KM 6(1982)
pag. 53-64

AN APPLICATION OF DIAGNOSTIC CHECKS TO A LINEAR REGRESSION MODEL

WILLEM FABER AND ARJEN E. RONNER

Econometric Institute
University of Groningen
The Netherlands

Key Words and Phrases: Regression diagnostics, influential points, partial leverage plots.

ABSTRACT

The use of regression diagnostics is recommended when larger models are used. We show that application of the criteria e.g. mentioned in Belsley e.a. (1980) clearly indicates that "there is something wrong" with the data in the same book. Specially the partial leverage plots give remarkable results.

1. INTRODUCTION

The robustness literature provides many solutions to the problem of detecting and handling outliers. In econometrics robust methods are of special interest, since models are always an approximation of the phenomenon under study. Moreover, "with the introduction of high-speed computers and the frequent use of large-scale models, ..., the researcher has become even more detached from intimite knowledge of this data", Belsley e.a. (1980). Recently much attention has been paid to the problem of irregular observations in the matrix of explaining variables; we refer to Krasker (1980), Krasker and Welsch (1979) and Maronna e.a. (1979). In Belsley e.a. (1980) several criteria are formulated in order to detect multivariate influential observations. The detection procedure often starts from single row effects. What is the influence of one single row of the X-matrix on e.g. the estimated parameter or on the predicted endogenous variable? Later on multiple row diagnostics will be introduced.

The value of regression diagnostics can nicely be illustrated by considering the savings equation given in Belsley e.a. (1980). It turns out that our estimates differ remarkably from the authors results. By using the criteria, derived in Belsley e.a. we are able to locate the source of the error.

Section 2 is devoted to the introduction of various diagnostic criteria. In Section 3 we present an example and the conclusions.

2. THE MODEL AND THE CRITERIA

We start from the linear regression model

$$y_i = x_i\beta + e_i \qquad i = 1,\ldots,n \tag{2.1}$$

where y_i is the i-th observation on the dependent variable, x_i is

a p-row vector of explanatory variables and $\beta \in \mathbb{R}^p$ is the vector of unknown regression coefficients. The disturbances $e_1,\ldots,e_n$ are assumed to be i.i. normally distributed with zero mean and variance σ^2. Let $\hat{\beta}$ be the LS estimator for β. In order to detect possible "influential points" we have used five criteria, which are mostly based upon the principle of deleting one of the observations. Belsley e.a. (1980).

The first criterion measures the influence of one observation, say (x_i, y_i), upon the estimate $\hat{\beta}$. Let $\hat{\beta}[i]$ be the LS estimator for β when (x_i,y_i) is deleted. In order to standardize the difference $\hat{\beta}-\hat{\beta}[i]$ we divide by its estimated variance. We then get

$$\text{DFBETAS}_{ij} := \frac{\hat{\beta}_j-\hat{\beta}_j[i]}{s[i]\sqrt{(X^TX)^{-1}_{jj}}} \tag{2.2}$$

where

$$s^2[i] := (n-p-1)^{-1}\sum_{k\neq i}(y_k-x_k\hat{\beta}[i])^2 \tag{2.3}$$

The second criterion particularly deals with outliers in the explanatory x_i variables. Let X be the $n \times p$ matrix of explanatory variables. We consider the diagonal elements h_i of the projection matrix $H := X[X^TX]^{-1}X^T$. The elements h_i indicate "the distance" between x_i and $\bar{x} := n^{-1}\sum_{i=1}^{n}x_i$. So we shall calculate

$$h_i := x_i(X^TX)^{-1}x_i^T \tag{2.4}$$

The third criterion is based upon the fit $\hat{y}_i := x_i\hat{\beta}$. Define

$$\text{DFITS}_i := \frac{x_i\hat{\beta}-x_i\hat{\beta}[i]}{s[i]\sqrt{h_i}} \tag{2.5}$$

The influence of outliers in the residuals can be detected by the estimated residuals $\hat{e}_i := y_i-x_i\hat{\beta}$. Define

$$\text{RSTUDENT}_i := \frac{\hat{e}_i}{s[i]\sqrt{1-h_i}} \tag{2.6}$$

In case of non-stochastic regressors $RSTUDENT_i$ has a t_{n-p-1} distribution.

