Economics Letters 38 (1992) 133–137
North-Holland

Model adequacy and influential observations

Philip Hans Franses
Econometric Institute, Erasmus University Rotterdam, Rotterdam, Netherlands

Guido Biessen
Department of Microeconomics, University of Amsterdam, Amsterdam, Netherlands

Received 25 September 1991
Accepted 18 November 1991

A common characteristic of diagnostic measures on influential observations is the assumption that all relevant regressors are included in the model, and that none of them can be deleted. We review and illustrate a method to detect data points which are influential enough to establish the empirical (in)significance of regressors.

1. Introduction

It has become increasingly regular to report linear regression estimation results together with diagnostics on influential observations and outliers, such as those proposed in, e.g., Belsley, Kuh and Welsch (1980), Cook and Weisberg (1982) and others. A common characteristic of these measures is that the correctness of the estimated model is assumed. This usually means that it is supposed that all relevant regressors are included in the model, and that none of them can be deleted. One can however imagine that there may be observations which are influential to such an extent that the empirical significance of a certain regressor is determined by one data point only. Hence, although the inclusion of a regressor may have been prescribed by some economic theory, its deletion from the model is only prevented by a single influential observation. Given that such data points can affect one's confidence in the adequacy of the theoretical model, the empirical model and/or the measurement of the variables, it seems important to check for their absence.

Procedures for the detection of influential observations, in case the model may be simplified by the deletion of regressors, are considered in, e.g., Chatterjee and Hadi (1988) and Weisberg (1981). Recently, a method was proposed that extends these two procedures by considering more than one redundant regressor and more than one model selection criterion, see Franses (1991). In section 2, this approach will be reviewed briefly. In section 3 it is applied to two economic examples. Section 4 concludes.

Correspondence to: Ph.H. Franses, Econometric Institute, Erasmus University Rotterdam, P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands

2. Detecting influential observations for regressor redundancy

Suppose there are two models M_1 and M_2, involving $k_1 + k_2$ and k_2 regressors, such that M_2 is nested in M_1. Usually, the choice between M_1 and M_2 is made using criteria belonging to the class of criteria that are functions of the estimated residual sums of squares (RSS). A typical form of these criteria is to choose M_1 if $RSS_2 > q \cdot RSS_1$, where q is a function of n, k_1 and k_2. A summary of q values for several criteria is given in Franses (1991). It is easy to recognize that a larger q reduces the probability of choosing M_1.

Two cases can be distinguished with respect to the influence of single observations on model selection: (1) M_1 is chosen with n observations and M_2 with $n-1$ observations, or (2) vice versa. Denote $z_{(i)}$ as the scalar that is computed without using the ith row of the observations matrix, z_i as the ith element of a vector, h_{1ii} as the (i,i)th element of the matrix $X(X'X)^{-1}X'$, where X contains the regressors of M_1, and h_{2ii} as the corresponding element for M_2. It is easy to see that case (1) applies for observation i when $RSS_2 > q \cdot RSS_1$ and $RSS_{2(i)} < q_{(i)} RSS_{1(1)}$. Since $q_{(i)}$ and q exceed 1, and $RSS_{j(i)} = RSS_j - \hat{\epsilon}_{ji}^2/(1-h_{jii})$, for $j = 1,2$, where the $\hat{\epsilon}_{ji}^2$ denotes the squared estimated ith residual for model j, these inequalities can be combined as

$$RSS_2 > qRSS_1 > (q/q_{(i)})RSS_2 + qf_{1i} - (q/q_{(i)})f_{2i}, \tag{1}$$

where $f_{ji} = \hat{\epsilon}_{ji}^2/(1-h_{jii})$. This can not hold when RSS_2 does not exceed the expression entirely on the right-hand side of (1), or

$$f_{2i} < \big(1 - (q_{(i)}/q)\big)RSS_2 + q_{(i)} f_{1i}. \tag{2}$$

A similar result can be obtained for case 2, where the f_{2i} should exceed the expression on the right-hand side of (2).

The line $L = \{(f_1/f_2) \in (\mathbb{R} \times \mathbb{R}) \mid f_2 = (1 - (q_{(i)}/q))RSS_2 + q_{(i)} f_{1i}\}$ can be drawn. For case 1, all points that lie under this line L, i.e. those that satisfy (2), are not influential for model choice. All (f_{1i}, f_{2i}) lying above L deserve attention for they may be influential. For these it will be true that observations with the largest distance to L will have the largest probability of being influential. For case 2 the reverse argument holds. Of course, one could just calculate these distances and skip the graphical part, but we feel that a graphical device is often insightful, and that it may prevent the calculation of too many regressions.

3. Illustrations

The examples to illustrate the procedure in the previous section are taken from Biessen (1991) and Reiss (1990). In both examples the linear regression model contains a constant and two

Table 1
Two model selection criteria and their q values.

