Heredity (1976), **37** (2), 299-303

BIAS IN THE ESTIMATION OF REGRESSION COEFFICIENTS IN THE ANALYSIS OF GENOTYPE-ENVIRONMENTAL INTERACTION

A. J. WRIGHT

Plant Breeding Institute, Trumpington, Cambridge

Received 28.v.76

Summary

Potential bias in estimates of regression coefficients when the environmental index in joint regressional analysis is the mean of all genotypes is examined in relation to assumptions of random and fixed genotypic effects and the error structure of the experiment. A modification is suggested.

1. Introduction

There are two possible sources of bias in the estimation of regression coefficients according to the method of Perkins and Jinks (1968), in which an environmental index composed of the mean value of all genotypes is used as the independent variable. The first arises because the environmental index represents only an estimate of the true environmental effect, and so its variance contains an error component presumed to be uncorrelated with the dependent variable (Tai, 1971). This is in fact a well-known phenomenon called the attenuation effect (Sprent, 1969), and while reducing absolute differences among estimated coefficients, cannot disturb their ranking (Hardwick and Wood, 1972). The second alleged bias results from the presence in the index of the genotypic values which are to be regressed on to it (Freeman and Perkins, 1971; Freeman, 1973), presumably giving rise to a spurious element of correlation which may differ from genotype to genotype. Hardwick and Wood (1972) have shown that only the first type of bias applies in the case of fixed genotypic effects, although they did not explicitly define or draw attention to the possibility of the second type. The purpose of this note is to compare the effects of the two assumptions, and to show that both types of bias can arise when genotypic effects are random. A modified method of calculation of coefficients is suggested which removes these biases.

2. Estimation under the basic model

The simplest model by which data from a set of m genotypes grown in n environments in each of r blocks is described is:

$$y_{ijk} = \mu + g_i + \varepsilon_j + b_k + f_{ij} + e_{ijk}, \qquad (1)$$

where y_{ijk} is the value of the ijth combination in the kth block, μ is the general mean, g_i the effect of the ith genotype, ε_j that of the jth environment, f_{ij} their interaction, b_k the kth block effect, and e_{ijk} the error associated with this plot. For the means of genotype-environment combinations over all r blocks, this can be reduced to

$$y_{ij} = \mu + g_i + \varepsilon_j + f_{ij} + \bar{b} + e'_{ij}, \tag{2}$$

where

$$e'_{ij} = \sum_k e_{ijk}/r, \quad \text{and} \quad \bar{b} = \sum_k b_k/r.$$

This model is not appropriate for the case when the blocks are distinct from one environment to another (Tai, 1971), nor when higher order interactions among terms in (1) are present. The effects of these modifications will be considered later.

A factor of major importance in the calculation of regression coefficients is the behaviour of the f_{ij} terms under summation over individual genotypes to generate the environmental index. This depends on the nature of the set of genotypes included in the experiment. If these represent the entire population about which inferences are to be made (fixed effects), then over this range the total of interaction effects must be zero (*i.e.* $\sum_i f_{ij} = 0$). If, on the other hand, the genotypes are a sample which is very small in comparison with the population from which they are drawn (random effects, Kempthorne, 1957), then this summation property no longer holds.

The influence of these two types of effects on the expectation of β_i calculated according to the normal procedure described by Perkins and Jinks (1968) will now be examined. In the first place, random genotypic effects will be assumed, the correct results for fixed effects being simply derived by the deletion of certain terms.

The environmental index used is the mean of the values for all m genotypes in each environment. For the jth, from (2):

$$x_j = \sum_i y_{ij}/m = \mu + \sum_i g_i/m + \varepsilon_j + \sum_i f_{ij}/m + \bar{b} + \sum_i e'_{ij}/m.$$

