Scientia Agricola

Analyzing genotype-by-environment interaction using curvilinear regression

Dulce Gamito Santinhos Pereira[1]*, Paulo Canas Rodrigues[2,4], Iwona Mejza[3], Stanislaw Mejza[3], João Tiago Mexia[2]

[1]Universidade de Évora/Escola de Ciências e Tecnologia CIMA-UE – Depto. de Matemática, R. Romão Ramalho, 59 – 7000-671 – Évora – Portugal.
[2]FCT/UNL – Centro de Matemática e Aplicações – 2829-516 – Caparica – Portugal.
[3]Poznan University of Life Sciences – Dept. of Mathematical and Statistical Methods, ul. Wojska Polskiego 28 – 60-637 – Poznań – Poland.
[4]ISLA Campus Lisboa, Laureate International Universities, Estrada da Correia, 53 – 1500-210 – Lisboa, Portugal.
*Corresponding author <dgsp@uevora.pt>

Edited by: Thomas Kumke

Received June 30, 2011
Accepted May 04, 2012

ABSTRACT: In the context of multi-environment trials, where a series of experiments is conducted across different environmental conditions, the analysis of the structure of genotype-by-environment interaction is an important topic. This paper presents a generalization of the joint regression analysis for the cases where the response (e.g. yield) is not linear across environments and can be written as a second (or higher) order polynomial or another non-linear function. After identifying the common form regression function for all genotypes, we propose a selection procedure based on the adaptation of two tests: (i) a test for parallelism of regression curves; and (ii) a test of coincidence for those regressions. When the hypothesis of parallelism is rejected, subgroups of genotypes where the responses are parallel (or coincident) should be identified. The use of the Scheffé multiple comparison method for regression coefficients in second-order polynomials allows to group the genotypes in two types of groups: one with upward-facing concavity (i.e. potential yield growth), and the other with downward-facing concavity (i.e. the yield approaches saturation). Theoretical results for genotype comparison and genotype selection are illustrated with an example of yield from a non-orthogonal series of experiments with winter rye (*Secalecereale* L.). We have deleted 10 % of that data at random to show that our meteorology is fully applicable to incomplete data sets, often observed in multi-environment trials.
Keywords: Scheffé multiple comparison method, joint regression analysis, test for parallelism, test of coincidence

Introduction

Farmers and scientists aim to identify superior performing genotypes across a wide range of environmental conditions. Here, by environments we mean combinations of locations and years. The main source of differences between genotypes in their yield stability is the fact that the genotype and environment effects are not additive, i.e. genotype-by-environment interaction (GEI) is present in the data. This interaction can be due to contrasting drought stress levels, winter low temperature stress, abiotic stresses, growing cycle duration, availability of nutrients, etc. The GEI can be expressed either as crossovers, when two different genotypes change in rank order of performance when evaluated in different environments, or inconsistent responses of some genotypes across environments without changes in rank order. The study and understanding of these interactions is a major challenge for breeders and agronomic researchers attempting to improve complex traits (e.g. yield) across environmental conditions.

Various techniques have been used to analyze the interaction in general and GEI in particular. Readers interested in those methods are referred to e.g. Aastveit and Mejza (1992); Annicchiarico (2002); Gauch (1992); Kang and Gauch (1996); Romagosa et al. (2009).

Regression is one of the most popular and most applicable methods used for inference about genotype comparison and selection in the context of multi-environmental experiments. In regression analysis two sets of variables are used, the first characterizing genotypes, and the second environments. The so-called adjusted means (or some other genotypic characteristic) for genotypes usually constitute observations of the dependent variable. In our illustration we take the original observations of the phenotypic variable as a realization of each dependent variable. The independent variable is defined by environmental indexes, which represent a measurement of productivity. Although Finlay and Wilkinson (1963) defined these environmental indexes as the average over all environments for every genotype, in this study we compute them with an iterative zigzag algorithm (Mexia et al., 1999; Pereira and Mexia, 2010) which leads to the best linear unbiased estimators of the joint regression parameters. In joint regression analysis (JRA), after selecting the variable of interest (e.g. yield), the joint regression model adjusts a linear regression per genotype across all environments (Pereira et al., 2011; Rodrigues et al., 2011), on a synthetic variable measuring productivity, the environmental index. Several variants of JRA have been proposed. The one in which we will be interested here was first proposed by Gusmão (1985), who showed that the precision in analyzing series of randomized block experiments was increased by considering environment indexes for individual blocks instead of only one environmental index per environment. Mexia et al. (1999) introduced the L_2 environmental indexes obtained by minimizing the sum of sums of squares of residuals, both in order to the coefficients of the regressions and to the environmental indexes.

