Open Access Library Journal
2022, Volume 9, e8425
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Correlation and Simultaneous Linear Regression

Khaled Fattah Sheet, Khawla Musta Sadiq*

Department of Mathematics, College of Education for Pure Sciences, University of Mosul, Mosul, Iraq
Email: *Ph.Khawla@uomosul.edu.iq

How to cite this paper: Sheet, K.F. and Sadiq, K.M. (2022) Correlation and Simultaneous Linear Regression. *Open Access Library Journal*, **9**: e8425.
https://doi.org/10.4236/oalib.1108425

Received: February 5, 2022
Accepted: May 24, 2022
Published: May 27, 2022

Abstract

Hirschfled (1935) posed the question. Is it always possible to introduce new variates for the rows and the columns of the contingency-table such that both regressions are linear. In reply, he derived the formulas of dual sealing. This approach was later employed by Lingoes (1963, 1968) who was obviously unaware of Hirschfeld's study, but noted that the approach would use the basic theory and equation worked out by Guttman (1941). We have to use a graphic with linear regression to find optimal weight which has good results by using a correlation as a new step to adjusting the spacing of rows and columns after quantification is linear, the condition under which correlation attains its maximum. It shall present here merely an example to illustrate the date have a certain $\rho = 0.65277$ between x and y which increases to reach the maximum value then the relation becomes a straight line which illustrates the maximum value of ρ.

Subject Areas

Mathematical Statistics

Keywords

Dual Scaling, Contingency-Table, Linear Regression, Optimal Weights

1. Introduction

Let us see the data in Table 1, suppose that those option weights and subject scores are simultaneously assigned to the responses (*i.e.*, I's) in Table 2, resulting in the table of weighted responses. Can you figure out how this table is prepared?

You can see the weight of option of each item chosen by subjects in Table 2 which constitutes the second term of each pair of Table 3. The left-hand side of each pair is nothing but the corresponding subject's score. As Guttmann (1941) reasoned, you can say that the two unknown equations in each pair in Table 4

Table 1. Data in terms of weights for options.

Subject	Item			Total for each subject
1	X_1	X_4	X_7	$X_1 + X_4 + X_7$
2	X_2	X_3	X_6	$X_2 + X_3 + X_6$
3	X_1	X_5	X_7	$X_1 + X_5 + X_7$
4	X_1	X_3	X_6	$X_1 + X_3 + X_6$
5	X_2	X_5	X_6	$X_2 + X_5 + X_6$
6	X_2	X_4	X_7	$X_2 + X_4 + X_7$
Grand total = $3X_1 + 3X_2 + 2X_3 + 2X_4 + 2X_5 + 3X_6 + 3X_7$				

Table 2. Data in terms of scores for subjects.

Item	1		2			3	
Option	1	2	1	2	3	1	2
	Y_1	Y_2	Y_2	Y_1	Y_3	Y_2	Y_1
	Y_3	Y_5	Y_4	Y_6	Y_5	Y_4	Y_3
	Y_4	Y_6				Y_5	Y_6

Table 3. Simultaneously weighted data.

	Item		
	1	2	3
1	(Y_1, X_1)	(Y_1, X_4)	(Y_1, X_7)
2	(Y_2, X_2)	(Y_2, X_3)	(Y_2, X_6)
3	(Y_3, X_1)	(Y_3, X_5)	(Y_3, X_7)
4	(Y_4, X_1)	(Y_4, X_3)	(Y_4, X_6)
5	(Y_5, X_2)	(Y_5, X_5)	(Y_5, X_6)
6	(Y_6, X_2)	(Y_6, X_4)	(Y_6, X_7)

Table 4. Multiple-choice data (categorical data).

