REGRESSION METHODS FOR STUDYING GENOTYPE-ENVIRONMENT INTERACTIONS

R. C. HARDWICK and J. T. WOOD
National Vegetable Research Station, Wellesbourne, Warwick

Received 20.vii.71

1. Introduction

When the performance of a set of genotypes is compared over a number of environments it is frequently found that genotypic effects are not fixed, but that they vary between environments. In biometrical investigations the problem then arises of how to characterise genotypic effects. A number of authors (*e.g.* Yates and Cochran, 1938; Finlay and Wilkinson, 1963; Eberhart and Russell, 1966; Perkins and Jinks, 1968*a*, *b*) have shown that in many such cases the performance of an individual genotype can be expressed as a linear function of the environmental index (the additive environmental effect). The slope of this regression, a dimensionless quantity, is a measure of the sensitivity of the genotype to the totality of environmental factors. Where the regressions are found to account for a substantial part of the genotype environment interaction variance, this empirical approach to the problem of interactions has proved to have considerable value. An identical technique has been used by Mandel (1959) and Mandel and Lashof (1959) to compare the result of tests at several laboratories on a number of materials, and the theoretical aspects of the statistical analysis have been discussed in a series of papers by Mandel (1961, 1969, 1970, 1971) and Mandel and McCrackin (1963). There has not, to our knowledge, been any reference to this work by biologists, nor as Mather (1971) has pointed out, any discussion of the underlying reasons why the regression technique works for plant material, although various authors have either denied that any such *a priori* reason exists (Wright, 1971), or have attributed the success of the method to the linearising nature of the transformation (Knight, 1970).

In contrast to the empirical methods mentioned above, the classical approach of plant physiologists to the problem of genotypic performance has been by multifactorial experiments and growth analysis. This approach has led to the formulation of various regression models which relate plant performance to environmental variables, and more recently to computer simulation models of the same relationship (*e.g.* de Wit and Brouwer, 1968). Freeman and Perkins (1971) have suggested on statistical grounds that it may be better to use this approach to analyse genotype-environment interactions, by regression on environmental variables, rather than to use regression on the environmental mean. This paper considers the two regression methods and their interrelationship from a different point of view.

First we show that there is a bias in the estimates of the coefficients of regression on the environmental mean when they are derived by the usual method. After considering the relations of this approach to various other techniques, we turn to the interpretation of the values of the coefficients of regression and of the deviations from the regressions on the environmental mean. It will be shown that it is possible to account for their values in terms of the coefficients of an " underlying " regression model. Finally a

method of analysis which is intermediate between regression on environmental variables and regression on the environmental mean is suggested.

2. The empirical approach

Suppose the performance of m genotypes is measured in n environments, and y_{ij} is the performance of the ith genotype in the jth environment, and suppose

$$y_{ij} = \mu + \rho_i + \gamma_j + g_{ij} + e_{ij} \tag{1}$$

where $$\sum_i \rho_i = 0, \sum_j \gamma_j = 0, \sum_i g_{ij} = 0, \sum_j g_{ij} = 0,$$

and e_{ij} is a normally distributed random variable with mean zero and variance σ^2. The more important symbols are summarised in the accompanying table. It is assumed that all effects are fixed. The model considered by

Table of the more important symbols

Symbol	Interpretation	See equation
$\mathbf{a}_i$	Vector whose elements are a_{ih}	(6)
a_{ih}	Partial regression coefficient of y_{ij} on x_{jh}, j varying	(6)
$\mathbf{d}_i$	Perpendicular vector from $\boldsymbol{a}_i$ on to $\mathbf{a}$.	(8)
e_{ij}	Normal random variable, mean 0, variance σ^2	(1)
g_{ij}	Interaction between genotype and environment	(1)
m	Number of genotypes, indexed by $i = 1, m$	
n	Number of environments, indexed by $j = 1, n$	
p	Number of environmental variables, indexed by $h = 1, p$	
x_{jh}	Orthonormalised measure of the hth environmental variable in the jth environment	(6)
y_{ij}	Observation of the ith genotype in the jth environment	(1)
β_i	Coefficient of regression of g_{ij} on γ_j	(2)
β_i'	$1+\beta_i$	(4)
γ_j	Effect of jth environment	(1)
δ_{ij}	Deviation from regression of g_{ij} on γ_j	(8)
μ	Overall mean	(1)
ρ_i	Effect of ith genotype	(1)
σ	Standard deviation of e_{ij}	(1)

Perkins and Jinks (1968*a*) assumes that the genotype-environment interaction, g_{ij}, is linearly related to the environmental effect, γ_j, *i.e.*

$$g_{ij} = \beta_i \gamma_j \tag{2}$$

where $$\sum \beta_i = 0.$$

Mandel and Lashof (1959) consider the identical model formulated in terms of laboratories and materials under test at the laboratories.

Estimates of μ, ρ_i, and g_{ij} are:

$$\begin{aligned}\hat{\mu} &= y_{..}\\ \hat{\rho}_i &= y_{i.} - y_{..}\\ \hat{\gamma}_j &= y_{.j} - y_{..}\\ \hat{g}_{ij} &= y_{ij} - y_{i.} - y_{.j} + y_{..}\end{aligned}$$

A dot replacing a subscript indicates that an arithmetic mean has been taken over the entire range of the subscript.

