Use of Regression Diagnostics in Political Science Research
Author(s): Sangit Chatterjee and Frederick Wiseman
Source: *American Journal of Political Science*, Vol. 27, No. 3 (Aug., 1983), pp. 601-613
Published by: Midwest Political Science Association
Stable URL: https://www.jstor.org/stable/2110986
Accessed: 07-01-2020 22:34 UTC

Use of Regression Diagnostics in Political Science Research

Sangit Chatterjee and Frederick Wiseman, *Northeastern University*

In a regression analysis there may be certain data points (which may or may not be outliers) that are influential in the sense that their presence or absence significantly influences the obtained values of the estimated regression coefficients. The nature of these effects needs to be analyzed in order to determine which, if any, data points should be removed from the data set in order to improve coefficient estimates. A relatively new technique for identifying influential data points is called regression diagnostics. In this presentation, the technique is discussed and its potential usefulness demonstrated by an application on a data set previously analyzed by Tufte (1974).

Regression models are frequently used to help explain and predict voting behavior. However, routine application of a regression model without regard to its assumptions and goodness of fit can lead to errors in inference. In the past, one particular area that received a considerable amount of attention was the analysis of the residuals of an estimated regression equation. Such an analysis enabled the researcher to identify outliers and may have led to the reestimation of the equation after the exclusion of all outlying points from the data set.

More recently, a second type of analysis, known as regression diagnostics, has been employed by statisticians in identifying those data points (which may or may not be outliers), which if excluded from the data set, brought about a significant change in the estimates of the unknown parameters of the specified model (Belsley, Kuh, and Welsch, 1980; Gunst and Mason, 1980).

The purpose of this presentation is to bring to the attention of opinion researchers the technique of regression diagnostics and to indicate its potential usefulness by applying the methodology to a data set previously analyzed by Tufte (1974).

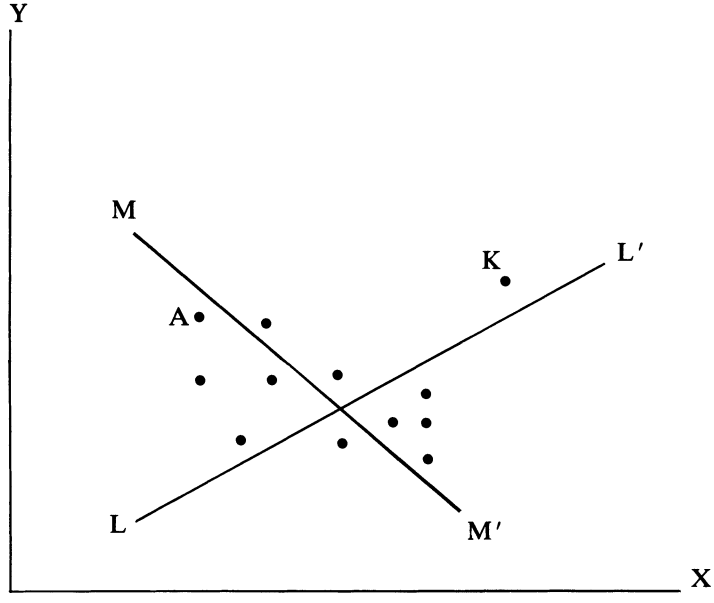
Regression Diagnostics

Regression diagnostics can be described in the following way. It is easy to understand that data points exist that have large residuals that significantly alter the estimates of the coefficients of a regression equation. However, there can be other data points that do not have large residuals, but whose inclusion in the data set may also significantly alter the coefficient estimates. This phenomenon is illustrated in Figure 1.

Figure 1 is a scatterplot for two variables (X, Y) where the line LL' is shown to represent the least squares regression line. Point A might be considered an outlier. Exclusion of A from the data set may or may not

FIGURE 1

Illustration of Outlier and Influential Data Points



alter the estimated regression line significantly from LL'. Thus, in this case, point A, though an outlier, may not be an influential point even though it has a large residual. If, on the other hand, point K, a point which is not an outlier were excluded, the least squares line could shift to MM'. In such a situation, K would be considered influential because its presence has a major effect on the estimated regression line in that it brings about a significant change in one or more of the estimated coefficient values.