The observed covariance of values of the ith genotype with the corresponding x_j is

$$E(W_i) = \sigma_\varepsilon^2 + (\sigma_{\varepsilon f})_i + ((\sigma_f^2)_i + (\sigma_{e'}^2)_i)/m,$$

where $(\sigma_f^2)_i$ is the variance of interactions involving the ith genotype, $(\sigma_{\varepsilon f})_i$ is their covariance with environmental effects, and $(\sigma_{e'}^2)_i$ is the variance of error effects on that genotype. The variance of the index is

$$E(V_x) = \sigma_\varepsilon^2 + (\sigma_f^2 + \sigma_{e'}^2)/m$$

so that

$$E(1+\beta_i) = \frac{\sigma_\varepsilon^2 + (\sigma_{\varepsilon f})_i + ((\sigma_f^2)_i + (\sigma_{e'}^2)_i)/m}{\sigma_\varepsilon^2 + (\sigma_f^2 + \sigma_{e'}^2)/m}$$

and

$$E(\beta_i) = \frac{(\sigma_{\varepsilon f})_i + ((\sigma_f^2)_i - \sigma_f^2 + (\sigma_{e'}^2)_i - \sigma_{e'}^2)/m}{\sigma_\varepsilon^2 + (\sigma_f^2 + \sigma_{e'}^2)/m}. \tag{3}$$

Here

$$\sigma_{e'}^2 = \sum_i (\sigma_{e'}^2)_i/m \quad \text{and} \quad \sigma_f^2 = \sum_i (\sigma_f^2)_i/m.$$

It is clear from (3) that potential bias is present in the numerator of the ratio used to calculate β_i, and depends on heterogeneity of $(\sigma_f^2)_i + (\sigma_{e'}^2)_i$ from genotype to genotype. Since differences in genotypic stability are being postulated the possibility of such heterogeneity cannot be ignored, and estimates of β_i must be regarded as potentially biased. The fixed effects case can be derived by deleting all terms involving σ_f^2 or $(\sigma_f^2)_i$, since these arise from the assumption that $\sum_i f_{ij} \neq 0$ in x_j. Bias in this case will therefore arise only with heterogeneity of $(\sigma_{e'}^2)_i$, and will therefore disappear under the assumptions usual in analysis of variance. Note that the attenuation effect is also present in both cases, but can be removed by the substitution of an estimate of σ_ε^2 for the denominator, as shown by Tai (1971).

It is possible to derive a modified estimate of β_i which is freed from both sources of bias with random effects. As the variance of values of the *i*th genotype is

$$E(V_i) = \sigma_\varepsilon^2 + 2(\sigma_{\varepsilon f})_i + (\sigma_f^2)_i + (\sigma_{e'}^2)_i$$

then, using (3):

$$W_i' = (m/(m-1))(W_i - V_i/m) = \sigma_\varepsilon^2 + (m-2)/(m-1)(\sigma_{\varepsilon f})_i$$

and

$$E(\overline{W}') = \sigma_\varepsilon^2.$$

W_i' is in fact the same as the covariance of the *i*th genotype on to an index based only on the $(m-1)$ other genotypes. Hence,

$$E((m-1)/(m-2)(W_i' - \overline{W}'/(m-1)) = \sigma_\varepsilon^2 + (\sigma_{\varepsilon f})_i,$$

as required. An estimate of σ_ε^2 for use as the denominator in $\hat{\beta}_i'$ is provided by $\overline{W}'$ which is preferred to the usual estimate offered by variance analysis because the latter will lead to spurious departures of the $\hat{\beta}_i'$ from a mean of zero. It follows that

$$\hat{\beta}_i' = \frac{(m-1)}{(m-2)}\left(\frac{W'}{\overline{W}'} - 1\right).$$

The ranking of these coefficients is in fact identical to that of the W_i'.

3. Other models

As stated earlier, the model (1) upon which the estimation procedures have so far been based may be an oversimplification for many experimental situations. In the first place, the blocks may be distinct from one environment to another, as is commonly the case in plant breeding practice when the environmental factors are locations or seasons. Then (1) and (2) become

$$y_{ijk} = \mu + g_i + \varepsilon_j + f_{ij} + b_{jk} + e_{ijk} \quad \text{(Tai, 1971)},$$

and

$$y_{ij} = \mu + g_i + \varepsilon_j + f_{ij} + b'_j + e'_{ij},$$

where b_{jk} is a random effect, and

$$b'_j = \sum_k b_{jk}/r.$$

These b'_j effects are expected to be independent of the ε_j but in so far as they are effectively environmental differences with causative effects on the phenotype, they may be correlated with the f_{ij}. The only way of dealing with this problem is to accept that the environmental index as defined and estimated contains these mean block effects, and then the foregoing developments with regard to fixed and random genotypic effects still hold good. The more blocks are grown, then the smaller will these block contributions be. Note also that these arguments will apply in the case of orthogonality of blocks and environments when these two effects interact to generate $(b\varepsilon)_{jk}$ terms in (1). The second complication is that any interactions of blocks with f_{ij} terms have to be incorporated into the e_{ijk} terms of (1). The effects of any heterogeneity of $(\sigma^2_{e'})_i$ caused by their occurrence have already been described, but it is further necessary to assume that they, like the pure error deviations, are independent of other effects in the model.