In practice the β_i's are estimated by

$$\hat{\beta}_i = \sum_j \hat{g}_{ij}(y_{.j}-y_{..})/\sum_j (y_{.j}-y_{..})^2.$$

This can be rewritten as

$$\hat{\beta}_i = \sum_j (\beta_i\gamma_j+e_{ij}-e_{.j})(\gamma_j+e_{.j})/\sum_j (\gamma_j+e_{.j})^2,$$

and it can be shown that to a first approximation

$$E(\hat{\beta}_i) = \beta_i\{1-(n-1)\sigma^2/(m\sum\gamma_j^2)\}. \tag{3}$$

Thus there is a tendency to underestimate the absolute value of β_i. The bias arises (Sprent, 1969, p. 33) because the assumption has to be made in regression analysis that the independent variable, in this case the environmental mean, is measured without error. The bias depends on m and the ratio of the between environments variation to the error mean square. In most well designed experiments this ratio will be large, and the bias will be small, but not necessarily negligible. The bias will also tend to be small for large m. This is illustrated in fig. 1 for $m = 2, 5, 12$, $\beta = 0{\cdot}5$ and 1. In the case where $m = 2$, $\beta_1 = -\beta_2$ and it is usual to estimate $\beta = \beta_1-\beta_2$ (Bucio Alanis, 1966), but the biases remain.

Equations (1) and (2) can also be written as

$$y_{ij} = \mu+\rho_i+\beta_i'\gamma_j+e_{ij} \tag{4}$$

where $\beta_i' = 1+\beta_i$. This is the form used by Mandel and Lashof (1959) and Finlay and Wilkinson (1963). The use of this approach as a basis for an analysis of variance and associated tests of hypothesis has been discussed by Freeman and Perkins (1971), and also by Mandel (1961), who shows it to be an extension of Tukey's "One degree of freedom for non-additivity" (Tukey, 1949; Scheffé, 1959, pp. 129-34).

An equivalent model has been considered by Williams (1952), who shows, in effect, that least squares estimation of the β_i's is equivalent to extracting the first principal component of the genotypic performances. The validity of the model can then be studied by extracting further components, or, more informally, by inspection of the residual correlation matrix after extracting the first component. A procedure similar to the latter approach has been suggested by Perkins and Jinks (1968*b*).

Gollob (1968) and Mandel (1969) consider the principal components approach further using the model

$$g_{ij} = \lambda_1 u_{i1}v_{j1}+\lambda_2 u_{i2}v_{j2}+\ldots e_{ij} \tag{5}$$

where $$\sum_i u_{ik} = \sum_j v_{jk} = 0,\ \sum_i u_{ik}^2 = \sum_j v_{jk}^2 = 1$$

and successive terms of (5) are chosen to maximise the sum of squares removed from $\sum g_{ij}^2$ at each stage. If $\mathbf{G}$ is the matrix (g_{ij}), u_{ik} and v_{jk} are elements of the eigenvectors of $\mathbf{G}.\acute{\mathbf{G}}$ and $\acute{\mathbf{G}}.\mathbf{G}$, and the λ_k's are the square roots of the associated eigenvalues. Mandel (1969) discusses the application of analysis of variance to this model.

