REGRESSION POST-MORTEM*

Dedicated to the memory of Shri J.M. Sengupta (Sankar Babu)

Sujit Kumar Mitra
Indian Statistical Institute
New Delhi - 110 016, India

I am conscious of the honour done to me by asking me to preside over the 1991 annual meeting of the Indian Society of Agricultural Statistics. When Dr. Prem Narain conveyed to me over the telephone the decision taken by the executive council in this regard, I pointed out to him my physical difficulties but eventually I succumbed to his persuasions. However, it did not take much time thereafter for me to realise the mistake that I have committed. This was when I started thinking about the topic of my presidential address, which I was warned, will in course of time be published in the Society's journal. With barely two months left for my retirement, I thought I have earned my place among the old men and I can definitely be excused if I talk about my reminiscences. To appreciate what I have to say, one has to turn back the clock little more than 30 years. We were then reasonably fresh from the university and were having an exciting time otherwise, basking in the creative atmosphere of the Research and Training School, Indian Statistical Institute. Our leader was a brilliant young man who had already earned an international reputation for himself. There were besides a few other senior colleagues I am proud to have had the opportunity to work with**. Everywhere around, there was general expectation that we shall be able to do something important though not necessarily something great. Creative work of all types were encouraged. This need not have to be writing a research paper. We used to enjoy preparing good questions for our students. Something which is not routine, which tested application more than memory, required some thinking on the part of the student, but was not

* 'Technical Address' delivered at the 45th Annual Conference of the Society at NDRI, Karnal on 27 November, 1991.

** My present talk is dedicated to the memory of one of them. Sankar Babu did not have even a Master's degree in any subject and did not satisfy any of the norms that would have earned him any of the so called 'honours'. I hope some day researchers digging into the history of the Indian Statistical Institute will make a proper evaluation of this man's contribution to the success story the institute is today.

difficult in the sense of qualifying for a research problem. I remember I devoted considerable time with our leader setting a question which eventually centred around a cheatercock. The cheatercock claimed that he made a fair selection of his share of half the plots growing a commercial crop. However from 10 pairs of plots, he was found to have selected the better half in 9 pairs. The students were required to comment on the cheatercock's claim, giving proper justification for the same. This question would have been trivial if the students were already exposed to a course in tests of hypotheses. The target batch in this case however had gone through the first course in Probability theory and our object was to find out whether some of them could give a proper argument from purely common sense point of view. Linear models always created some difficulties. It was given considerable weight in the teaching on account of its connection with analysis of variance. There were essentially one or two theorems and it was a challenging experience, innovating good questions in this area. Once the examiners had no option but to break a costly stone of known weight. The different pieces that could be traced were weighed in different combinations. From the recorded weights, it was necessary to conclude if all the pieces were recovered. But in a poor country like ours, how many costly stones can we afford to break and lose? When searching for other alternative questions it occured to me that it may be a nice idea to find out how to correct the best linear estimate, its variance and other steps in the statistical analysis of such a model if one decides to delete an observation from or add a fresh observation to the statistical analysis. There is of course the dual problem of having to delete one parameter or add a fresh one. But why should one delete an observation? May be the motivation for doing this came from the suspicion that it was a fake observation or an outlier and so on. I derived the expressions myself using elementary arguments and since then generations of students who sat in my course on this subject must have gone thro' this exercise. Soon it occured to me that the results though simple are quite interesting by themselves and may deserve journal publication as a small note. An off-shoot of this investigation was a paper[3] I published in 'Sankhya' entitled "Some remarks on the Missing plot Analysis" which made use of some of these results. I gathered the main group of results in the form of a note which I submitted to Biometrics. I had suggested missing plot analysis as a possible area of application. The referee report said that good techniques are already available in literature for this problem. Results were not sufficiently motivated and therefore the paper was not publishable. The referee otherwise appeared to have liked the results and so did the editor who even suggested that I could consider resubmitting the paper once I have found a more convincing motivation. About a decade later, when P. Bhimasankaram was working with me on his doctoral investigation, we derived the same results in terms of generalised inverses of the design matrix and published these results in 'Sankhya' in a paper [4] entitled "Generalized inverses of partitioned matrices and recalculation of least squares estimates for data or model changes".