	Item	**1**		**1**	**2**	**3**	**1**	**2**	
	Option	**1**	**2**	**1**	**2**	**3**	**1**	**2**	**Score**
Subject	1	1	0	0	1	0	0	1	Y_1
	2	0	1	1	0	0	1	0	Y_2
	3	1	0	0	0	1	0	1	Y_3
	4	1	0	1	0	0	1	0	Y_4
	5	0	1	0	0	1	1	0	Y_5
	6	0	1	0	1	0	0	1	Y_6
Weights		X_1	X_2	X_3	X_4	X_5	X_6	X_7	

are assigned to the same response, that is, they are the common descriptions of a single response, and therefore, the two unknowns should be given as similar values as possible (Block & Jones, 1968) [1].

One of the popular measures of the relationship between a pair of variables is the so-called product-moment correlation or Pearson tan correlation. This measure indicates the degree of linear relationship, which is the tendency that as one variable increases, the other increases, too. Let us indicate this correlation by ρ. To simplify the expression for ρ, let's choose the units and the origins of y's and x's as follows:

(The sum of squares of responses weighted by y_i) = (The sum of squares of responses weighted by X_i) = d, and (the sum of responses weighted by y_i) = (the sum of responses weighted by x_j) = 0.

Don't worry about these conditions on y_i and x_j because they will not alter the value of ρ or η^2. Now ρ can be expressed simply as:

$$\rho = \frac{\text{the sum of products of paired weights}}{d} = \frac{\sum\sum f_{ij} Y_i X_j}{d} \tag{1}$$

where f_{ij} = 1 or 0 as shown in **Table 1**.

Dual scaling is also a technique to determine Y_i and X_j in such a way that ρ is a maximum (Block & Jones, 1968) [1]. Note again that these subject scores, Y_i, and option weights, X_j, are identical to those obtained by the methods discussed so far. In addition, you should note that:

$$\rho = \eta \text{ , that is, } \eta^2 = \rho^2 \tag{2}$$

In statistics, the squared product-moment correlation is not equal to the squared correlation ratio generally. The equality between them as shown in Equation (2) is strictly a result of the duality of this scaling method (Fisher, 1940) [2].

We here try to illustrate the ordinary correlation and then go to what you mention in another research.

It is important to look at this approach to dual scaling as applied to the contingency table, because it will offer you another opportunity to see the distinction between continuous data and categorical data in analysis (Guttman, 1946) [3]. Let us consider a contingency table which is typically obtained by asking two multiple-choice questions consider the following questions:

Q1. How do feel about taking sleeping pills?

( ) strongly for, ( ) for, ( ) neutral, ( ) against, ( ) strongly against.

Q2. Do you sleep well every night?

( ) never, ( ) rarely, ( ) some nights, ( ) usually, ( ) always.

Suppose you obtain the data from 140 subjects as shown in **Table 5**.

The important distinction between continuous and categorical data, referred to previously, can now be explained as follows. Suppose you assign weight y_1 to

Table 5. Sleeping and sleeping pills.

Q1 \ Q2						
Subject weight	**−2**	**−1**	**0**	**+1**	**+2**	
	Never	**Rarely**	**Some**	**Usually**	**Always**	**Total**
Strongly for	15	8	3	2	0	28
For	5	17	4	0	2	28
R' Neutral	6	13	4	3	2	28
Against	0	7	7	5	9	28
Strongly against	1	2	6	3	19	28
Total	27	47	24	13	32	
Subject weight	−2	−1	0	+1	+2	

the "strongly for" option of Q1.

This is a step to illustrate the weights when you have more than one way to find the optimal solution, this is true for y_1, y_2, y_3, y_4, y_5 as same as.

The element in the first row (strongly for) and the first column (never) in the table. That is, 15, is now given weight y, so that the weighted response is $15y$, so far is the same for both types of data once you consider the sum of squares of weighted response, however, you will recognize the difference in the meaning of the expression $15y_1$ between the two types. In continuous data, 15 is a single number or a quantity (Nishisato & Clavel, 2003) [4]. Then by using the matrix including the variables of rows and columns, we find the other values for the remained question.

