First Draft

ANALYSIS OF COVARIANCE

by

W. G. Cochran
Institute of Statistics
University of North Carolina

Institute of Statistics
Mimeo Series # 6
For limited distribution

1. Introduction. Analysis of covariance differs from analysis of variance in that, instead of analyzing the variation of a quantity y, we analyze the deviations from its linear regression on some quantity x. For illustration, consider the two-way (row, column) classification with a single observation in each cell. If the suffix (ij) denotes the i^{th} row and the j^{th} column, and if $\bar{x}_{..}$ is the mean of the x's, the model appropriate to the analysis of covariance is

$$(1) \quad y_{ij} - \beta (x_{ij} - \bar{x}_{..}) = \mu + \xi_i + \gamma_j + e_{ij}$$

where β, μ, ξ_i and γ_j are population parameters, with the usual interpretations. The e_{ij} are random variables assumed to be normally and independently distributed with zero means and variance σ_e^2. As a rule all quantities β, μ, ξ_i, γ_j and σ_e^2 are unknown.

The principal objects of the analysis are usually to estimate the ξ_i and γ_j and to test certain null hypotheses about them. These quantities are often described as adjusted row or column effects, to emphasize the fact that the original data y_{ij} have been adjusted so as to remove their linear regression on the x_{ij}. Since the parameters ξ_i and γ_j are needed only to measure differences between the effects of different rows or columns, we may assume that $\Sigma \xi_i = \Sigma \gamma_j = 0$. With this convention, the mean value of the right hand side of (1), taken over the i^{th} row, is $(\mu + \xi_i)$. This quantity may therefore be interpreted as the true (or population) adjusted mean of the i^{th} row.

2. Uses of the technique. Analysis of covariance was first applied by Fisher (1) in agricultural experimentation, where the y's were plot yields and the rows (say) corresponded to the experimental treatments. The technique has proved valuable in two sets of circumstances. First, the x's may be variables which are correlated with the y's although not themselves affected by the treatments. In other words the x's measure some environmental factor which influences the results of the experiment. For instance, in the application first suggested by Fisher, the y's were the yields of 16 plots of tea bushes in a hypothetical experiment and the x's were the yields of the same 16 plots during a period, before the beginning of the

experiment, in which all plots were being uniformly treated. Thus the x's measured, in a sense, the inherent productiveness of the 16 plots. Clearly, one would expect to obtain a more accurate comparison of the effects of the treatments by adjusting the experimental yields y so as to correct for the effects of such differences in productiveness. In this way, covariance may reduce the experimental error by adjusting for the effects of some disturbing environmental factor.

Competent experimenters have, of course, always attempted to nullify such disturbing effects by careful control of the experimental technique. The introduction of covariance places a new weapon at their disposal. If it is not feasible or convenient to control such factors, it may still be possible to measure them. Their effects on the experimental results are then estimated in the covariance analysis and the appropriate corrections are made. It need not be stressed that the "linear" regression can be extended to include non-linear effects in the usual way. Numerous applications of covariance have been made in biological experimentation with striking success. There seems no doubt that equally fruitful uses could be made in other types of experimentation where covariance at present appears to be little known.

Secondly, a covariance analysis sometimes helps to clarify the interpretation in experiments where the x's <u>are</u> affected by the treatments. The numerical example to be presented later is a case in point. This experiment tested the effects of certain soil fumigants on eelworms, which live in the soil and attack some English farm crops. After the treatments had been applied, the numbers of eelworm cysts per plot and the yields of the crop (spring oats) were both recorded. The treatments produced significant effects both on eelworms and on the oats. It is clearly of interest to discover whether the effects on the crop can be regarded as a reflection of the effects on the worms. This can be done by finding whether the treatment effects on the oats persist or disappear after adjusting for the differences in eelworm numbers. In this way a covariance analysis may throw some light on the mechanism by which the treatments produce their results.

As Bartlett (8) has pointed out, the interpretation of this second use of covariance must be considered carefully, since a rather dangerous extrapolation may be involved. Suppose in an experiment that the plot plant numbers for one variety vary from 400 to 600 with a mean of 500, and that for another variety the range is 800 to 1,000 with a mean of 900. If yields are adjusted for differences in stand, the adjusted yields for each variety would formally correspond to a plant number of 700, this being the mean of 500 and 900. Unless there is additional information not supplied by the experiment, however, it would be risky to interpret these adjusted yields literally, since neither variety has any plots with plant numbers near 700. It is proper to state whether the differences in yield can or cannot be attributed to a linear regression on plant number, but the possibility exists of curvilinear effects that the data at hand are incapable of detecting.

3. Computational procedure. Before describing the theory of the method we shall outline the computational procedure for the two-way classification. The notation to be used is given below.

	Columns			Row means
	y_{11}	y_{12}	y_{1q}	$\bar{y}_{1\cdot}$
	y_{21}	y_{22}	y_{2q}	$\bar{y}_{2\cdot}$
Rows	- -	- - -	- - -	-
	y_{p1}	y_{p2}	y_{pq}	$\bar{y}_{p\cdot}$
Column means	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\bar{y}_{\cdot q}$	$\bar{y}_{\cdot\cdot}$

The information wanted from the statistical analysis is usually the following:

(i) An estimate b of the regression coefficient β and a test of significance of the null hypothesis $\beta = 0$.