For the Perkins and Jinks model, since the main interest lies in genotypic

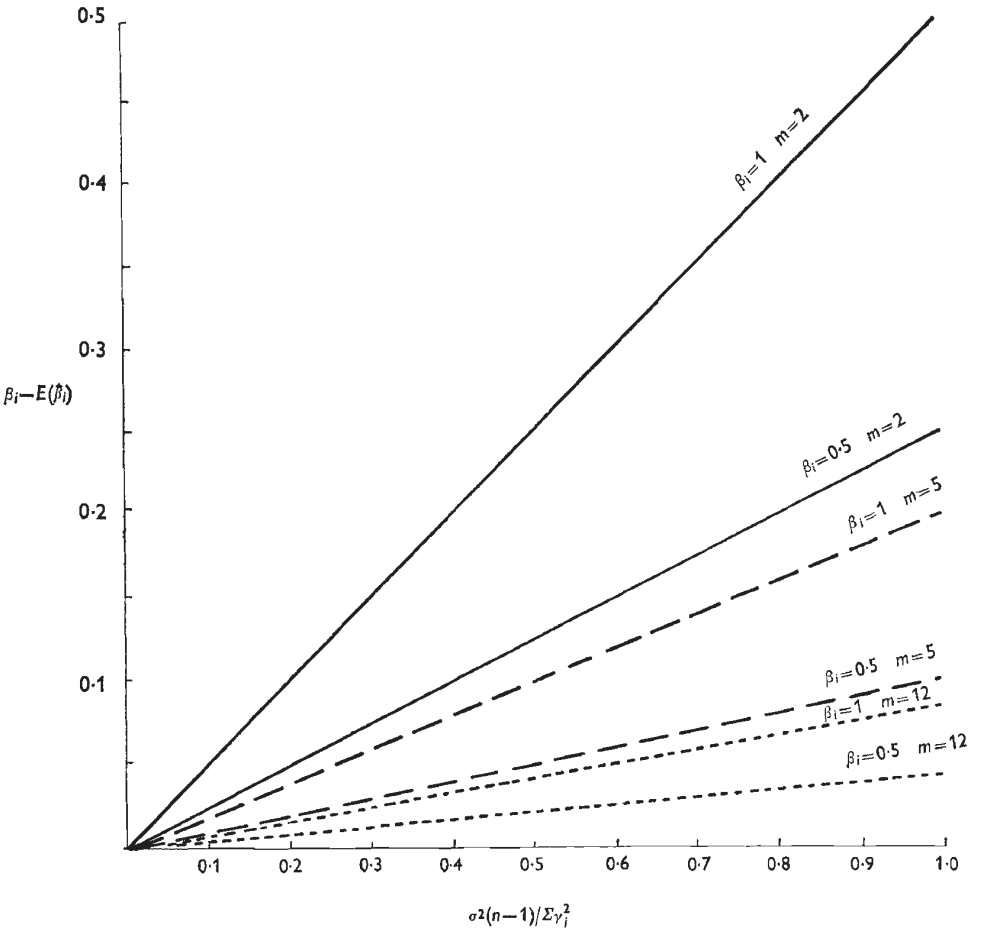

FIG. 1.—Bias in Perkins and Jinks' (1968*a*) estimate of β_i, versus $\sigma^2(n-1)/\Sigma\gamma_j^2$ for m (number of genotypes) = 2, 5 and 12, β = 0·5 and 1.

performance, a more logical extension of this approach is to assume, by analogy with equation (4), that

$$y_{ij} = \mu + \rho_i + \lambda_1 u_{i1} v_{j1} + \lambda_2 u_{i2} v_{j2} + \ldots e_{ij}$$

without the restriction $\sum_i u_{ik} = 0$, but the basic idea, of chosing successive terms to maximise the sums of squares which each removes, remains the same.

3. USE OF MULTIPLE REGRESSION

We now consider the situation when both phenotypic performance and also a number of relevant environmental variables have been measured. It is convenient to assume that the response to these variables is linear or can at least be represented by a polynomial of low degree. The model is then

$$y_{ij} = \mu + \rho_i + a_{i1} x_{j1} + a_{i2} x_{j2} + \ldots + a_{ip} x_{jp} + e_{ij} \qquad (6)$$

where the x's are measures of the environments, the a's are regression coefficients and e_{ij} is a random normal deviate with mean zero and variance σ^2.

The x's may measure completely different aspects of the environment or they may be terms of a polynomial. The x variates can always be transformed so that they are orthogonal, each with mean zero and sum of squares equal to one, and, because it makes the interpretation of variance components easier, in the following it is assumed that this orthonormalisation has been done.

All the parameters in equation (6) can be estimated without difficulty using least squares. After removing genotypic effects the sum of squares can be partitioned into four terms; the between environments sum of squares can be partitioned into a regression component and a residual component, and a similar partitioning can be applied to the sum of squares for genotypes × environments. This is shown in table 1. These partitioned terms will be referred to as the overall regression and deviations, and the heterogeneity of regressions and of deviations. Since the partitioning is orthogonal the two terms for deviations from regression can be combined to provide a single estimate of variation about the regression lines. This combined component which is part of the sum of squares for variation within genotypes, has $m(n-p-1)$ degrees of freedom, and may be expressed as $\sum_i \sum_j (e_{ij}-e_{..})^2$. It has been partitioned by making the customary assumption that all errors are independent. Thus

$$\sum_i \sum_j (e_{ij}-e_{..})^2 = \sum_i \sum_j (e_{ij}-e_{.j})^2 + m\sum_j (e_{.j}-e_{..})^2.$$

The first component has $(m-1)(n-p-1)$ degrees of freedom and the second has $(n-p-1)$ degrees of freedom. These terms, called here " heterogeneity of deviations " and " overall deviation " are identical with those for " interaction residual " and " environments residual " in tables 2 and 3 of Freeman and Perkins (1971). The derivation given here emphasises the nature of these residuals (see test 2 below).

TABLE 1

Regression on environmental variables

Item	d.f.	Expectation of mean square
Genotypes	$m-1$	$n\sum_i \rho_i^2/(m-1)+\sigma^2$
Environments	$n-1$	$m\sum_k a_{.k}^2/(n-1)+\sigma^2$
Regression	p	$m\sum_k a_{.k}^2/p+\sigma^2$
Deviations from regression	$n-p-1$	σ^2
Genotypes × environments	$(m-1)(n-1)$	$\sum_{i,k}(a_{ik}-a_{.k})^2/((m-1)(n-1))+\sigma^2$
Regression	$(m-1)p$	$\sum_{i,k}(a_{ik}-a_{.k})^2/((m-1)p)+\sigma^2$
Deviations from regression	$(m-1)(n-p-1)$	σ^2
Total	$mn-1$	

The following tests can be made in table 1.

(1) If either the overall regression mean square or the heterogeneity of

regression mean square is significantly greater than the total deviations from regression mean square, then this is evidence that the model identifies some of the environmental variables associated with variation in performance.

(2) If there is replication the mean square for the overall deviations can be compared with the residual mean square, testing for the existence in the deviations e_{ij} of a systematic element $(e_{.j}-e_{..})$ which is common to the deviations of all genotypes about their respective regressions. Such deviations may arise in two ways; either because a variable to which all genotypes respond similarly has been neglected, or because the underlying relationships with the included variables are not (as assumed) linear.

(3) If the latter is the case then similar patterns of deviations will be expected to occur every time the experiment is repeated. If on the other hand the deviations arise through the action of a neglected variable, then similar values of this variable, and hence similar patterns of deviations, are unlikely to recur. Hence by repeating the experiment in a different site or season it is possible tentatively to distinguish between the alternative origins of $(e_{.j}-e_{..})$.