Here a generalization of the joint regression analysis is presented for cases where the response (e.g. the yield) is not linear across environments and can be written as a second (or higher) order polynomial or

another non-linear function. In our considerations we shall start with the estimation of the regression functions (linear or curvilinear) independently for all genotypes. In the second step the hypothesis of parallelism of regressions for all genotypes is tested. In general it is expected that the hypothesis of parallel regression lines will be rejected because of the presence of GEI in the data. If we reject the parallelism of regression curves, the next step to investigate GEI is to try to find subgroups of genotypes with similar responses to different environmental conditions. The genotypes can then be divided into groups based on the Scheffé multiple comparison method for regression coefficients, i.e. the genotypes with similar behavior will be grouped together and the calculations performed for each group separately.

The methodology will be illustrated by using a data set for winter rye (*Secalecereale* L.) from multi-environment trials carried out in the years 1997 and 1998 in Słupia Wielka, Poland (52°13′ N, 17°13′ E).

Materials and Methods

The zigzag algorithm

For convenience, let us consider data arranged in a two-way array with I rows and b columns. Suppose Y_{ij} is a continuous response variate (e.g. yield) for genotype i in block j if present. The joint regression model discussed here is an extension of that of Finlay and Wilkinson (1963) where the environmental indexes are computed for each block instead of for each environment. Assuming that the yield vectors are independent, normal, homoscedastic and that genotype i is present in block j (or replicate), the joint regression model can be written as:

$$Y_{ij} = \alpha_i + \beta_i \chi_j + \varepsilon_{ij} \quad (i=1,\ldots, I;\ j = 1,\ldots, b) \qquad (1)$$

with α_i the intercept and β_i the slope for genotype i, χ_j the block environmental index and ε_{ij} the residuals. These environmental indexes represent the averages over a block/superblock and can be seen as a (spatial) measure of productivity.

Gusmão (1985, 1986) showed that the precision in analyzing series of randomized block experiments was improved considering environmental indexes for individual blocks of only one environmental index per experiment instead of one environmental index per environment. This proposal results in K experiments each with b blocks, i.e. Kb supporting points per regression instead of only K such points used by the classical Finlay-Wilkinson joint regression model (Finlay and Wilkinson, 1963).

To estimate the model parameters, we wish to minimize

$$S\left(\boldsymbol{\alpha}^J, \boldsymbol{\beta}^J, \mathbf{x}^b\right) = \sum_{i=1}^{I} \sum_{j=1}^{b} p_{ij}\left(Y_{ij} - \alpha_i - \beta_i x_j\right)^2, \qquad (2)$$

where p_{ij} is the weight of genotype i in block j. If the genotype is absent we take $p_{ij}=0$. When the genotype occurs we take $p_{ij} = p_j$, j = 1, 2,..., b. These weights may differ from block to block to express differences in the representativeness of the blocks. If there are several blocks in the same location, their weights will be the same. In the illustration presented in this paper we use 1 and 0 for the weights, because no information was available about the relevance of the importance of blocks.

The zigzag algorithm (Pereira and Mexia, 2010) is used to minimize the loss function (2) iteratively, with respect to α_i , to β_i and to the environmental index x_j . For the complete case (i.e. all the genotypes are present in each environment) the average yield per block can be a good initial value for searching the environmental indexes (Gusmão, 1985). When incomplete blocks are used one may take the average yields for the corresponding superblock as the initial values. In the worst case any initial values may be taken, since the computation time does not increase much.

We assume that the yield vectors have components normally and independently distributed, so that the zigzag algorithm will lead to maximum likelihood estimators and enable us to make inferences while comparing genotypes.