4. Discussion

The use of an environmental index which explicitly excludes the genotype being regressed on to it has been explored empirically by Snoad and Arthur (1975), and is also a common feature of the method of modification of regression coefficients suggested here for the random effects model and that given by Mather and Caligari (1974). This resemblance is in fact superficial, however, since the latter take no account of the nature of the genotypic effects, and their treatment of the values of genotype-environmental combinations as mathematical variables with complete functional interdependence is valid only with complete linearity of all regressions, with no deviations due either to true interaction or even to experimental error (Sprent, 1969; and see also Hill, 1975). This can be of practical use in only the most restricted circumstances.

In the majority of cases the genotypic values will be properly regarded as fixed, and, provided that the error variance is homogeneous, there will be no systematic bias in the normal estimates of regression coefficients. This conclusion is in agreement with those of Hardwick and Wood (1972) and Freeman (1973). Perkins and Jinks (1973) devised and applied several methods for the estimation of a statistically independent index. They recognised that the regression of members of one group of genotypes on to an index derived from another is likely to be biased by differential interaction of the two groups with the environment. As an alternative, they used replicate individuals from each genotype to assess the environment.

Although this approach removes the statistical dependence due to error effects, it can do nothing to counteract the bias which can arise from the interactions themselves in the case of random effects.

The present paper represents a theoretical rather than an empirical approach to questions of possible bias in estimates and their appropriate modification. Whether the application of the modification suggested here is likely to result in substantial alterations in estimated rates of response to environmental change and even to changes in ranking is beyond the scope of this note. It is however apparent that the importance of the modification increases as the size of the genotypic sample decreases. The omission of the regressed genotype from the environmental index ensures that the rate of response of any genotype is always assessed in relation to a random sample from the reference population, thus providing a common basis of assessment even for genotypes grown in different experiments. The use of a different index for each genotype in a single experiment is analogous to the omission of progenies resulting from selfing in the estimation of general combining ability in a diallel crossing scheme (Griffing, 1956).

5. References

FREEMAN, G. H. 1973. Statistical methods for the analysis of genotype-environment interactions. *Heredity*, *31*, 339-354.

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.

GRIFFING, B. 1956. Concepts of general and specific combining ability in relation to diallel crossing systems. *Aust. J. Biol. Sci.*, *9*, 463-493.

HARDWICK, R. C., AND WOOD, J. T. 1972. Regression methods for studying genotype-environment interactions. *Heredity*, *28*, 209-222.

HILL, J. 1975. Genotype-environment interactions—a challenge for plant breeding. *J. Agric. Sci.*, *85*, 477-494.

KEMPTHORNE, O. 1957. *An Introduction to Genetic Statistics.* John Wiley & Sons Inc., London.

MATHER, K., AND CALIGARI, P. D. S. 1974. Genotype × environment interactions. I. Regression of interaction on overall effect of the environment. *Heredity*, *33*, 43-59.

PERKINS, J. M., AND JINKS, J. L. 1968. Environmental and genotype-environmental components of variability. III. Multiple lines and crosses. *Heredity*, *23*, 339-346.

PERKINS, J. M., AND JINKS, J. L. 1973. The assessment and specificity of environmental and genotype-environmental components of variability. *Heredity*, *30*, 111-126.

SNOAD, B., AND ARTHUR, A. E. 1975. The use of regression techniques for predicting the response of peas to environment. *Theoret. Appl. Genet.*, *47*, 9-20.

SPRENT, P. 1969. *Models in Regression and Related Topics.* Methuen, London.

TAI, G. C. C. 1971. Genotypic stability analysis and its application to potato regional trials. *Crop Sci.*, *14*, 587-590.