Criterion	q [a]	$n = 27$	$n = 26$	$n = 25$	$n = 24$
a. F-test [b]	$1 + (ck_2)/(n - k_1 - k_2)$	1.178	1.186	1.196	1.206
b. $\bar{R}^2$	$1 + k_2/(n - k_1 - k_2)$	1.042	1.044	1.046	1.048

[a] See Franses (1989) for the derivations for these and other criteria. It is easy to show that the F-test and the $\bar{R}^2$ are those with the highest and the lowest q value given the current n, k_1 and k_2.

[b] The c denotes the 5% critical value of the $F(k_2, n - k_1 - k_2)$ distribution.

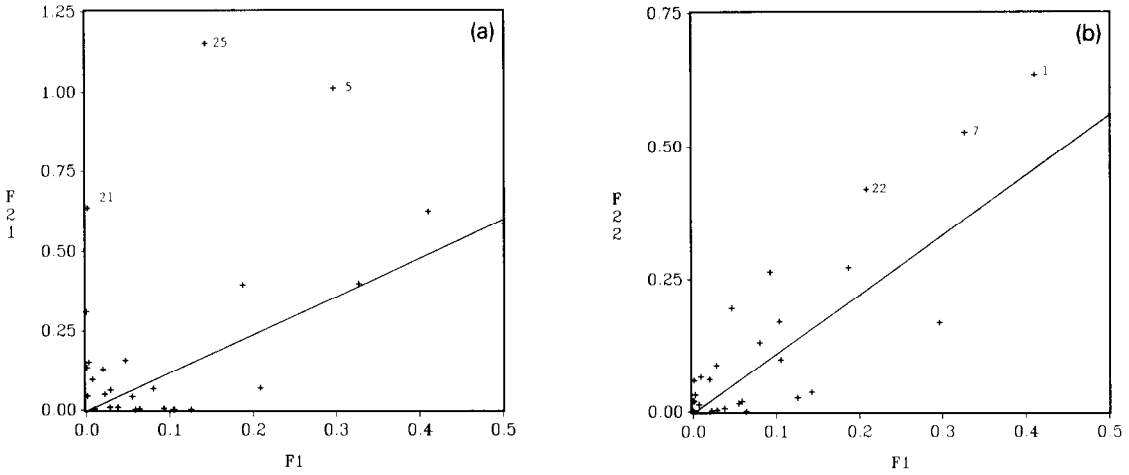

Fig. 1. Model selection and influential observations in example from Biessen (1991). F1, F21 and F22 correspond to models M_1, M_{21} and M_{22}, respectively. The straight line corresponds to the F-test.

regressors, i.e. k_1 is 2 and k_2 is 1, and are estimated with $n = 27$ observations. In table 1, the q values of the two most relevant criteria are displayed. Model choice is established by calculating $\tilde{q} = RSS_2/RSS_1$, and comparing it with the q in this table. When $\tilde{q}$ exceeds q, the model M_1 with all regressors is chosen. The restricted models are denoted by M_{21} and M_{22}, in which, except for the constant, only the first or the second regressor are included, respectively.

The model in Biessen (1991) considers a regression of the export ratio (ER) on a constant, income per capita (IC) and population (N), all measured in logs, for 27 countries. The estimation results are (with standard errors in parentheses)

$$ER = 3.123 + 0.335IC - 0.267N, \tag{3}$$
$$(0.983)\ (0.108)\quad (0.049)$$

and it is clear that all variables are highly significant. An investigation to the effect of single or several observations on this estimation result can be worthwhile given, e.g., the large differences between countries like the United States and the Soviet Union, which both have a large value for N, but substantially different IC values. The two (f_1, f_2) graphs are given in fig. 1. They suggest that for the selection between M_1 and M_{21} there are some influential observations, but that the selection between M_1 and M_{22} is not likely to be effected. From the results in table 2 it appears indeed that there are no significant effects of observations on either model choice. This leads to an increased confidence in the model proposed in Biessen (1991).

The model proposed in Griliches and Lichtenberg (1984) is taken as an illustration of detecting multiple outliers in Reiss (1990). It considers the regression of total factor productivity growth, $TFPG$, on a constant, and on private and federal expenditures on R&D, $PRIV$ and FED, for 27 industries. Estimation of the model yields

$$TFPG = -0.579 + 0.346PRIV + 0.010FED, \tag{4}$$
$$(0.295)\ (0.086)\qquad (0.233)$$

Table 2
Model selection results when some observations are deleted.