A section was also added in the monograph on generalized inverses that I co-authored with Prof. C.R. Rao. Even though it is a diversion, I am tempted to record here an incident which happened in early 80's. I received an invitation from a professor of Electrical Engineering in West Virginia University who was editing a volume of papers devoted to "Adaptive Linear Filtering". The letter said that he was inviting me on account of my (considerable?) contributions to this area. Since I was pretty sure that I have not done any work in Electrical Engineering, I politely excused myself thinking the Editor of the volume must have written to me on account of some mistaken identity. From wiser men I met later, I have learnt that recursive least squares would come under 'adaptive linear filtering'. Of course by then it was too late for me to retract my steps and write to the editor accepting the invitation.

My purpose in digging out all these events from my memory bank is definitely not that I urgently needed some publicity for my failures. These would perhaps be lying dormant in some cells of the brain had not my attention been drawn to some techniques which have gained considerable popularity among data analysts in recent years [1, 2] under the name "regression diagnostics" and "sensitivity analysis". When I read thro' the relevant literature, I started lamenting why did I not think of these problems in the late 1950's when I submitted my paper to 'Biometrics'. This surely would have been found by the referee as a worthwhile motivation and my name would have been associated with an important technique in data analysis. The answer is not very hard to find. It can be summarised in just one word, "computers". The techniques of regression diagnostics and sensitivity analysis are so computer intensive that naturally it would not have occured to anybody in the late 1950's, when it was difficult to conceive that one day the computer would be a household word almost everywhere. In the rest of my talk I shall introduce these techniques by applying them on a historical data reported in the book "Statistical methods for agricultural workers [5]" by Panse and Sukhatme. It would not be fair to call it "Regression diagnostics". I have taken your indulgence for granted in giving the title of my talk as "regression post mortem".

Deleting an Observation

I have no inclination of recalling here all the formulae that were derived in my Biometrics submission. However I shall present here a result which I never attempted to derive earlier, mainly to give some flavour of the proofs that I am fond of using. What my friends do not realise is that I am mortally afraid of long matrix computations which have prompted me often to search for escapes through statistical arguments. Let P be the orthogonal projector onto the column span of the design matrix X of order n x m. How to recalculate the orthogonal projector (of one dimension less) if one decides to exclude the nth observation Y_n and consequently the last row of the X matrix? Let p_{ij} be the (i,j)*th* element of P. If $p_{nn}=1$, from the symmetry and idempotency of P, it follows that $p_{nj} = p_{jn} = 0$ for all $j=1, 2, \ldots, (n-1)$. The orthogonal projector to the column span of the truncated design matrix X is therefore obtained simply by deleting the last row and last column of P. Let Y′ denote the vector $(Y_1, Y_2, \ldots, Y_n)$. It is well known that PY is the Best Linear Unbiased Estimator (BLUE) of $E(PY) = PX\beta = X\beta = E(Y)$, if $D(Y) = \sigma^2 I$. Notice that here, Y_n is the BLUE of its expectation.

If $0 < p_{nn} < 1$, $p_{n1}Y_1 + p_{n2}Y_2 + \ldots + p_{nn}Y_n$ is the BLUE of $E(Y_n)$. Therefore $(1-p_{nn})Y_n - p_{n1}Y_1 \ldots - p_{n(n-1)}Y_{n-1}$ is an error function, that is a linear function of observations having identically a zero expectation. Since $1-p_{nn} > 0$, one can subtract suitable multiples of this error function from the BLUE of $E(Y_i)$, $i = 1, 2, \ldots, (n-1)$ so as to eliminate Y_n. What remains is thus

(a) a linear function of $Y_1, Y_2, \ldots, Y_{(n-1)}$

(b) an unbiased estimator of $E(Y_i)$, indeed

(c) the BLUE of $E(Y_i)$ based on the first (n–1) observations,

the last one on account of its zero covariance with error functions based only on the first (n–1) observations. Such an error function is individually uncorrelated with $\sum_j p_{ij} Y_j$ for each i and also with Y_n. Why? I leave it to you to ponder.