The purpose of regression diagnostics is to identify influential data points and to determine the nature and extent of their effects on the estimation of the regression coefficients. There are three major indicators of potential influential data points, and each is described in the next section.

Indicators of Potential Influential Points

To simplify the discussion, we assume that there are two independent variables, X_1 and X_2, with associated regression coefficients β_1 and β_2 and estimated coefficients b_1 and b_2. Let b_1^* and b_2^* be the estimated coefficients using the complete data set minus one particular observation point K (X_{1K}, X_{2K}, Y_K). One indicator of a potential influential point is the relative measure of distance BB_K^* in Figure 2. If BB_k^* is large, then it means that the elimination of data point K from the data set brings about a substantial shift in the parameter estimates. This relative size measure known as Cook's statistic (C_K) is one indicator of a potential influential data point (Cook, 1979). Frequently, a standardized multivariate measure of distance is used for computational purposes, a practice which has been followed in this paper.

A second measure frequently referred to as Q_k is a measure of the impact of the excluded data point on the incremental sum of squares of the errors. That is, Q_k is the extra sum of squares of the errors due to the K^{th} data point, that is:

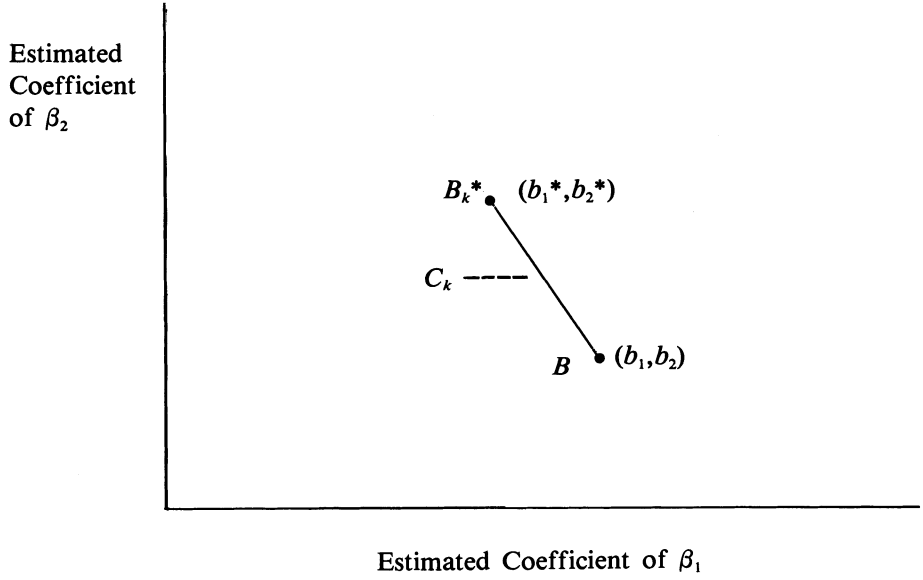
$$Q_k = \sum_{i=1}^{n} e^2_i - \sum_{\substack{i=1 \\ i \neq k}}^{n} e_i^{*2}$$

where e^2_i and e_i^{*2} are residuals based on n and $n - 1$ data points, respectively.

A high value of Q_k implies a large residual contribution associated with the excluded data point and therefore an indication that the excluded data point may be either an influential or an outlier point or both.

FIGURE 2

Measurement of Cook's Statistic—C_k



The third measure, the Andrews Pregibon statistic (AP_k), is defined in the following manner:

$$AP_k = \frac{\text{Elliptical volume } (V_2) \text{ of data set excluding a particular observation}}{\text{Elliptical volume of total data set}} = V_2 / (V_1 + V_2)$$

It is illustrated in Figure 3 using the data presented in Figure 1. The closer the ratio is to 1, the less likely it is that the excluded point is influential. The closer the ratio is to 0, the more likely the excluded point is influential (Daniel and Wood, 1980).