(ii) Estimates of the true adjusted row means $(\mu + \xi_i)$, and a test of significance for any given linear combination of these means. Probably the most common linear function is simply the difference between two adjusted row means.

(iii) A test of the null hypothesis that all adjusted row means are equal.

Similar tests may, of course, be required for the column effects.

The first step in the computations is to carry out separate analyses of variance for y and x, and a corresponding analysis of the cross-products (yx).

Table I. Sums of squares and products

	d.f.	(x^2)	(yx)	(y^2)
Columns	(q-1)	C_{xx}	C_{yx}	C_{yy}
Rows	(p-1)	R_{xx}	R_{yx}	R_{yy}
Error	(p-1)(q-1)	E_{xx}	E_{yx}	E_{yy}
Rows + Error	(p-1)q	S_{xx}	S_{yx}	S_{yy}

The only unusual feature is that a line is added at the foot giving totals for rows and error. Thus $S_{yy} = R_{yy} + E_{yy}$, etc.

The sum of squares for y in the error line is now divided into two parts: the sum of squares for regression on x (1.d.f.) and the sum of squares of deviations. The same subdivision is carried out for the "Rows plus Error" line.

Table II. Sums of squares for regression and deviations.

	Regression		Deviations		
	d.f.	SS	d.f.	SS	M.S.
Error	1	E^2_{yx}/E_{xx}	(p-1)(q-1)-1	$E_{yy} - \frac{E^2_{yx}}{E_{xx}}$	s^2_e
Rows + Error	1	S^2_{yx}/S_{xx}	(p-1)q-1	$S_{yy} - \frac{S^2_{yx}}{S_{xx}}$	
Rows (by subtraction)			(p-1)	$R_{yy} - \frac{S^2_{yx}}{S_{xx}} + \frac{E^2_{yx}}{E_{xx}}$	s^2_r

Finally, the adjusted sum of squares for rows, with (p-1) d.f., is obtained by subtracting the deviations S.S. for error from that for rows plus error.

The information mentioned above is now obtainable as follows.

(i) The regression coefficient β is estimated from the error line; thus $b_e = E_{yx}/E_{xx}$. The significance of b_e is tested by taking the ratio of the regression to the deviations mean square; i.e. $E^2_{yx}/s^2_e E_{xx}$. This is distributed as F with 1 and $\{(p-1)(q-1)-1\}$ d.f.

(ii) The adjusted mean of the i^{th} row is estimated by $R_i = \bar{y}_{i.} - b_e(\bar{x}_{i.} - \bar{x}_{..})$

If the x's are regarded as fixed, any specified linear function $\Sigma g_i R_i$ is normally distributed with mean $\Sigma g_i(\mu + \xi_i)$ and variance

$$\sigma_e^2 \left[\frac{\Sigma g_i^2}{q} + \frac{\{\Sigma g_i(x_{i.} - \bar{x}_{..})\}^2}{E_{xx}} \right]$$

Hence a t-test of the hypothesis $\Sigma g_i(\mu + \xi_i) = 0$ is obtained by substituting s_e for σ_e in the usual way. In particular, for testing the difference between the adjusted means of the i^{th} and j^{th} rows, we have

$$(1)' \qquad t = \frac{(\bar{y}_{i.} - \bar{y}_{j.}) - b_e(\bar{x}_{i.} - \bar{x}_{j.})}{s_e\sqrt{\frac{2}{q} + \frac{(\bar{x}_{i.} - \bar{x}_{j.})^2}{E_{xx}}}}$$

with $\{(p-1)(q-1)-1\}$ d.f.

(iii) For a test of the hypothesis that all adjusted row means are equal, we write $F = s_r^2/s_e^2$, where s_r^2 is the adjusted mean square for rows, as given in Table II. This F variate has (p-1) and $\{(p-1)(q-1)-1\}$ d.f.

If the corresponding test is wanted for the columns, Table II is re-calculated with columns in place of rows. Similarly, for testing a particular subset of the (p-1) d.f. for rows, Table II is re-computed with rows replaced by the subset in question.

4. The adjusted sum of squares for rows. When analysis of covariance is first encountered, the feature which seems least familiar is the procedure for obtaining the adjusted sum of squares for rows (ASSR). It may be helpful to study the ASSR in more detail.

As shown in Table II, the ASSR is obtained by subtraction. By analogy with ordinary analysis of variance, one might expect instead to compute it as some kind of sum of squares of deviations of the adjusted row means $R_i = \{\bar{y}_{i.} - b_e(\bar{x}_{i.} - \bar{x}_{..})\}$. However, the ASSR is not equal to the sum of squares of deviations of the adjusted

row means, as will now be shown. From Table II the ASSR is given by

$$(2)\quad \text{ASSR} = R_{yy} - \frac{(R_{yx} + E_{yx})^2}{R_{xx} + E_{xx}} + \frac{E^2_{yx}}{E_{xx}}$$