The zigzag algorithm may be described as follows:

(i) Calculate the initial values for the environmental indexes $\mathbf{x}_0^b$, which range within the interval $[a_0, b_0]$, where $a_0 = Min\{x_{01},\ldots,x_{0b}\}$ and $b_0 = Max\{x_{01},\ldots,x_{0b}\}$;

(ii) Minimize the function $S\left(\boldsymbol{\alpha}^J, \boldsymbol{\beta}^J \middle| \mathbf{x}_0^b\right)$ and obtain $\tilde{\alpha}\left(\mathbf{x}_0^b\right)$ and $\tilde{\beta}\left(\mathbf{x}_0^b\right)$;

(iii) To minimize $S\left(\mathbf{x}^b \middle| \tilde{\alpha}\left(\mathbf{x}_0^b\right), \tilde{\beta}\left(\mathbf{x}_0^b\right)\right)$, minimize the functions:

(iv) $h_j\left(x \middle| \boldsymbol{\alpha}^J, \boldsymbol{\beta}^J\right) = \sum_{i=1}^{I} p_{ij}\left(Y_{ij} - \tilde{\alpha}_i - \tilde{\beta}_i x_j\right)^2, j = 1, 2,\ldots, b,$

to obtain the new vector $\mathbf{x}_0'^b$ of new environmental indexes;

(v) Standardize the vector of environmental indexes to keep the range unchanged. With $a_0' = Min\{x_{01}',\ldots,x_{0b}'\}$, $b_0' = Max\{x_{01}',\ldots,x_{0b}'\}$ take $x_{1j} = a_0 + \frac{b_0 - a_0}{b_0' - a_0'}\left(x_{0i}' - a_0'\right)$; to obtain the vector $\mathbf{x}_1^b$, the new environmental indexes.

Repeat steps (ii) to (iv) until successive sums of sums of squares of weighted residuals differ by less than a fixed value.

At the end of each iteration, a standardization of the adjusted environmental indexes is carried out so that the range does not change from iteration to iteration. The procedure is carried out until the goal function stabilizes.

The environmental indexes adjusted in this way are called L_2 environmental indexes, because the L_2 norm was used. The described zigzag algorithm is a version of the iterative algorithms existing in the literature; see for example, Digby (1979); Gabriel and Zamir (1979) and Ng and Williams (2001). Pereira and Mexia (2010) proved the convergence of the zigzag algorithm and that the adjusted parameters could be seen as maximum likelihood estimators. The alternative algorithms only considered the numerical adjustment.

Test for coincidence

Considering as before Y_{ij}, the phenotypic observation for genotype i in block j, $j = 1,\ldots,b$; $i = 1,\ldots, I$, and X_j, the environmental index for environment j, can be written as

$$Y_{ij} = \beta_{i0} + \beta_{i1}z_{1j} + \ldots + \beta_{it}z_{tj} + e_{ij} \quad (3)$$

where β_{ik} ($k = 0, 1,\ldots, t$) are the $t+1$ unknown regression coefficients, z_{kj} ($=f_k(x_j)$) are known functions of the environmental indexes x_j, e_{ij} are independent and identically distributed random variables following normal distribution with $E(e_{ij}) = 0$, for all i, j, and

$$\mathrm{Cov}(e_{ij}, e_{i'j'}) = \begin{cases} \sigma_e^2, & i = i',\ j = j', \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

After identifying the regression functions for all genotypes, a test of coincidence for regressions can be used to check whether the yield responses for genotypes are similar with respect to environmental indexes. This test can be performed in two stages (Kleinbaum et al., 2008; Williams, 1967): (i) a test for parallelism of regression curves; and (ii) a test of coincidence for those regressions.

Equation (3) can be rewritten as

$$Y_{ij} = \mu_i + \beta_{i1}(z_{1j} - \overline{z}_1) + \beta_{i2}(z_{2j} - \overline{z}_2) + \ldots + \beta_{it}(z_{tj} - \overline{z}_t) + e_{ij},\ j = 1, \ldots, b;\ i = 1, \ldots I \quad (5)$$

After centering the observations we have

$$\beta_{i0} = \mu_i - \beta_{i1}\overline{z}_1 - \beta_{i2}\overline{z}_2) - \ldots - \beta_{it}\overline{z}_t. \quad (6)$$

The null hypothesis to test the parallelism of regression functions can be written as

$$H_0 : \beta_{ik} = \beta_{ck},\ i = 1,\ldots, I;\ k = 1,\ldots, t, \quad (7)$$

where β_{ck} denotes the common k^{th} regression coefficient, equal for all genotypes.