In summary, three measures are used for detecting potential influential data points—C_k, Q_k, and AP_k. Each measures a different aspect of the same data set, and thus they must be used together in order to properly evaluate the influential nature of a given point. Analysis of the relative sizes of these measures for all data points will indicate whether or not a particular point is likely to be an influential point and, hence, a point whose effects should be studied more closely. Application of regression diagnostics should enable researchers to analyze their data more effectively and efficiently and to provide improved estimates of regression coefficients.

It should be emphasized that the primary objective of regression diagnostics is not to exclude data points in order to maximize R^2 or to ensure support of a particular hypothesis, but rather it is to make analysts aware of particular data points that may be distorting the estimation process of the true underlying relationship between the dependent and independent variables under investigation. In addition, it should also be noted that one could experiment with more than one data point being excluded at a time. However, the corresponding combinatorial problem is difficult.

An Application

Regression diagnostics were applied to a data set previously analyzed by Tufte (1974). In this particular study, Tufte examined the relationship between

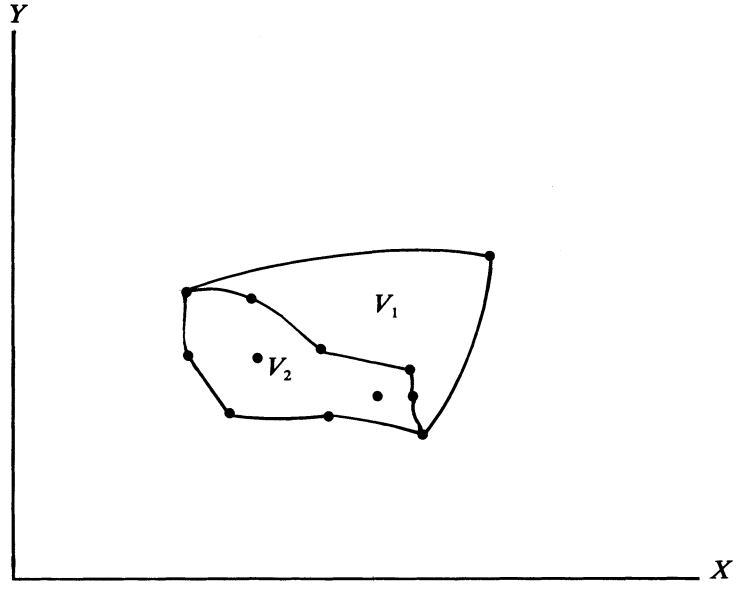
Y = current president's party standardized percentage vote loss in midterm congressional elections

and two independent variables

X_1 = current president's popularity at the time of the midterm congressional election as measured by the Gallup poll ques-

FIGURE 3

Measurement of the Andrews Pregibon Statistic



tion: "Do you approve or disapprove of the way President _____is handling his job as President?"

X_2 = economic conditions as measured by the current yearly change in real disposable income per capita.

Those interested in the rationale and development of the model should see Tufte. The data that he presents and upon which the present analysis is based are given in Table 1—data for the seven midterm congressional elections beginning in 1946 and ending in 1970.

Using standard multiple regression techniques, the regression equation is estimated to be

$$\begin{array}{l}\text{Standardized} \\ \text{percentage vote} \\ \text{loss}\end{array} = 10.37 - .123\ (\text{Gallup rating}) - .024 \begin{array}{l}(\text{Change in} \\ \text{real} \\ \text{disposable} \\ \text{income})\end{array}$$

For the above, $R^2 = .79$, the standard error of estimate = 1.30, and significant t values of 2.93 and 4.78 were obtained for the coefficients of the two independent variables, respectively.

Applying regression diagnostics to the data set produced the results that are given in Table 2. An analysis of them reveals observations 1 (1946) and 4 (1958) to be potential influential data points. The 1946 data point has a high C_k value (meaning the regression coefficients change substantially when this point is excluded) and a low AP_k value (meaning that the data point is somewhat away from the other data points). On the other hand, the 1958 data point shows an extremely high Q_k value (meaning the point has a large residual error associated with it) and a relatively high C_k value.

