Case Influence Statistics Available in SAS Version 5

John T. Pohlman
Southern Illinois University, Carbondale

Abstract

Case influence statistics are a useful diagnostic tool for identifying high leverage cases in a sample. A case's influence on a solved regression model depends on that case's residual and its location in the distribution of the predictor variables. Cases with large residuals and located in extreme ranges of the predictor variables' distributions will be most influential. Case influence is illustated with an SAS analysis of a simple data set.

The REG program in version 5 of the Statistical Analysis System (SAS) provides a collection of case influence statistics described by Belsley, Kuh and Welsch (1980), and Freund and Littell (1986). Influence statistics are designed to aid in the detection of cases which are highly influential in the estimation of the regression coefficients. A case's influence on the regression solution is estimated by deleting that case from the sample and recomputing the coefficients. If the coefficients change considerably upon deleting a case, that case is deemed influential. Generally, cases which have large residuals and are in extreme ranges of the predictor variables' distributions will be most influential.

Figure 1 presents a scatter diagram which illustrates case influence for a simple linear regression model in which a dependent variable (Y) is regressed on one predictor (X). The ten data points denoted with the symbol (•) yield the regression equation

$$Y' = 1 + 1X.$$

The ten data points denoted with the letters A to J are then used, one at a time, to augment the original sample of ten observations. Ten augmented samples of size 11 are thus created. The first augmented sample is composed of the 10 original data points plus point A. The second augmented sample consists of the 10 original observation plus point B, and so on to the tenth augmented sample using case J along with the original observations. The influence of the ten lettered data points is determined by comparing the regression coefficients obtained when the lettered data point is included in the analysis with the coefficients obtained after deleting that data point. Table 1 shows the results of this analysis.

Insert Figure 1 About Here

The second and third columns in Table 1 contain the regression coefficients obtained when cases A to J augment the original sample of 10 cases. The last two columns of the table show the change in the regression coefficients due to the presence of each lettered case. Note that the largest change in the slope coefficient occurs for cases F and J. Cases F and J have the largest deleted residuals and are the most disparate cases in the distribution of X. Cases F and J are the most influential cases. Case J has a strong positive influence on the slope coefficient, since case J's presence in the sample causes the slope coefficient to be .231 units higher than it would be if case J were not in the sample. Case F, to the contrary, has an identically strong negative influence on the slope coefficient.

Insert Table 1 About Here

INFLUENCE STATISTICS AVAILABLE IN PROC REG

The influence statistics described here are available in the SAS REG procedure as options. SAS provides the statistics HAT DIAG H, DFBETA and DFFITS. For this illustration assume that the general linear model is fit to a data set, namely

$$Y = XB + E$$

where Y is a vector of values on the response variable, X is an $n \times (p+1)$

matrix of values on the independent variables with a leading unit vector, B is the vector of regression coefficients and E is a residual vector. Letting X^T denote the transpose of X, the ordinary least squares regression coefficients are given by

$$B = (X^TX)^{-1}X^TY,$$

and the predicted values of Y are produced by

$$Y' = XB$$

$$= X(X^TX)^{-1}X^TY$$

Letting $H = X(X^TX)^{-1}X^T$, then

$$Y' = HY.$$

The matrix H is the projection matrix for the predictor space in that it operates on Y to yield Y', and is termed the hat matrix. H is of order n×n and of the same rank as X. The main diagonal values of H, h_{ii}, are measures of the dispersion of case i from the centroid of the predictor variable space. Two cases with the same value of h_{ii} are on the same probability contour of the multivariate distribution of the predictor variables. In fact, h_{ii} is a linear transformation of the Mahalanobis distance of case i from the centroid of X (Weisberg, 1980, p. 105). The h_{ii} values are labeled HAT DIAG H by the REG program. The h_{ii} values measure the potential for a case to be influential. The actual influence exerted by a case will also depend on that case's residual.