(4) The heterogeneity of deviations term may be tested against the residual mean square. A significant heterogeneity of deviations term indicates failure of the model for individual genotypes and as before, this may be due either to inadequacies in the chosen regression function, or to the operation of an undetected variable. Again, the two alternatives can tentatively be distinguished by repeating the experiment.

4. Regression on the environmental mean

Where regression is performed not on independent measures of the environment, but on the environmental mean, the analysis of variance shown in table 2 can be constructed (Mandel and Lashof, 1959). In this case, the overall deviation from regression is zero, because the overall regression has been constrained to be a perfect fit to the mean environmental yields.

As already shown, equation (3), estimation of β_i by unweighted least

Table 2

Joint regression on the environmental mean

Item	d.f.	Expectation of mean square
Genotypes	$m-1$	$n\sum_i \rho_i^2/(m-1)+\sigma^2$
Environments	$n-1$	$m\sum_k a_{.k}^2/(n-1)+\sigma^2$
Heterogeneity of regressions	$m-1$	$\sum_k a_{.k}^2 \sum_i \beta_i^2/(m-1)+\sigma^2(1+\Delta)$
Residual	$(m-1)(n-2)$	$\left\{\sum_{i,k}(a_{ik}^2-a_{.k}^2)-\sum_k a_{.k}^2\sum_i \beta_i^2\right\}/\{(m-1)(n-2)\}$ $+\sigma^2(1-\Delta)/(n-2)$
Total	$mn-1$	

$$\Delta = \left\{\frac{\sum_{i,k}(a_{ik}^2-a_{.k}^2)}{\sum_k a_{.k}^2} - n\sum_i \beta_i^2\right\}/m.$$

squares leads to a bias. The unbiased estimate is given by Williams' (1952) method (see Appendix).

Some authors (for example Mandel, 1969; Wright, 1971) have used regression on the $y_{i.}$'s as well as on the $y_{.j}$'s. The condition for these regressions to be appropriate is that the regressions on $y_{.j}$ should be concurrent (Mandel, 1961).

The empirical coefficient β_i can be expressed in terms of the coefficients of the underlying model introduced in the previous section. By equating the two models (2) and (6) we find, using the orthonormality of the x variates, that

$$\beta_i = \sum_h \{(a_{ih} - a_{.h})a_{.h}\} / \sum_h a_{.h}^2. \tag{7}$$

Equation (7) and results which follow from it, can be written in a more convenient form by using vector notation. If $\mathbf{a}_i'$ is the vector $(a_{i1}, a_{i2} \ldots a_{ip})$ and $\mathbf{a}'_.$ the vector $(a_{.1}, a_{.2} \ldots a_{.p})$ then

$$\beta_i = (\mathbf{a}_i'\mathbf{a}_./\mathbf{a}'_.\mathbf{a}_.) - 1$$

and thus $(1+\beta_i)$ is proportional to the projection of the vector $\mathbf{a}_i$ on to $\mathbf{a}_.$ (see, for example, Lanczos, 1957, p. 360). If $\mathbf{a}_i$ is a multiple of $\mathbf{a}_.$ for all i the data will conform to the simple model of equation (2);

$$y_{ij} = \mu + \rho_i + (1+\beta_i)\gamma_j + e_{ij}.$$

Deviations from this model will occur if $\mathbf{a}_i$ is not a multiple of $\mathbf{a}_.$ for all i; *i.e.* if there is variation between genotypes in more than one dimension. These deviations from the linear model are given by

$$\begin{aligned} \delta_{ij} &= g_{ij} - \beta_i\gamma_j \\ &= \mathbf{d}_i'\mathbf{x}_j \end{aligned} \tag{8}$$

when $\mathbf{x}_j' = (x_{j1}, x_{j2} \ldots x_p)$

and $\mathbf{d}_i$ is given by

$$\mathbf{d}_i = \mathbf{a}_i - \beta_i\mathbf{a}_.$$

Since $\mathbf{d}_i$ and the projection $\beta_i\mathbf{a}_.$ represent an orthogonal partitioning of the vector $\mathbf{a}_i$ it follows that the deviations δ_{ij}, the sum of squares of which have been proposed by Eberhart and Russell (1966) as a second parameter of " stability ", are not independent of the regression on the environmental mean but are rather a necessary adjunct of the line fitting procedure. The deviations sum of squares will vanish only for the case $p = 1$. Equation (8) can be written in terms of the coefficients of the model (6) as

$$\delta_{ij} = \left\{ \sum_{s=2}^{p} \sum_{r=1}^{s-1} (a_{.r}x_{js} - a_{.s}x_{jr})(a_{.s}a_{ir} - a_{.r}a_{is}) \right\} \Big/ \sum_{h=1}^{p} a_{.h}^2. \tag{9}$$

The derivation of equations (7), (8) and (9) has been based on expectations; in practice the deviations sum of squares will also contain a random error component. If in addition there are significant deviations from the assumed underlying model, then β_i and δ_{ij} will not be accurately predicted by these equations.

