilon}_1'\hat{\varepsilon}_1 > (q/q_{(i)})\cdot\hat{\varepsilon}_2'\hat{\varepsilon}_2 + q\cdot f_{1i} - (q/q_{(i)})\cdot f_{2i} \tag{1.5}$$

where $f_{1i}=(\hat{\varepsilon}_{1i}^2/(1-h_{1ii}))$ and $f_{2i}=(\hat{\varepsilon}_{2i}^2/(1-h_{2ii}))$. This inequality can not hold when

$$f_{2i} < \left(1-(q_{(i)}/q)\right)\cdot\hat{\varepsilon}_2'\hat{\varepsilon}_2 + q_{(i)}\cdot f_{1i}. \tag{1.6}$$

An analogous result applies to case 2 which occurs when

$$\hat{\varepsilon}_2'\hat{\varepsilon}_2 < q\cdot\hat{\varepsilon}_1'\hat{\varepsilon}_1$$
$$\text{and} \tag{1.7}$$
$$[\hat{\varepsilon}_2'\hat{\varepsilon}_2]_{(i)} > q_{(i)}\cdot[\hat{\varepsilon}_1'\hat{\varepsilon}_1]_{(i)}.$$

Rewriting this as above gives that this is not possible when

$$f_{2i} > \left(1-(q_{(i)}/q)\right)\cdot\hat{\varepsilon}_2'\hat{\varepsilon}_2 + q_{(i)}\cdot f_{1i}. \tag{1.8}$$

The line L, with

$$L=\left\{(f_1, f_2)\in\mathbb{R}\times\mathbb{R} \mid f_2=\left(1-(q_{(i)}/q)\right)\cdot\hat{\varepsilon}_2'\hat{\varepsilon}_2 + q_{(i)}\cdot f_1\right\}.$$

can be drawn in the (f_1, f_2)-space. For case 1, all points which lie under this L are certainly not influential for model choice. And all combinations (f_{1i}, f_{2i}) lying above L deserve our attention for they may be influential. For case 2 the reverse argument will hold. Anyway, for both cases it will be true that observations with the largest distance from L will have the largest probability of being influential. Furthermore, clusters of such observations may be recognized as will be seen in forthcoming examples, although the impact of the

masking phenomenon car blur correct recognition. The method needs only two regressions, the computation of the q and $q_{(l)}$ values and the points in the (f_1, f_2)-space, and these are all straightforward calculations. The final scatterplot may then contain the lines L belonging to the extreme q valued criteria.

2. Examples

Three examples for the proposed method to detect possibly influential data points will be given in this section. Extensive computational results are omitted for the sake of readability, but they can be obtained from the author on request. The first example is taken from Belsley et al. (1980) and deals with an intercountry life-cycle savings function. The model relates the aggregate personal savings rate in country i (SR_i) to the percentages of the population under 15 and over 75 ($POP15_i$ and $POP75_i$), to the level of real per-capita disposable income (DPI_i) and to the percentage growth rate of DPI_i (ΔDPI_i), all averaged over the period 1960–1970. The observations for 50 countries are given in Belsley et al. (1980, p. 41). For model H_1, where SR is explained by a constant and the four explanatory variables, the estimated parameters for POP75 and DPI appear to be insignificant, and hence the deletion of these variables results in model H_2. The choice between the models is made with criteria as in (1.3). The q and $q_{(l)}$ values of these criteria for this case, where $k_1 = 3$, $k_2 = 2$, can be calculated from Table 1. It can be inferred that the Schwarz criterion and the adjusted R^2 obtain the highest and lowest values respectively. Comparing the residual sums of squares of both models it is clear that model H_2 is preferred with all nine criteria (see Table 2, first row).

To investigate if there are any observations influencing this choice, 50 pairs of (f_{1l}, f_{2l})-values have been calculated. In Figure 1 these are plotted and also the line L is drawn for the two extreme valued criteria.

