USE OF REGRESSION APPROACH IN THE ANALYSIS OF GENOTYPE-ENVIRONMENTAL INTERACTION

BY

G.K. SHUKLA

Indian Institute of Technology, Kanpur-208 016

(Received : September, 1981)

SUMMARY

In the last one and a half decades a lot of research workers have used the method of regressing genotypic mean on the environmental mean for analysing genotype-environmental interaction with varying degree of success. While some workers have found a considerable amount of interaction being accounted by these regression parameters, the others have found only a small amount being accounted by these parameters. Moreover, the regression parameters have been found varying considerably from trial to trial. In the present paper an attempt has been made to look at the reasons for the above types of anomalies analytically.

1. INTRODUCTION

In the last one and a half decades a considerable amount of work has been done to analyse and interpret genotype-environmental (G-E) interaction. Yates and Cochran [9] regressed the genotype means on the environmental means, calculated by taking the average of all genotypes in that environment, and partitioned G-E interactions into two-components. The same approach was later used by Finlay and Wilkinson [2] and since then it has been applied and used in many circumstances. Comprehensive reviews on this subject have been given by Freeman [3] and Hill [6]. Yates and Cochran [9] used this approach for gaining further insight into the relative behaviour of genotypes in different environments, in the absence of any

This paper was presented in the 32nd annual conference of the Indian Society Agricultural Statistics, held from 21st-32rd December, 1978, at Ludhiana.

knowledge of the underlying environmental factors affecting the yields at different environments (sites). However, in the recent past this technique has also been used to analyse G-E interaction when the environments were known levels of controllable factors, as often the case in planned experiments. It has been pointed out by Freeman and Perkins [4] that site means may give a better measurement of environment as they give combined effect of all relevant factors operating in that environment. It has been emphasised by several authors that the conclusions drawn from such studies are valid only for the population from which the particular set of environments is a random sample yet no particular care has been taken to examine whether there can by any realistic population corresponding to environments chosen in a planned experiment.

Knight [7] has emphasised, through his studies on published data, that the regression coefficients thus obtained are dependent on the genotypes included in the trial. Fatunla and Frey [1] have shown that the regression coefficients based on two different sets of environmental factors, are not the same, thus showing that the regression estimates may not be repeatable. These findings have led us to look into this regression approach analytically and examine whether the above observations can be explained.

2. Notation and Model

To keep approach simple we shall work with a linear regression model. Consider t genotypes taken in n environments and let $y_{ij}(i=1,\ldots,t\,;j=1,\ldots,n)$ represent the yield of the ith genotype at the jth environment. In cases where more than one replication has been taken in each environment y_{ij} may be taken as the mean over all replications We shall further assume that there are p environmental factors $X_1, X_2, \ldots, X_p$, which can affect the yield at any environment. These environmental factors may be different nutrients, humidity, temperature, sunlight etc., which may affect the yield. They may be functionally related and may not be independent of each other. Let us assume that X_k takes value x_{kj} $(k=1,\ldots,p)$ at the jth environment. We shall further assume that y_{ij} can be expressed as,

$$y_{ij}=\mu_i+\beta_{i1}x_{ij}+\ldots+\beta_{ip}x_{pj}+\theta_{ij}, \tag{1}$$

where μ_i is the expected yield of the ith variety at the jth environment when x_{1j}, x_{2j}, ..., x_{pj} all take zero values ; $\beta_{ik}(k=1, \ldots, p)$ is the regressoni coefficients corresponding to the ith variety and kth environmental factor, and θ_{ij} is the random error and remainder of the interaction. The model in (1) can be expressed more economically as