This analysis can be extended to the tests of significance in the analysis of variance of regression on the environmental mean (table 2). If σ^2 is small, the test of heterogeneity of regression against residual will give a " significant " result if

$$F = (n-2)(\mathbf{a}'_{\cdot}\mathbf{a}_{\cdot}\sum_i \beta_i^2)(\mathbf{a}_i'\mathbf{a}_i - \mathbf{a}'_{\cdot}\mathbf{a}_{\cdot} - \mathbf{a}'_{\cdot}\mathbf{a}_{\cdot}\sum_i \beta_i^2)^{-1}$$

$$= (n-2)(\sum_i \mathbf{a}_i'\mathbf{a}_{\cdot}\mathbf{a}_i'\mathbf{a}_{\cdot} - \mathbf{a}'_{\cdot}\mathbf{a}_{\cdot}\mathbf{a}'_{\cdot}\mathbf{a}_{\cdot})/(\mathbf{a}'_{\cdot}\mathbf{a}_{\cdot}\sum_i \mathbf{d}_i'\mathbf{d}_i) \qquad (10)$$

is sufficiently large, that is if $\sum_i \mathbf{d}_i'\mathbf{d}_i$ is small. If this term is not small the residual will be large compared with the error mean square, and then the model (equation (4)) is an inadequate description of the data and a more elaborate model, for example equations (5) or (6) is called for. Perkins and Jinks (1968*b*) have shown that in such cases the correlation between deviations for pairs of varieties can sometimes be used to identify groupings of similar genotypes. The correlation between the deviations of the ith and jth genotype is

$$(\mathbf{d}_i'\mathbf{d}_j)/(\mathbf{d}_i'\mathbf{d}_i\mathbf{d}_j'\mathbf{d}_j)^{\frac{1}{2}}$$

and this will be close to 1 or -1 as the varieties show similar or discrepant patterns of departure from their regressions on the overall mean. As mentioned previously, this form of analysis is an informal alternative to the extraction of further components, in the principal components approach.

It will now be shown that equations (7) and (8) can be used to account for the coefficients of joint regression analysis in a published experiment; that of Richards (1965) on the performance of three species of tropical grass in 48 nominated environments. Because the levels of the environmental variables in this experiment—a $3 \times 4 \times 4$ combination of cutting heights, cutting frequencies and fertiliser level—are known, the data are suitable for analysis by regression on the environmental mean and by regression on external variables. Following a preliminary analysis the data were transformed to logarithms. Regression on the environmental mean accounted for a significant part of the genotype-environment interaction, but the deviations from regression were large (table 3). The bias in the estimates of β_i was small because the estimated ratio of between plot variation to within plot variation was large.

TABLE 3

Joint regression on the environmental mean of Richards (1965) data

Item	d.f.	Mean square
Species	2	1·787 7
Environments	47	0·359 3
Species × environment	94	0·032 0
Regression	2	0·246 5
Residual	92	0·027 4
Total	143	

An estimate of the experimental error is provided by the mean square for the highest order interaction—0·004 with 36 degrees of freedom.

Linear models were fitted to the transformed data, by regression of yield on the levels of the three experimental variables. 54·78 per cent. of the within species variance was accounted for by first degree equations, and 55·88 per cent. by including quadratic terms for fertiliser and cutting frequency. The coefficients of these regressions were substituted in equations (7) and (8) to predict the values of the two stability parameters of Eberhart

and Russell (1966), and the entries in the correlation matrix of Perkins and Jinks (1968*b*). Observed and predicted values were in reasonable agreement (table 4).

TABLE 4

Prediction of coefficients of joint regression, and correlations between deviations from regression on the environmental variables, Richards (1965) data

	β_i	
	Observed	Predicted
Cynodon dactylon	0·156 3	0·153 2
Digitaria decumbens	0·081 4	0·135 4
Panicum maximum	−0·237 7	−0·288 6

	Correlation between deviations					
	Observed			Predicted		
C. dact.	1·0			1·0		
D. dec.	−0·033	1·0		0·074	1·0	
P. max.	−0·66	−0·73	1·0	−0·89	−0·52	1·0
	C.d.	*D.d.*	*P.m.*	*C.d.*	*D.d.*	*P.m.*

Bias in estimates of $\beta_i = 0{\cdot}00371\ \beta_i$ (from equation (3)).

5. WILLIAMS' METHOD

Williams (1952) considers the model

$$E(y_{ij}) = \mu + \rho_i + \theta c_i w_j \tag{11}$$

subject to the constraints $\sum c_i^2 = \sum w_j^2 = 1$. This is equivalent to reparameterising the model of equation (2). In terms of equation (2) w_j, c_i and θ in equation (11) are given by

$$w_j = \frac{\gamma_j}{(\sum \gamma_j^2)^{\frac{1}{2}}}$$

$$c_i = \frac{1+\beta_i}{\{\sum (1+\beta_i)^2\}^{\frac{1}{2}}}$$

and

$$\theta^2 = \sum \gamma_j^2 \sum (1+\beta_i)^2$$

Williams proposes estimating the parameters of the model (11) by minimising

$$\sum_{i,j} (y_{ij} - \mu - \rho_i - \theta c_i w_j)^2 + 2v_1 \sum \theta w_j + 2v_2 \sum \rho_i + v_3 \sum c_i^2$$

where the θw_j's are here treated as one set of constants and v_1, v_2 and v_3 are Lagrange multipliers.

He shows that, if $\mathbf{Y}$ is the matrix with elements $(y_{ij} - y_{i.})$, then θ^2 is the largest eigenvalue of $\mathbf{YY'}$ and $\mathbf{Y'Y}$ and $\{c_i\}$ and $\{w_j\}$ are the corresponding eigenvectors. Essentially the same derivation was used by Mandel (1969) in the formulation of equation (5).