The specific nature of the influence of these two data points is shown in Table 3, which presents the results of seven additional regression analyses, each with one of the original data points excluded. As seen in this table and as previously indicated by the regression diagnostics, the one point that is exerting the greatest influence on the estimated coefficients is the initial data point, that is, the one for the least recent year—1946. Eliminating this data point brings about substantial changes in the constant term as well as the two estimated coefficients. More specifically, the constant term increases by 70 percent, and the estimates of β_1 and β_2 decrease by 100 percent and 71 percent, respectively. Thus, the effect of a unit change in either presidential popularity or real disposable income would be estimated to be almost twice as large if a decision was made to eliminate this point from the data set.

TABLE 1

The Data Set

Year	Mean Congressional Vote for Party of Current President in Last Eight Elections (%)	Nationwide Congressional Vote for Party of Current President (%)	Standardized Vote Loss (Y) (%)	Gallup Poll Rating (X_1) (%)	Current Yearly Change in Real Disposable Income per Capita (1972 Dollars) (X_2)
1946	Democratic 52.57	45.27	7.30	32	−71
1950	Democratic 52.04	50.04	2.00	43	133
1954	Republican 49.79	47.46	2.33	65	−15
1958	Republican 49.83	43.91	5.92	56	−14
1962	Democratic 51.63	52.42	−.79	67	71
1966	Democratic 53.06	51.33	1.73	48	122
1970	Republican 46.66	45.68	.98	56	104
1974	Republican 48.29	44.15	4.14	55	−89
1978	Democratic 52.23	53.07	−.84	49	164

SOURCE: Tufte (1974), the Gallup organization, and the *Economic Report of the President* (1980). Our economic data differ slightly from Tufte's because for X_2 we use yearly change in real disposable income (1972 dollars).

TABLE 2

Results of Regression Diagnostics

Data Point Excluded	Indicators (High) C_k	(High) Q_k	(Low) AP_k
1 (1946)	28.80	.49	.37
2 (1950)	1.58	.16	.79
3 (1954)	.01	.00	.77
4 (1958)	15.80	6.54	.93
5 (1962)	11.88	1.08	.76
6 (1966)	.94	.23	.90
7 (1970)	.18	.08	.94

[a] C_k values are standardized.

The effect of the second potential influential data point (4), 1958, is that it increases the relative importance attached to presidential popularity (X_1) and decreases the relative importance associated with economic conditions (X_2). Also of interest is that eliminating this data point results in a substantial increase in R^2. R^2 without the data point jumps to .96, a 22 percent increase.

Discussion

The regression diagnostics have indicated that two points are exerting the greatest influence on the coefficient estimates. The diagnostics do not tell the researcher that coefficient estimates can be improved or better predictions can be made if either of these points were excluded, but rather they indicate that the influence of these points should be examined further to determine which data set should be used for estimation purposes. In the present example, it is possible to conduct an objective validity test by forecasting the standardized vote loss in the 1974 and the 1978 midterm congressional elections using the following three alternative data sets:

Data set 1: Includes all seven data points for years 1946–1970;

TABLE 3

Results of Regression Analyses Eliminating One Data Point

Data Point Excluded	Constant	b_1	b_2	R^2	Standard Error
None	10.37	−.123	−.024	.79	1.30
1	18.42	−.245	−.041	.90	.71
2	10.18	−.119	−.025	.76	1.50
3	9.93	−.111	−.026	.77	1.47
4	10.54	−.138	−.019	.96	.52
5	9.04	−.093	−.024	.78	1.22
6	10.24	−.121	−.025	.76	1.49
7	10.37	−.123	−.024	.75	1.50

Data set 2: Includes the six data points for years 1950–1970 (excludes data point 1—1946); and

Data set 3: Includes the six data points for years 1946–1954, 1962–1970 (excludes data point 4—1958).

