A DERIVATION OF THE COEFFICIENTS AND DIVISORS FOR SETS OF ORTHOGONAL COMPARISONS IN REGRESSION *

by

D. S. Robson

BU-11-M July 1950

R. A. Fisher has demonstrated the facility with which regression problems may be handled when the independent variable is equally spaced. He has developed a simple method of fitting successively higher degree polynomials to n pairs of observations (Section 14.6) and has computed a table of coefficients and divisors for determining the reduction in sum of squares of deviations due to the fitting of successively higher degree polynomials (Table 15.1). These two contributions are distinct in that the first enables one to compute each of the n-1 polynomials which may be fitted to the n pairs of observations while the second (Table 15.1), although it does not allow for an explicit expression of any but the first degree regression equations, is more general in the sense that it allows for k groups of n pairs of observations where the value of the independent variable is constant within any group and equally spaced between groups.

The mechanics of applying the coefficients and divisors of Table 15.1 to any appropriate set of data are relatively simple; however, in order to fully understand the results of the application it may be helpful to relate this procedure to the general least squares curvilinear regression problem (Section 14.3). Suppose we have kn pairs of observations on the variables X and Y and that these kn pairs fall into k groups of size n when classified according to the value of X in the pair. Suppose, further, that these k distinct values of

---

* The section and table numbers in this paper refer to "Statistical Methods" by George W. Snedecor.

X are equally spaced; i.e., the value of X for the i'th group, i = 1, 2, ..., k, is $X_i = X_1 + c(i-1)$ where c is the constant difference between successive values of X. Thus in group 1 we have n pairs of X and Y — (X_1, Y_{11}), (X_1, Y_{12}), (X_1, Y_{13}), ..., (X_1, Y_{1n}); in group 2, $X = X_2 = X_1 + c$ and $Y = Y_{2j}$, j = 1, 2, ..., n; in group i, $X = X_i$ and $Y = Y_{ij}$, j = 1, 2, ..., n. Since the values of X_i, i = 1, 2, ..., k, are equally spaced, they may be coded into any other set of equally spaced numbers; for example, to code them into successive digits beginning with unity it is necessary only to subtract X_1 from each value, divide the result by c, and add 1:

$$(1) \qquad X_i' = \frac{X_i - X_1}{c} + 1 = \frac{[X_1 + c(i-1)] - X_1}{c} + 1 = i \qquad i = 1, 2, \ldots, k$$

It will be convenient to code X into simple integral values in such a manner that the mean of the coded X is also an integral number; this will be exemplified in the following cases.

Case I k = 2 groups

The simplest case to consider is that covered by column 1 of Table 15.1, the case involving two groups of n pairs of observations:

	GROUP I		GROUP II	
	X	Y	X	Y
	X_1	Y_{11}	X_2	Y_{21}
	X_1	Y_{12}	X_2	Y_{22}
	X_1	Y_{13}	X_2	Y_{23}
	⋮	⋮	⋮	⋮
	X_1	Y_{1n}	X_2	Y_{2n}
TOTAL =	nX_1	$\sum_{j=1}^{n} Y_{1j}$ $(=S_1)$*	nX_2	$\sum_{j=1}^{n} Y_{2j}$ $(=S_2)$*

* Snedecor's notation $\sum_{j=1}^{n} Y_{ij} = S_i$

Since the independent variable X assumes only two distinct values the linear regression line is clearly the highest degree polynomial one would fit to the data. In view of the fact that one needs to compute the quantities

$$\sum_{i=1}^{2}\sum_{j=1}^{n} x_i^2 = n\sum_{i=1}^{2} x_i^2 = n\left[(X_1-\bar{x})^2 + (X_2-\bar{x})^2\right]$$

$$\sum_{i}\sum_{j} x_i y_{ij} = \sum_{i}\sum_{j} X_i Y_{ij} - \frac{\Sigma\Sigma X_i \Sigma\Sigma Y_{ij}}{2n}$$

$$= X_1 \sum_j Y_{1j} + X_2 \sum_j Y_{2j} - \bar{x}(\Sigma Y_{1j} + \Sigma Y_{2j})$$

$$= (X_1-\bar{x}) \sum_j Y_{1j} + (X_2-\bar{x}) \sum_j Y_{2j}$$

it is seen that a convenient method of coding the X_i would be to let $X_1' = 1$, $X_2' = 3$ so that $\bar{x}' = 2$,

$$(X_1' - \bar{x}') = -1,$$

$$(X_2' - \bar{x}') = +1,$$

$$n\left[(X_1' - \bar{x}')^2 + (X_2' - \bar{x}')^2\right] = n(1+1) = 2n,$$

and $(X_1' - \bar{x}') \sum_j Y_{1j} + (X_2' - \bar{x}') \sum_j Y_{2j} = -\sum_j Y_{1j} + \sum_j Y_{2j}$ $(=S_1+S_2)$*.