It will be realized that the sum of squares of deviations of the adjusted row means is the same as the rows sum of squares of $(y - b_e x)$. Hence

$$(3)\quad \text{SSD} = R_{yy} - 2b_e R_{yx} + b_e^2 R_{xx}$$

$$(4)\quad = R_{yy} - \frac{2E_{yx}R_{yx}}{E_{xx}} + \frac{E^2_{yx} R_{xx}}{E^2_{xx}}$$

Subtracting (2) from (4) and taking the common denominator, we find that the difference can be expressed as the single square

$$(5)\qquad \frac{(R_{yx} - b_e R_{xx})^2}{(R_{xx} + E_{xx})}$$

$$(6)\ \text{or alternatively}\qquad \frac{R^2_{xx}\,(b_r - b_e)^2}{(R_{xx} + E_{xx})}$$

where $b_r = R_{yx}/R_{xx}$, is the regression coefficient calculated from the rows line in the analysis of variance. Expression (5) was first given by Yates (2).

Consequently the SSD of the adjusted row means is always too large. Before considering why this is so, we may note that if x exhibits no significant effect of rows, the correction term (5) or (6) is usually small relative to the SSD or ASSR. Use of this fact to shorten the computations will be illustrated later.

The discrepancy between the SSD and the ASSR is due to the sampling error of b_e. If β were substituted for b_e, it is easy to see that the SSD would be distributed as a multiple of χ^2 (when the true adjusted row means were all equal) and would be appropriate for use as the numerator of the F-test. The sampling error in b_e complicates the issue in two ways. First, each adjusted row mean R_i has a different standard error, since

$$V(R_i) = \sigma^2_e \left\{ \frac{1}{q} + \frac{(\bar{x}_{i.} - \bar{x}_{..})^2}{E_{xx}} \right\}$$

Further, the adjusted means for any two rows are correlated, since b_o enters into both. A quadratic form in the R_i rather than a simple SSD might therefore be expected as the best measure of the variation among the adjusted row means.

This quadratic form is found as follows. From (5)

$$ASSR = SSD - \frac{(R_{yx} - b_o R_{xx})^2}{R_{xx} + E_{xx}}$$

$$(7) \qquad = q \sum_{i=1}^{p} (R_i - \bar{R})^2 - \frac{(R_{yx} - b_o R_{xx})^2}{R_{xx} + E_{xx}}$$

now $$R_{yx} - b_o R_{xx} = q \sum_{i=1}^{p} (\bar{x}_{i.} - \bar{x}_{..}) \left\{ (\bar{y}_{i.} - \bar{y}_{..}) - b_o (\bar{x}_{i.} - \bar{x}_{..}) \right\}$$

$$= q \sum_{i=1}^{p} (\bar{x}_{i.} - \bar{x}_{..})(R_i - \bar{R})$$

Substituting this expression in (7) we find that

$$ASSR = \sum_{i,j=1}^{p} a_{ij} (R_i - \bar{R})(R_j - \bar{R})$$

where $$a_{ii} = q \left\{ 1 - \frac{q(\bar{x}_{i.} - \bar{x}_{..})^2}{R_{xx} + E_{xx}} \right\}$$

$$a_{ij} = a_{ji} = - \frac{q^2 (\bar{x}_{i.} - \bar{x}_{..})(\bar{x}_{j.} - \bar{x}_{..})}{R_{xx} + E_{xx}}$$

As might be expected from general theory, the inverse of the matrix a_{ij} is the variance - covariance matrix of the R_i (apart from the common factor σ_o^2).

It is also instructive to compare the ASSR with the sum of squares of deviations of the row means $\bar{y}_{i.}$ from the rows regression on x, or in other words with the rows sum of squares of $(y - b_r x)$. This sum of squares is

$$(8) \qquad R_{yy} - R^2_{yx}/R_{xx}$$

and has (p-2) d.f. By comparison with (2), it will be found that this SSD is always less than the ASSR by the amount

$$(9) \qquad \frac{R_{xx} E_{xx} (b_r - b_o)^2}{R_{xx} + E_{xx}}$$

Consequently, the adjusted sum of squares for rows divides into two parts: a single d.f., as shown in (9), which compares the regression coefficients from the rows and error lines in the analysis of variance, and (p-2) d.f., as shown in (8), representing the SSD from the rows regression. This result is occasionally useful in interpreting analyses of data. It sometimes happens that the ASSR is large solely because of the difference between b_r and b_e. In many problems, however, no simple meaning can be attached to b_r, and this approach is not then helpful.

To complete the circle of comparisons, we may note from (6) and (9) that the rows sum of squares for $(y - b_e x)$ exceeds that for $(y - b_r x)$ by

$$(9a) \qquad R_{xx}(b_r - b_e)^2$$

Consequently, the rows sum of squares for $(y - b_e x)$ contains (p-2) properly weighted d.f. The remaining d.f. is inflated by the factor $(R_{xx} + E_{xx})/E_{xx}$, as is seen by comparing (9a) with (9).