In 1974,

Y_{1974} = Standardized percentage vote loss = 4.14

X_1 = Gerald Ford's popularity rating = 55

X_2 = Change in real disposable income = -89

and in 1978,

Y_{1978} = Standardized percentage vote loss = -.84

X_1 = Jimmy Carter's popularity rating = 49

X_2 = Change in real disposable income = 164.

TABLE 4

Standardized Percentage Vote Loss

Data Set	1974			1978			Average Absolute Deviation
	Predicted	Actual	Absolute Deviation	Predicted	Actual	Absolute Deviation	
All data points	5.74	4.14	1.60	.41	−.84	1.25	1.42
All data points except 1946	8.59	4.14	4.45	−.31	−.84	.53	2.49
All data points except 1958	4.64	4.14	.50	.66	−.84	1.50	1.00

Substituting the above values into the appropriate regression equations (see Table 2) produces the results shown in Table 4. Examination of Table 4 indicates that no clear-cut "winner" was obtained among the three data sets. Exclusion of the 1958 data point provided the best prediction for 1974, but the worst prediction in 1978. The opposite was true for the 1946 data point. The data set using all data points placed second in both instances. With respect to average absolute deviation, exclusion of 1958 resulted in the lowest average error.

Which model should be used for 1982? For this purpose, regression diagnostics was reapplied to the extended data set for the years 1946–1978. Results were virtually identical, as in the previously discussed case, the 1958 and 1946 points once again being identified as potential influential points. The analyses conducted indicate that the Tufte model performs well and that the only question is upon which data set the estimates should be based. The choice is left to the analyst. In this instance, we would prefer to use the data set that excludes the 1958 point because (1) it yields the highest R^2 value, (2) the point has a large residual error associated with it, and (3) its elimination produces predictions that have the smallest average absolute error. Thus, using the data set 1946–1954 and 1962–1978, we obtain the estimated regression function to be

$$\begin{array}{lll} \text{Standardized} & & \text{(Change in} \\ \text{percentage vote} & = 10.20 - .135\ \text{(Gallup rating)} - .021 & \text{real} \\ \text{loss} & & \text{disposable} \\ & & \text{income)} \end{array}$$

Conclusion

Regression models based on data points that do not include outlier points, influential points, or both are more robust models. Outliers can be detected by large residuals, and several methods can be used to detect potential influential data points. In the example presented, two observations were found to exert considerable leverage on the coefficient estimates. Their suggested treatment in the model was guided by an independent validation test. The example illustrates the need and value of regression diagnostics, and we believe its use by political science researchers will result in improved data analysis procedures.

Finally, this discussion warrants a word on computation techniques of the regression diagnostics. The cost of computation with large data sets can increase dramatically if one does not use the techniques of sequential updating of regression parameters (Plackett, 1950). A program utilizing this techniques for computing regression diagnostics has been written by the authors and will soon be available. In addition, it should be noted that some statistical packages are now providing a menu of

diagnostics in order to better evaluate estimated regression models. For example, MINITAB now prints out the presence of possible influential data points.

Manuscript submitted 30 September 1982
Final manuscript received 18 December 1982

REFERENCES

Belsley, David A., Edwin Kuh, and Roy E. Welsch. 1980. *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Cook, R. Dennis. 1979. "Detection of influential observations in linear regression." *Journal of the American Statistical Association*, 74 (March): 169–174.

Daniel, Cuthbert, and Frederick Wood. 1980. *Fitting equations to data*. 2d ed. New York: John Wiley & Sons.

Economic Report of the President. 1980. Washington, D.C.: U.S. Government Printing Office.

Gunst, Richard F., and Robert L. Mason. 1980. *Regression analysis and its applications: A data oriented approach*. New York: Marcel Dekker.

Plackett, Ronald L. 1950. "Some theorems in least squares." *Biometrika*, 37 (1): 149–157.

Tufte, Edward R. 1974. *Data analysis for politics and policy*. Englewood Cliffs, N.J.: Prentice-Hall.