Biometrical Letters
Vol. 43 (2006), No. 2, 79-89

A comparison of some methods of detecting influential observations in studies of aphidofauna

Anna Budka[1], Maria Kozłowska[1], Barbara Wilkaniec[2]

[1]Department of Mathematical and Statistical Methods, Agricultural University of Poznań, Wojska Polskiego 28, 60-637 Poznań, e-mail: markoz@au.poznan.pl
[2]Department of Entomology, Agricultural University of Poznań, Dąbrowskiego 159, 60-594 Poznań, e-mail: wilk@au.poznan.pl

SUMMARY

Four methods for the detection of influential observations are described. The identification of observations having an influence on the linear regression model of the structure of the aphid population and their subsequent elimination from the dataset are non-trivial tasks, when the main purpose is to determine the significance of the regression coefficient for precipitation. In this paper a comparison of the effectiveness of the described methods for studies of aphidofauna is presented and discussed.

Key words: Influential observation, multiple polynomial regression, aphidofauna

1. Introduction

The scientific planning of each of the various operations in entomology trials is based on proper experimentation that yields statistically valid and easily verifiable results. Analyses of these trials usually involve multiple regression procedures to identify relationships between several independent variables and a dependent variable. The term 'multiple regression' was first used by Pearson (1908). The general computational problem that needs to be solved in multiple regression analysis is to fit a straight line or curve to a number of experimental points of each two dimensional subspace of a k-dimensional space. A 'best' regression model is sometimes developed in stages. The stepwise procedure or the backward removal procedure may be applied to the building of the regression model. After applying one of these simple model-building procedures, it is necessary to make an analysis in order to detect influential observations.

This paper has been written to assist researchers concerned with aphidofauna trials in applying appropriate methods for the detection of influential

observations. The general purpose of the analysis is to determine the significance of the regression coefficient for precipitation in the multiple regression model of the structure of the aphid population.

2. Linear model

Let $\mathbf{y}$ be the n-dimensional vector of observations of the dependent variable, $\mathbf{X}$ be the (n x p) full-column rank matrix of observations of the k ($k \leq p < n$) independent variables and $\boldsymbol{\beta}$ be the p-dimensional vector of the structural parameters. Using this notation we can write

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

where $\boldsymbol{\varepsilon}$ is the n-dimensional vector of random variables having normal distribution with expectation zero and the same standard deviation σ. From the Gauss-Markov theorem we know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is the unique linear unbiased estimator of $\boldsymbol{\beta}$ having minimum variance. This estimator is called the best linear unbiased estimator. The vector of ordinary residuals $\mathbf{e}$=[e1,e2,...,en] is given by $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, where

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \tag{2.2}$$

denotes the $(n \times n)$-dimensional orthogonal projector. Moreover, the estimator of variance σ^2 is of the form $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/(n-p) = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}/(n-p)$ (Chatterjee, Hadi 1986).

It is well known that there may be one or more influential observations. The research problem concerns the identification of the influential observations in the regression model. The methods of the detection of the influential observations are numerous. In next section we will describe the most common ones. We will consider methods based on measures defined by residuals or diagonal elements of the orthogonal projector matrix.

3. Influential observations

The chosen model may be inappropriate because of the occurrence of large values of residuals. Inferences made on the basis of such a model can sometimes lead to incorrect conclusions. In this case we need to consider the influence of certain vectors of observations on the values of residuals.

The problem concerns the one or more vectors of observations. Let $\mathbf{Z}$ denote an (n x (p+1))-dimensional matrix. The rows of this matrix are p+1 dimensional vectors and its elements are formed by observations of the independent variables and the dependent variable. The matrix $\mathbf{Z}$ has the following form

$$\mathbf{Z} = [\mathbf{X} \mid \mathbf{y}] = [(\mathbf{x}_1', y_1)', (\mathbf{x}_2', y_2)', ..., (\mathbf{x}_n', y_n)']', \tag{3.1}$$

where vector $\mathbf{z}_i = (\mathbf{x}_i', y_i)' \in \mathbf{R}^{p+1}$ for i = 1, 2, ..., n, and the vector $\mathbf{z}_i$ is called the vector observation. The one vector observation or the set of some number of vector observations (say m, where m<n) we will call the influential observation or the set of influential observations, respectively, when they significantly contribute to changes in the values of the adjustment measures of the considered regression model (2.1).