It appears that observation number 24 might be important, and as demonstrated in Table 2 it is indeed. Furthermore, it is worth mentioning that

Table 2
Model selection results when observations are deleted

Example [a]	n	k_1	k_2	Observation(s) deleted	$\tilde{q}$ [b]	Model choice [c] (# of criteria)
Belsley, Kuh and Welsch (1980) [d]	50	3	2	none	1.0214	H_2 (all)
				7	1.0323	H_2 (all)
				46	1.0148	H_2 (all)
				24	1.1010	H_2 (2)
						H_1 (7)
Weisberg (1981) [e]	20	4	2	none	1.2634	H_2 (4)
						H_1 (5)
				1, 20	2.5362	H_1 (all)
				3, 17	1.0063	H_2 (all)
Chatterjee and Hadi (1988) [f]	40	4	2	none	2.1326	H_1 (all)
				2, 15, 31, 40	1.8721	H_1 (all)
				2, 7, 8, 15, 16, 31, 40	1.1414	H_2 (5)
						H_1 (4)

[a] The examples and the model selection of interest are more extensively discussed in the text.
[b] For notational convenience, the $\tilde{q}$ is defined as $\hat{\varepsilon}_2'\hat{\varepsilon}_2/\hat{\varepsilon}_1'\hat{\varepsilon}_1$.
[c] Model H_1 is chosen when $\tilde{q} > q$. The q values for the several criteria can be found using Table 1.
[d] For $n = 50(49)$, $k_1 = 3$ and $k_2 = 2$, the q values for Schwarz to adj.R^2 range from 1.1694(1.1722) to 1.0444(1.0455).
[e] For $n = 20(18)$, $k_1 = 4$ and $k_2 = 2$, the q values for F-test to Shibata range from 1.5343(1.6467) to 1.1429(1.1538), although the highest q value for $n = 18$ is 1.6667 and is obtained for the Rice criterion.
[f] For $n = 40(33)$, $k_1 = 4$ and $k_2 = 2$, the q values for Schwarz to adj.R^2 range from 1.2026(1.2360) to 1.0588(1.0741), although the highest q value for $n = 33$ is 1.2489 and is obtained for the F-test.

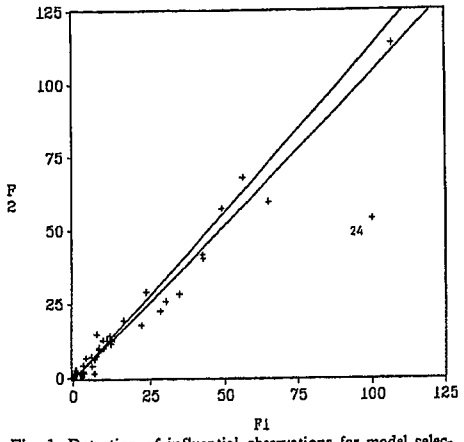

Fig. 1. Detection of influential observations for model selection. Example from Belsley, Kuh and Welsch (1980).

the observations 7 and 46, both having large estimated residuals for model H_1 (and also for H_2), and which have been found to be influential from the regression diagnostics in Belsley et al. (1980), do not seem to be decisive in model selection (see Table 2). The impact of the various criteria in combination with single observations can also be depicted, because with the Schwarz-criterion and the F-test model H_2 is still preferred when observation 24 is deleted, while the other criteria indicate another choice.

The second example is taken from Weisberg (1981, Section 4), where a certain variable Y might be explained by a constant and explanatory variables X_1 through X_5. Suppose that the model selection of interest is whether this model can be simplified by deleting X_1 and X_3. The result for all 20 observations, which are numbered 1 through 20 here for convenience, is displayed in Table 2. It can be concluded that with e.g. the F-test model H_2 is preferred, while with e.g. the AIC (and, of course, the C_p (see Weisberg, 1981) model H_1 is chosen. To investigate whether some observations determine these model choices, consider Figure 2.

Now, possibly influential observations may lie at both sides of the lines. The observations (1, 20) and (3, 17) can be important for they may stimulate a preference for model H_2 and H_1, respectively. From Table 2 one can conclude that these suggestions are confirmed. Moreover, subset model H_2 would also be chosen with the C_p statistic in case observations (3, 17) are deleted, and hence our method yields an additional insight to those already presented in Weisberg (1981).

In Chatterjee and Hadi (1988, p. 236) the data for the last example can be found. Again, an endogenous variable Y is explained by 5 exogenous variables $X_1, \ldots, X_5$ and a constant, and now model selection considers the simultaneous deletion of X_2 and X_3. Using all 40 observations, the model with all explanatory variables is chosen with all criteria (see Table 2). In Chatterjee and Hadi (1988, p. 244) it is concluded that observations 2, 15, 31 and 40 are influential with respect to the omission of one variable. Deleting these observations simultaneously, however, yields for our case that model selection is not altered. So, although they individually are influential for one variable, they jointly do not have an effect on the omission of more than one variable.