The coding may be accomplished by letting

$$(2) \qquad X_i' = \frac{2(X_i - X_1)}{c} + 1 = \frac{2[X_1 + c(i-1) - X_1]}{c} + 1 = 2(i-1) + 1$$

$$= 2i - 1 \qquad , i = 1, 2, 3, \ldots, k$$

$$= 1, 3, \ldots, 2k - 1 .$$

---

*Ibid.

It is this characteristic of equally spaced numbers that so greatly simplifies the computations required in estimating regression equations; they may be coded down to any desired set of equally spaced numbers, and by selecting the correct set of coded values one minimizes the amount of work involved. Thus, in the case under consideration the only computations required to arrive at a coded estimate of the linear regression coefficient are

$$(3) \qquad b' = \frac{\Sigma\Sigma x'y}{\Sigma\Sigma x'^2} = \frac{-S_1 + S_2}{2n} \quad .$$

The sum of squares of deviations due to regression is, in the coded data,

$$(4) \qquad b'\Sigma\Sigma x'y = \frac{(-S_1 + S_2)^2}{2n} \quad .$$

One could, of course, decode to the original scale of measurement and would do so if he were interested in estimating the population regression coefficient; however, table 15.1 was designed for the type of problem in which one is interested only in tests of significance, not in estimation. Since probabilities are in no way altered by changing the scale of measurement there is no point in decoding the data for the purpose of tests of significance - and in practice, of course, one would not take the trouble of devising an expression such as (1) or (2) but would, instead, automatically select the correct coding system by applying the appropriate coefficients and divisors from table 15.1. The source of the coefficients and divisor found in column 1 of table 15.1 is now readily seen by examining the steps leading up to (3) and (4).

Case II $k = 3$ groups

With three groups, or three distinct values of the independent variable X, the simplest set of coded* X values is $X_1 = 1$, $X_2 = 2$, $X_3 = 3$. This is seen

* Since actual decoding never takes place, the prime on the X-variable will be dropped.

from the fact that

$$\bar{x} = \sum_{i=1}^{3} \sum_{j=1}^{n} X_i / 3n = \frac{n\sum X_i}{3n} = \frac{n(1+2+3)}{3n} = 2,$$

$$(X_1 - \bar{x}) = (1 - 2) = -1,$$

$$(X_2 - \bar{x}) = (2 - 2) = 0,$$

and $$(X_3 - \bar{x}) = (3 - 2) = +1 .$$

The sum of squares due to linear regression is then computed in the usual manner:

$$\sum_i \sum_j x_i^2 = n[(X_1 - \bar{x})^2 + (X_2 - \bar{x})^2 + (X_3 - \bar{x})^2] = n[(-1)^2 + 0 + (+1)^2] = 2n$$

$$\sum_i \sum_j xy = (X_1 - \bar{x}) \Sigma Y_{1j} + (X_2 - \bar{x}) \Sigma Y_{2j} + (X_3 - \bar{x}) \Sigma Y_{3j}$$

$$= -\Sigma Y_{1j} + 0\Sigma Y_{2j} + \Sigma Y_{3j}$$

$$= -S_1 + S_3$$

$$(5) \qquad b \Sigma \Sigma xy = \frac{(\Sigma \Sigma xy)^2}{\Sigma \Sigma x^2} = \frac{(-S_1 + S_3)^2}{2n}$$

with coefficients and divisor as given by column 2 of table 15.1 .

With three groups it is possible also to fit a quadratic regression line through the sample points. As indicated earlier, however, this is not the type of problem in which one is interested in actually estimating regression equations; rather, the chief interest here lies in testing whether or not the fitting of a second degree polynomial will account for a significant <u>additional</u> part of the variation. The preceding steps indirectly give us an estimate of

the linear regression line, $\hat{Y} = \bar{y} + b(X - \bar{x})$, and the sum of squares of deviations accounted for by this line; the scattering of points about the line represents the unaccounted for variation. The question, then, is -- would a second degree regression line through the sample points (X, Y) account for a significantly greater part of the variation than did the linear regression; i.e., is the <u>additional</u> sum of squares of deviations due to quadratic regression significantly large. To answer this question one could compute directly the sum of squares of deviations due to fitting a quadratic and subtract from it the already computed sum of squares due to fitting the linear; this difference could then be tested for significance. A simpler procedure is followed, however, by the use of table 15.1. The difference, itself, is directly obtainable by the method described below.