The influential vector observation may be detected by examining the diagonal elements hii of the prediction matrix $\mathbf{H}$ of the form (2.2), i=1,2,...,n. The values of the diagonal elements of this matrix are contained within the interval $\langle 1/n; 1)$. Assuming a certain threshold value, say h0, we can detect the influential observations, which means to find the set of vector observations $\mathbf{x}_i$ for which the values of diagonal elements of $\mathbf{H}$ are greater than h_0. It is known that we can assume

$$h_0 = \frac{2(p+1)}{n} \tag{3.2}$$

(Fox 2005). We can distinguish the influential observations among those which are called leverage points or high-leverage points.

Let us denote by $\mathbf{H}^*$ the (n x n)-dimensional matrix of orthogonal projection on the column space of matrix $\mathbf{Z}$ of the form (3.1). The values of the diagonal elements h_{ii}^* of this matrix are related to the diagonal elements hii of matrix $\mathbf{H}$ and residuals ei. If the residual ei, for a certain i, is high then it may be an outlier observation.

We consider now the measure defined by Cook (1977) of the form

$$D_i = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{p\hat{\sigma}^2} = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})' (\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{p\hat{\sigma}^2}, \tag{3.3}$$

i =1,2,...,n, where

$$\hat{\boldsymbol{\beta}}_{(i)} = \hat{\boldsymbol{\beta}} - \frac{e_i}{1 - h_{ii}} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{x}_i \text{ and } \hat{\mathbf{y}}_{(i)} = \mathbf{X}\hat{\boldsymbol{\beta}}_{(i)}$$

denotes the estimator of the vector of the structural parameters and the estimator of the vector **y**, respectively, in the 1-cut regression model. This measure is called the Cook distance. Now we can assume the threshold value D0 and we can detect the influential observations. The assumption has sometimes been made (Fox 2005) that the threshold value is of the form

$$D_0 = \frac{4}{n - p - 1} \tag{3.4}$$

but Fox (2002, p.198) is among many authors who recommend 4/(n-2) as "a rough cutoff for noteworthy values of D_i".

A different influence measure is the measure described by Belsley, Kuh and Welsch (1980). This is the measure called the DFFITS distance and it is the prime indicator of influence. This statistic has the form

$$DFFITS_i = \frac{\hat{\mathbf{y}}_\mathbf{i} - \hat{\mathbf{y}}_{\mathbf{i(i)}}}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}}, \tag{3.5}$$

where $\hat{\sigma}^2_{(i)} = \frac{1}{n-p-1}[(n-p)\hat{\sigma}^2 - \frac{e_i^2}{1 - h_{ii}}]$.

The numerator of the statistic (3.5) gives the difference between the predicted value for the i-th observation obtained by the model (2.1) using all observations and the model estimated without that observation. The difference is standardized, using the residual standard deviation estimate from all other observations. Belsley, Kuh and Welsch (1980) suggest that the absolute values of $DFFITS_i$ values exceeding

$$2\sqrt{\frac{p+1}{n}} \text{ (say } DFFITS_0) \tag{3.6}$$

and $DFFITS_0$ provide a convenient criterion for identifying influential observations.

The set of selected single influential observations is the starting point for investigation of the 2-dimensional and the larger dimensional joint influential observations.