$$y_{ij}=\mu+\underline{\beta}_i'\underline{x}_j+\epsilon_{ij}, \tag{2}$$

where $\underline{\beta}_i'(\beta_{i1},...,\beta_{ip})$ and $\underline{x}_j' = (\bar{x}_{ij},...,x_{pj})$, are p component vectors representing regression coefficients and levels of environmental factors, respectively. We shall further assume that $X_k's$ are fixed variables but their actual values at any environment may by unknown. The model in (1) is quite general as it can take polynomial as well as cross-product, terms in $X_k's$ into account by suitably defining $X_k's$ - Let us further denote by

$$\underline{\bar{x}}' =(\bar{x}_{1.},...,\bar{x}_{p.}),$$

the mean vector of environmental variables, where mean is taken over environments. $\underline{\bar{\beta}}' =(\bar{\beta}_1,...,\bar{\beta}_p)$ denotes the mean vector of regression coefficients, where mean is taken over genotypes.

In the usual genotype-environmental interaction model, considered by several authors, the mean yield of the ith variety at the jth environment, y_{ij}, has been expressed as

$$y_{ij}=m+v_i+s_j+\eta_{ij}+\varepsilon_{ij}, \tag{3}$$

where m is the general mean, v_i is the effect of the ith genotype, s_j is the effect of the jth environment, η_{ij} being the G-E interaction component of ith genotype with jth environment, and ε_{ij} is the mean random error component. In the usual regression on environmental mean approach η_{ij} has been further partitioned as

$$\eta_{ij}=b_i s_j+\eta_{ij}' ,$$

and thus model (3) becomes

$$y_{ij}=m+v_i+(1+b_i)s_j+\eta_{ij}'+e_{ij}, \tag{4}$$

where η_{ij}' is the remainder component of interaction left after removihg the regression component on the environmental mean.

Now we have represented y_{ij} in two alternative ways in (2) and (4). Had we known the values of all $x_{kj}'s$ we would have worked with the model in (2) but in the absence of this knowledge we work with model (4) and try to interpret the results with the help of model (2).

3. Results

In the usual methodology s_j's are estimated as $\hat{s}_j=\bar{y}_{.j} - \bar{y}_{..}$ and b_i's are estimated as

$$\hat{b}_i = \frac{\sum\limits_{j} (y_{ij}-\bar{y}_{.j})(\bar{y}_{.j}-\bar{y}_{..})}{\sum\limits_{j} (\bar{y}_{.j}-\bar{y}_{..})^2}, \qquad (5)$$

under the condition that $\sum\limits_i \hat{b}_i=0$. The expression in (5) is a ratio of two correlated random variables and no simple expression for its expectation exists. The estimator $\hat{b}_i$ is not a consistent estimator of b_i and this has been pointed out by many authors (Shukla, [8]. By substituting y_{ij}, from (2) and then taking expectations independently of numerator and denominator of (5), it is not difficult to see for large n,

$$E(\hat{b}_i) \to \frac{(\underline{\beta}_i-\bar{\underline{\beta}})' S_{xx}\, \bar{\underline{\beta}}}{\bar{\underline{\beta}}' S_{xx} \bar{\underline{\beta}}+(n-1)\,\sigma^2/t} \qquad (6)$$

where,

$$S_{xx} = (s_{kk'})\ ;\ k, k'=1,\dots,p$$

$$s_{kk'} = \sum_{j=1}^{n} (x_{kj}-\bar{x}_{k.})\,(x_{k'j}-\bar{x}_{k'.}).$$

When t is large the expression in (6) tends to b_i^* as follows :

$$E(\hat{b}_i) \to \frac{(\underline{\beta}_i-\bar{\underline{\beta}})'\, S_{xx}\bar{\underline{\beta}}}{\bar{\underline{\beta}}'\, S_{xx}\bar{\underline{\beta}}} = b_i^* \qquad (7)$$

This amounts to substituting for $(\bar{y}_{.j}-\bar{y}_{..})$ its expected value $\bar{\underline{\beta}}'(\underline{x}_j-\bar{\underline{x}})$, in the expression of $\hat{b}_i$. From now onwards whenever we shall consider expectation we shall consider $\bar{y}_{.j}-\bar{y}_{..}$ as a constant quantity equal to $\bar{\underline{\beta}}'(\bar{\underline{x}}_j-\bar{\underline{x}})$ in the expression of $\hat{b}_i$, and denote this by $E^*(u)$, rather than $E(u)$ in the usual notation of expected value of u.