Consider the variable $d_{y \cdot x} = Y - \hat{Y}$, which represents the deviations from the linear regression line $\hat{Y} = \bar{y} + b(x - \bar{x})$. If we attempted to run a linear regression of the variable $d_{y \cdot x}$ on X we would, of course, find the slope <u>b</u> of the resulting regression line to be zero; this is true by definition or may be seen from the fact that crossproducts sum to zero. If we compute the sum of squares of deviations of $d_{y \cdot x}$ from the regression of $d_{y \cdot x}$ on X we would arrive at

$$\Sigma \Sigma d^2_{y \cdot x} = (1 - r^2_{xy}) \Sigma \Sigma y^2 .$$

This latter sum of squares may, however, be partitioned into two parts by fitting a parabola to the points $(d_{y \cdot x}, X)$; the one part, the sum of squares of deviations due to the quadratic regression of $d_{y \cdot x}$ on X, represents the <u>additional</u> sum of squares due to the quadratic regression of Y on X, and is the difference which is to be tested for significance; the other part, the sum of squares of deviations from quadratic regression, is zero in the case

where $k = 3$ but would be ≥ 0 when $k > 3$.

In order to fit a quadratic to the deviations from linear regression it is necessary to compute the three partial correlation coefficients. The computations would proceed as follows:

$$\text{Since } d_{y\cdot x} = Y - \hat{Y} = Y - \bar{y} - b(X - \bar{x})$$

$$= Y - \frac{\Sigma\Sigma Y}{kn} - \frac{\Sigma xy}{\Sigma x^2}(X - \bar{x})$$

$$= Y - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}(X - \bar{x})$$

the three variables to be considered are as listed in Table 1.

$$\Sigma\Sigma(X_2 - \bar{x}_2)^2 = n\Sigma X_1^4 - \frac{(n\Sigma X_1^2)^2}{3n} = 98n - \frac{(14n)^2}{3n} = \frac{98}{3}n$$

$$\Sigma\Sigma x_1 d_{y\cdot x} = 0$$

$$\Sigma\Sigma x_2 d_{y\cdot x} = S_1 + 4S_2 + 9S_3 - \frac{14}{3}(S_1 + S_2 + S_3) - 4(-S_1 + S_3)$$

$$= \frac{1}{3}(S_1 - 2S_2 + S_3)$$

$$\Sigma\Sigma d_{y\cdot x}^2 = (1 - r_{yx}^2)\Sigma\Sigma y^2$$

$$r_{x_1 d_{y\cdot x}} = \frac{\Sigma\Sigma x_1 d_{y\cdot x}}{\sqrt{\Sigma\Sigma x_1^2 \Sigma\Sigma d_{y\cdot x}^2}} = 0$$

$$r_{x_2 d_{y\cdot x}} = \frac{\Sigma\Sigma x_2 d_{y\cdot x}}{\sqrt{\Sigma\Sigma x_2^2 \Sigma\Sigma d_{y\cdot x}^2}} = \frac{S_1 - 2S_2 + S_3}{3\sqrt{\frac{98}{3}n\,\Sigma\Sigma y^2(1 - r_{xy}^2)}}$$

$$r_{x_1 x_2} = \frac{\Sigma\Sigma x_1 x_2}{\sqrt{\Sigma\Sigma x_1^2 \Sigma\Sigma x_2^2}} = \frac{8n}{14n\sqrt{\frac{1}{3}}} = \frac{4\sqrt{3}}{7}$$

Table 1.