Let us consider now the case when some of the observations Y_{ij} are missing. Then, let n_i ($\leq b$) be the number of environments in which the genotype i is observed, and $N = \sum_{i=1}^{I} n_i$.

Classical regression techniques are used to estimate the parameters by the least squares method, independently for each genotype. Then we have

$$\hat{\mu}_i = \frac{1}{n_i}\sum_j Y_{ij}; \qquad \mathbf{b}_i = (\mathbf{Z}_i'\mathbf{Z}_i)^{-1}\mathbf{Z}_i'\mathbf{Y}_i$$

$$\mathrm{SS}_{i,e} = \mathbf{Y}_i'\mathbf{Y}_i - \mathbf{b}_i'\mathbf{Z}'\mathbf{Y}_i, \qquad i = 1, 2,\ldots, I,$$

where b_i is the vector of estimators of regression parameters for the genotype i, Z_i is an $(n_i \times t)$ matrix of centered values of explanatory variables, $Y_i = [Y_{i1}, Y_{i2}, \ldots, Y_{in_i}]'$, $i = 1, 2, \ldots, I$. The $\mathrm{SS}_{i,e}$ has $n_i - t - 1$ degrees of freedom.

Table 1 gives the analysis of variance to test the parallelism of regression functions. The statistic $F_{C,I}$, under H_0, follows an F central distribution with $(I - 1)t$ and $N - It - I$ degrees of freedom. After rejecting the hypothesis that all intercepts are equal we can test the hypothesis of the form:

$$H_{01} : \beta_{10} = \beta_{20} = \ldots \beta_{I0}. \quad (8)$$

To test H_{01} it is necessary to calculate the estimator of common regression coefficients b_C (under hypothesis H_0). These estimators can be obtained by solving normal equations of the form

$$\mathbf{Z}'\mathbf{Z}\mathbf{b}_C = \mathbf{Z}'\mathbf{Y} \quad (9)$$

where, $\mathbf{Y} = \sum_{i=1}^{I}\mathbf{Y}_i$, $\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{I}\mathbf{Z}_i'\mathbf{Z}_i$, $\mathbf{Z}'\mathbf{Y} = \sum_{i=1}^{I}\mathbf{Z}_i'\mathbf{Y}_i$, $\mathbf{Y}'\mathbf{Y} = \sum_{i=1}^{I}\mathbf{Y}_i'\mathbf{Y}_i$.

Table 1 – Analysis of variance for parallelism of the regression lines.

Source of variation	d. f.	Sum of squares	Mean square	F-ratio
Combined Regression	t	$\mathrm{SS}_{R,C} = \mathbf{b}_C'\mathbf{Z}'\mathbf{Y}$	$\mathrm{MS}_{R,C} = \frac{\mathrm{SS}_{R,C}}{t}$	
Between regressions	$(I-1)t$	$\mathrm{SS}_{C,I} = \sum_{i=1}^{I}\mathbf{b}_i'\mathbf{Z}'\mathbf{Y}_i - \mathrm{SS}_{R,C}$	$\mathrm{MS}_{C,I} = \frac{\mathrm{SS}_{C,I}}{(I-1)t}$	$\mathrm{F}_{C,I} = \frac{\mathrm{MS}_{C,I}}{\mathrm{MS}_e}$
Combined Residuals	$N - It - I$	$\mathrm{SS}_e = \sum_{i=1}^{I}\mathbf{Y}_i'\mathbf{Y}_i - \sum_{i=1}^{I}\mathbf{b}_i'\mathbf{Z}'\mathbf{Y}_i$	$\mathrm{MS}_e = \frac{\mathrm{SS}_e}{N - It - I}$	
Total within genotypes	$N - I$	$\mathrm{SS}_{Y,I} = \sum_{i=1}^{I}\mathbf{Y}_i'\mathbf{Y}_i$		

By $SS_C = \mathbf{Y'Y} - \mathbf{b}'_C\mathbf{Z'Y} = \sum_{i=1}^{I}\mathbf{Y}'_i\mathbf{Y}_i - \mathbf{b}'_C\mathbf{Z'Y}$ we denote the comon sum of squares for deviation from regression with $\nu_e = N - I - t$ degrees of freedom.