Uniqueness of the BLUE under the stated assumptions now prompts us to conclude that the orthogonal projector onto the column span of the truncated design matrix is a matrix with (i, j)*th* element given by

$$p_{ij} + p_{in}p_{nj} \; / \; (\,1 - p_{nn})$$

$$i, j = 1, 2, \ldots, (n-1)$$

Regression Diagnostics

The original data from Panse and Sukhatme's book (Page 124) is reproduced in Table 2. Under regression diagnostics, one is required to determine individually the influence that each single observation exerts in the statistical analysis of the entire data. This is done by redoing the analysis after dropping observations one at a time and computing afresh the various statistics that are relevant for the analysis. Depending on the outcome the procedure is repeated by dropping observations two at a time and so on which will essentially identify the sets of observations which had the maximum influence on the entire analysis. Note that one does not necessarily prescribe omission of this set of observations. The purpose of the entire exercise is often to issue a warning that if you did include this set, you must also be aware of the extent the statistical analysis of the data has become vulnerable on this account, so that the limitations of the statistical analysis are clearly understood.

The regression of the progeny mean (Y) on parental plant value (X_1) and parental plot mean (X_2) is given by

$$Y = b_1 + b_2X_1 + b_3X_2.$$

The values of b_1, b_2 and b_3 were computed as follows:

$$b_1 = 6.3838 \qquad b_2 = 0.4452 \qquad b_3 = 0.2399$$

Let (p-1) be the number of regressors, and n, the number of observations on the p-tuple consisting of single regressand and the (p-1) regressors. In the given example p=3, n=25. Further, let X denote the n x p matrix formed by the i^{th} recorded observation on the regressors preceded by a unity constituting the i^{th} row of the matrix. As before, Y denotes the vector of observations on the regressand. Define

$$b = (X'X)^{-1}\; X'Y, \quad e=(e_1, e_2, \ldots, e_n)' = \; Y - Xb$$

$s^2 = \mathbf{e}'\,\mathbf{e}/\,(n-p)$, $h_i = x_i'\,(X'X)^{-1}\,x_i$, where x'_i is the i^{th} row of X. Also, let $X_{(i)}$ be the matrix obtained by deleting the i^{th} row of X and $Y_{(i)}$, the vector obtained by deleting the i^{th} component of Y. Let

$$b_{(i)} = (X'_{(i)} X_{(i)})^{-1} X'_{(i)} Y_{(i)}, \qquad e_{(i)} = Y_{(i)} - X_{(i)} b_{(i)}$$

$$s^2_{(i)} = \frac{e'_{(i)} e_{(i)}}{(n-p-1)}$$

The statistics $b_{(i)}$, $e_{(i)}$ and $s^2_{(i)}$ are respectively the least square estimators of the regression coefficients, the residual vector and the unbiased estimator of error variance when the i^{th} observation is deleted. In general when we are deleting m rows, e.g., those belonging to a set of indices D_m and the corresponding observations in Y, let $b_{(D_m)}$ and $X_{(D_m)}$ be defined in an analogous manner. The different diagnostic elements computed from the above quantities are reported in Table 1.

The various diagnostic calculations done on the data on fibre length of cotton progenies and parental plants are described in the Tables 1-7. These tables are self-explanatory. You may like to go through these tables and draw your own conclusions. We have not used the entire arsenal that is available to a data analyst in executing this task. For example, collinearity aspects have not been studied at all. The interested reader may refer to a forthcoming follow up paper by P. Bhimasankaram and his associates.

Acknowledgements

The author wishes to thank K. Balasubramanian for his help in computing the value in tables 3-7, which he did on his own personal computer using MATLAB. When one is physically handicapped, one needs logistic support from his colleagues and friends to undertake such a responsibility as preparing a presidential address. Aloke Dey has always been generous with his time in providing me the required support. I have the added advantage of discussing with him the various technical points that crop up in preparing a technical paper. I am most thankful to him for his support.

References

[1] Belsley, D. A., Kuh, E. and Welsh, R. E. Regression Diagnostics : Identifying Influential Observations and Sources of Collinearity, Wiley, New York (1980).

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

[3] Mitra, S. K. Some remarks on the missing plot analysis, Sankhya **21** (1959), 337-344.