$$G\text{—}E \text{ interaction } S\,S_q=\sum_i \sum_j (y_{ij}-\bar{y}_{i.}-\bar{y}_{.j}+\bar{y}_{..})^2,$$

$$E\ (\text{Interaction } S.S_q.)=\sum_i (\underline{\beta}_i-\bar{\underline{\beta}})'\, S_{xx}(\underline{\beta}_i-\bar{\underline{\beta}})+(t-1)\,(n-1)\,\sigma^2, \qquad (8)$$

$S.S_q.$ due to Regression $=\sum_i b_i^2 \sum_j (\bar{y}_{.j}-\bar{y}_{..})^2.$

$$E^* \,(S.S_q. \text{ due to regression}) = \underline{\bar{\beta}}' S_{xx} \underline{\bar{\beta}}. \sum_{i=1}^{t} (K_i-1)^2+(t-1)\,\sigma^2, \tag{9}$$

where

$$K_i = \underline{\bar{\beta}}_i' \, S_{xx} \underline{\bar{\beta}} / (\underline{\bar{\beta}}' \, S_{xx} \underline{\bar{\beta}}). \tag{10}$$

$S.S_q.$ due to deviation from regression is $\Sigma\, \delta_i^2$, where

$$\delta_i^2 = \sum_{j=1}^{n} \{y_{ij}-\bar{y}_{i.}-(1+\hat{b}_i)\,(\bar{y}_{.j}-\bar{y}_{..})\}^2.$$

$$E^*(\delta_i^2 = (\underline{\beta}_i - K_i \underline{\bar{\beta}})' S_{xx} (\underline{\beta}_i - K_i \underline{\bar{\beta}}) + (t-1)\,(n-2)\,\sigma^2/t \tag{11}$$

4. Interpretation of the Results

4.1 Interpretation of b_i :

We shall consider the case of $p=1$ and $p>1$ separately.

(*i*) Case with $p=1$:

Let us consider the case when only one x takes into account all the variation present in the yield in the range of interest, and can be approximated by model (1) with $p=1$. Let the regression coefficient corresponding to the *i*th genotype be β_{i1}. From (7) we see that $\hat{b_i}$ estimates $E^*(\hat{b_i})$ given by

$$E^*(\hat{b_i}) = \frac{\beta_{i1}-\bar{\beta}}{\bar{\beta}} = \frac{\beta_{i1}}{\bar{\beta}} - 1. \tag{8}$$

From (8) it is apparent that $\hat{b_i}$ does not estimate β_{i1} but a quantity $(\beta_{i1}-\bar{\beta})/\bar{\beta}$, which is a relative measure-relative to other genotypes in the trial. This implies that two investigators working with some common genotypes and some uncommon genotypes may arrive at different set of $E^*(\hat{b_i})$ for common varieties as they are estimating different parameters. Infact, it is quite possible that a genotype in a trial may give a value of $\hat{b_i}=-1$ and thus may be called a stable variety while it may turn out to be unstable when working with other set of genotypes. However, when a large number of genotypes are considered for calculating the stability parameters b_i' s, as in the case of Finlay and Wilkinson [2], the effect of deleting a few genotypes may not be that marked as it would be in case when only a small number of genotypes are considered. This indeed is the case with the example of Knight [7], that the removal of a single variety can affect the estimates of the parameters considerably.

(ii) Case when $p>1$:

Even if there is only one environmental factor which affects the yield, it is quite possible that over the whole range of variability the relationship between yield and that factor may be described by a suitable degree polynomial in $x's$ rather than only one term of model (1). In general, more than one environmental factor may affect the yield and thus the relationship may better be described by taking $p>1$, which may include polynomial as well as cross-product terms in $X_k's$.