The fifth criterion deals with "multiple row effects". Sometimes two outliers occur together and single row deletion methods are not decisive. The partial-leverage plots give the opportunity to detect more influential points at the same time. As well-known the k-th regression coefficient can be calculated in two steps. First calculate the regressions of y resp. the k-th column of X on X[k], X[k] being the $n \times (p-1)$ matrix obtained from X by deleting the k-th column. Let u_k and v_k be the columns of estimated residuals, respectively. Then $\hat{\beta}_k$ equals the simple regression coefficient obtained from the regression of u_k on v_k. The scatter diagram of u_k and v_k is called the "partial leverage plot" for the estimate $\hat{\beta}_k$.

3. THE SAVINGS RATIO DATA RECONSIDERED

The model describes the savings ratio as a linear function of per-capita disposable income, the percentage rate of change in per-capita disposable income, and two demographic variables: the percentage of population less than 15 years old and the population over 75 years old. The cross-sectional data are averaged over the decade 1960-1970 to remove the business cycle or short-term fluctuations. The regression equation and variable definitions are then

$$SR_i = \beta_1+\beta_2 POP15_i+\beta_3 POP75_i+\beta_4 DPI_i+\beta_5 \Delta DPI_i+\varepsilon_i, \qquad (3.1)$$

where

SR_i = the average aggregate personal savings rate in country i over the period 1960-1970,

$POP15_i$ = the average percentage of the population under 15 years of age over the period 1960-1970 in country i,

$POP75_i$ = the average percentage of the population over 75 years

of age over the period 1960-1970 in country i,

DPI_i = the average level of real per-capita disposable income in country i over the period 1960-1970 measured in U.S. dollars,

ΔDPI_i = the average percentage growth rate of DPI_i over the period 1960-1970.

Table I summarizes the results given in Belsley e.a. (1980).

TABLE I

The estimated coefficients of equation (3.1) according to Belsley e.a. (1980).

coefficient	estimate	estimated standard deviation
β_1	28.56	7.345
β_2	-0.4611	0.1446
β_3	-1.691	1.083
β_4	-0.000337	0.000931
β_5	0.4096	0.1961

$R^2 = 0.33$.

When we apply LS to the data we end up with the following estimates.

TABLE II

The correct estimates of the coefficients of equation (3.1).

coefficient	estimate	estimated standard deviation
β_1	21.56	6.966
β_2	-0.3239	0.1377
β_3	-0.844	1.053
β_4	-0.000188	0.000969
β_5	0.4141	0.2055

$R^2 = 0.28$.

How to explain the differences? Obviously the results can be analysed by the tools provided in the textbook Belsley e.a. (1980) itself. The following table shows the figures for the four criteria of Section 2, where we have used the standard critical values. We have only listed the influential points.

TABLE III

The $DFBETAS_{ij}$, h_i, $DFITS_i$, $RSTUDENT_i$ criteria for the savings-ratio data.

Criterion $DFBETAS_{ij}$ *) (critical value 0.28)

according to (1)		our results	
observation	value	observation	value
10	0.2842		
21	0.4815	21	0.303
23	-0.6739	23	-0.703
		24	-1.258
33	-0.2871		
		44	-0.338
46	-0.3389	46	-0.303
47	-0.1954	47	-0.306
49	-1.0244	49	-1.277

*) We have listed the observations with the largest absolute values of $DFBETAS_{ij}$.

Criterion h_i (critical value: 0.20)

according to (1)		our results	
observation	value	observation	value
21	0.2122		
23	0.2233		
		24	0.3138
44	0.3336	44	0.3283
49	0.5314	49	0.5375

Criterion $DFITS_i$ (critical value: 0.63)

according to (1)		our results	
observation	value	observation	value
23	0.8596	23	0.940
		24	-1.406
46	0.7482	46	0.729
49	1.1601	49	-1.441

Criterion $RSTUDENT_i$ (critical value: 2)