X_1	$d_{y \cdot x}$	$X_2 = X_1^2$
1	$Y_{11} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (1-2)	1
1	$Y_{12} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (1-2)	1
⋮	⋮	⋮
1	$Y_{1n} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (1-2)	1
2	$Y_{21} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (2-2)	4
2	$Y_{22} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (2-2)	4
⋮	⋮	⋮
2	$Y_{2n} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (2-2)	4
3	$Y_{31} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (3-2)	9
3	$Y_{32} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (3-2)	9
⋮	⋮	⋮
3	$Y_{3n} - \frac{(S_1 + S_2 + S_3)}{3n} - \frac{(-S_1 + S_3)}{2n}$ (3-2)	9
TOTAL=n+2n+3n	$\Sigma \Sigma Y_{ij} - \Sigma \Sigma Y_{ij} - 0 = 0$	n+4n+9n= 14n

$$R^2 = \frac{r^2_{x_1 d_{y\cdot x}} + r^2_{x_2 d_{y\cdot x}} - 2 r_{x_1 d_{y\cdot x}} r_{x_2 d_{y\cdot x}} r_{x_1 x_2}}{1 - r^2_{x_1 x_2}}$$

$$= \frac{(S_1 - 2S_2 + S_3)^2}{6n\Sigma\Sigma y^2(1-r^2_{xy})}$$

The sum of squares of deviations due to the quadratic regression of $d_{y\cdot x}$ on X is then

$$R^2 \Sigma\Sigma d^2_{y\cdot x} = R^2 \Sigma\Sigma y^2(1-r^2_{xy}) = \frac{(S_1 - 2S_2 + S_3)^2}{6n}$$

with coefficients and divisors as given in column 3, table 15.1, and this quantity is likewise the additional sum of squares of deviations due to fitting the quadratic regression of Y on X.

THE GENERAL CASE

This argument can be extended to show that the coefficients for the i'th degree polynomial may be gotten by fitting an i'th degree polynomial to the deviations from the (i-1)'th degree polynomial which in turn was fitted to the deviations from the (i-2)'th degree polynomial, and so on.

DISCUSSION

The numbers appearing in table 15.1 were probably not derived in the preceding manner; this derivation does, however, give one some understanding of the logic behind the calculations.

One additional feature of the preceding discussion which should be noted is that it could as well have been labelled "A Derivation of the Coefficients and Divisors for Sets of Orthogonal Comparisons in the Analysis of Variance." In this sense it would be applicable only to the analysis of variance problem involving k equal sized groups, each group, of course, corresponding to some particular treatment.

A convenient example is to reconsider Case II in the light that X_1, being a characteristic common to all n elements in the first group, now corresponds to treatment 1, while X_2 corresponds to treatment 2 and X_3 to treatment 3. Let us assume a randomized block design and the following analysis:

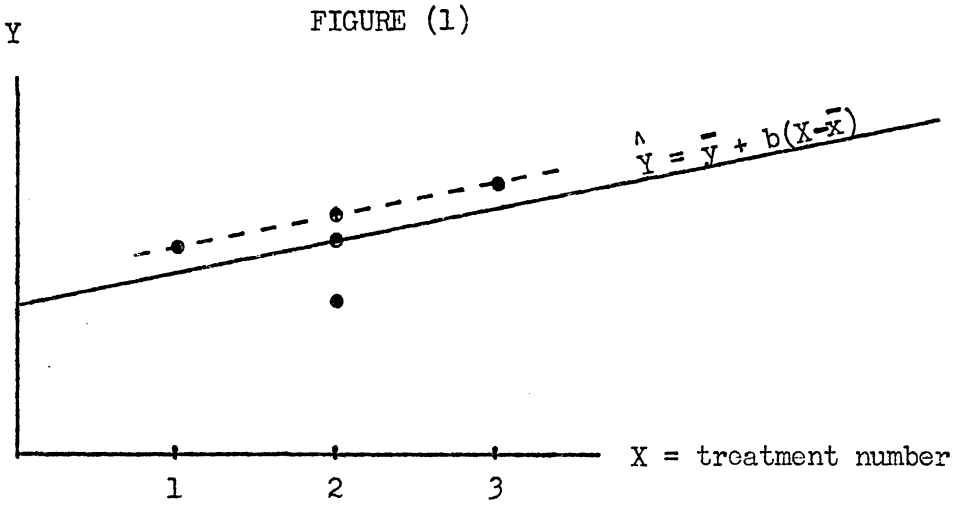
	Treatment 1	Treatment 2	Treatment 3
	Y_{11}	Y_{21}	Y_{31}
	Y_{12}	Y_{22}	Y_{32}
	⋮	⋮	⋮
	Y_{1n}	Y_{2n}	Y_{3n}
TOTAL	S_1	S_2	S_3