Generally we can consider an m-cut regression model. In this case we detect the m- dimensional influential vector observation applying the Cook measure for the m-cut model. The measure has the form (Cook, Weisberg 1980):

$$D_I = \frac{(\hat{\boldsymbol{\beta}}_{(I)} - \hat{\boldsymbol{\beta}})' \mathbf{X}' \mathbf{X} (\hat{\boldsymbol{\beta}}_{(I)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} = \frac{1}{p\hat{\sigma}^2} \mathbf{e}_I' (\mathbf{I} - \mathbf{H}_I)^{-1} \mathbf{H}_I (\mathbf{I} - \mathbf{H}_I)^{-1} \mathbf{e}_I,$$

where the symbol I expresses here the distinguished numbers of influential vector observations; there are m distinguished numbers. For m=2 we have two distinguished numbers, $I=\{i;j\}$, $1 \le i < j \le n$. The Cook measure has now the form (Gray, Ling 1984):

$$D_I = D_{\{i;j\}} = B_1 + B_2 + B_3,$$

where

$$B_1 = (D_i + D_j)(1 + \frac{h_{ij}^2}{d_{ij}})^2 \ , \quad B_2 = \frac{h_{ij}^2}{2\hat{\sigma}^2 d_{ij}^2}[e_i^2(2 - h_{jj}) + e_j^2(2 - h_{ii})],$$

$$B_3 = \frac{2e_i e_j h_{ij}}{2\hat{\sigma}^2 d_{ij}^2}[1 + h_{ij}^2 - h_{ii}h_{jj}] \quad \text{and} \quad d_{ij} = (1 - h_{ii})(1 - h_{jj}) - h_{ij}^2.$$

If for $I=\{i;j\}$ the element h_{ij} of the matrix **H** is greater than zero, $h_{ij}>0$, then for $B_3 > 0$ and $e_i e_j > 0$ we say that these vector observations are joint influential, while for $B_3 < 0$ and $e_i e_j < 0$ we say that these vector observations are not influential. On the other hand if $h_{ij}<0$, then for $B_3 > 0$ and $e_i e_j < 0$ we say that these vector observations are jointly influential, while for $B_3 < 0$ and $e_i e_j > 0$ we say that these vector observations are not influential. When observations are not joint influential, but these observations are single influential, then we can say that the influence of one is masked by the influence of the other (Lawrance 1995).

4. Research problem

The present analysis is based on the results of research into aphidofauna. This was carried out in a green area of Poznań and has already been published in part (Wilkaniec 2005). The experiment began in 1998 and finished in 2004. The purpose of the research was to study the influence of weather conditions on the extent of aphid occurrence. Detailed observations of numerous aphids and the changing of the weather conditions were made from the beginning of May until the end of October, at intervals of ten days. Several meteorological factors were observed: the global daily temperature, the global minimum and maximum temperature and the global precipitation in mm, and the number of days with temperature > 30^0C per ten days. Altogether 126 vectors of observations were received.

The analysis began with a study of the correlation between the independent variables and the log-transformed dependent variable. The logarithm transformation was applied to the dependent variable. Transformation of data was carried out to achieve the following objectives: the equalization of variances and the normalization of observations. A fundamental problem in the multiple regression analysis is to eliminate insignificant variables. The backward removal procedure was applied to the building of the regression model. After this selection, a set of three independent variables was obtained: the date of catches, the global maximum and the daily temperature. The sensitivity of aphids to precipitation was taken into consideration. It is known that the structure of the aphid population is related to precipitation (Dixon 1985, Fink and Völkl 1995). The global precipitation was added to the regression model.

The following regression model was obtained:

$$y = -2.9658 + 0.14x_1 - 0.0347x_1^2 + 0.0019x_1^3 + 0.0515x_2 + \\ -0.0001x_2^2 - 0.0118x_3 + 0.0035x_4$$

where y denotes the logarithm of the number of aphids, x_1 denotes the date of catches (first ten-day interval,...,18^{th} ten-day interval), x_2 the global maximum temperature, x_3 the global daily temperature and x_4 the global precipitation. The adjustment measures for the above model are presented in the first row of Table 2.