Therefore, the square of this weighted response is $(15y_1)^2 = 225Y_1^2$. In contrast, $15y_1$, in dual scaling means that each of 15 responses is given y_1, hence the sum of squared responses being equal to $Y_1^2 + Y_1^2 + \cdots + Y_1^2 = 15Y_1^2$.

Do you see this distinction? When you derive formulas for categorical data, this is of the utmost important importance, because it is one of the main distinctions in the formulation of categorical data analysis that of continuous data.

Dual scaling of the data in Table 5 determines five weights Y_i for the options of Q1, and five weights X_j for the options of Q2 in such a way that statistic ρ is maximum. In the formula:

$$\eta Y_i = \sum \frac{f_{ij} X_j}{f_i}, \quad \eta x_j = \sum \frac{f_{ij} X_j}{f_{.j}} \tag{3}$$

f_{ij} is no longer 1 or 0 but the frequency of row i and column j of Table 5. You may wonder what the above operation of maximizing ρ really means. Let us start with a case of (nonoptimal weights) (Nishisato, 2014) [5].

2. Numrical Example

Suppose that you decide, as most people do, to use your subjective, or common-

sense, weights of −2, −1, 0, +1, +2 for the options "never, rarely", some nights, usually always respectively, for Q2. Using these weights, calculate the mean weighted response for "strongly for" of Q1 Thus, mean,

$$m_1(\text{strongly for Q1}) = \frac{15x_2(-2) + 8x_2(-1) + 3x_2(0) + 2x_2(1) + 0x_2(2)}{28} = -1.3$$

using the same way for all m's of Q1, Q2 and **Table 6** and **Table 7** show the values.

How good are these common sense weights in explaining the data? One way to check is to construct a graph where you plot these means. Just calculated, against your say subject weights (**Figure 1**). Assuming as before that you assign weight y_i for row i (Q1) and X_i for column j (Q2). Let us call the plot of n_i against the subjective column weights. The "regression of Y on X" and the plot of n_j against the row weights the "regression of X on Y". This graph alone does not tell us much, So, just wait until you see the corresponding results when you use instead of your subjective weights, optimal weights (Guttman, 1946) [3]. obtained from dual scaling, that is, those weights that maximize ρ as is given in Nishisa (1980a, pp 66-68), we show you only the graph obtained without computation, but using dual scaling weights (**Figure 2**). Can you see that this is a remarkable plot? Both lines are straight and their slopes are identical! Those optimal weights had the effect of adjusting the spacing of rows and columns in such a way that the relation between rows and columns after quantification is linear, the condition under which ρ attains its maximum. This remarkable characteristic was termed simultaneous linear regression by Lingoes (1964) indeed, it served as the criterion in Hirsch Feld's (1935) formulation of this quantification method specifically. Hirschfield posed the questions: Is it always possible to introduce new variates for the rows and the columns of a contingency table such that both regressions are linear?

You can see Hirschfield's results in the expressions of dual relations or transition Formulas (4) As you recall ρ in formula (4) is the same as ρ in the:

$$\rho = \frac{\text{The sum of products of paired weight}}{d}$$

$$\rho^2 = \sum\sum \frac{f_{ij} Y_i X_j}{d} \tag{4}$$

Table 6. Means for Q1.

m_i	m_1	m_2	m_3	m_4	m_5
Value	−1.3	−0.8	−0.8	0.6	1.1

Table 7. Means for Q2.

n_i	n_1	n_2	n_3	n_4	n_5
Value	1.2	0.5	0.4	0.5	1.3

where $\rho = \eta$ that is $\eta^2 = \rho^2$ which is the key quantity, called the parameter, in linear regression. There is one more approach which is also obvious now that you know the dual relations (Nishisat, 2014) [5]; Nishisato and Shen, 1984 [6]). This approach provides a simple method of calculating "optimal weights" (Table 8).