The DFBETA statistics are measures of the influence each case has on each of the regression coefficients. For each case the will be a separate DFBETA value for each regression coefficient in the model, including the intercept. The DFBETA for case i on coefficient j is

$$DFBETA_{j(i)} = \frac{b_j - b_{j(i)}}{\left[S^2_{(i)} (X^TX)^{ii}\right]^{1/2}}$$

where b_j is the regression coefficient for predictor j estimated from the total sample, $b_{j(i)}$ is the regression coefficient for variable j estimated in the sample with case i deleted, $S^2_{(i)}$ is the error variance estimate from the sample with case i deleted and $(X^TX)^{ii}$ is the i-th diagonal element of $(X^TX)^{-1}$.

The DFFITS statistic is a scaled measure of the influence of case i on the predicted value of Y. Since all of the regression coefficients are used to produce a predicted Y value, DFFITS becomes an aggregate measure of the influence of case i on the entire regression equation. The DFFITS statistic for case i is given by

$$DFFITS(i) = \frac{Y'_i - Y'_{i(i)}}{[S2(i)\ h_{ii}]^{1/2}}$$

where Y'_i is the predicted Y for case i based on the total sample, $Y'_{i(i)}$ is the predicted Y based on the regression equation estimated without case i in the sample, and hii is the i-th diagonal value of H. The DFFITS statistic is very similar to Cook's D (Cook, 1979), another measure of influence available in the REG program and also in the SPSSx regression program. Cases with DFFITS values greater than $2[(p+1)/n]^{1/2}$ are considered to be high leverage cases (Belsley et al., 1980, p. 28).

ILLUSTRATION WITH A DATA SET

Appendix A provides a SASLOG and LISTING for a sample regression model based on 24 cases. Page 1 in Appendix A contains the model statement (SASLOG line 30) which requests the regression of attitudes toward school (ATTSCH) on INCOME and IQ. The INFLUENCE option is requested for the model.

Page 2 in the Appendix contains the parameter estimates for the model, followed by the influence statistics. The studentized residuals (RSTUDENT) and the HAT DIAG H present the two important sources of case influence. Case 6 has the largest studentized residual (2.9823) and case 14 also has a large studentized residual (-1.5497). The DFFITS value for case 14 is (-1.5747), and this is the largest value, in absolute terms, in the sample. The negative value of DFFITS for case 14 means that the predicted Y for case 14 is increased when case 14 is deleted from the sample. Conversely, the presence of case 14 in the sample causes that case's predicted value to be reduced.

The DFBETA statistics are then presented for each regression coefficient, for each case. Case 14 is also the most influential case for estimating each of the regression parameters individually: INTERCEP DFBETA = -.5455, INCOME DFBETA = -1.4997 and IQ DFBETA = .9250. As with the DFFITS statistic, the sign of the DFBETAs indicate the direction of influence on the regression coefficients for case 14. Case 14's presence in the sample causes the y-intercept to decrease, the regression coefficient for INCOME to decrease and the coefficient for IQ to increase. On page 5 of the Appendix the regression equation is estimated with case 14 deleted from the sample, and indeed the changes in the coefficients are as suggested by the DFBETA diagnostics for case 14.

HANDLING INFLUENTIAL CASES

Once the influential cases have been identified the analyst must decide what to do with them. The first step should be to determine if the influential cases are correctly coded. Typographical errors made while entering the data can produce highly influential cases. If data errors are detected, clearly the proper course of action is to correct the data values. If the correct data values are not available then deletion of such cases is reasonable.

However, if the analyst determines that a case is correctly coded and still highly influential, three alternatives are available: 1. delete the case from the sample, 2. retain the case in the sample but note that the case is influential, or 3. revise the model to accommodate the influential case.

It is a questionable practice to delete cases from a sample simply because they are unusual. In fact, unusual cases often point to weaknesses in our models and may suggest improvements in our theories. For example, if a researcher fit a linear model to a nonlinear relationship many of the data points would be found to have large residuals and therefore might be highly influential. Deletion of unusual cases in this example would lead to the interpretation of an incorrect model. When a case is deleted from a sample it is presumed that the model is correct and the offending case is invalid. Our models should be burdened to fit our data; our data should not be obliged to fit our models. Data should not be deleted to better fit our models unless we have compelling evidence that the data is wrong.