When $p>1$, then $b_i^* s$ not only depend upon the parameters of other genotypes present in the trial but also on the actual level of environmental factors through S_{xx}, as is apparent from [7]. Thus, one working with different range of the same environmental factors may be estimating different $b_i^* s$ if S_{xx} in two cases are different. This is the case considered by Fatunla and Frey [1], where by considering two different sets of environmental factors with the same set of genotypes they obtain different values of $b_i's$. However, if the number of environment (sites) is large and can be considered to be selected randomly from the population of environments then $n^{-1} S_{xx}$ may be estimating population variance-covariance matrix Σ_{xx} and one would then expect that $b_i's$ are estimating the same parameters. This, however, is not the case when only a small number of factors are considered in a controlled experiment, as indeed the case with many reported papers.

4.2 Interpretation of Deviation from Regression :

When $p=1$ then $K_i=\beta_{i1}/\beta$ and thus

$$E(\delta_i^2)=(t-1)(n-2)\sigma^2/t,$$

and the departure from linear regression would be insignificant. When $p>1$ and $\underline{\beta}_1=\underline{\beta}_2=\ldots=\underline{\beta}_t=\bar{\underline{\beta}}$, say, *i.e.* regression parameters of all the genotypes are equal, then the sum of squares due to interaction, regression and due to deviation from regression are all insignificant and $b_i^*'s$ are all zeroes. In cases when δ_i^2 is insignificant for a particular variety this means that for that particular variety the regression vector is equal to average regression vector. This, however, does not necessarily mean that there is a linear increase or decrease in the yield with the increase or decrease in the level of environmental factor. For exploiting this relationship to any use one has to estimate the parameters of the relationship in (1).

In literature one sees that some workers have found that most of the variation due to interaction is accounted by regression parameters b_i's in the model (4) whereas other workers have found only a modest or very little variation being accounted by the regression parameters. From the expression of interaction sum of squares it is evident that when $\underline{\beta}_i$'s are equal for all genotypes then there is no interaction. In case $\underline{\beta}_i$'s are unequal then interactions are present, but the sum of squares due to deviation from regression is insignificant, when β_i's are proportional *i.e.*

$$\frac{\underline{\beta}_1}{K_1}=\frac{\underline{\beta}_2}{K_2}\cdots\frac{\underline{\beta}_t}{K_t}=\bar{\underline{\beta}}\text{, say.} \tag{9}$$

In general the relationship in (9) is unlikely to hold good but for $p=1$ this relationship always holds. This fact was also noted by Hardwick and Wood [5] who have also used a similar approach. This means that in trials where most of the variation is produced by one factor, or only one factor is dominating, and the range of variation is such that $p=1$ in the model (1) gives an adequate representation of variation, the regression sum of squares may account for most of the interaction. However, when there are more than one dominating factors, and thus p is likely to be greater than one, it is unlikely that any major amount of interaction would be accounted by the regression parameter b_i.

5. EXAMPLE

For explaining the above results with the help of a worked out example we have used the data from Knight [7] on the response of six cultivars (genotypes) of grass to levels of temperature (environments). The yields are given in Table I and are only approximate as they have been read from the plotted figure.

One sees from the ANOVA Table I that only 8.8 percent of the total interaction is accounted by heterogeneity among regressions when all six temperatures are considered, whereas 46.5 percent of the interaction is accounted (ANOVA Table *ii*) when the data of

TABLE 1

Yield of six cultivars of grass for different temperatures

Varieties Temp. in °F	*45*	*55*	*65*	*75*	*85*	*95*
1. Cocksfoot	8.0	15.0	18.0	19.0	15.0	7.5
2. Paspalum	5.0	8.0	17.0	22.5	26.5	12.0
3. Ryegrass 1	7.5	16.0	18.0	16.0	12.0	3.0
4. Ryegrass 2	8.0	16.0	21.0	20.5	15.0	3.0
5. Browntop	7.5	16.0	18.0	20.0	14.0	5.0
6. Yark Fog	8.0	15-0	19.0	22.5	15.0	1.5

Table 2 gives the analyses of variance (ANOVA) for (*i*) all six temperatures, and (*ii*) highest three temperatures.