Let $\tilde{\mathbf{Y}}$ be the vector of all N observations corresponding to the vector $\hat{\beta}$ of t regression parameters and an $N \times t$ matrix $\tilde{\mathbf{Z}}$ of observed values of $\mathbf{Z}$. Using the standard regression technique we obtain the sum of squares for the residual $SS_E = \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}} - \tilde{\mathbf{b}}\tilde{\mathbf{Z}}'\tilde{\mathbf{Y}}$.

The overall analysis of variance to test the hypotheses H_0 and H_{01} is presented in Table 2. The statistic $F_{I'}$ under $H_{01'}$ follows an *F*-distribution with $I - 1$ and $N - It - I$ degrees of freedom.

Plant materials

To illustrate the described methodology, we use the yield data from a winter rye (*Secalecereale* L.) experiment, obtained in multi-environment trials carried out at Słupia Wielka (Poland) in 1997 and 1998. In each design there were four superblocks, with four blocks of four plots each. Each genotype occurred in one plot per superblock. The data set used in this illustration is a subset with five genotypes (CHD_296, RAH_596, RAH_697, RAH_797 and URSUS) and 32 blocks. From this two-way table with 32 rows and five columns, 10 % of the data values were deleted to simulate the possibility of having missing values due to pests, animals or other likely factors. The removal of missing values was performed so as to produce an identical number of missing cells per genotype. The summary of the two-way data is presented in Table 3.

Results and Discussion

After applying the zigzag algorithm to the non-orthogonal series of experiments described in the previous section, the environmental indexes are obtained and used as independent variable for the regression curves. The response function (i.e. yield) with respect to environmental index was estimated by several functions and the adjusted coefficient of determination, R^2, obtained. Table 4 shows the adjusted R^2 for all the considered functions and all five genotypes.

Although all the adjusted R^2 are high and very similar, we have decided to use quadratic regression to express the responses of genotypes because this model is that for all genotypes it gives the best fit of the data, as can be seen from Table 4. The adjusted regression coefficients of the quadratic model, R^2 and *p*-values arepresented in Table 5. Figure 1 represents the adjusted quadratic regressions for the five winter rye genotypes in study.

After rejecting the hypothesis that the regression is not a quadratic curve for each of the genotypes ($p < 0.001$, Table 5) we are led to test whether the regression curves are parallel for all five genotypes (cf. Table 1). The hypothesis that the regression lines are parallel is rejected (Table 6), as expected after analyzing Figure 1 and from a preliminary analysis where GEI was found in this data.

In the next step of investigating GEI we tried to find subgroups of genotypes where responses are parallel (or coincident). By simple inspection of the coefficients (Table 5), we can consider two groups of genotypes: (1) CHD_296; RAH_797 and URSUS; (2) RAH_596 and RAH_697. In this case this is in accordance with the preliminary analysis presented in Figure 1.

To quantify the differences we can use multiple comparison methods such as Scheffé (Scheffé, 1959; Miller, 1991). When using the Scheffé method, representing by $f_{1-\alpha,\ r,\ g}$ the 1 – α quantile of the central F distribution with r and g degrees of freedom, and $s^2_{b_{im}}$, $i = 1,...,I$, $m = 0,1,2$ the variance of the regression coefficient $b_{im'}$ the pairs of quadratic regression coefficients which satisfy the condition $|b_{im} - b_{i'm}| > \sqrt{(J-1) f_{1-\alpha,J-1,N-3J} MSE\left(s^2_{b_{im}} + s^2_{b_{i'm}}\right)}$, m = 0, 1, 2, are different at the significance level α.

Since the regression coefficients b_1 and b_2 do not differ among environments ($p < 0.05$), we present in Table 7 only the results for the Scheffé multiple comparison method of the b_0 coefficients.

Table 3 – Descriptive statistics for the five genotypes and the environmental index.

Genotypes	n_i	Mean	Std. Dev.	Min. value	Max. value
CHD_296	29	8.24	2.26	5	12.5
RAH_596	29	9.13	2.80	5.8	13.8
RAH_697	29	9.75	2.71	5.9	14
RAH_797	29	9.62	3.00	5.4	13.6
URSUS	28	10.45	3.57	5.5	15.8
Environmental index	32	9.68	2.96	5.73	14.13

Table 2 – Overall analysis of variance to test the coincidence of the regression lines.

