

SENSITIVITY ANALYSIS OF THE MSAE REGRESSION RESULT: POST-OPTIMALITY ANALYSIS OF THE LHS COEFFICIENTS

John F. Wellington

Richard T. Doermer School of Business and Management Sciences
Indiana University – Purdue University Fort Wayne
2101 E. Coliseum Blvd., Fort Wayne, IN 46805
260-481-6061 (voice), wellingj@ipfw.edu

ABSTRACT

In this paper, we present a technique for assessing the sensitivity of the optimal solution to a widely encountered problem in data analysis. The problem of estimating the parameters of the single equation linear regression model under the minimum sum of absolute errors (MSAE) criterion has a well known linear programming (LP) formulation. We propose post-optimality analysis of the MSAE solution that allows the investigator to assess impact of unforeseen changes in the data used for formulating the LP problem (fitting the model). It is a sensitivity analysis of the technical or left-hand side (LHS) coefficients of the non-binding constraints of the LP formulation for the MSAE problem.

Keywords: Curve fitting, linear programming, linear regression, mathematical programming/optimization, minimum sum of absolute errors regression

INTRODUCTION

The need to assess the sensitivity of the widely used minimum sum of absolute errors (MSAE) solution to the single equation linear regression problem drove the analysis proposed in this paper. In some experimental situations in which the linear regression model is used for data modeling purposes, a subset of the data within a given observation may have been collected and/or recorded using a device or methodology subject to error. For a variety of reasons, the occurrence, magnitude, or impact of the error may not be known at the time of data collection or model estimation. Consequently, in these situations it is useful to assess the tolerance of the modeling result to variations in the data used for fitting. Ideally, the analyst would like the modeling result to be robust for a wide variety of alternate values of the data used for fitting. When only a very narrow interval of alternate values is possible, the analyst's attention should be drawn to the circumstances of data collection and recording for the affected observation(s). An allowable maximal variation of say 0.2% in the value of each datum of a certain observation may move the analyst to investigate the experimental conditions under which the data were collected whereas 10% may not. The ability to identify data in the sample that has the potential to significantly change the MSAE solution is valuable to the analyst. Unlike the MSAE result, the popular least squares regression solution has no tolerance to variation in the data used for model fitting. In the linear programming (LP) formulation of the MSAE problem, each observation is

treated as a constraint in which its left hand side (LHS) coefficients are the data values of the independent (predictor, x) variables and its right hand side (RHS) is the value of the corresponding dependent (response, y) variable. See Figure 1 (b) below.

The following illustration makes use of data from Draper and Stoneman [1] that is further analyzed in Narula and Wellington [2]. The data appear in Table 1.

TABLE 1
Wood Beam Data

Obs. i	1	2	3	4	5	6	7	8	9	10
x_{i1} ¹	1	1	1	1	1	1	1	1	1	1
x_{i2}	0.499	0.558	0.604	0.441	0.550	0.528	0.418	0.480	0.406	0.467
x_{i3}	11.10	8.90	8.80	8.90	8.80	9.90	10.70	10.50	10.50	10.70
y_i	11.14	12.74	13.13	11.51	12.38	12.60	11.13	11.70	11.02	11.41

¹ Indicates that the intercept or constant term is included in the model.

In these studies, the data was fitted to the linear regression model

$$\hat{y}_i = b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} \quad (1)$$

under the MSAE criterion with the result

$$\hat{y}_i = 9.084 x_{i1} + 9.189 x_{i2} - 0.171 x_{i3} \quad (2)$$

where \hat{y}_i denotes the predicted value of y_i corresponding to values of the three predictor variables x_{i1} , x_{i2} , and x_{i3} , $i=1, \dots, 10$. The MSAE result, parameter estimates 9.084, 9.189, and -0.171, were obtained from the solution to the following LP problem

$$\text{Minimize } e_1 + e_2 + e_3 + e_4 + e_5 + e_6 + e_7 + e_8 + e_9 + e_{10} \quad (3)$$