Example model selection		Observations deleted [a,b]	$\tilde{q}$ [c]	Model selected (with criterion)
Biessen (1991)	M_1 vs. M_{21}	none	2.306	M_1 (a,b)
		5, 21, 25	1.467	M_1 (a,b)
	M_1 vs. M_{22}	none	1.399	M_1 (a,b)
		1, 7, 22	1.278	M_1 (a,b)
Reiss (1990)	M_1 vs. M_{21}	none	1.008	M_{21} (a,b)
		2	1.135	M_1 (a), M_{21} (b)
		2, 18	1.202	M_1 (a,b)
	M_1 vs. M_{22}	none	1.683	M_1 (a,b)
		2, 18	1.160	M_1 (b), M_{22} (a)

[a] The numbers of the observations in the first example correspond to Ireland (5), Soviet Union (21), U.S. (25), Belgium (1), Netherlands (7) and Greece (22). Those of the second example are missiles (2) and computers (18).
[b] When some observations are considered jointly this means that they individually did not have any relevant effects.
[c] Model M_1 is selected when $\tilde{q}$ exceeds the value of q given in table 1.

and it appears that the *FED* variable is not significant. Given that the data include missiles and spacecraft industries (number 2), as well as the computer industry (18), both having high values on either the *FED* or the *PRIV*, it seems useful to consider the impact of single or several observations on the inclusion of these regressors. The graphs in fig. 2 show that the data points 2, 18, and possibly 15 (farms), can be influential. With respect to 2 and 18, this seems to be confirmed by the results displayed in table 2. The incorporation of observation 15 does not change this. Summarizing the results in table 2 for this example gives that the significance of *PRIV* and the insignificance of *FED* is established by only two observations. When these two are deleted, the estimation results show a significant parameter for *FED* and an insignificant one for *PRIV*. This

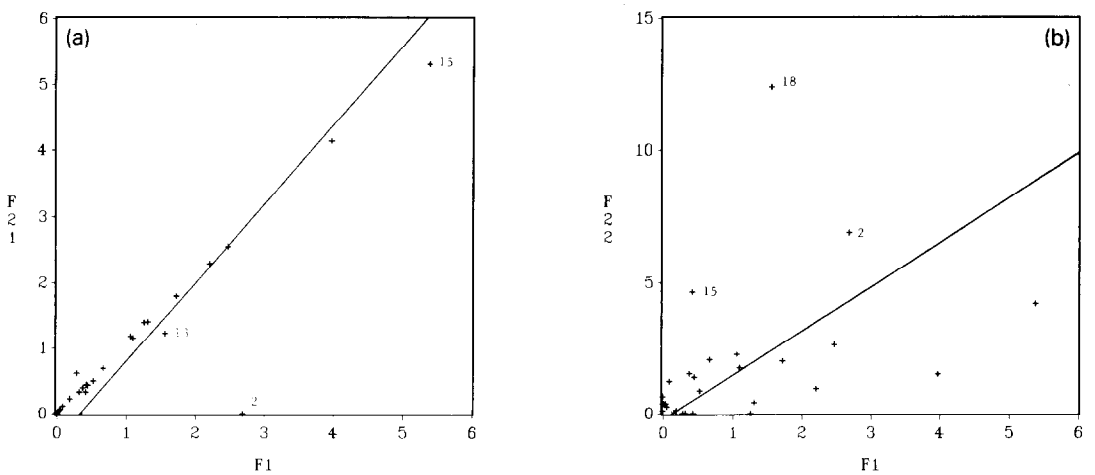

Fig. 2. Model selection and influential observations in example from Reiss (1990). F1, F21 and F22 correspond to models M_1, M_{21} and M_{22}, respectively. The straight line corresponds to the F-test.

curious outcome suggests that any conclusions drawn from the estimated model in (4) should be treated with great caution.

4. Concluding remarks

A simple method to detect influential observations for the adequacy of the inclusion of variables in a linear regression model is briefly reviewed and applied to two recent economic examples. It turns out that the model in Biessen (1991) does not seem to be affected by such data points, and hence an increased confidence in his model is gained. However, the estimation results of a model for R&D and productivity growth proposed in Griliches and Lichtenberg (1984) are entirely dominated by two such data points.

References

Belsley, D.A., Kuh, E. and R.E. Welsch, 1980, Regression diagnostics: Identification of influential data and sources of collinearity (Wiley, New York).

Biessen, G., 1990, Is the impact of central planning on the level of foreign trade really negative?, Journal of Comparative Economics 15, 22–44.

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).

Franses, P.H., 1989, The distance between regression models and its impact on model selection, Applied Mathematics and Computation 34, 1–16.

Franses, P.H., 1991, The detection of observations possibly influential for model selection, Statistics & Probability Letters 11, 321–325.

Griliches, Z. and F. Lichtenberg, 1984, R&D and productivity growth at the industry level: Is there still a relationship?, in: Z. Griliches, ed., R&D, patents, and productivity (Chicago University Press, Chicago).

Reiss, P.C., 1990, Detecting multiple outliers with an application to R&D productivity, Journal of Econometrics 43, 293–315.

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