5. Theory of the technique. A completely general proof, starting from first principles, would be rather lengthy. We will assume some acquaintance with the method of least squares as applied in the analysis of variance. The mathematical model is

$$(10) \quad y_{ij} = \mu + \rho_i + \gamma_j + \beta(x_{ij} - \bar{x}_{..}) + e_{ij}$$

where the x's are fixed and the e's are normally and independently distributed with zero means and variances σ_e^2. Also the sums of the ρ's and the γ's are both zero. By the method of least squares, the unknown parameters are to be estimated by minimizing

$$(11) \quad \sum_{ij} \{y_{ij} - m - r_i - c_j - b(x_{ij} - \bar{x}_{..})\}^2$$
$$\text{subject to} \quad \Sigma r_i = \Sigma c_j = 0.$$

Calculation of the minimum is facilitated by means of a slight transformation. The prediction equation

$$(12) \quad Y_{ij} = m + r_i + c_j + b(x_{ij} - \bar{x}_{..})$$

may be made identical with the prediction equation

$$(13)\quad Y_{ij} = m' + r'_i + c'_j + b(x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..})$$

if we write

$$(14)\quad \begin{cases} m = m' \\ r_i = r'_i - b(\bar{x}_{i.} - \bar{x}_{..}) \\ c_j = c'_j - b(\bar{x}_{.j} - \bar{x}_{..}) \end{cases}$$

Moreover, since $\Sigma r_i = \Sigma c_j = 0$, it is clear that $\Sigma r'_i = \Sigma c'_j = 0$.

Hence, writing $x'_{ij} = x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x}_{..}$, we may minimize, instead of (11),

$$(15)\quad \sum_{ij} (y_{ij} - m' - r'_i - c'_j - bx'_{ij})^2 .$$

Those familiar with the analysis of variance will recognize x' as the error component of x in its analysis of variance.

The reason for introducing x'_{ij} is that its sum is zero over any row or column. Thus (15) may be written

$$(16)\quad \sum_{ij} (y_{ij} - m' - r'_i - c'_j)^2 - 2b \sum_{ij} y_{ij}x'_{ij} + b^2 \sum_{ij} x'^2_{ij} .$$

Differentiating with respect to b, we find

$$(17)\quad b = \frac{\Sigma y_{ij}x'_{ij}}{\Sigma x'^2_{ij}} = \frac{E_{yx}}{E_{xx}} = b_e$$

since x'_{ij} is the error component of x_{ij}.

Now the minimization of the first term in (16) is equivalent to an ordinary analysis of variance of y(without covariance). The solutions are

$$(18)\quad m' = \bar{y}_{..},\ r'_i = \bar{y}_{i.} - \bar{y}_{..},\ c'_j = \bar{y}_{.j} - \bar{y}_{..} .$$

Hence, by (14), the least squares estimates for the covariance problem are

$$(19)\quad \begin{cases} m = \bar{y}_{..} \\ r_i = (\bar{y}_i - \bar{y}_{..}) - b_e(\bar{x}_{i.} - \bar{x}_{..}) \\ c_j = (\bar{y}_{.j} - \bar{y}_{..}) - b_e(\bar{x}_{.j} - \bar{x}_{..}) \end{cases}$$

The value $(m + r_i)$ will be found to agree with the expression given in section 3 for the estimate of the adjusted row mean.

Further, from (16), the residual sum of squares is

$$(20) \qquad E_{yy} - \frac{2E_{yx}^2}{E_{xx}} + \frac{E_{yx}^2}{E_{xx}} = E_{yy} - \frac{E_{yx}^2}{E_{xx}} .$$

with $\{(p-1)(q-1) - 1\}$ d.f. Hence the mean square is s_e^2 as defined in Table II.

With the x's fixed, b_e, r_i and c_j are normally distributed, being linear functions of the y's. Also, by the general theory of least squares, they are unbiased estimates of the corresponding population parameters. The tests of significance given in section 3 for these quantities are established by the usual methods and will not be given in detail.

With regard to the F-test of the adjusted row means, it is perhaps best to quote the general theorem on the F-test in least squares, as applied to this problem.

Theorem: With the assumptions specified in (10), let

$$D_a = \sum_{ij} \{y_{ij} - m - r_i - c_j - b(x_{ij} - \bar{x}_{..})\}^2$$

$$D_r = \sum_{ij} \{y_{ij} - m'' - c_j'' - b''(x_{ij} - \bar{x}_{..})\}^2$$

where the constants are chosen in each case so as to minimize the corresponding sum of squares. Then if all ρ_i are zero, the quantity

$$\frac{(D_r - D_a)}{(p-1)} \Big/ \frac{D_a}{\{(p-1)(q-1)-1\}}$$

is distributed as F with $(p-1)$ and $\{(p-1)(q-1)-1\}$ d.f.

General proofs of this useful result, which is not as well known as it deserves to be, have been given by Yates (3) and Wald (4).

It has already been shown in (20) that D_a is the SSD from the error regression and that the corresponding mean square is s_e^2. By the same approach, D_r will be the SSD from the "error" regression when only column effects are eliminated in the analysis of variance. But in that event the "error" will be equal to "rows plus error" in the present analysis. Hence the numerator of F, $(D_r - D_a)/(p-1)$, is equal to s_r^2 as defined in Table II. This completes the proof.

An alternative and more direct proof may be developed from (9). In this it was shown that the adjusted sum of squares for rows divides into 1 d.f. which compares b_r and b_e, and (p-2) d.f. representing deviations from the rows regression. If the null hypothesis holds, each sum of squares may be shown to be distributed independently as $\chi^2\sigma^2_e$ and to be independent of s^2_e. The appropriate F test follows.