The question arises whether the insignificance of the regression coefficient for precipitation is caused by the occurrence of atypical observations. We applied the four previously described methods for detection of influential observations (Table 1). As a result we found that the vector observation of rank 126 is a high-leverage point, because the proper element hii of the matrix **H** is greater than h0=0.43, see (3.2). The set of influential vector observations contains points of ranks 16, 108, 68, 107, 109, 62, 55, 67 and 35, because for each of them DFFITSi>0.535, see (3.6). If the Cook measure (3.3) is used, the number of influential observations is greater. All observations for which Di>0.034 are called influential observations, see (3.3) and (3.4) – all are presented in Table 1.

Table 1. Ordered values of the measures of the influence and ranks of vector observations

rank	h_{ii}	rank	h_{ii}^{*}	rank	Di	rank	|DFFITS|
126	0.54	126	0.54	16	0.156	16	0.869
27	0.41	27	0.41	108	0.110	108	0.722
108	0.35	108	0.35	68	0.099	68	0.704
16	0.16	16	0.19	107	0.093	107	0.673
99	0.13	99	0.13	109	0.079	109	0.616
54	0.13	54	0.13	62	0.076	62	0.605
90	0.12	109	0.13	55	0.075	55	0.605
47	0.12	107	0.13	67	0.067	67	0.567
73	0.11	67	0.13	35	0.065	35	0.562
109	0.11	90	0.13	19	0.057	19	0.521
67	0.11	19	0.12	27	0.052	27	0.495
36	0.11	68	0.12	113	0.046	113	0.478
18	0.10	47	0.12	33	0.045	33	0.465
37	0.10	73	0.12	36	0.037	36	0.420
19	0.10	55	0.12	126	0.036	115	0.416
107	0.10	36	0.12	84	0.036	84	0.414
91	0.10	62	0.11	115	0.035	126	0.412
14	0.09	18	0.11	26	0.034	26	0.404
72	0.09	35	0.10	70	0.033	70	0.396
55	0.09	14	0.10	32	0.032	32	0.392
1	0.09	37	0.10	114	0.031	114	0.385
89	0.09	72	0.10	64	0.029	64	0.382
84	0.09	84	0.10	90	0.027	90	0.357
62	0.09	91	0.10	14	0.026	14	0.351
83	0.08	89	0.09	34	0.023	34	0.331
94	0.08	83	0.09	20	0.021	20	0.317
53	0.08	32	0.09	12	0.021	12	0.315
71	0.08	1	0.09	88	0.019	88	0.301
32	0.08	94	0.09	72	0.018	65	0.293
97	0.08	71	0.09	89	0.018	72	0.293
35	0.07	53	0.09	65	0.018	89	0.292
12	0.07	113	0.08	83	0.017	83	0.284
100	0.07	33	0.08	69	0.017	69	0.282
105	0.07	12	0.08	94	0.016	9	0.272
28	0.07	70	0.08	9	0.016	94	0.272
74	0.07	97	0.08	66	0.015	66	0.268

In the next step of our analysis we removed different selected sets of vector observations. For each choice we built the regression model and calculated diagnostic measures. The diagnostic measures were useful and played an important part in the analysis. We observed that by rejecting observations which according to (3.4) are called single influential observations, we obtained a regression model of the form:

$$y = -2.643 + 0.1368x_1 - 0.0338x_1^2 + 0.0018x_1^3 + 0.0532x_2 + \\ -0.0001x_2^2 - 0.0179x_3 + 0.006x_4 .$$

We obtained this model after removing vector observations whose ranks belong to the following set: {16, 108, 68, 107, 109, 62, 55, 67, 35, 19, 27, 113, 33, 36, 126, 84} (see Table 2). For each regression coefficient we calculated the p-value, and it was smaller than 0.05 in each case, including for precipitation (p=0.042).