The scatterplot in Figure 3 may confirm this finding by indicating that there are possibly some more influential observations. Deleting data points 2, 7, 8, 15, 16, 31, and 40 provides that with e.g. the F-test the simplified model is chosen, although deletion of some subsets of these 7 observations

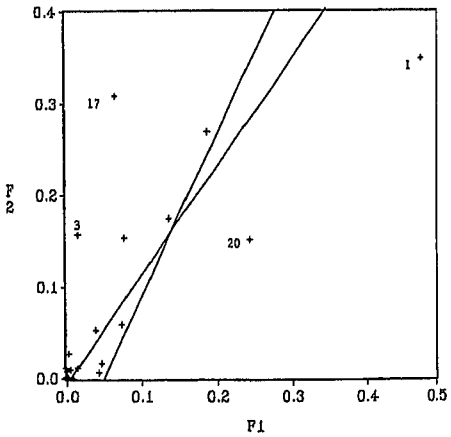

Fig. 2. Detection of influential observations for model selection. Example from Weisberg (1980).

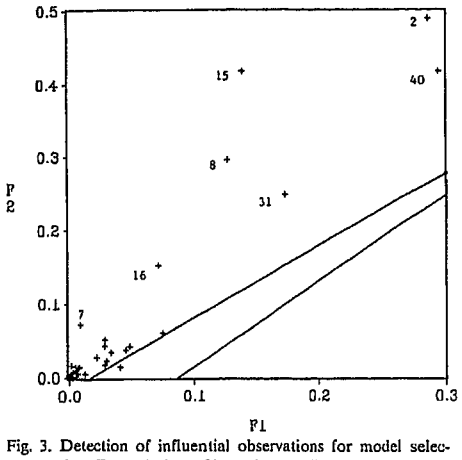

Fig. 3. Detection of influential observations for model selection. Example from Chatterjee and Hadi (1988).

yields comparable results. Moreover, from Figure 3 one can observe that most points lie above the lines, which is an indication of the fragility of the intended model selection.

3. Concluding remarks

The method to detect observations possibly influential for model choice, as is developed and illustrated in this paper, seems to meet its purpose. In principle, it seems possible to detect the impact of the simultaneous omission of more than one observation and more than one variable on a regression model. Together with the aspect that model selection can be done with several criteria, the proposed method seems to extend the approaches developed in Weisberg (1981) and Chatterjee and Hadi (1988). Application of our method to examples presented in these studies seems to confirm its potentials.

Finally, some inference can be made with respect to the robustness and preferable use of model selection criteria (see also Jungeilges, 1989, for a related study). From Example 1 it can be seen that in case H_2 is chosen with all observations, the highest q valued criterion is most robust in the sense that model choice is the least affected by influential observations. Analogously, this applies to the choice of H_1 and the lowest q valued criterion, as can be seen from Example 3. Intuitively, one would be more confident in model selection results when a certain model is chosen with all criteria (see also Franses, 1989), and when there are no observations influential enough to change this. So, in case k possibly influential observations can be depicted, then it seems preferable to choose the same model on the basis of n and $n-k$ observations with all criteria.

Acknowledgement

The suggestions from an anonymous referee are gratefully acknowledged.

References

Akaike, H. (1974), A new look at the statistical model identification, *IEEE Trans. Automat. Control* **AC-19**, 716–723.

Amemiya, T. (1980), Selection of regressors, *Internat. Econom. Rev.* **21**, 331–354.

Belsley, D.A., E. Kuh and R.E. Welsch (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity* (Wiley, New York).

Chatterjee, S. and A. Hadi (1988), *Sensitivity Analysis in Linear Regression* (Wiley, New York).

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

Craven, P. and G. Wahba (1979), Smoothing noisy data with spline functions, *Numer. Math.* **31**, 377–403.

Engle, R.F. and S.J. Brown (1986), Model selection for forecasting, *Appl. Math. Comput.* **20**, 313–327.

Franses, P.H. (1989), The distance between regression models and its impact on model selection, *Appl. Math. Comput.* **34**, 1–16.

Hocking, R.R. (1976), The analysis and selection of variables in linear regression, *Biometrics* **32**, 1–49.

Jungeilges, J. (1989), On the robustness of a class of model selection rules designed for simplification searches, Paper presented at the ESEM89, Munich.

Rice, J. (1984), Bandwidth choice for nonparametric regression, *Ann. Statist.* **12**, 1215–1230.

Schwarz, G. (1978), Estimating the dimension of a model, *Ann. Statist.* **6**, 461–464.

Shibata, R. (1981), An optimal selection of regression variables, *Biometrika* **68**, 45–54.

Weisberg, S. (1981), A statistic for allocating C_p to individual cases, *Technometrics* **23**, 27–31.