Source of variation	d. f.	Sum of squares	Mean square	F-ratio
Overall regression	t	$SS_R = \tilde{\mathbf{b}}'\tilde{\mathbf{Z}}\tilde{\mathbf{Y}}$	$MS_R = \frac{SS_R}{t}$	
Between intercepts	$I - 1$	$SS_I = SS_E - SS_C$	$MS_I = \frac{SS_I}{I-1}$	$F_I = \frac{MS_I}{MS_e}$
Between regressions	$(I - 1)t$	$SS_{C,I}$	$MS_{C,I}$	$F_{C,I}$
Residual-combined	$N - It - I$	SS_e	MS_e	
Total within genotypes	$N - 1$	$\tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$		

Table 4 – Adjusted R^2 values for several response functions.

Function \ Genotype	CHD_296	RAH_596	RAH_697	RAH_797	URSUS
Linear $\left(\hat{Y} = b_0 + b_1 x\right)$	0.924	0.958	0.954	0.978	0.983
Logarithmic $\left(\hat{Y} = b_0 + b_1 \ln(x)\right)$	0.931	0.928	0.919	0.985	0.985
Inverse $\left(\hat{Y} = b_0 + {}^{b_1}/_{x}\right)$	0.919	0.881	0.865	0.972	0.968
Quadratic $\left(\hat{Y} = b_0 + b_1 x + b_2 x^2\right)$	0.931	0.972	0.972	0.985	0.986
Compound $\left(\hat{Y} = b_0 b_1^{\,x}\right)$	0.893	0.966	0.959	0.952	0.950
Power $\left(\hat{Y} = b_0 x^{b_1}\right)$	0.914	0.947	0.935	0.975	0.977
S-Curve $\left(\hat{Y} = e^{b_0 + {}^{b_1}/_{x}}\right)$	0.918	0.909	0.891	0.981	0.985
Growth $\left(\hat{Y} = e^{b_0 + b_1 x}\right)$	0.893	0.966	0.959	0.952	0.950
Exponential $\left(\hat{Y} = b_0 e^{b_1 x}\right)$	0.893	0.966	0.959	0.952	0.950
Logistic $\left(\hat{Y} = \left(\frac{1}{16} + b_0 b_1^{\,x}\right)^{-1}\right)$	0.920	0.942	0.932	0.982	0.907

Table 5 – Regression coefficients, coefficient of determination and *p*-value for the five genotypes.

Genotype	Regression coefficients			R^2	*p*-value
	b_0	b_1	b_2		
CHD_296	-2.049	1.486	-0.037	0.93	<0.001
RAH_596	6.021	-0.354	0.064	0.97	<0.001
RAH_697	6.900	-0.440	0.068	0.97	<0.001
RAH_797	-4.194	1.968	-0.049	0.99	<0.001
URSUS	-3.964	1.901	-0.037	0.99	<0.001

Table 6 – ANOVA for parallelism of all five quadratic regression lines.

Source of variation	d.f.	Sum of squares	Mean square	F-ratio	*p*-value
Combined regression	2	1100.25	550.12		
Difference of regressions	8	33.23	4.15	17.74	< 0.001
Combined residuals	129	30.20	0.23		
Total within groups	139	1163.67			

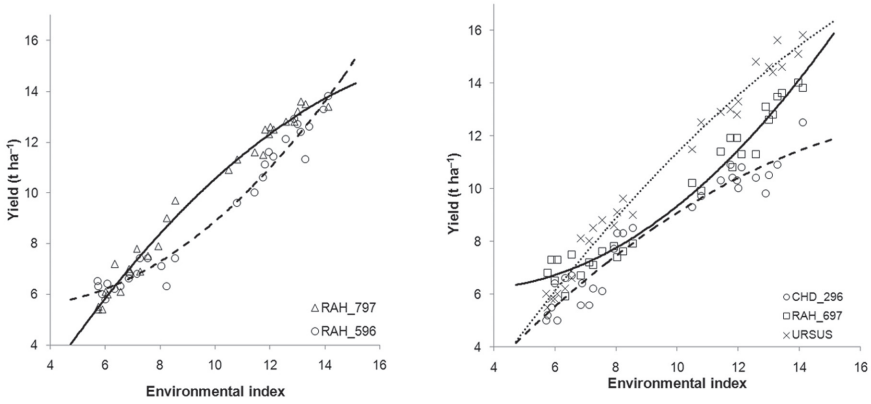

Figure 1 – Adjusted quadratic regressions for the five winter rye genotypes in study. The abscissa corresponds to the yield and the ordinate to the environmental index. The dots represent the genotypes and the solid lines the adjusted quadratic regressions.