While the theory above has been developed for a particular type of classification, the methods apply without difficulty to other classifications.

6. Numerical example. This example is intended to illustrate the computations and to present two short cuts that are sometimes useful. The experiment was conducted at the Rothamsted Experimental Station. The plan and yields are given in the 1935 Report, pp 176-178. There were four fumigants, chlorodinitrobenzene (CN), carbon disulphide jelly (CS), "Cymag" (CM) and "Seekay" (CK), each applied in a single and double dressing. The ninth treatment, no fumigant, was replicated four times within each block and the experiment comprized four randomized blocks of 12 plots each (plots 1/80 acre).

Two weeks before the fumigants were applied, an estimate was made from soil samples of the number of eelworm cysts on each plot. It was planned to use these data as an x-variable in covariance in the hope of removing the influence of natural variations in the degree of eelworm infestation. The oats were drilled five days after the fumigants had been ploughed in. To measure the treatment effects on the worms, a second count was made on each plot after harvest. Treatment totals for the two counts are shown in Table III. The lines marked 'L' and 'Q' will be explained later. It will be noted that the cysts increased from the first to the second counts under all treatments. This does not mean that treatments were necessarily ineffective since they may have kept down the rate of increase.

Table III. Treatment totals for numbers of cysts.*

	Before fumigation (x)					After fumigation (y)				
Level	CN	CS	CM	CK	Total	CN	CS	CM	CK	Total
0	1975				1975	5858				5858
1	402	417	513	570	1902	1066	928	1431	892	4317
2	389	554	568	778	2289	1265	877	1241	1122	4505
L	1180	1525	1649	2126	6840	3596	2682	3913	3136	13327
Q	415	280	458	362	1515	867	979	1621	662	4129

*Numbers of cysts per plot were the numbers found in 400 gms. of soil. Thus the totals in the table represent 1600 gms of soil, except for the no-fumigant totals, which represent 6400 gms.

The experiment constitutes a two-way classification with 48 plots and is analyzed into replicates (3 d.f.), treatments (8 d.f.) and error (36 d.f.). The 36 d.f. for error contain 24 which represent the interactions of treatments and replicates and 12 which represent variations within the sets of 4 no-fumigant plots in each replicate.

The simplest hypothesis to consider about the effects of the fumigants is that the response is a linear function of the level. Under this hypothesis we represent the number of cysts for the u^{th} fumigant as $(\alpha - \delta_u L)$ where L (0,1,2) is the level in question and δ_u is the decrease in cysts for each increase in level. The parameter α (response at zero level) may be presumed the same for all fumigants. If this hypothesis is tenable, the quantities δ_u (or estimates of them) are appropriate for comparing the effectiveness of the fumigants.*

With this approach in mind, the 8 d.f. for treatments may be separated into four orthogonal parts. Consider first the totals over all four fumigants, e.g. the figures 5858, 4317 and 4505 for the 'after fumigation' data. These three figures provide 2 d.f. as follows:

$$\text{Average linear response} = (5858-4505)^2/32 = 57{,}207$$
$$\text{Average curvature} = \{5858 + 4505 - 2(4317)\}^2/\ 96 = 31{,}140.$$

*The least squares solution for this problem (in a more general case) has been given recently by Bliss (5).

Next, there are 3 d.f. which compare the differences in the linear responses to the four fumigants. First we add the total at the single level to twice the total at the double level, obtaining the figures in row L of table III. The zero level, being common to all fumigants, does not enter. For the 'after fumigation' data, the 3 d.f. are then given by $\frac{1}{20}\{(3596)^2 + (2682)^2 + (3913)^2 + (3136)^2\} - \frac{1}{80}(13327)^2$ = 43,409.

Finally, there are 3 d.f. which measure the differences in the curvatures of the four response curves. To obtain these we subtract the double level from twice the single (line Q in Table III). Again, the zero level is not used. The 3 d.f. are $\frac{1}{20}\{(867)^2 + (979)^2 + (1621)^2 + (662)^2\} - \frac{1}{80}(4129)^2\} = 25,693$.

One reason for making this subdivision is that if curvature effects are significant the simple model will have to be abandoned.

The sums of squares and products are shown in Table IV.

Table IV. Sums of squares and products. x = before, y = after fumigation

	d.f.	(x^2)	(yx)	(y^2)
Replications	3	159,597	175,873	289,427
Treatments	8	29,141	-9,222	157,449
Average linear response	1	3,081	-13,276	57,207
Average curvature	1	2,204	8,285	31,140
Differences in linear	3	22,975	-6,837	43,409
Differences in curvature	3	881	2,606	25,693
Error	36	121,429	189,278	544,690

The regression coefficient b_e is (189,278)/121,429 = 1.558754. We wish first to make an F-test of the adjusted sum of squares for each of the four components of treatments. Strictly speaking, it is necessary to carry out the procedure of section 3 separately for each component. Since this is rather laborious when there are numerous components, we first make approximate F-tests from an analysis of $(y-b_e x)$.