The least squares criterion can itself be the cause of an influence problem. A case's influence is proportional to the square of its residual when OLS estimation is used. A researcher might try fitting a model using a criterion other than OLS. The SAS version 5 package has a procedure that fits models using the least absolute value error (PROC LAV). Unfortunately, this procedure is not available in version 6 of SAS. This program minimizes the sum of the absolute deviations from the model, thereby tempering the influence of high residual cases. If the coefficients estimated with OLS and LAV criteria are comparable, the model may be considered sufficiently robust for interpretation. Page 4 in the Appendix shows the LAV solution for the same model estimated earlier using OLS. The only coefficient that is changed markedly is the y-intercept. The coefficients for INCOME and IQ are approximately the same as their OLS counterparts. One might, therefore, conclude that the OLS estimates are fairly robust in this sample.

REFERENCES

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). Regression diagnostics. New York: John Wiley and Sons, Inc.

Cook, R. D. (1979). Influential observations in linear regression. Journal of the American Statistical Association, 74, 169-174.

Freund, R. J. and Littell, R. C. (1986). SAS system for regression 1986 edition. Cary, N. C.: SAS Institute Inc.

Weisberg, S. (1980). Applied linear regression. New York: John Wiley and Sons, Inc.

Figure 1. Scatter Diagram Illustating Influence

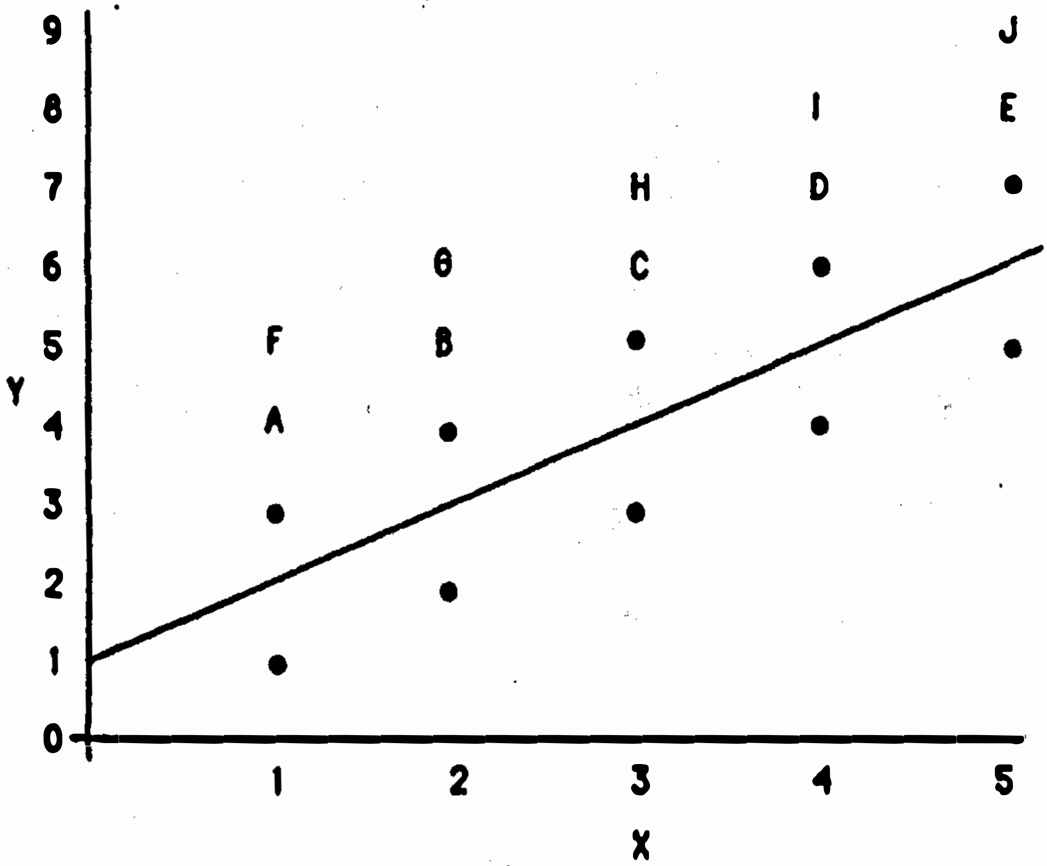

Table 1. Influence of Cases A-J on Model Coefficients

Case	Regression Coeficients		Influence of Case on	
	Intercept	Slope	Intercept	Slope