A similar approach may be used to reduce the large number of parameters in equation (6). Analogy with equations (1) and (11) suggests putting

$$a_{ih} = \theta \phi_i a_h \tag{12}$$

where

$$\begin{aligned} \sum_i \phi_i^2 &= 1 \\ E_j &= \sum_h \alpha_h x_{hj} \\ \sum_j E_j &= 0 \end{aligned} \tag{13}$$

and

$$\sum_j E_j^2 = 1.$$

E_j provides a single measure of the jth environment. The parameters of this model are estimated by minimising

$$S = \sum_{i,j} (y_{ij} - \mu - \rho_i - \theta \phi_i E_j)^2$$

subject to the constraints. This minimisation is considered in the Appendix. There it is shown that, if $\mathbf{X}$ is the matrix whose elenents are $\{x_{hj}\}$ then

$$(\mathbf{X'Y'YX} - \hat{\theta}^2 \mathbf{X'X})\alpha = 0 \tag{14}$$

where $\hat{\theta}$ takes the largest value possible in equation (14), and α is the vector whose elements are the estimates of the α_h's. Since the x variates are orthogonal $\mathbf{X'X}$ is the $p \times p$ identity matrix $\mathbf{I}$, and

$$(\mathbf{X'Y'YX} - \theta^2 \mathbf{I})\alpha = 0.$$

Thus this procedure is equivalent to projecting the yields for each variety on to the space spanned by the x variates, and then extracting the first principal component of the projections, so it is a natural extension of Williams' method. The application of significance tests when this method is used, and the possibility of extracting more than one component, requires further investigation.

6. Discussion

Until it was discovered that a significant part of the interaction may often be accounted for by regression on the additive environmental component (Yates and Cochran, 1938; Finlay and Wilkinson, 1963), interaction between genotype and environment presented an intractable problem to plant breeders (see, for example, the discussions at some symposia before 1963; Warren, 1955; Hanson and Robinson, 1963). With the regression method it became possible for the previously unaccountable part of a variety's performance to be expressed by two empirical parameters—the slope of the regression line, and the sum of squares of deviations from regression (Eberhart and Russell, 1966). These parameters have proved to have considerable utility in genetics and plant breeding. Evidence has been presented by Jinks and Perkins (1970) that predictions of the slope parameter can be made both across environments and across generations. Perkins and Jinks (1968*b*) have shown that the deviations from regression can be used to identify groupings of related genotypes. Many breeders have successfully applied regression to their material as an empirical measure of genotypic sensitivity. But as Mather (1971) has pointed out, this work lacks a theoretical foundation.

In this paper we have tried to show how joint regression analysis is related to various other statistical techniques and why it should have been successful in so many instances. If the variation in genotypic performance

between environments can be described by a linear function of environmental variables, and if the coefficients of this relationship obey the condition expressed in equation (10), then it has been shown that significant linear relationships between individual and mean performance will be found. From the frequency with which such relationships have been observed it appears that this requirement is usually satisfied; *i.e.* in most experiments variation in the response coefficients $\mathbf{a}_i$ between genotypes is not very great.

The important assumption is made here that the " underlying " response to environmental variables is linear. Many relationships between performance and environmental variables have been postulated, some linear (*e.g.* Putter, Yaron and Bielorai, 1966), some non-linear (*e.g.* Richards, 1959; Nelder, Austin, Bleasdale and Salter, 1960), and recently some have been advanced which are only expressible by computer simulation (*e.g.* Paltridge, 1970; de Wit and Brouwer, 1968). We assume that these relationships are in accord with reality, and that they can be approximated for practical purposes by a polynomial, that is by a linear function of environmental variables. It is recognised that the resulting coefficients will not necessarily be easy to interpret in physiological terms, but as a working hypothesis the assumption of linearity seems reasonable. It will be seriously wrong only in cases (as perhaps susceptibility to pest or disease) where performance is a discontinuous function of some environmental variable. The errors which arise from inadequate specification of either the response variables, or of the response function, can be distinguished from one another by repeating the experiment.

It has been shown that there is a bias in the usual estimate of β_i, though in the example chosen to illustrate this paper this bias was negligible. It has also been shown that the values of β_i and the correlations between deviations which were observed in the example can be accounted for by the relations with known environmental variables and equations (7) and (8). The deviations from regression are not functionally independent of the slope of the line in either an algebraic or a biological sense. They are rather an inescapable result of fitting lines to data which can only be properly represented in several dimensions. In the quoted experiment the deviations were relatively large, and this was because there were substantial response differences between genotypes for more than one environmental factor. In such situations a more elaborate model is required to describe the data; for example a multiple regression analysis with either linear or non-linear models, provided that estimates of the levels of the operative environmental variables are available. The advantages of multiple regression are that the regression coefficients for each genotype are independent of the number of genotypes in the experiment, and they may also be independently estimated. Alternatively a more economic parameterisation of the data is offered by the extension of Williams' principal component (1952) method, but the economies only become valuable when there is a larger number of genotypes than in the example considered here.