Table 2. The diagnostic measures: the determination coefficient R2, the adjusted determination coefficient $\overline{R}^2$, the standard deviation $\hat{\sigma}$, the p-value for testing of hypotheses concerning the regression coefficients

Rank of removed observation	R^2 (%)	$\overline{R}^2$ (%)	$\hat{\sigma}$	p-value						
				x_1	x_1^2	x_1^3	x_2	x_2^2	x_3	x_4
---	59.1	56.7	0.483	0.191	0.007	<0.001	<0.001	<0.001	0.001	0.249
16	60.3	58.0	0.477	0.119	0.003	<0.001	<0.001	<0.001	0.001	0.297
108	60.4	58.0	0.478	0.117	0.003	<0.001	<0.001	<0.001	0.001	0.331
68	61.7	59.3	0.466	0.186	0.007	<0.001	<0.001	<0.001	0.001	0.240
107	63.3	61.1	0.457	0.158	0.006	<0.001	<0.001	<0.001	<0.001	0.369
109	64.1	61.9	0.452	0.060	0.002	<0.001	<0.001	<0.001	<0.001	0.514
62	64.5	62.3	0.447	0.048	0.001	<0.001	<0.001	<0.001	<0.001	0.252
55	65.2	63.1	0.442	0.124	0.004	<0.001	<0.001	<0.001	<0.001	0.267
67	66.0	63.9	0.436	0.190	0.009	<0.001	<0.001	<0.001	<0.001	0.109
35	65.8	63.6	0.431	0.197	0.009	<0.001	<0.001	<0.001	<0.001	0.083
19	66.4	64.2	0.427	0.423	0.032	<0.001	<0.001	<0.001	<0.001	0.090
27	65.5	63.3	0.429	0.381	0.027	<0.001	<0.001	<0.001	0.001	0.077
113	67.1	64.9	0.417	0.550	0.048	0.001	<0.001	<0.001	0.001	0.060
33	67.5	65.4	0.409	0.447	0.033	<0.001	<0.001	<0.001	0.001	0.059
36	67.2	65.0	0.407	0.334	0.019	<0.001	<0.001	<0.001	<0.001	0.063
126	67.0	64.7	0.409	0.279	0.016	<0.001	<0.001	<0.001	0.008	0.055
84	67.8	65.6	0.405	0.211	0.010	<0.001	<0.001	<0.001	0.007	0.042
115	69.5	67.4	0.393	0.399	0.025	<0.001	<0.001	<0.001	0.015	0.057
26	69.6	67.5	0.384	0.323	0.016	<0.001	<0.001	<0.001	0.012	0.055

These removed influential observations are marked in a two dimensional space (x_4;log(y+1)) (Fig. 1). It is very interesting that these influential observations lie inside the area of experimental points.

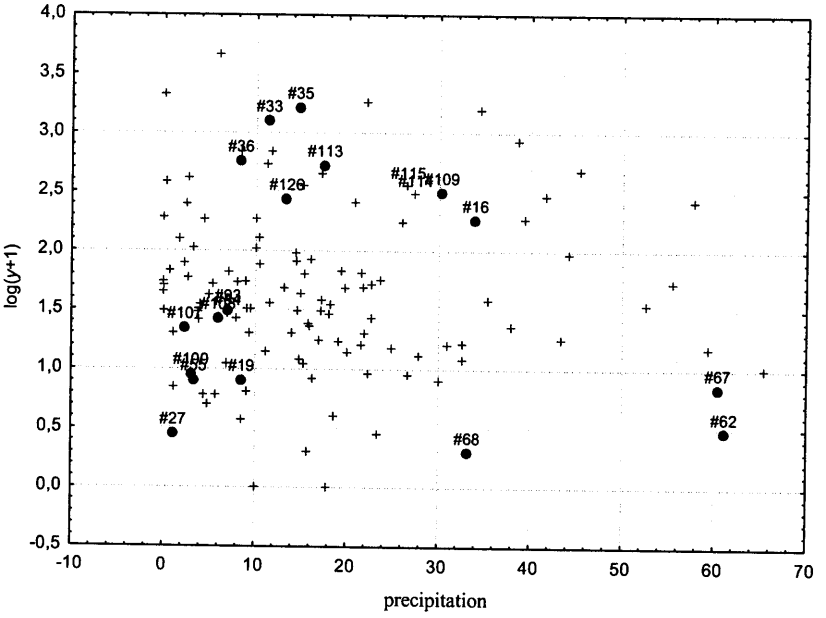

Fig. 1. The observations presented in a two dimensional space (x_4;log(y+1)) and the removed influential observations (marked points)

In this analysis we also calculated the Cook measures for 2-cut models. We investigated each pair of the considered single influential observations (Table 3). We observed that there were no opposing effects. The influence of one observation was not masked by the influence of the other. This completed the analysis.