A	1.625	.846	.625	-.154
B	1.435	.913	.435	-.087
C	1.182	1.000	.182	.000
D	.913	1.087	-.087	.087
E	.692	1.154	-.308	.154
F	1.920	.769	.920	-.231
G	1.652	.870	.652	-.130
H	1.273	1.000	.273	.000
I	.870	1.130	-.130	.130
J	.538	1.231	-.462	.231

Note: The regression equation for the original 10 cases is Y' = 1 + 1X.

Appendix Page 1

·SASLOG FOR THE INFLUENCE ILLUSTRATION

```
   1 DATA ONE;
   2 OPTIONS LS = 70 NUMBER;
   3 INPUT SUBID GENDER IQ HEALTH GRADE INCOME ATTSCH;
   4 CARDS;

NOTE: DATA SET WORK.ONE HAS 24 OBSERVATIONS AND 7 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.07 SECONDS AND 64K.

  29 PROC REG;
  30 MODEL ATTSCH = INCOME IQ /INFLUENCE;
NOTE: THE PROCEDURE REG USED 0.15 SECONDS AND 416K
      AND PRINTED PAGES 1 TO 2.

  31 PROC LAV;
  32 MODEL ATTSCH = INCOME IQ;

NOTE: LAV IS NOT SUPPORTED BY THE AUTHOR OR BY SAS INSTITUTE INC.
NOTE: THE PROCEDURE LAV USED 0.16 SECONDS AND 3020K
      AND PRINTED PAGE 3.

  33 DATA TWO;
  34 SET ONE;
  35 IF SUBID NE 14;

NOTE: DATA SET WORK.TWO HAS 23 OBSERVATIONS AND 7 VARIABLES.
NOTE: THE DATA STATEMENT USED 0.04 SECONDS AND 424K.

  36 PROC REG;
  37 MODEL ATTSCH = INCOME IQ;
NOTE: THE PROCEDURE REG USED 0.10 SECONDS AND 440K
      AND PRINTED PAGE 4.
```

Appendix Page 2

DEP VARIABLE: ATTSCH
ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	4964.02906	2482.01453	42.469	0.0001
ERROR	21	1227.30427	58.44306054		
C TOTAL	23	6191.33333			

ROOT MSE	7.644806	R-SQUARE	0.8018
DEP MEAN	34.66667	ADJ R-SQ	0.7829
C.V.	22.05239		

PARAMETER EST IMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	-0.31262577	8.23978083	-0.038	0.9701
INCOME	1	1.13895467	0.16087102	7.080	0.0001
IQ	1	0.11410777	0.08624277	1.323	0.2000

