MATHÉMATIQUES ET SCIENCES HUMAINES

J. K. LINDSEY

An application of analysis of covariance to response surface

Mathématiques et sciences humaines, tome 50 (1975), p. 31-38

<http://www.numdam.org/item?id=MSH_1975__50__31_0>

Math. Sci. hum. (13e année, n°50, 1975, p.31-38)

AN APPLICATION OF ANALYSIS OF COVARIANCE TO RESPONSE SURFACE

J.K. LINDSEY *

A useful refinement of multiple regression, when few independent variables are present, exists which is called response surface analysis (Lindsey, 1972). This method has a number of advantages : (1) the relationship between the dependent and independent variables may be non-linear, (2) with the use of power transformations, the dependent variable need not necessarily be distribution according to a normal distribution, and (3) the relationship between the mean of the dependent variable and various values of the independent variables can be presented visually in terms of a "contour map". For example, one may study how mean achievement on some test at school varies with class size and hours of instruction per week.

However, this response surface model is limited by the necessity that all variables be metric. By a combination of this response surface methodology with analysis of covariance procedures, this limitation can be overcome. The blocks of the analysis of covariance become the values of a discrete independent variable. Hence, we can study the dependence of mean achievement on social class (as discrete categories) simultaneously with class size and hours of instruction. Hence, for this statistical model, we have one continuous dependent variable, which may or may not be normally distributed, and a small number of independent variables, some of which are continuous and some nominal (discrete).

---

* G.E.M.A.S., Maison des Sciences de l'Homme, 54 bd Raspail 75006 PARIS.
Present address : International Institute of Educational Planning, 7-9, rue Eugène Delacroix 75016 PARIS.

This method may find many uses in the social sciences where so many variables are in fact nominal. Examples are given in Cherkaoui and Lindsey (1974), and Lindsey (1974). The traditional scale construction can in this way be avoided. In addition, the flexibility of the model allows a much more close representation both of the theoretical and the empirical reality.

Suppose that the mathematical model for a response surface with two independent variables is

$$\mu = E(Y) = \beta_0 + \sum_{k=1}^{5} \beta_k x_k \qquad (1)$$

where $x_3 = x_1^2$, $x_4 = x_2^2$ and $x_5 = x_1 x_2$ and μ is the mean of a normal distribution. The extension to power transformations, as described by Lindsey (1972), is direct in all that follows. Let i label the I blocks and j the N_i observations in each block i. Three possible models are available : (A) all β coefficients (and power transformations) vary with the block i :

$$\mu_{ij} = E(Y_{ij}) = \beta_{i0} + \sum_{k=1}^{5} \beta_{ik} x_{ijk} \qquad (2)$$

(B) only β_0 varies with the block so that the same shape of response surface is found in each block, but the height varies (no interaction between blocks and surfaces) :

$$\mu_{ij} = E(Y_{ij}) = \beta'_{i0} + \sum_{k=1}^{5} \beta'_k x_{ijk} \qquad (3)$$

and (C) no parameters vary among blocks so that the response variable does not depend on the block variable, at least when the response surface variables are allowed for :

$$\mu_{ij} = E(Y_{ij}) = \beta_0 + \sum_{k=1}^{5} \beta_k x_{ijk} \qquad (4)$$

The usual sums of squares and cross-products for estimating the coefficients are, for each block of model (A)

$$XX_{Aikl} = \sum_j (x_{ijk} - \bar{x}_{i.k})(x_{ijl} - \bar{x}_{i.l})$$

and

$$XY_{Aik} = \sum_j (x_{ijk} - \bar{x}_{i.k})(y_{ij} - \bar{y}_{i.})$$

for all blocks of model (B),

$$XX_{Bkl} = \sum_i XX_{Aikl}$$

and $$XY_{Bk} = \sum_i XY_{Aik}$$

and for model (C),

$$XX_{Ckl} = \sum_i \sum_j (x_{ijk} - \bar{x}_{..k})(x_{ijl} - \bar{x}_{..l})$$

and $$XY_{Ck} = \sum_i \sum_j (x_{ijk} - \bar{x}_{..k})(y_{ij} - \bar{y}_{..})$$

The following quantities must be calculated for the F tests among the three models :

$$Q_1 = \sum_i \sum_{k=1}^{5} \hat{\beta}_{ik} XY_{Aik}$$

$$Q_2 = \sum_{k=1}^{5} \hat{\beta}'_k XY_{Bk}$$

$$Q_3 = \sum_{k=1}^{5} \hat{\beta}_k XY_{Ck}$$

$$SS_1 = \sum_i \sum_j (y_{ij} - \bar{y}_{i.})^2$$

$$SS_2 = \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2$$

where $\hat{\beta}_{ik}$ $\hat{\beta}'_k$ and $\hat{\beta}_k$ are the maximum likelihood estimates for models A, B, and C, respectively. The F tests, which are derivable directly from the likelihood ratios, as shown by Lindsey (1972), are

$$F_{AB} = \{N_. - I(K+1)\} \{Q_1 - Q_2\} / \{K(I-1)\} \{SS_1 - Q_1\}$$

for testing if model A may be replaced by the simpler model B, and

$$F_{BC} = (N_. - I - K)(SS_2 - Q_3 - SS_1 + Q_2) / (I-1)(SS_1 - Q_2)$$

for testing if model B may be replaced by C. The degrees of freedom are as given in the two expressions.