[4] Mitra, S. K. and Bhimasankaram, P. Generalized inverse of partitioned matrices and recalculation of least squares estimates for data or model changes, Sankhya, A **33** (1975), 395-410.

[5] Panse, V. G. and Sukhatme, P. V. Statistical Methods for Agricultural Workers, Indian Council of Agricultural Research, New Delhi (1954).

Table 1. Explanation of various diagnostic calculations

Element	Formula	Critical values
h_i	$x_i' (X' X)^{-1} x_i$	2p/n *
$s^2_{(i)}$	$s^2_{(i)} = \frac{n-p}{n-p-1} s^2 - \frac{e_i^2}{(n-p-1)(1-h_i)}$	—
$DFBETA_{j(i)}$ **	$\frac{b_j - b_{j(i)}}{s_{(i)} \sqrt{(X' X)^{-1}_{jj}}}$	$2/\sqrt{n}$
$DFFIT_i$	$\hat{Y}_i - \hat{Y}_i (i) = \frac{h_i e_i}{1 - h_i}$	—
DFFITS	$\left(\frac{h_i}{1-h_i}\right)^{1/2} \frac{e_i}{s_{(i)} \sqrt{1-h_i}}$	$2 \sqrt{p/n}$
RSTUDENT (e_i^*)	$\frac{e_i}{s_{(i)} \sqrt{1-h_i}}$	upper 5% point of $\mid t \mid$ with n − p − 1 d.f.
COVRATIO ***	$\left[\left(\frac{n-p-1}{n-p} + \frac{e_i^{*2}}{n-p}\right)^p (1-h_i)\right]^{-1}$	1 ± 3p/n
MDFFIT	$(b - b_{(D_m)})' X'_{(D_m)} X_{(D_m)} (b - b_{(D_m)})$	

* One may find it hard to associate a probability interpretation with this critical value. In most applications the h_i values are nonstochastic. When h_i exceeds the relevant critical value the corresponding data point is called a 'leverage point' indicating that the observation Y_i here has a substantial contribution to the final answer on account of this fact alone. It may therefore be worthwhile subjecting this observation to stringent scrutiny.

** b_j and $b_{j(i)}$ are respectively the j^{th} component of b and $b_{(i)}$

*** COVRATIO is the ratio of determinants of the estimated dispersion matrix of coefficients b_1, b_2 and b_3. The numerator corresponds to a situation where the i^{th} observation is deleted while the denominator relates to complete data.

Table 2. Fibre length of cotton progenies and parental plants

Progeny No.	Progeny mean (m.m)	Parental plant value (m.m)	Parental plot mean (m.m)
1	24.3000	26.0000	25.5000
2	24.4800	25.8000	25.5000
3	23.4100	25.2000	25.5000
4	21.6000	23.4000	25.0000
5	22.4900	26.6000	25.0000
6	23.6200	25.4000	24.6000
7	22.7500	23.4000	24.6000
8	24.4000	27.6000	23.6000
9	22.6000	24.4000	23.6000
10	25.3600	24.0000	24.4200
11	23.2100	24.2000	24.4200
12	24.7600	26.0000	24.4200
13	21.5300	22.8000	22.5600
14	21.3200	20.8000	22.5600
15	22.8100	24.8000	22.5600
16	25.4100	26.2000	24.9000
17	24.3000	27.2000	24.9000
18	23.6500	26.6000	24.9100
19	24.3100	25.0000	24.9100
20	21.8800	23.4000	24.0500
21	24.1000	25.6000	24.0500
22	21.9100	23.0000	24.0500
23	22.2400	25.4000	24.5700
24	23.4500	23.4000	24.5700
25	22.1000	24.2000	24.5700
Total	581.9900	623.4000	609.3200

Page 124 : Statistical Methods for Agricultural workers, V.G. Panse and P.V. Sukhatme, Published by Indian Council of Agricultural Research, 1957.