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITS	INTERCEP DFBETAS
1	9.5264	1.2285	0.0864	0.9832	0.4086	-0.1528
2	-2.4677	-0.3309	0.0728	1.2295	-0.0894	-0.0266
3	-10.2098	-1.4342	0.0892	0.9476	-0.4490	-0.1037
4	14.4111	2.0796	0.0462	0.6760	0.4680	-0.0780
5	-1.3487	-0.1798	0.0811	1.2837	-0.0894	-0.0137
6	18.9446	2.9823	0.0500	0.4041	0.6841	0.0167
7	-6.4506	-0.6926	0.1181	1.1696	-0.3219	-0.0821
8	5.0227	0.6926	0.1246	1.2314	0.2616	0.2274
9	-4.0811	-0.5718	0.1360	1.2864	-0.2601	0.1846
10	2.6169	0.3459	0.0620	1.2123	0.0850	0.0619
11	-2.6141	-0.3441	0.0541	1.2023	-0.0820	-0.0190
12	2.8244	0.3846	0.1080	1.2694	0.1338	0.1125
13	-12.9959	-1.8651	0.0880	0.7771	-0.0850	0.3092
14	-8.0459	-1.5407	0.5080	1.6743	-1.5747	-0.8455
15	-3.0565	-0.4636	0.2842	1.8661	-0.2921	0.1620
16	-1.8220	-0.2494	0.1266	1.3131	-0.0850	-0.0821
17	1.4782	0.1939	0.0404	1.2110	0.0441	0.0134
18	-3.0891	-0.4588	0.2536	1.5029	-0.2674	0.1657
19	-1.9172	-0.2540	0.0746	1.2367	-0.0723	-0.0343
20	3.3009	0.4419	0.0818	1.2246	0.1319	0.1080

OBS	RESIDUAL	RSTUDENT	HAT DIAG H	COV RATIO	DFFITS	INTERCEP DFBETAS
21	7.3604	1.1207	0.2530	1.2909	0.6522	-0.3998
22	-1.5951	-0.2199	0.1402	1.3367	-0.0888	-0.0779
23	-3.6772	-0.5110	0.0495	1.1714	-0.1166	-0.0628
24	-1.1594	-0.1528	0.0605	1.2279	-0.0388	0.0140

OBS	INCOME DFBETAS	IQ DFBETAS
1	-0.2449	0.2673
2	0.0560	-0.0075
3	0.3081	-0.0658
4	-0.0875	0.1724
5	0.0345	-0.0061
6	0.2325	0.0099
7	0.2376	-0.0408
8	-0.0245	-0.1678
9	-0.0158	-0.1629
10	-0.0059	-0.0401
11	0.0363	-0.0121
12	0.0554	-0.1049
13	0.0481	-0.3650
14	-1.4997	0.9250
15	-0.1526	-0.1098
16	-0.0410	0.0778
17	-0.0150	0.0001
18	0.1655	-0.2399
19	0.0369	0.0092
20	0.0088	-0.0830
21	0.2792	0.3027
22	-0.0398	0.0745
23	-0.0249	0.0463
24	0.0111	-0.0216

Appendix Page 4

LAV REGRESSION PROCEDURE FOR DEPENDENT VARIABLE ATTSCH

VARIABLE	LAV COEFFICIENT
INTER	-0.23255814
INCOME	1.13953488
IQ	0.09302326

(NOTE: THE COEFFICIENT ESTIMATES ARE UNIQUE.)

RESIDUAL SUM OF ABSOLUTE VALUES = 120.74418605
ADJUSTED TOTAL SUM OF ABSOLUTE VALUES = 272.00000000
NUMBER OF OBSERVATIONS IN DATA SET = 24

DEP VARIABLE: ATTACH
ANALYSIS OF VARIANCE

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	PROB>F
MODEL	2	4425.92197	2212.96099	40.392	0.0001
ERROR	20	1095.73020	54.7865[illegible]		
C TOTAL	22	5521.65217			

ROOT MSE	7.401791	R-SQUARE	0.8016
DEP MEAN	33.56522	ADJ R-SQ	0.7817
C.V.	22.05197		

PARAMETER ESTIMATES

VARIABLE	DF	PARAMETER ESTIMATE	STANDARD ERROR	T FOR H0: PARAMETER=0	PROB > |T|
INTERCEP	1	4.03922071	8.45765999	0.478	0.6381
INCOME	1	1.37254205	0.21674666	6.332	0.0001
IQ	1	0.0[illegible]42	0.09724365	0.379	0.7086