Table V Analysis of variance of $(y - b_e x)$

	d.f.	S.S.	M.S.	F
Treatments	8	257,003	32,125	
Average linear response	1	106,081	106,081(-2,625)	14.87
Average curvature	1	10,667	10,667(- 191)	1.50
Differences in linear	3	120,546	40,182(-4,199)	5.63
Differences in curvature	3	19,709	6,570(-4)	0.92
Error	35	249,652	7,133	

It will be recalled from section 4 that these F-values are all too large. Since, however, the average curvature and the differences in curvature in Table V are both non-significant, there is no need for the exact tests of those components. Further, the two linear terms are both significant at the 1 percent level. To one experienced in such analyses, it would also be evident that the exact tests would not change these verdicts materially. As a matter of interest the correction terms for the mean squares, as calculated from (5), are shown in parentheses in Table V. It should again be emphasized, however, that where treatments affect the x-variable the correction terms may be very substantial, and this approximate method may give misleading results.

The adjusted treatment means are shown in Table VI. For example, for the untreated plots the mean numbers of worms are 123.4 (first count) and 366.1 (second count). Hence the adjusted mean is

$$366.1 - 1.56(123.4 - 128.5) = 374$$

the figure 128.5 being the general mean at the first count.

Table VI Adjusted treatment means (cysts per 400 gms)

Level	CN	CS	CM	CK
0		374		
1	310	270	358	201
2	365	203	289	178

For t-tests of the differences between pairs of treatment means, it is rather inconvenient to have to calculate a separate standard error for every pair, though for exactness this is required because of the term $(\bar{x}_{i.} - \bar{x}_{j.})^2/E_{xx}$. Finney (6) and

Cochran (7) have suggested as an approximation the use of an average derived as follows. The estimated variance of the difference between two adjusted treatment means (averaged over q columns) is

$$s^2_e \left\{ \frac{2}{q} + \frac{(\bar{x}_{i.} - \bar{x}_{j.})^2}{E_{xx}} \right\}$$

The average over all pairs of treatments is found to be $\frac{2s^2_e}{q} \left\{ 1 + \frac{t_{xx}}{E_{xx}} \right\}$ where t_{xx} is the treatments mean square for x. This suggests that we may regard $s^2_e \left(1 + \frac{t_{xx}}{E_{xx}}\right)$ as the estimated variance per plot, to be used for all t-tests based on the adjusted means. As Finney points out, this device cannot be used with safety unless the treatments produce no significant effects on x.

From Tables V and IV, the effective error variance per plot in this example is found to be 7,133 x 1,030 = 7,347. For the figures in Table VI which are based on 4 replicates, the standard error of the difference between two means would therefore be taken as $\sqrt{(2 \times 7{,}347)} / \sqrt{4} = 60.6$*

Since it takes account of the sampling errors of b_e, the effective error variance 7,347, is also the appropriate error to assign to the experiment for answering the question: how much has the error been reduced by covariance? If the first counts had not been taken, the error variance would have been 544,690/36, or 15, 130, (from Table IV). It may therefore be estimated that the use of covariance was equivalent to doubling the number of replicates.

*If we accept the hypothesis of linear responses to the fumigants, the results of this experiment would be summarized in terms of the estimates a, d_1, ... d_4 (after adjustment for the first count), rather than from Table VI. The reader who is interested in the details may wish to verify that (i) the weighted mean of the two levels, giving double weight to the second level, (e.g. (310 + 2 x 365)/3 for CN) is the least-squares estimate $(a - \frac{5}{8}d_1)$. This result implies, incidentally, that t-tests of the differences between these weighted means are identical with t-tests of the differences between the corresponding d's. (ii) while the mean for no-fumigant, 374, is an estimate of α, it is not the least-squares estimate. This happens because quantities like (2 x 310 - 365) are also estimates of α on the null hypothesis. The solution assigns a weight of 20 to the estimate 374 and a weight of 1 to the estimate from each fumigant. The resulting estimate of α is 363. (iii) the estimates of the d's (i.e. the decreases in eelworms to the single level of fumigant) are 10 for CN, 83 for CS, 31 for CM and 106 for CK.

As mentioned previously, this experiment also serves to illustrate the second use of covariance, in which the cysts at the second count are taken as x, and the oats yields as y. Since the error regression of y on x was not significant, no details will be given.

7. Multiple covariance. The extension to the case where there is more than one independent variable presents no essential difficulty, though the computations naturally become more involved. With two variates x and z, the regression coefficients are obtained from the normal equations

$$E_{xx}b_x + E_{xz}b_z = E_{yx}$$

$$E_{xz}b_x + E_{zz}b_z = E_{yz}$$

If t-tests of the adjusted row means are wanted, it is advisable to compute the inverse of the E_{xz} matrix, say w_{xz}.