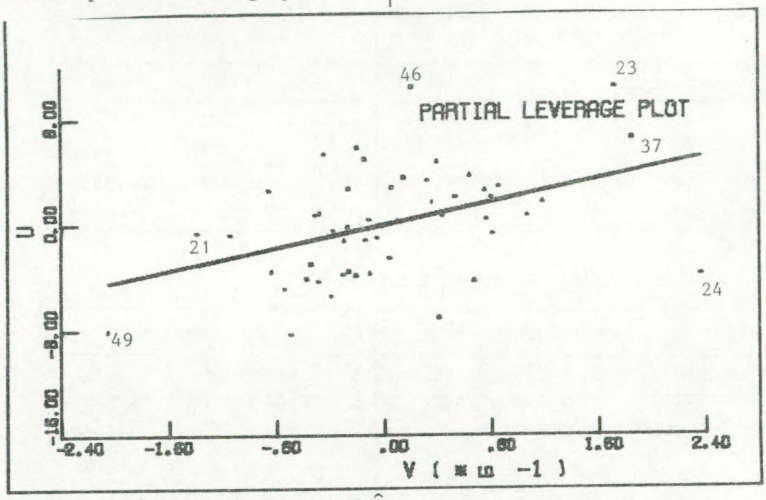
according to (1)		our results	
observation	value	observation	value
7	-2.3134	7	-2.111
		24	-2.695
46	2.8535	46	2.798

The striking difference between our calculations and (1) is the 24-th observation. This observation is not discovered in (1). The partial leverage plots also illustrate the special meaning of observation 24.

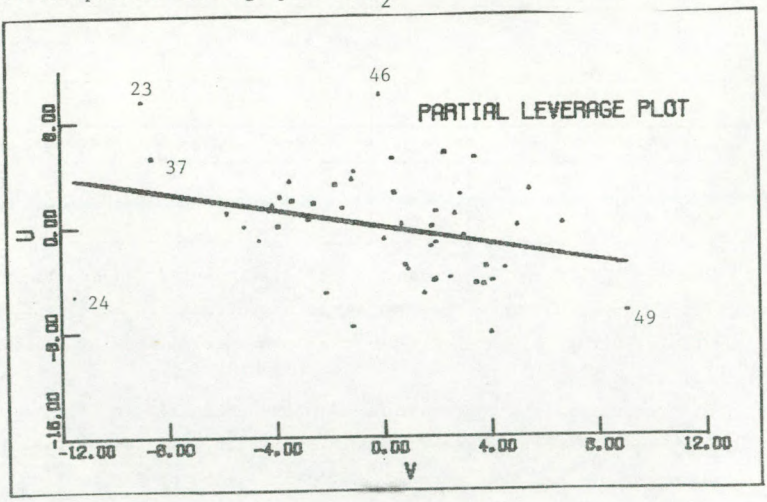
It is remarkable that observation 24, not recognized in (1), occurs as an highly influential point in Table III. The partial leverage plots are even more decisive: in all cases observation 24 has a high marginal leverage.

Since the h_i criterion only depends on the explanatory variables and clearly recognizes observation 24, we expect an error in the

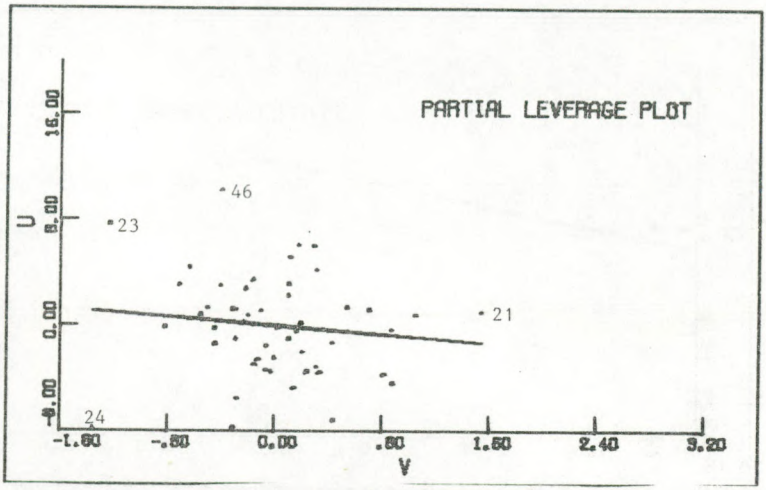
1. The partial leverage plot of $\hat{\beta}_1$.