Analysis of Variance

Source	d.f.	s.s.
Treatments	2	$(S_1^2 + S_2^2 + S_3^2)/3 - (S_1 + S_2 + S_3)^2/3n$
1 vs. 3	1	$(S_1^2+S_3^2)/n - (S_1+S_3)^2/2n = (-S_1+S_3)^2/2n$
1+3 vs. 2	1	$(S_1+S_3)^2/2n + S_2^2/n - (S_1+S_2+S_3)^2/3n = (S_1-2S_2+S_3)^2/6n$
Error	2(n-1)	

It is seen that the coefficients and divisors of the individual treatment comparison sums of squares (or mean squares) are precisely those obtained earlier under Case II. This is not surprising when one looks upon the comparisons as regression problems; if we assign X values of 1, 2, 3 to the treatments 1, 2, 3, we may depict the comparison graphically. Let the solid points represent the treatment means:

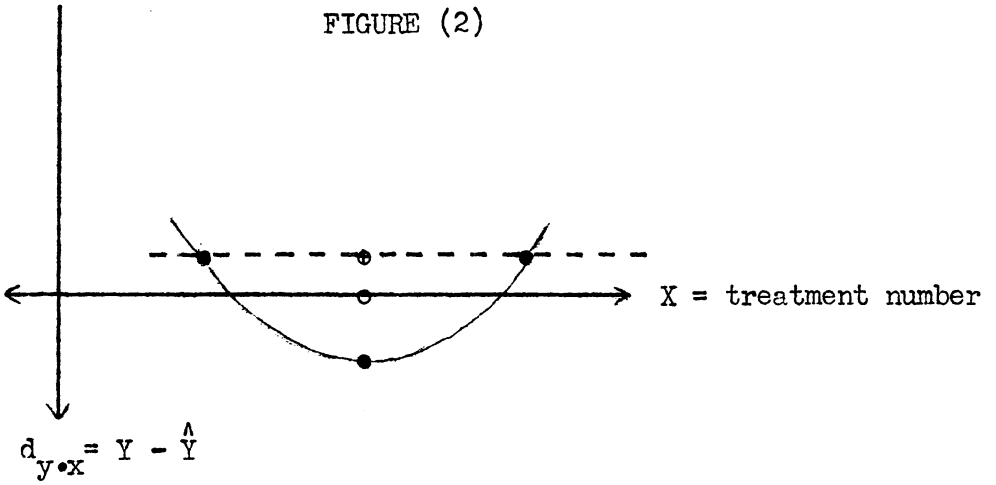
FIGURE (1)



The solid points from left to right are defined by the coordinates ($X=1$, $Y=\bar{y}_1$), ($X=2$, $Y=\bar{y}_2$), and ($X=3$, $Y=\bar{y}_3$), respectively; the point ⊕ is defined by ($X=2$, $Y=\frac{\bar{y}_1+\bar{y}_3}{2}$) and the point ⊖ is defined by ($X=\bar{x}=2$, $Y=\bar{y}$).

Using these arbitrarily assigned treatment numbers as the X-values enables one to compute the equation of the least squares regression line having the slope $b = \frac{-S_1+S_3}{2n} = \frac{-\bar{y}_1+\bar{y}_2}{2}$; b is determined by the difference between the means of treatments 1 and 3, regardless of the value of $\bar{y}_2$, and is identical to the slope of the dotted line which passes through the points (X_1, $\bar{y}_1$), (X_3, $\bar{y}_3$). The sum of squares (or M.S.) due to linear regression is $\frac{(-S_1+S_2)^2}{2n}$ = mean square for the comparison treatment 1 vs. treatment 3, and if this mean square is significant then the slope of the regression line is significantly different from zero and the difference between the means, which determines the slope, is likewise different from zero.

FIGURE (2)



A quadratic fitted to the deviations from linear regression must pass through all three points [figure (2)] and attain its minimum (or maximum) at the point $(X=2,\ d_{y \cdot x} = \bar{y}_2 - \hat{Y}_2)$, which is the solid point in the center of figure (2). If the additional sum of squares of deviations due to quadratic regression $= \frac{(S_1 - 2S_2 + S_3)^2}{6n}$ is significant then the distance from the point $(X=2,\ d_{y \cdot x} = \bar{y}_2 - \hat{Y}_2)$ to the dotted line in figure (2) is significantly different from zero, and this distance represents the difference $\frac{\bar{y}_1 + \bar{y}_2}{2} - \bar{y}_3$ as seen from figure (1).

If one wished a different set of treatment comparisons it would simply necessitate a different assignment of the numbers 1, 2, 3 to the three treatments. In practice the researcher automatically makes this assignment when he selects a particular set of orthogonal comparisons.