The adjusted mean of the i^{th} row is estimated by

(21) $\bar{y}_{i.} - b_x(\bar{x}_{i.} - \bar{x}_{..}) - b_z(\bar{z}_{i.} - \bar{z}_{..})$ while the variance of the difference between the i^{th} and j^{th} rows is estimated by

(22) $$s^2_e \left\{ \frac{2}{q} + (\bar{x}_{i.} - \bar{x}_{j.})^2 w_{xx} + 2(\bar{x}_{i.} - \bar{x}_{j.})(\bar{z}_{i.} - \bar{z}_{j.})w_{xz} + (\bar{z}_{i.} - \bar{z}_{j.})^2 w_{zz} \right\}$$

The average over all pairs of rows is

(23) $$\frac{2}{q} s^2_e \left\{ 1 + r_{xx}w_{xx} + 2r_{xz} w_{xz} + r_{zz}w_{zz} \right\}$$

where r_{xx} is the rows mean square for x, etc. In experiments where the rows do not affect x or z, this expression may be used to derive an average standard error for t-tests.

Calculation of the exact F-test is laborious, since two sets of regression equations must be solved for each F. If rows do not affect x or z, it is usually best to begin with an approximate F-test based on the analysis of variance of $(y - b_x x - b_z z)$. In cases where this F is highly significant or non-significant it will not be necessary to obtain the correct F. If, however, rows affect x or z the safest procedure is to compute the correct F.

8. More complex classifications - the split-plot design. The procedures also extend readily to more complex classifications. A new situation arises, however, with classifications where the mathematical model requires more than one error variance. Although thorough discussion of such cases is beyond the scope of this paper, the nature of the problem will be illustrated for the split-plot design with a single independent variate. In this case, owing to the positive correlation between sub-plots in the same whole-plot, treatments applied to whole-plots have in general a larger error than those applied to sub-plots.

In the analysis of variance, two independent estimates of the regression are obtained: one, b_w, from the whole-plot error and one, b_s, from the sub-plot error. Often there will be a priori reasons for believing that β_w and β_s are equal. If this hypothesis is not contradicted by the data, what is wanted is a pooled estimate. As might be guessed, the maximum likelihood estimate assigns to each b a weight that is the corresponding E_{xx}/s^2_e. Since the denominators s^2_e depend on the value of b, successive approximation is needed and the resulting sampling theory is not simple. Often, however, it will be evident that b_s supplies most of the information and can be used satisfactorily as the estimate. An example in which this was done has been presented by Bartlett (9).

If the whole-plot regression differs from the sub-plot regression, b_w may be used to adjust the whole-plot yields and b_s to adjust the sub-plot yields. Construction of the table of adjusted treatment means requires a little care. The adjustment to the mean $\bar{y}_t$ for any treatment combination is

$$- b_s(\bar{x}_t - \bar{x}_w) - b_w(\bar{x}_w - \bar{x}_g)$$

where $\bar{x}_t$, $\bar{x}_w$ and $\bar{x}_g$ are respectively the x means for the treatment combination, the corresponding whole-plot treatment and the whole experiment. The standard error required for any t-test of the difference between adjusted treatment means can be worked out from the algebraic structure of the difference.

Even with a relatively simple classification, it may occasionally be considered that β will vary in different parts of the data. Thus with a two-way classification, β might vary from column to column. If it seems worthwhile, a separate b may

be fitted for each column. Usually there is no convenient simplification of the normal equations in these cases, which are best treated by straight-forward regression methods.

9. The estimation of components of variance. Just as in the analysis of variance, the mathematical model for covariance may be appropriate in cases where some or all of the unknowns ξ_i, γ_j etc. are regarded not as parameters, but as random variates whose variances we wish to estimate. In other words, the function of the analysis is to estimate certain components of the variance of $(y - \beta x)$. The statistical problems involved appear to have been little discussed. Only an introductory account will be attempted.

Consider for simplicity a one-way classification, where y_{ij} is the j^{th} member of the i^{th} row ($i = 1,2, ..p$; $j = 1,2, ...q$).

The model is

$$(24) \quad y_{ij} = \mu + \xi_i + \beta x_{ij} + e_{ij}$$

where ξ_i, e_{ij} are all normally and independently distributed with zero means and variances σ^2_r, σ^2_e respectively, while μ, β and the x's are fixed.

By familiar transformations the log. of the likelihood is expressed by the equation

$$-2L = \sum_{i,j} \frac{\{(y_{ij}-\bar{y}_{i.}) - \beta(x_{ij} - \bar{x}_{i.})\}^2}{\sigma^2_e} + q \sum_i \frac{\{(\bar{y}_{i.} - \bar{y}_{..}) - \beta(\bar{x}_{i.} - \bar{x}_{..})\}^2}{\sigma^2_g}$$

$$+ pq \frac{\{(\bar{y}_{..} - \mu - \beta \bar{x}_{..})^2\}}{\sigma^2_g} + 2p \log \sigma_g + 2p(q-1) \log \sigma_e$$

where $\sigma^2_g = \sigma^2_e + q\sigma^2_r$.

The close parallel with the problem of the split-plot design will be noted. The maximum likelihood estimate of β is a weighted mean of the b's derived from the rows and error lines in the analysis of variance, and successive approximation is involved. If, however, the information about β from the rows line can be ignored without too much loss of information, the procedure becomes easier.