Table 3. RSTUDENT and hat-matrix diagonals

RSTUDENT	hi
.2476	.1140
.5494	.1134
–.3476	.1220
–1.4209	.1479
–2.0929*	.0950
.0277	.0461
.0509	.1000
.0853	.3374*
–.3366	.0762
3.1834*	.0561
.2073	.0495
1.0453	.0684
–.5003	.2329
.3427	.3532*
–.0340	.2864*
1.5946	.0740
–.1882	.1352
–.6115	.0934
.8989	.0586
–.7608	.0735
.5997	.0711
–.5350	.0966
–1.5167	.0458
.8420	.0974
–1.0504	.0559

* Exceeds cut off values : RSTUDENT = 2.08 ; hi = 0.24

Table 4. Table of DFBETAS

DFBETAS		
b1	b2	b3
–.0697	.0066	.0592
–.1533	–.0026	.1396
.0951	.0345	–.1046
.1882	.4492*	–.4202*
.3441	–.3954	–.0992
–.0015	.0016	.0006
–.0016	–.0128	.0088
.0162	.0537	–.0430
–.0662	–.0144	.0639
.0153	–.4145*	.2328
.0005	–.0206	.0121
–.0181	.1822	–.0773
–.2542	.0324	.2022
.1988	–.1498	–.0901
–.0170	–.0096	.0200
–.2140	.2221	.0764
.0283	–.0569	.0051
.0872	–.1231	–.0131
–.1099	–.0409	.1244
–.0760	.1311	–.0097
.0452	.0985	–.0899
–.0565	.1265	–.0229
.0695	–.0930	–.0195
–.0208	–.2081	.1373
.0346	.1264	–.1059

* Exceeds cutoff value : DFBETAS = 0.4

Table 5. COVRATIO and DFFITS

COVRATIO	DFFITS
1.2864	.0914
1.2424	.2029
1.2871	–.1284
1.0245	–.5247
.7197	–.7729*
1.2052	.0102
1.2771	.0188
1.7334*	.0315
1.2246	–.1243
.3738*	1.1755*
1.2023	.0765
1.0600	.3860
1.4465	–.1848
1.7481*	.1266
1.6109*	–.0126
.8812	.5888
1.3229	–.0695
1.2028	–.2258
1.0906	.3319
1.1438	–.2809
1.1763	.2215
1.2220	–.1976
.8822	–.5601
1.1530	.3109
1.0444	–.3879

Exceeds cut off values : COVRATIO = 1+-0.36 ; DFFITS = 0.6928

Table 6. Extreme values from internal scaling

Item	Hat-matrix diagonals	DFFIT	DFBETA			
			b1	b2	b3	Row count
4	–	–	–	U	L	2
5	–	L	–	–	–	–
8	U	–	–	–	–	–
10	–	U	–	L	–	1
14	U	–	–	–	–	–
15	U	–	–	–	–	–
Column Count L	0	1		1	1	
Column Count U	3	1		1	0	
Total	3	2		2	1	

Subset of potentially influential points : 4, 5, 8, 10, 14, 15

Table 7. MDFFIT
For m ≤ 3 = max m

m	subset	MDFFIT	Rel index to max
1	4	.2498	.7099
	5	.3158	.8974
	8	.0022	.0063
	10	.3519	1.0000
	14	.0378	.1074
	15	.0003	.0009
2	4,5	.6216	.9946
	4,8	.3511	.5618
	4,10	.0833	.1413
	4,14	.2180	.3488
	4,15	.2651	.4242
	5,8	.3386	.5418
	5,10	.4842	.7747
	5,14	.5409	.8654
	5,15	.3201	.5122
	8,10	.3475	.5560
	8,14	.0361	.0578
	8,15	.0015	.0024
	10,14	.6250	1.0000
	10,15	.3514	.5622
	14,15	.0386	.0618
3	4,5,8	.6233	.5967
	4,5,10	.2735	.2618
	4,5,14	.7193	.6887
	4,5,15	.6289	.6021
	4,8,10	.1249	.1196
	4,8,14	.3068	.2937
	4,8,15	.4249	.4068
	4,10,14	.2352	.2252
	4,10,15	.1009	.0966
	4,14,15	.2642	.2357
	5,8,10	.5390	.5160
	5,8,14	.5871	.5621
	5,8,15	.3575	.3423
	5,10,14	1.0445	1.0000
	5,10,15	.4865	.4652
	5,14,15	.5422	.5791
	8,10,14	.6241	.5975
	8,10,15	.3471	.3323
	8,14,15	.0390	.0373
	10,14,15	.6368	.6097