Student's t test may be used to determine if individual parameters in a given model are significantly different from zero. For the three models, they are respectively

$$t_{Aik} = |\hat{\beta}_{ik}| / \sqrt{XX^{I}_{Aik}(SS_{1i} - Q_{1i}) / (N_i - K - 1)}$$

$$t_{Bk} = |\hat{\beta}_{k}| / \sqrt{XX^{I}_{Bk}(SS_{1} - Q_{2}) / (N_{\cdot} - K - I)}$$

$$t_{Ck} = |\hat{\beta}_{k}| / \sqrt{XX^{I}_{Ck}(SS_{2} - Q_{3}) / (N_{\cdot} - K - 1)}$$

where XX^{I}_{Aik} is the k^{th} diagonal element of the inverse of the matrix XX_{Ai}, etc., and SS_{1i} and Q_{1i} are the i^{th} terms in the sums SS_1 and Q_1.

To illustrate one way in which the procedure may be applied, an example using non-experimental data will be given. A reanalysis, reported in Lindsey (1974), has been made of the data of Husén (1967) with particular attention being made to the dependence of mathematics score for 13 year olds on the number of students in the class and the hours of mathematics instruction per week. Three other discrete variables, social class with five values, programme in the school with three values, and sex with two values, are to be introduced to determine under what conditions this dependence changes. Thus, we start with model A and if it is significantly better than B, we examine the regression coefficients within each block and choose two sets of blocks within which the coefficients are most similar, and which are also sociologically meaningful. The analysis is reapplied to each set and the process continued until sets of blocks are found in which either model B or C is acceptable.

Since the variable for sex is binary (with 1 for boys, 2 for girls), it may be introduced directly into the regression model as x_3

$$\mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \beta_6 x_3 + \beta_7 x_1 x_3 + \beta_8 x_2 x_3$$
$$+ \beta_9 x_1^2 x_3 + \beta_{10} x_2^2 x_3 + \beta_{11} x_1 x_2 x_3$$

Although a two-way analysis of covariance might be applied with the two remaining discrete variables, to simplify the analysis, a one-way analysis using a fifteen value composite variable is used. When a set of blocks has been found for which either model B or C is acceptable, a step-wise procedure has been applied to eliminate non-significant coefficients. In addition, since the data have been found to be extremely non-normal, a transformation of the response variable,

$$(y + \lambda_1)^{\lambda_2}$$

has been estimated. These parameters are re-estimated at each step. No transformation has been applied to the independent variables because of limitations of computing time available. Of the seven countries analyzed, the final results for England are given in Table 1. Both the response surfaces and the (non-normal) distributions have been plotted in Lindsey (1974).

The first division of the blocks is between the academic and the two non-academic programmes. With the latter, the shape of the surface is the same for all blocks, but the height varies both with programme (vocational and general) and with social class (model B). The next split, within the academic programme, is between the executive/administrative and professional/technical social classes and the small proprietor, white collar, and manual worker classes. Within the former, model B holds, but the latter must be split, with manual workers separate. Small proprietor and white collar children have exactly the same surface (model C). One notes that the power transformation is the same ($\hat{\lambda}_2 = 1.8$) for all sets within the academic programme, although the variance changes considerably, but that it is much different ($\hat{\lambda}_2 = 0.6$) for the other two programmes. Although it is not evident from the parameter values, the shape of the response surface is similar for manual workers in the academic programme and for all social classes in the other two programmes. In all cases a significant interaction between sex and the response surface appears, so that, in effect, we have not four but eight sets of blocks, i.e. eight differently shaped response surfaces.

The parameter values given in Table 1. are interpreted in the same way as with traditional multiple regression. However, care must be taken since more than one coefficient refers to the same independent variable because

of non-linearity and interaction effects. For example, in Table 1a, coefficients β_1, β_3, β_5 and β_{11} all refer to the number of students in the class. For this reason, it is extremely difficult to interpret the model without the visual representation given in Lindsey (1974).

Acknowledgements

The author would like to thank the International Association for the Evaluation of Educational Achievement for allowing access to the raw data used in the example and the members of the Groupe d'Etude des Méthodes d'Analyse sociologique, especially M. Cherkaoui, for their support.

REFERENCES

CHERKAOUI, M. & LINDSEY, J.K., "Le poids du nombre dans la réussite scolaire", Rev. fr. socio., 15 (1974), 201-215.

HUSEN, T., International Study of Achievement in Mathematics, vol. I & II, New York, Wiley, 1967.

LINDSEY, J.K., "Fitting response surfaces with power transformations", Jr. Roy. Statist. Soc. C, 21 (1972), 234-247.