The value of Mandel's (1971) method (equation (5)) as compared with the methods involving regression on environmental variables is that it indicates the number of dimensions necessary to contain the genotypic variation and gives estimates of the corresponding coefficients, without any prior requirement of knowing what factors these dimensions represent. The method may prove particularly valuable in analysing data which show

substantial residuals from the unidimensional model but for which there are no concomitant measurements available of environmental variables. As in other applications of principal components analysis, it may subsequently be found that the environmental parameters of Mandel's model can be equated with particular environmental variables, but this question, together with that of the biometrical interpretation of the genotypic parameters, is a problem which requires further study.

7. Appendix

In section 5 we wish to find the values, $\hat{\mu}$, $\hat{\rho}_i$, $\hat{\theta}$, $\hat{\phi}_i$, $\hat{a}_h$ of μ, ρ_i, θ, ϕ_i, a_h which minimise

$$S = \sum_{\cdot\cdot}(y_{ij}-\mu-\rho_i-\theta\phi_iE_j)^2$$

subject to the constraints

$$\Sigma\rho_i = 0,\ \Sigma\phi_i^2 = \Sigma E_j^2 = 1, \text{ where } E_j = \sum_h a_h x_{hj},\ \sum_j E_j = 0.$$

Consider

$$S = \sum_{i,j}(y_{ij}-\mu-\rho_i-\theta\phi_iE_j)^2 + 2v_1\sum_i\rho_i + v_2(\sum_j E_j^2 - 1) \tag{A1}$$

where v_1 and v_2 are Lagrange multipliers. Taking partial derivatives

$$\begin{aligned}\frac{\partial S}{\partial \mu} &= 2\sum_{i,j}(y_{ij}-\mu-\rho_i-\theta\phi_iE_j)\\ &= 2\sum_{i,j}y_{ij} - 2mn\mu. \end{aligned}\tag{A2}$$

$$\frac{\partial S}{\partial \rho_i} = 2\sum_j(y_{ij}-\mu-\rho_i) - 2v_1. \tag{A3}$$

$$\begin{aligned}\frac{\partial S}{\partial \phi_i} &= 2\theta\sum_j E_j(y_{ij}-\mu-\rho_i-\theta\phi_iE_j)\\ &= 2\theta\sum_j E_j(y_{ij}-\mu-\rho_i) - 2\theta^2\phi_i. \end{aligned}\tag{A4}$$

$$\begin{aligned}\frac{\partial S}{\partial a_h} &= 2\theta\sum_{i,j}\phi_i x_{hj}(y_{ij}-\mu-\rho_i-\theta\phi_iE_j) - 2v_2\sum_j x_{hj}\\ &= 2\theta\sum_{i,j}\phi_i x_{hj}(y_{ij}-\mu-\rho_i) - 2\theta^2\sum_j x_{hj}E_j. \end{aligned}\tag{A5}$$

$\hat{\mu}$, $\hat{\rho}_i$, $\hat{\theta}$, $\hat{\phi}_i$, $\hat{a}_h$ are the values of μ, ρ_i, θ, ϕ_i, a_h for which these derivatives are zero.

From equation (A2)

$$\hat{\mu} = \sum_{i,j}y_{ij}/(mn). \tag{A6}$$

Summing equations (A3) over i gives $v_1 = 0$, so

$$\hat{\rho}_i = \sum_j(y_{ij}-\hat{\mu})/n. \tag{A7}$$

In section 5, $\mathbf{Y}$ and $\mathbf{X}$ were defined as the matrices whose elements are $\{(y_{ij}-y_{i\cdot})\}$ and $\{x_{hj}\}$ respectively. If $\mathbf{E}$ is the vector of estimates of the E_j's and $\boldsymbol{\phi}$ and $\boldsymbol{\alpha}$ the vectors with elements $\{\hat{\phi}_i\}$ and $\{\hat{a}_h\}$ equations (A4) and (A5) can be written

$$\mathbf{YE} = \hat{\theta}\boldsymbol{\phi}. \tag{A8}$$

$$\boldsymbol{\phi}'\mathbf{YX} = \hat{\theta}\mathbf{E}'\mathbf{X}. \tag{A9}$$

Hence $$\mathbf{E}'\mathbf{Y}'\mathbf{YX} - \hat{\theta}^2\mathbf{E}'\mathbf{X} = 0.$$

From equation (13)

$$\mathbf{E} = \mathbf{X}\boldsymbol{\alpha}$$

or $$(\mathbf{X}'\mathbf{Y}'\mathbf{YX} - \hat{\theta}^2\mathbf{X}'\mathbf{X})\boldsymbol{\alpha} = 0.$$

Thus $\hat{\theta}^2$ is an eigenvalue of this equation, $\boldsymbol{\alpha}$ is the corresponding eigenvector, and $\boldsymbol{\phi}$ can be obtained from equation (A8). Then

$$S = \text{trace } (\mathbf{Y}'\mathbf{Y}) - 2\theta\boldsymbol{\phi}_i'\gamma\mathbf{E}_j + \mathbf{E}_j'\boldsymbol{\phi}_i'\boldsymbol{\phi}_i\mathbf{E}_j.$$

From equation (A8)

$$\boldsymbol{\phi}_i'\mathbf{YE}_j = \theta$$

or $$S = \text{trace } (\mathbf{Y}'\mathbf{Y}) - \theta^2$$

and the largest eigenvalue must be chosen to minimise S.

8. Summary

1. The method of investigating interactions in two-way tables by regression of the entries on row or column means, which was introduced by Yates and Cochran (1938) is discussed with reference to subsequent work by Tukey (1949), Williams (1952), Mandel (1959, 1961, 1963, 1969, 1970, 1971) and Gollob (1968).