$$\begin{aligned} \text{subject to} \quad & b_1 + 0.499b_2 + 11.1b_3 + e_1 = 11.14 \\ & b_1 + 0.558b_2 + 8.9b_3 + e_2 = 12.74 \\ & b_1 + 0.604b_2 + 8.8b_3 + e_3 = 13.13 \\ & b_1 + 0.441b_2 + 8.9b_3 + e_4 = 11.51 \\ & b_1 + 0.550b_2 + 8.8b_3 + e_5 = 12.38 \\ & b_1 + 0.528b_2 + 9.9b_3 + e_6 = 12.60 \\ & b_1 + 0.418b_2 + 10.7b_3 + e_7 = 11.13 \\ & b_1 + 0.480b_2 + 10.5b_3 + e_8 = 11.70 \\ & b_1 + 0.406b_2 + 10.5b_3 + e_9 = 11.02 \\ & b_1 + 0.467b_2 + 10.7b_3 + e_{10} = 11.41 \end{aligned}$$

b_1 , b_2 , b_3 and e_1 , e_2 , e_3 , e_4 , e_5 , e_6 , e_7 , e_8 , e_9 , e_{10} unrestricted in sign.

The MSAE errors, $e_i = y_i - \hat{y}_i$, associated with (2) and available from the solution to (3) are e_i , $i=1, \dots, 10\} = \{-0.632, 0.050, 0, -0.105, -0.254, 0.356, 0.034, 0, 0, -0.136\}$.

It is well known that the MSAE regression plane passes through at least k observations where k is the number of parameters to be estimated in the single equation linear regression model. These observations have zero error or residual value and are known as the defining observations; those with non-zero residual values are known as the non-defining observations; see Figure 1 (a) below. The MSAE result (b_1, b_2, b_3) is completely determined by the defining observations. Any variation in their x -data or y -data will change the MSAE result. Consequently, they are not addressed in the methodology proposed in this paper. However, the MSAE result is tolerant to variation among the data of the non-defining observations. The post-optimality analysis is also known as ranging or interval analysis, see Narula and Wellington [2]. For the wood beam data, $k=3$.

The distinction between defining and non-defining observations is central to the analysis that follows. For the wood beam data, the defining observations are denoted by indices $i=3,8,9$ and the non-defining by indices $i=1,2,4,5,6,7,10$. With knowledge of the defining observations and because their residuals are zero, the MSAE values of b_1, b_2, b_3 for the wood beam data may be determined from

$$\begin{aligned} b_1 + 0.604b_2 + 8.8b_3 &= 13.13 & (i=3) \\ b_1 + 0.480b_2 + 10.5b_3 &= 11.70 & (i=8) \\ b_1 + 0.406b_2 + 10.5b_3 &= 11.02 & (i=9). \end{aligned} \quad (4)$$

CURRENT PRACTICE - VARIATION IN ONE DATUM (LHS COEFFICIENT)

Table 2 is a display of the allowable variation in each x -datum of each non-defining observation that maintains the solution to (3) assuming no other changes in the original data – the *ceteris paribus* qualification.

For x_{13} , the value of predictor variable 3 in observation 1, if an alternate value of interest under any experimental condition were in the interval $[10.956, 11.832]$, the MSAE model (2) would be unchanged, *ceteris paribus*. However, if the impact of alternate value $x_{13} = 10.6$ was interesting to the analyst, the results of Table 2 show that the MSAE model (2) under this variation would not hold. In this case, the optimal solution to (3) would change and the MSAE model would become

$$\hat{y}_i = 8.243 x_{i1} + 9.864 x_{i2} - 0.122 x_{i3} \quad i = 1, \dots, 10 \quad (5)$$

(9.084) (9.189) (-0.171)

where the (\bullet) denotes the original value of the indicated MSAE estimate. Note that the variation of -0.5 in the original value of x_{13} ($=11.1$) changed the MSAE model whereas the variation of $+0.5$ ($=11.6$) would leave the result given in (2) unchanged. In general, the intervals are not symmetric with respect to the original value of the datum. Observe also that the value ($=9.9$) of predictor variable 3 in observation 6 is close to the upper bound ($=10.044$) of allowable variation in x_{63} that maintains the result given in (2). The analyst may want to confirm its accuracy; otherwise the MSAE model could be different.