It may be shown by standard methods that the residual error sum of squares $(E_{yy} - E^2_{yx}/E_{xx})$ is an unbiased estimate of $p(q-1)\sigma^2_e$ and, with normality assumed,

is distributed as $\chi^2 \sigma^2_e$ with $p(q-1)$ d.f. The adjusted sum of squares for rows, calculated by the usual subtraction procedure, will now be considered. As in the case of the two-way classification, this divides into two parts.

$$(25)\quad q\Sigma\left\{(\bar{y}_{i.} - \bar{y}_{..}) - b_r(\bar{x}_{i.} - \bar{x}_{..})\right\}^2 + (b_r - b_e)^2 \frac{R_{xx}E_{xx}}{R_{xx}+E_{xx}}$$

where b_r is the regression coefficient from the rows line. Now from (24)

$$(26)\quad \bar{y}_{i.} = \mu + \varphi_i + \beta\bar{x}_{i.} + \bar{e}_{i.}$$

so that the quantities $\bar{y}_{i.}$ are independent and have a linear regression on the $\bar{x}_{i.}$, with residual variance $(\sigma^2_r + \sigma^2_e/q.)$ Hence by ordinary regression theory the first tern in (25) is distributed as $\chi^2(\sigma^2_e + q\sigma^2_r)$ with $(p-2)$ d.f.

For the second term, we have

$$b_r = \frac{q\sum_i \bar{y}_{i.}(\bar{x}_{i.} - \bar{x}_{..})}{R_{xx}} = \beta + \frac{q\Sigma(\varphi_i + \bar{e}_{i.})(\bar{x}_{i.} - \bar{x}_{..})}{R_{xx}}$$

$$b_e = \frac{\sum_{i,j} y_{ij}(x_{ij} - \bar{x}_{i.})}{E_{xx}} = \beta + \frac{\Sigma e_{ij}(x_{ij} - \bar{x}_{i.})}{E_{xx}}$$

Consequently $(b_r - b_e)$ is normally distributed with mean zero and variance

$$\frac{(\sigma^2_e + q\sigma^2_r)}{R_{xx}} + \frac{\sigma^2_e}{E_{xx}}$$

Hence the second term

$$(b_r - b_e)^2 \frac{R_{xx}E_{xx}}{R_{xx}+E_{xx}} = \chi^2\left\{\sigma^2_e + \left(\frac{E_{xx}}{R_{xx}+E_{xx}}\right)(q\sigma^2_r)\right\}$$

with 1 d.f.

To summarize

$$\text{ASSR} = (p-1)s^2_r = (\sigma^2_e + q\sigma^2_r)\chi^2_{(p-2)} + (\sigma^2_e + \lambda q\sigma^2_r)\chi^2_1$$

$$\text{Residual error SS} = \left\{p(q-1) - 1\right\}s^2_e = \sigma^2_e\chi^2_{\{p(q-1)-1\}} \quad \text{where } \lambda = E_{xx}/(R_{xx} + E_{xx}).$$

Further, by customary analysis of variance theory, it is easy to show that the three χ^2 are independent.

It follows that s^2_e is an unbiased estimate of σ^2_e; while an unbiased estimate of σ^2_r is

$$\frac{(s^2_r - s^2_e)}{q}\,\frac{(p-1)}{(p-2+\lambda)}$$

This estimate involves some loss of information at two stages (i) information in the rows about β is ignored (ii) the single d.f. in the ASSR is presumably given slightly too much weight. It seems likely that the loss of information will often be trivial.

If the estimate of β from the error line is used, the results for the two-way classification are exactly parallel.

The best method for obtaining confidence limits for the ratio $v = \sigma^2_r / \sigma^2_e$ is not quite clear. One procedure would be to use the fact that

$$\left\{ \frac{(\text{ASSR})_{(p-2)}}{(1+qv)} + \frac{(\text{ASSR})_{(1)}}{(1+\lambda qv)} \right\} \Big/ (p-1)s^2_e$$

is distributed as F with (p-1) and $\left[p(q-1)-1\right]$ d.f., where the suffixes distinguish the two components of the ASSR. The upper and lower limits can then be found by trial and error.

References

(1) R. A. Fisher. Statistical Methods for Research Workers. Edinburgh, Oliver and Boyd, 4th ed., (1932), § 49.1

(2) F. Yates. A complex pig-feeding experiment. Jour. Agr. Sci., Vol 24 (1934), p 519.

(3) F. Yates. Orthogonal functions and tests of significance in the analysis of variance. Jour. Roy. Stat. Soc. Suppl., Vol 5, (1938), pp. 177-180.

(4) A. Wald. Lectures on the analysis of variance and covariance. Columbia University. (1946).

(5) C. I. Bliss. An experimental design for slope-ratio assay. Ann. Math. Stat., Vol 17, (1946), pp 232-237.

(6) D. J. Finney. Standard errors of yields adjusted for regression on an independent measurement. Biom. Bull. Vol 2, (1946), pp 53-55.

(7) W. G. Cochran. The analysis of lattice and triple lattice experiments in corn varietal tests. II. Mathematical theory. Iowa Agr. Exp. Sta. Res. Bull. 281, (1940), pp 64-65.

(8) M. S. Bartlett. A note on the analysis of covariance. Jour. Agr. Sci., Vol 26, (1936), pp 488-491.

(9) M. S. Bartlett. Some examples of statistical methods of research in agriculture and applied biology. Jour. Roy. Stat. Soc. Suppl., Vol 4, (1937), pp 142-146.