2. It is shown that there is a bias in the estimate of the slope of the regression when this is derived by the usual means.

3. An alternative method of analysis, using multiple regression of performance, on the levels of environmental variables is considered and methods for investigating deviations from regression discussed.

4. It is shown, using data from an experiment due to Richards (1965), that the slopes of regression on the environmental mean can be expressed in terms of the coefficients of regression on environmental variables. The deviations from regression are not independent of the slopes, but can be expressed in terms of the same coefficients.

5. When the deviations from regression on the environmental mean are substantial, a more elaborate model is required. Two alternatives are briefly considered; Mandel's (1970, 1971) extension of the empirical method, and a development of Williams' (1952) approach is proposed for use when measurements of environmental variables are available.

9. References

BUCIO ALANIS, L. 1966. Environmental and genotype-environmental components of variability. *Heredity, 21*, 387-397.

DE WIT, C. T., AND BROUWER, R. 1968. Über ein dynamisches Modell des vegetativen Wachstum von Pflanzenbestanden. *Angew. Bot., 42*, 1-12.

EBERHART, S. A., AND RUSSELL, W. A. 1966. Stability parameters for comparing varieties. *Crop Sci., 6*, 36-40.

FINLAY, K. W., AND WILKINSON, G. N. 1963. The analysis of adaptation in a plant-breeding programme. *Aust. J. agric. Res., 14*, 742-754.

FREEMAN, G. H., AND PERKINS, J. M. 1971. Environmental and genotype-environmental components of variability. VIII. Relations between genotypes grown in different environments and measures of these environments. *Heredity, 27*, 15-23.

GOLLOB, H. F. 1968. A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika, 33,* 73-116.

HANSON, W. D., AND ROBINSON, H. F. (ed.). 1963. Statistical genetics and plant breeding. Washington, D.C.: National Academy of Sciences.

JINKS, J. L., AND PERKINS, J. M. 1970. Environmental and genotype-environmental components of variability. VII. Simultaneous prediction across environments and generations. *Heredity, 25,* 475-480.

KNIGHT, R. 1970. The measurement and interpretation of genotype-environment interactions. *Euphytica, 19,* 225-235.

LANCZOS, C. 1957. *Applied analysis.* Pitman, London.

MANDEL, J. 1959. The measuring process. *Technometrics, 1,* 251-267.

MANDEL, J. 1961. Non-additivity in two-way analysis of variance. *J. Am. statist. Ass., 56,* 878-888.

MANDEL, J. 1969. A method for fitting empirical surfaces to physical or chemical data. *Technometrics, 11,* 411-429.

MANDEL, J. 1970. The partitioning of interaction in analysis of variance. *J. Res. natn. Bur. Stand. (U.S.), 73*B, 309-328.

MANDEL, J. 1971. A new analysis of variance model for non-additive data. *Technometrics, 13,* 1-18.

MANDEL, J., AND LASHOF, T. W. 1959. The interlaboratory evaluation of testing methods. *Bull. Am. Soc. Test. Mater., 239,* 53-61.

MANDEL, J., AND MCCRACKIN, F. L. 1963. Analysis of families of curves. *J. Res. natn. Bur. Stand. (U.S.), 67*A, 259-267.

MATHER, K. 1971. On biometrical genetics. *Heredity, 26,* 349-364.

NELDER, J. A., AUSTIN, R. B., BLEASDALE, J. K. A., AND SALTER, P. J. 1960. An approach to the study of yearly and other variation in crop yield. *J. hort. Sci., 35,* 73-82.

PALTRIDGE, G. W. 1970. A model of a growing pasture. *Agric. Meteorol., 7,* 93-130.

PERKINS, J. M., AND JINKS, J. L. 1968*a*. Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. *Heredity, 23,* 339-356.

PERKINS, J. M., AND JINKS, J. L. 1968*b*. Environmental and genotype-environmental components of variability. IV. Non-linear interactions for multiple inbred lines. *Heredity, 23,* 525-535.

PUTTER, J., YARON, D., AND BIELORAI, H. 1966. Quadratic equations as an interpretative tool in biological research. *Agron. J., 58,* 103-104.

RICHARDS, F. J. 1959. A flexible growth function for empirical use. *J. exp. Bot., 10,* 290-300.

RICHARDS, J. A. 1965. Effects of fertilisers and management on three promising tropical grasses in Jamaica. *Expl. Agric., 1,* 281-288.

SCHEFFÉ, H. 1959. *The analysis of variance.* John Wiley, New York.

SPRENT, P. 1969. Models in regression and related topics. Methuen, London.

TUKEY, J. W. 1949. One degree of freedom for non-additivity. *Biometrics, 5,* 232-242.

WARREN, K. B. (ed.). 1955. *Population genetics: The nature and causes of genetic variability in populations.* The Biological Laboratory, Cold Spring Harbor, New York.

WILLIAMS, E. J. 1952. The interpretation of interactions in factorial experiments. *Biometrika, 39,* 65-81.

WRIGHT, A. J. 1971. The analysis and prediction of some two factor interactions in grass breeding. *J. agric. Sci., Camb., 76,* 301-306.

YATES, F., AND COCHRAN, W. G. 1938. The analysis of groups of experiments. *J. agric. Sci., Camb., 28,* 556-580.