TABLE 2
Alternate Values of the Predictor Variables (x) for the Non-defining Observations That Maintain the MSAE Fit for the Wood Beam Data – Variation in Individual Datum

Obs. i	Lower Bound	x_{i2} Datum	Upper Bound	Lower Bound	x_{i3} Datum	Upper Bound	y_i Datum
1	0.488	0.499	0.584	10.956	11.1	11.832	11.14
2	0.473	0.558	0.563	8.609	8.9	9.044	12.74
3 ¹	-	0.604	-	-	8.8	-	13.13
4	0.431	0.441	0.526	8.756	8.9	9.515	11.51
4	-	-	-	11.285	8.9	12.038	11.51
5	0.539	0.550	0.635	8.656	8.8	9.532	12.38
6	0.443	0.528	0.539	9.168	9.9	10.044	12.60
7	0.333	0.418	0.422	10.502	10.7	10.844	11.13
8 ¹	-	0.480	-	-	10.5	-	11.70
9 ¹	-	0.406	-	-	10.5	-	11.02
10	0.326	0.467	0.361	10.556	10.7	11.432	11.41
10	0.456	0.467	0.552	-	-	-	11.41

¹Defining observation. Hence, variation in any datum would change the MSAE parameter estimates.

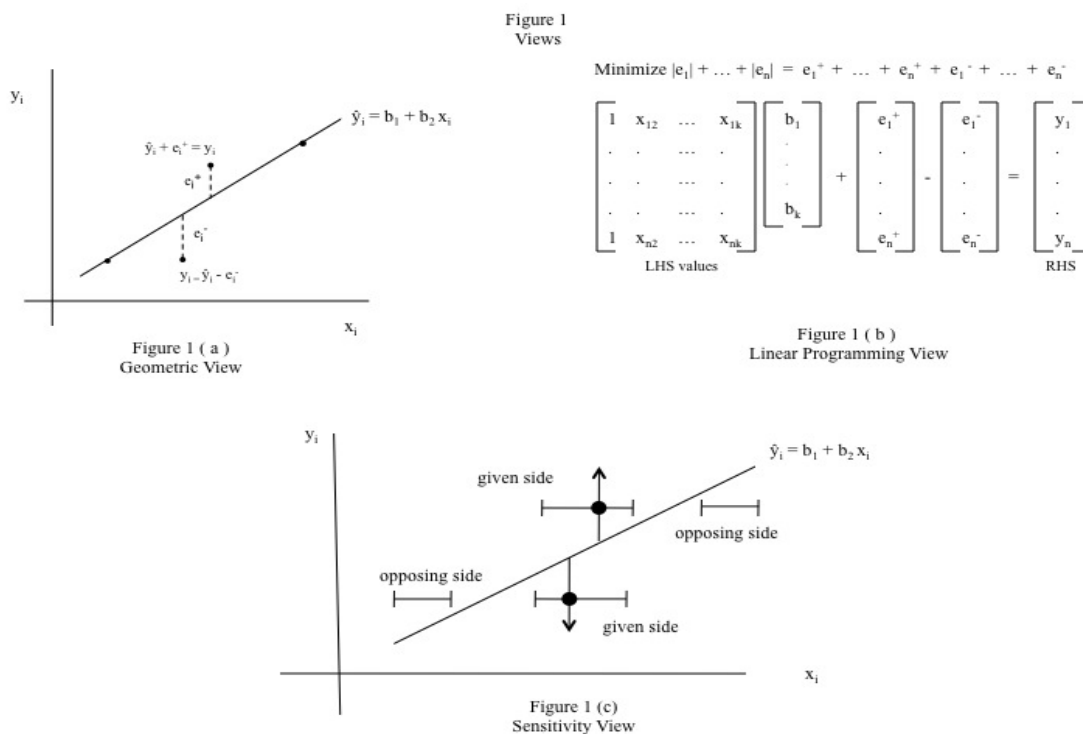
For alternate values of predictor variable 3 in observation 4, there are two disjointed intervals. Ceteris paribus, an alternate value for x_{43} could be in the intervals [8.756, 9.515] or [11.285, 12.032] and produce the MSAE model given in (2). Alternate values for predictor variable 2 in observation 10 also occur in two disjointed intervals. This is an unusual feature of sensitivity analysis for a LP problem.