TABLE 2

Analyses of Variance

	(i) ANOVA for all six temperatures			*(ii) ANOVA for the highest three temperatures*		
Source	*D.F.*	*S.S.*	*M.S.*	*D.F.*	*S.S.*	*M.S.*
Varieties	5	29.47	5.89	5	170.61	34.12
Temp.	5	1087.22	217.44	2	702.86	351.43
Interaction	25	273.62	10.94	10	65.81	6.68
Heterogeneity among Regressions	5	24.04	4.81	5	31.08	6.22
Deviation from Regression	20	249.58	12.48	5	35.73	7.15
Total	35	1390.31		17	940.28	

Regression coefficients $(1+\hat{b})$

Varieties	1	2	3	4	5	6
(*i*) All six temperatures	0.810	0.945	0.888	1.154	0.967	1.237
(*ii*) Highest three temperature	0.760	0.876	0.869	1.167	0.975	1.383

three highest temperatures are analysed. If one sees the the data in Table I it appears that for complete set of six temperatures the yield of each variety can he approximated by a parabola with $p=2$ in model (1), whereas for the last three temperatures the decrease in yield is approximately linear in temperature and thus could be approximated by $p=1$ in model (1). This explains the difference in the percentage of interaction accounted by two sets of data, as discussed in Section 4.2.

For explaining the repeatability of the regression coefficients (Section 4.1) we have used the data of Yates and Cochran [9] and calculated the regression coefficients by including all the data and then deleting the data of the variety 'Tribi', which is very sensitivc to changes in environment. The values of the regression coefficients $(1+\hat{b})$ are given as follows :

Varieties	*Manchusia*	*Svansota*	*Velvet*	*Tribi*	*Peatland*
All five varieties	0.844	0.986	0.946	1.609	0.615
Excluding 'Tribi'	0.997	1.176	1.140	—	0.715

One sees that the varieties Svansota and Velvet, which have sensitivity below average in the presence of Tribi, have become more sensitive than the average when Tribi is excluded. This shows that the regression calculated above is a relative measure-relative to other varieties present in the trial.

References

[1] Fatunla, K. and Frey, K.J. (1976) : Repeatability of regression stability indexes for grain yields of Oats. *Euphytica* **25**, 21-28.

[2] Finlay, K.W., and Wilkinson, G.N. (1963) : The analysis of adaptation in plant breeding programme. *Aust. J. Agric. Res.* **14**, 742-54.

[3] Freeman, G.H. (1973) : Statistical methods for the analysis of genotype-environment interactions *Heredity* **31**, 339-54.

[4] Freeman, G.H. and Perkins, J.M. (1971) : Environmental and Genotype-environmental Components of Variability VIII Relations between genotypes grown in different environments and measure of these environments. *Heredity* **27**, 15-23,

[5] Hardwick, R.C. and Wood, J,T. (1972) ; Regression methods for studying genotype-environment interactions *Heredity* **28**, 209-222.

[6] Hill, J. (1975) : Genotype-environment interaction -- a challange for plant breeding, *J. Agric. Sci. Camb.* **85b**, 477-93.

[7] Knight, R. (1970) : The measurement and interpretation of genotype-environment interactions *Euphytica* **19**, 225-35.

[8] Shukla, G.K. (1972) : Some statistical aspects of partitioning genotype-environmental components of variability *Heredity* **29**, 237-45.

[9] Yates, F. and Cochran, W.G. (1938) : The analysis of groups of experiments *J. Agric. Sci.* **28**, 556-80.