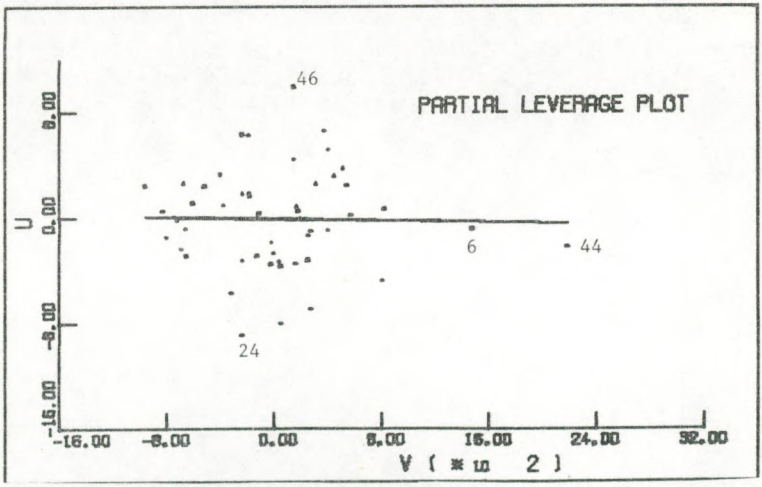
2. The partial leverage plot of $\hat{\beta}_2$.



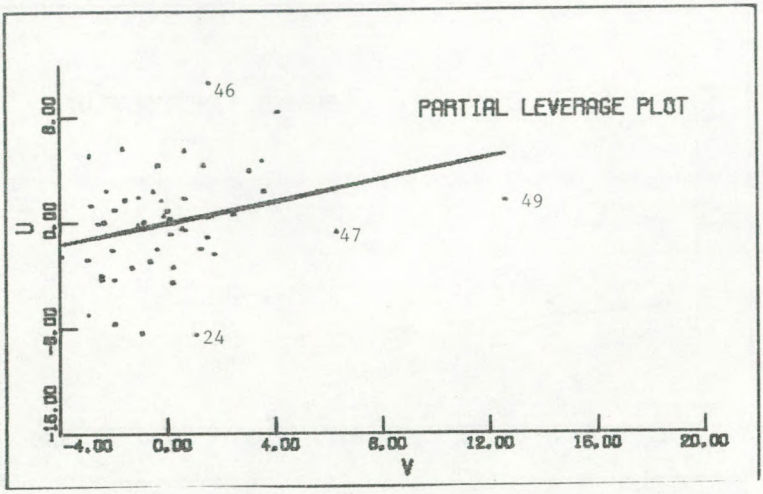
3. The partial leverage plot of $\hat{\beta}_3$.



4. The partial leverage plot of $\hat{\beta}_4$.



5. The partial leverage plot of $\hat{\beta}_5$.



row x_{24}, the Korean observations. The Korean data are

		SR	POP15	POP75	DPI	ΔPI
24	Korea	3.91	31.74	0.91	207.60	5.81

By changing the Korean POP15 value 31.74 into 41.73 we found the estimates of Table IV. Moreover, comparing the POP15 variable for Korea with the percentage of the population under 15 years of age of Taiwan (44.75), India (41.31), Malaysia (47.20) and Philippines (46.26) we also expect a POP15 value for Korea of at least 40. The Demographic Yearbook (1970) confirmed our expectations. On the basis of the revised data we have the following estimation results.

TABLE IV

Regression coefficients of model 1. Revised data.

coefficient	estimate	estimated standard deviation
β_1	28.55	7.314
β_2	-0.4610	0.1439
β_3	-1.689	1.077
β_4	-0.000343	0.000929
β_5	0.4137	0.1959

$R^2 = 0.34$.

The small differences between the results of the Tables I and IV are due to rounding errors. So we conclude that the calculations made in Belsley e.a. (1980) are correct, but that the presented data contain a printing error in the Korean data.

BIBLIOGRAPHY

1. Belsley, D.A., Kuh, E. and Welsch, R.E. (1980). "Regression Diagnostics, Identifying Influential Data and Sources of Collinearity", John Wiley and Sons.
2. Krasker, W.S. (1980). Estimation in linear regression models with disparate data points, Econometrica 48, 1333-1346.
3. Krasker, W.S. and Welsch, R.E. (1979). Efficient bounded influence regression estimation using alternative definitions of sensitivity. Techn. Report MIT.
4. Maronna, R., Bustor, O. and Yohai, V. (1979). Bias- and efficiency-robustness of general M-estimators for regression with random carriers. In: Smoothing Techniques for Curve Estimation, Springer-Verlag Berlin.
5. Demographic Yearbook, United Nations, New York (1970).