The groups obtained with the Scheffé method at significance level 1 % are the same as those given by the simple inspection of the coefficients or by analysis of Figure 1.

The multiple comparison method of Scheffé made it possible to divide the genotypes into two groups: one group with upward-facing concavity (i.e. potential yield growth) and other with downward-facing concavity (i.e. the yield approaches saturation). Inspecting the coefficients, especially b_2, it is possible to see the form of the yield curves. If b_2 is positive, the curve will be convex, otherwise concave.

Table 8 shows the common regression coefficients for all genotypes together and each of the two groups obtained using the Scheffé multiple comparison method, while Table 9 gives the ANOVA to test the parallelism of regression lines for each of the two groups of genotypes. The hypotheses of parallelism between the quadratic regressions were rejected for the first group of genotypes (Table 9). However, we do not reject the same hypothesis for the second group. Hence, in this case, we can go one step further and test whether the regressions in the second group are coincident (hypothesis H_{01}). From the ANOVA presented in Table 10 we reject that hypothesis, i.e. although the quadratic regression lines are parallel they are distinct. Therefore the adjusted regression functions for the genotypes RAH_596 and RAH_697 can be written as:

$$\hat{Y}_{RAH\_596} = 6.229 - 0.397x + 0.066x^2;$$

$$\hat{Y}_{RAH\_697} = 6.717 - 0.397x + 0.066x^2,$$

Table 7 – Scheffé multiple comparison tests of the b_0 coefficients.

$\lvert b_{i0} - b_{i'0} \rvert$	RAH_596	RAH_697	RAH_797	URSUS
CHD_296	8.07**	8.949**	2.145*	1.915^{NS}
RAH_596	-	0.879^{NS}	10.215**	9.985**
RAH_697	-	-	11.094**	10.864**
RAH_797	–	-	-	0.23^{NS}

*Significant at the 0.05 probability level; **Significant at the 0.001 probability level; ^{NS}not significant at the 0.05 probability level.

Table 8 – Common regression coefficients, coefficient of determination and *p*-values for all genotypes together and each of the two groups (centered data).

Groups of genotypes	Regression coefficients			R^2	*p*-value
	b_0	b_1	b_2		
All	0.000	0.852	0.005	0.95	<0.001
Group 1	0.000	1.628	-0.033	0.95	<0.001
Group 2	0.000	-0.397	0.066	0.97	<0.001

Table 9 – ANOVA to test the parallelism of regression lines for each of the two groups of genotypes.

Group	Source of variation	d.f.	Sum of squares	Meansquare	F-ratio	*p*-value
1	Combined Regression	2	699.09	349.54		
	Difference of Regressions	4	21.13	5.28	22.03	< 0.001
	Combined Residuals	77	18.46	0.24		
	Total within groups	83	738.68			
2	Combined Regression	2	413.25	206.62		
	Difference of Regressions	2	0.008	0.004	0.018^{NS}	0.98
	Combined Residuals	52	11.73	0.23		
	Total within groups	56	424.99			

^{NS}not significant at the 0.05 probability level.

Table 10 – ANOVA to test the coincidence of regression functions in the second group (RAH_596 and RAH_697).

Source of variation	d.f.	Sum of squares	Mean square	F-ratio	*p*-value
Overall regression	2	415.40	207.70		
Between intercepts	1	3.44	3.44	15.26*	0.0003
Between regressions	2	0.008	0.004	0.18^{NS}	0.98
Residual-combined	52	11.73	0.23		
Total-within genotypes	57	430.58			

*Significant at the 0.001 probability level; ^{NS}not significant at the 0.05 probability level.

where the b_1 and b_2 are the common regression coefficients from Table 8 and the b_0 coefficients are calculated according to expression (6).