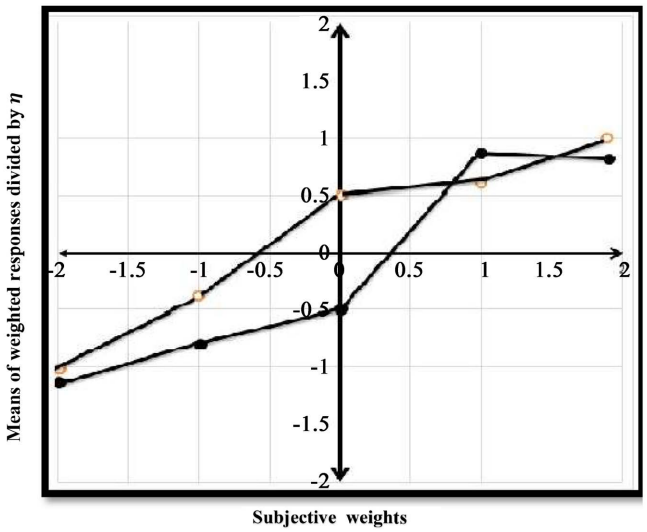

Figure 1. Graph for subjective weights.

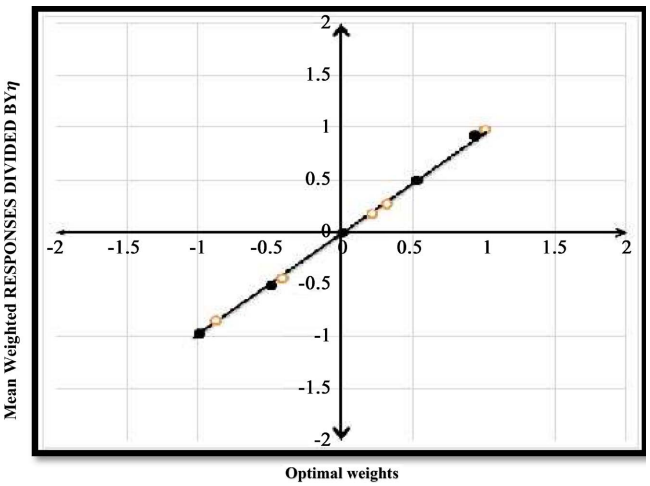

Figure 2. Optimal weights.

Table 8. Optimum weight values.

	Optimal mean X	weight X^*		Optimal mean X	weight X^*
Never	−1.30	−0.84	Strongly against	−1.20	−0.78
Rarely	−0.59	−0.38	Against	−0.64	−0.42
Some times	0.43	0.28	Neutral	−0.49	−0.32
Often	0.58	0.38	For	0.87	0.56
Always	1.55	1.00	Strongly for	1.47	0.95

3. Conclusions

We have to use a graphic with linear regression to find optimal weight as a new method instead of the iterative method.

Correlation and simultaneous linear regression is a good and interesting process to find the optimal solution for any problem we have, and we need to find correlation as we have in our paper.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Block, R.D. and Jones, L.V. (1968) The Measurement and Prediction of Judgemen-tand Choices. Holden-Day, San Francisco.

[2] Fisher, R.A. (1940) The Precision of Discriminant Functions. *Annals of Eugenics*, **10**, 422-429. https://doi.org/10.1111/j.1469-1809.1940.tb02264.x

[3] Guttman, L. (1946) An Approach for Quantifying Paired Comparisons and Rank Order. *The Annals of Mathematical Statistics*, **17**, 144-163. https://doi.org/10.1214/aoms/1177730977

[4] Nishisato, S. and Clavel, J.G. (2003) A Note on Between-Set Distances in Dual Scal-ing and Correspondence Analysis. *Behaviormetrika*, **30**, 87-98. https://doi.org/10.2333/bhmk.30.87

[5] Nishisato, S. (2014) Elements of Dual Scaling: An Introduction to Practical Data Analysis. Psychology Press, New York. https://doi.org/10.4324/9781315806907

[6] Nishisato, S. and Sheu, W.J. (1984) A Note on Dual Scaling of Successive Categories Data. *Psychometrika*, **49**, 493-500. https://doi.org/10.1007/BF02302587