The multiple regression determination coefficient, R^2, explains how much of the variability in the y's (logarithm of the number of aphids) can be explained by the fact that they are related to x_1 (date of catches) and to x_2 (global maximum temperature), x_3 (global daily temperature) and x_4 (global precipitation). For our model this coefficient is equal to 67.8% and the adjusted coefficient of determination equals 65.6%. The difference of these values is small, showing the model to be well-chosen for practical purposes.

Table 3. Ordered values of the Cook measure for the two-cut model for chosen pairs of observations I={i;j}

Rank i	Rank j	D_I
108	107	0.773
16	35	0.495
68	67	0.390
16	27	0.374
55	19	0.331
108	33	0.300
16	109	0.299
62	67	0.299
108	90	0.283
16	14	0.250
108	68	0.248
108	109	0.240
68	107	0.240
16	84	0.240
16	126	0.234
16	55	0.233
68	62	0.232
68	19	0.228
107	109	0.223
108	36	0.222
16	36	0.221
.....	.....	.....

5. Conclusions

The paper contains descriptions of four methods for detecting influential observations. The Cook measure was considered, as was DFFITS, as well as methods based on analysis of the elements of operators of the orthogonal projection on the space of columns of the observation matrix and the space of columns of the extended observation matrix. An evaluation was made of the usefulness of these methods for showing that, apart from the sum of daily temperatures and the sum of maximum temperatures, the sum of precipitation also (together with those mentioned) has an influence on the size of the aphid population. Different sets of influential observations were obtained. In the studies presented, by removal of the influential observations identified using the Cook

measure, a uniform dataset was obtained. A dataset is uniform if it does not contain atypical data, i.e. data which interfere with the real relations between the variables being studied. The other three methods for detecting influential data proved ineffective here. The occurrence of 16 atypical observations in a set of 126 observations probably results from the confounding of many uncontrolled factors in the process of catching aphids in a green area of Poznań. An attempt was also made to reduce the number of influential observations by use of the Cook measure for an m-cut model. For the identified influential observations, however, no masking effect was found to occur.

REFERENCES

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

Chatterjee S., Hadi A. S. (1986): Influential observations, high leverage points, and outliers in linear regression. Statistical Science 1/3, 376-416.

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

Cook R. D., Weisberg S. (1980): Characterizations of an Empirical Influence Function for Detecting Influential Cases in Regression. Technometrics 22, 495-507.

Gray J. B., Ling R. F. (1984): K – Clustering as a Detection Tool for Influential Subsets in Regression. Technometrics 26, 305-318.

Dixon A.F.G. (1985). Structure of Aphid Populations. Annual Review of Entomology 30, 155-174.

Fink U., Völkl W. (1995). The effect of abiotic factors on foraging and oviposition success of the aphid parasitoid, *Aphidius rosae*. Oecologia 103, 371-378.

Fox J. (2002). An R and S-PLUS Companion to Aplied Regression. Sage, California.

Fox J. (2005). Overview of Regression Diagnostics. Sociology 740. http://socserv. mcmaster.ca /jfox/Courses/soc740/diagnostics-overview.pdf

Lawrance J. (1995): Deletion Influence and Masking in Regression. J. R. Statist. Soc. B, 57, 181-189.

Pearson K. (1908). On the generalized probable error in multiple normal correlation. Biometrika, 6, 59-68.

Wilkaniec B. (2005): Many years dynamics of aphid recurrence in urban green spaces of Poznań. Aphids and other hemipterous insects, 11. 203-211.