LINDSEY, J.K., "A reanalysis of class size and achievement as interacting with four other critical variables from the IEA mathematics study", Comp. Educ. Rev., 18 (1974), 314-326.

Table 1. Response surface analysis of the data from England for the school year having the most thirteen year olds (Husén 1967). When β_0 has three subscripts (model B), the first is the social class (1 : executive/administrative, 2 : professional/technical, 3 : small proprietor, 4 : white collar, 5 : manual worker) and the second is the programme (1 : academic, 2 : general, 3 : vocational). The regression variables are : x_1 : number of students in the class, x_2 : hours of mathematics instruction per week, x_3 : sex.

Table 1a. Academic programme for executive/administrative and professional/technical social classes ($N_. = 502$)

Transformation : $\hat{\lambda}_1 = 15.0$, $\hat{\lambda}_2 = 1.8$ $\hat{\chi}^2_2 = 36.50$

Variance : $\hat{\sigma}^2 = 302630.2$

F tests : model A vs B ($\beta_{ik} = \beta'_k$) : $F_{7,486} = 0.529$

model B vs C ($\beta'_{i0} = \beta_0$) : $F_{1,493} = 4.114$

			t
$\hat{\beta}'_{110} = 994.942$	x_1 : $\hat{\beta}'_1$ =	104.391	3.45
$\hat{\beta}'_{210} = 1\,119.447$	x_1^2 : $\hat{\beta}'_3$ =	- 1.457	2.91
	x_2^2 : $\hat{\beta}'_4$ =	46.513	2.38
	x_1x_2 : $\hat{\beta}'_5$ =	- 13.091	3.01
	x_3 : $\hat{\beta}'_6$ =	-549.713	2.74
	$x_2^2x_3$: $\hat{\beta}'_{10}$ =	- 20.774	1.82
	$x_1x_2x_3$: $\hat{\beta}'_{11}$ =	5.036	2.28

Table 1b. Academic programme for small proprietor and white collar social classes ($N_. = 491$)

Transformation : $\hat{\lambda}_1 = 12.2$, $\hat{\lambda}_2 = 1.8$, $\chi^2_2 = 49.83$

Variance : $\hat{\sigma}^2 = 270628.7$

F tests : model A vs B ($\beta_{ik} = \beta'_k$) : $F_{4,481} = 1.006$

model B vs C ($\beta'_{i0} = \beta_0$) : $F_{1,485} = 0.395$

			t
$\hat{\beta}_0 = 260.206$	x_2 : $\hat{\beta}_2$ =	641.658	4.04
	x_1^2 : $\hat{\beta}_3$ =	1.569	4.28
	x_1x_2 : $\hat{\beta}_5$ =	-20.435	3.71
	x_2x_3 : $\hat{\beta}_8$ =	-68.124	4.86

Table 1c. Academic programme for the manual worker social class (N = 542)

Transformation : $\hat{\lambda}_1 = 23.4$ $\hat{\lambda}_2 = 1.8$ $\chi^2_2 = 38.69$

Variance : $\hat{\sigma}^2 = 374849.3$

			t
$\hat{\beta}_0 = -712.465$	x_2 : $\hat{\beta}_2$ =	1189.525	4.80
	x_2^2 : $\hat{\beta}_4$ =	-154.608	5.73
	x_1x_2 : $\hat{\beta}_5$ =	4.973	3.52
	x_3 : $\hat{\beta}_6$ =	685.964	2.01
	x_2x_3 : $\hat{\beta}_8$ =	-242.874	2.46

Table 1d. General and vocational programmes for all social classes (N = 1475)

Transformation : $\hat{\lambda}_1 = 8.7$, $\hat{\lambda}_2 = 0.6$, $\chi^2_2 = 116.01$

Variance : $\hat{\sigma}^2 = 6.0714$

F tests : model A vs B ($\beta_{ik} = \beta'_k$) : $F_{45,1415} = 1.295$

model B vs C ($\beta'_{i0} = \beta_0$) : $F_{9,1460} = 24.258$

				t
$\hat{\beta}'_{120} = 8.723$	$\hat{\beta}'_{130} = 7.809$			
$\hat{\beta}'_{220} = 7.720$	$\hat{\beta}'_{230} = 7.694$	x_1 : $\hat{\beta}'_1$ =	0.051	5.22
$\hat{\beta}'_{320} = 7.154$	$\hat{\beta}'_{330} = 6.937$	x_2 : $\hat{\beta}'_2$ =	1.090	3.69
$\hat{\beta}'_{420} = 7.403$	$\hat{\beta}'_{430} = 7.269$	x_2^2 : $\hat{\beta}'_4$ =	-0.210	4.40
$\hat{\beta}'_{520} = 5.376$	$\hat{\beta}'_{530} = 5.917$	x_3 : $\hat{\beta}'_6$ =	-1.467	5.25
		$x_2^2x_3$: $\hat{\beta}'_{10}$ =	0.045	2.46