We point out that we did not find groups of genotypes with identical regressions; this may be due to the high level of GEI. All that we can say is that genotypes RAH_596 and RAH_697 have yields parallel, with that of the second genotype a little higher.

Conclusions

The hypothesis of parallelism of regression curves was rejected, which is natural in multi-environment trials with interaction between genotype and environment. The main difference in the two subgroups of genotypes where the responses are parallel is that one group had upward-facing concavity (i.e. potential yield growth) and the other had downward-facing concavity (i.e. the yield approaches saturation), which can help breeders in their genotype selection. The approach proposed in this paper is general and applicable to any series of experiments conducted in multi-environment trials or simply to the case of two-way classified data.

Acknowledgments

Dulce G. Pereira is a member of the CIMA-UE, a research center financed by the Science and Technology Foundation, Portugal. The work was partially supported by Ministry of Science and Higher Education Grant N N310 447838, and by the Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through PEst-OE/MAT/UI0297/2011 (CMA). We thank the anonymous referees and the Associated Editor for their suggestions, which greatly improved the paper.

References

Aastveit, A.H.; Mejza, S. 1992. A selected bibliography on statistical methods for the analysis of genotype × environment interaction. Biuletyn Oceny Odmian 25: 83–97.

Annicchiarico, P. 2002. Genotype × Environment Interactions: Challenges and Opportunities for Plant Breeding and Cultivar Recommendations. Food and Agricultural Organization, Rome, Italy. (FAO Plant Production and Protection Paper, 174).

Digby, P.G.N. 1979. Modified joint regression-analysis for incomplete variety × environment data. Journal of Agricultural Science 93: 81–86.

Finlay, K.W.; Wilkinson, G.N. 1963. Analysis of adaptation in a plant-breeding programme. Australian Journal of Agricultural Research 14: 742–754.

Gabriel, K.R.; Zamir, S. 1979. Lower rank approximation of matrices by least-squares with any choice of weights. Technometrics 21: 489–498.

Gauch, H.G. 1992. Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs. Elsevier, Amsterdam, Netherlands.

Gusmão, L. 1985. An adequate design for regression analysis of yield trials. Theoretical and Applied Genetics 71: 314–319.

Gusmão, L. 1986. Inadequacy of blocking in cultivar yield trials. Theoretical and Applied Genetics 72: 98–104.

Kang, M.S.; Gauch, H.G., eds. 1996. Genotype-by-environment interaction. CRC Press, Boca Ratón, FL, USA.

Kleinbaum, D.G.; Kupper, L.L.; Nizam, A.; Muller, K.E. 2008. Applied Regression Analysis and Other Multivariable Methods. 4ed. Thompson Higher Education, Belmont, CA, USA.

Mexia, J.T.; Pereira, D.G.; Baeta, J. 1999. L_2 environmental indexes. Biometrical Letters 36: 137–143.

Miller, R.G. 1991. Simultaneous Statistical Inference. Springer, New York, NY, USA.

Ng, M.P.; Williams, E.R. 2001. Joint-regression analysis for incomplete two-way tables. Australian & New Zealand Journal of Statistics 43: 201–206.

Pereira, D.G.; Mexia, J.T. 2010. Comparing double minimization and zig-zag algorithms in joint regression analysis: the complete case. Journal of Statistical Computation and Simulation 80: 133–141.

Pereira, D.G.; Rodrigues, P.C.; Mejza, S.; Mexia, J.T. 2011. A comparison between joint regression analysis and AMMI: a case study with barley. Journal of Statistical Computation and Simulation 82: 193–207.

Rodrigues, P.C.; Pereira, D.; Mexia, J.T. 2011. A comparison between JRA and AMMI: the robustness with increasing amounts of missing data. Scientia Agricola 68: 679–686.

Romagosa, I.; van Eeuwijk, F.A.; Thomas, W.T.B. 2009. Statistical analyses of genotype by environment data. v.3. In: Phohens, J.; Nuez, F.; Carena, M.J. eds. Handbook of Plant Breeding. Elsevier, New York, NY, USA.

Scheffé, H. 1959. The Analysis of Variance. Wiley, New York, NY, USA.

Williams, E.J. 1967. Regression Analysis. Wiley, New York, NY, USA.