Table 3 is taken from Narula and Wellington [2] and displays the results of interval analysis under the ceteris paribus assumption for the values of the response variable (y) among the wood beam data. Note the relationship between the finite value of the lower or upper bounds and the value as well as sign of the corresponding MSAE residual for the non-defining observations. In each case, the finite bound is the predicted value of the response variable obtained from the fitted MSAE model (2). This feature holds in general for MSAE regression and is the basis for the statement that the value of the response variable (y) for each non-defining observation can change indefinitely as long as its residual value does not change sign, ceteris paribus. This means geometrically that the observation under the variation of interest in the y-direction does not cross to the opposing side of the fitted MSAE regression plane. Although the variation in the value of the response variable for each non-defining observation can change indefinitely in one direction, the change in the value of a predictor variable (x) for any non-defining observation always has real bounds. See Figure 1 (c) below. The insensitivity of the MSAE fit to variation in the x-data or y-data of the non-defining observations helps in understanding what statisticians call the robustness of the MSAE regression model.

TABLE 3
Alternate Values of the Response Variable (y) for the Non-defining Observations That Maintain the MSAE Fit of the Wood Beam Data - Variation in Individual Datum

Obs. i^1	Lower Bound	y_i Datum	Upper Bound	MSAE Residual
1	$-\infty$	11.14	11.772	-0.632
2	12.690	12.74	∞	0.050
4	$-\infty$	11.51	11.615	-0.105
5	$-\infty$	12.38	12.634	-0.254
6	12.244	12.60	∞	0.356
7	11.096	11.13	∞	0.034
10	$-\infty$	11.41	11.546	-0.136

¹Defining observations 3,8,9 omitted. Any variation in their values will change MSAE regression plane.



PROPOSED METHODOLOGY – THE CASE OF SIMULTANEOUS CHANGES

The results given in Table 2 apply for variation in one x-datum ceteris paribus. They are not suitable for simultaneous variations among the x-data of a non-defining observation of interest and as such are restrictive in many situations. Table 4 is a display of the allowable simultaneous percentage variation among the values of the predictor variables in each non-defining observation of the wood beam data set. Under the experimental conditions that prevailed when the data of observation 2 were collected and recorded, alternate values of x_{22} and x_{23} no smaller than 5.345% of the originals, 0.558 and 8.9 respectively, would leave model (2) unchanged. Alternate values no greater than 0.872% of the same would also leave the results unchanged.

TABLE 4

Alternate Values of the Predictor Variables (x) for the Non-defining Observations That Maintain the MSAE Fit for the Wood Beam Data – Simultaneous Percentage Variations

Obs. ¹	Original X ₂ Datum	Original X ₃ Datum	Percentage Simultaneous Changes	
			Allowable Decrease	Allowable Increase
1	0.499	11.1	0.805	4.757
2	0.558	8.9	5.345	0.872
4	0.441	8.9	0.966	5.769
5	0.550	8.8	0.883	5.412
6	0.528	9.9	5.070	0.842
7	0.418	10.7	5.122	0.879
10	0.467	10.7	0.844	4.975

¹ Defining observations 3,8,9 omitted. Any variation in their values will change the MSAE regression plane.

If the values for x_{22} (=0.558) and x_{23} (=8.9) were simultaneously recorded by a common device that given the prevailing experimental conditions could cause each datum to be inaccurately recorded, the analyst has a useful diagnostic to assess the model's tolerance for error from this source. Based on past history, if the analyst believes the device or any other circumstance of the data collection and recording have the capacity to overstate (understate) by more than 0.872% (5.345%), then caution should be exercised in using the model. In this case, the analyst may want to investigate the state of the experimental conditions at the time of data collection for observation 2 and any other observation with what is perceived to be small allowable variation.

SUMMARY

In the post-estimation analysis of the MSAE regression model, if it is found that the original data used for fitting should be revised, corrected, or otherwise different, the analyst now knows where (which observation) and to what extent (magnitude of allowable increase or decrease) the model may be affected (sensitive or insensitive). Details of proposed analysis to appear in future paper.

REFERENCES

- [1] Draper, N. R. and Stoneman, D. M. *Testing for the Inclusion of Variables in Linear Regression by a Randomization Technique*. Technometrics, 1966, 8, 695-699.
- [2] Narula, S. C. and Wellington, J. F. *Sensitivity Analysis for the Predictor Variables in the MSAE Regression*. Computational Statistics and Data Analysis, 2002, 40, 355-373.