Characterisation of grain yield performances of four sunflower varieties using heteroscedastic mixed factorial regression

V. FOUCTEAU

INRA, Unité de Biométrie, F-78026 VERSAILLES Cedex

()) foucteau@bmve01.versailles.inra.fr

26.06.2014

A genotype by environment table involving grain yield data of the four check varieties used in French official trials from 1993 to 1996 is investigated for genotype by environment interaction. The heteroscedastic mixed factorial regression model removed a large part of the interaction. Performances of the four varieties were characterised by genotypic responses to disease pressure and high temperatures as well as stability variances.

Keywords*: covariates, genotype by environment interaction, heteroscedasticity, REML (REstricted Maximum Likelihood), stability, sunflower*

Introduction

In plant breeding, candidate varieties are usually evaluated in a set of environments. These multi-environment trials generate genotype by environment (GE) data. When there is no genotype by environment interaction (GEI), genotypic means are sufficient for comparing genotypic performances. In the presence of GEI, which is the usual situation, other parameters are needed to characterise genotypes.

The environments are considered as location by year combinations. The interest of plant breeders is often more focused on the set of genotypes to evaluate, than on the environments, which are regarded only as providing information about the genotypes. In this context, the environmental effect must be considered as random.

Heteroscedastic mixed factorial regression (Denis *et al*, 1997) is a unified class of mixed models which allows genotypic regressions on environmental covariates and allows to deal with heteroscedasticity, i.e. differences in variance between genotypes. It provides useful parameters for describing genotype performances: responses to the main environmental characteristics for each genotype as well as a measure of their stability, a part of it being explained by genotypic characteristics.

This study presents the use of this model for the analysis of grain yield data of the four check varieties used in French official trials from 1993 to 1996.

Material and methods

Experimental data

The data set investigated in this study consists of a 4 by 130 genotype by environment table. It gathers grain yield of the four check varieties used in French official trials from 1993 to 1996. Because SANTAFE has been replaced by EUROSOL as check variety, there are missing data (nearly 10%). The traits measured in each plot of the trials were: flowering date, grain yield (GY) oil content (OC) and grain moisture content at harvest (MC). From the 2 GE tables of OC and MC, we derived 2 genotypic covariates (OC.G and MC.G) by calculating genotypic adjusted means.

Sunflower development cycle has been divided into 5 periods according to physiology of the plant. Period 1 runs from sowing to emergence. Sowing date was available for all the trials whereas emergence date was estimated by adding 90 degree Celsius days to sowing date. Period 2 goes from emergence to the B9 stage which corresponds to the emergence of the 9th leaf and was computed by adding 160 degree days to emergence date. The flowering period (period 4) was delimited using the flowering dates of each genotype in each trial. Flowering date minus 13 days and flowering date plus 12 were chosen as boundaries for period 4, in order to get a period of length comparable to the effective length of flowering for sunflower (approximately 20 days). Period 3 goes from the B9 stage to beginning of flowering as defined above. M3 is the maturity stage when grain moisture content decreases down to 15%; it matches with the end of physiological activity. The M3 stage was estimated from harvest date using the empirical relationship:

$$M3_{ij} = \text{HarvestDate}_j - (30 - 2MC_{ij}),$$

where M3 and Harvest Date are expressed in number of days and MC is a percentage. Period 5 goes from the end of flowering to the M3 stage.

Climatic covariates were computed using meteorological data from the national meteorological network of Météo France. Each location was characterised by the nearest meteorological station measuring temperatures, radiation, rainfall (R) or potential evapotranspiration (ETp). Water deficit (WD) in mm estimated as:

$$WD = ETm - (R + I),$$

where I is the amount of water provided from irrigation and ETm is the maximum evapotranspiration, was computed for each of the 5 periods described above and called WD1 to WD5. ETm was calculated as $ETm = k\,ETp$ with k a coefficient equal to 0.6, 0.9, 1.25, 1.3 and 0.7 for periods 1 to 5 respectively (following Merrien 1992). The influence of low temperatures during flower differentiation (LT3) was quantified by summing daily minimum temperatures below 5°C over the 20 days following the B9 stage. In the same way, effect of high temperatures during maturation (HT5) was appreciated by summing daily maximum temperatures above 30°C over period 5. The sum of daily radiation from emergence to the M3 stage was computed (RAD). The sum of degree days (ST) based on 6°C (sum of daily mean temperatures above 6°C) was computed over the same period for each genotype by environment combination; means by genotype and environment resulted in 2 additional covariates (ST.G and ST.E respectively).

Latitude (LAT) and longitude (LONG) of locations were used as environmental covariates. The year of experimentation (YEAR) and the possible application of fungicide treatments (FT) were used as environmental factors.

Symptoms of lodging, phomopsis (on stem), sclerotinia (on capitulum, neck, bud and stem), rhizopus or phoma were recorded in some trials. Adjusted environmental means were computed from these notations and are used as environmental covariates for diseases pressure (called LODG, PHOMO, SCC, SCN, SCB, SCS, RH and PHOMA respectively). For each

disease, we gave the value 0 for the environments where no notation was done, following the hypothesis that there were no symptoms for this disease in these environments.

Statistical methods

Heteroscedastic mixed factorial regression

Heteroscedastic factorial regression model (Denis *et al*, 1997) allows multiple regression on covariates depending on either genotypes or environments:

$$Y_{ij} = \mu + \alpha_i + B_j + \sum_{h=1}^{H} \rho_{ih} z_{ijh} + \sum_{k=1}^{K} T_{jk} x_{ik} + E_{ij}$$

Y_{ij} denotes the random variable Y for the genotype i and the environment j. Fixed parameters μ, α_i and ρ_{ih} are respectively the grand mean, the effect of genotype i and the regression coefficient on the h^{th} environmental covariate for genotype i. Random parameters B_j and T_{jk} are respectively the effect of the environment j and the regression coefficient on the k^{th} genotypic covariate for genotype i. x_{ik} is the value of the k^{th} genotypic covariate ($k = 1,\ldots,K$) for genotype i. z_{ijh} is the value of the h^{th} environmental covariate ($h = 1,\ldots,H$) for environment j; we allow some of the environmental covariates to depend on both genotypes and environments and thus we write z_{ijh} instead of z_{jh}. Covariates are centred. E_{ij} is the residual term. Expectation and variance of Y_{ij} are respectively:

$$\mathrm{E}(Y_{ij}) = \mu + \alpha_i + \sum_{h=1}^{H} \rho_{ih} z_{ijh} \text{ and } V(Y_{ij}) = \sigma_B^2 + \sum_{k=1}^{K} \sigma_{T_k}^2 x_{ik}^2 + \sigma_E^2,$$

where σ_B^2, $\sigma_{T_k}^2$ and σ_E^2 are respectively the variance components for the environment, the k^{th} genotypic covariate and the error. The K random terms involving environmental regressions on genotypic covariates produce heteroscedasticity, i.e. differences in variances between genotypes.
Alternatively, heteroscedasticity can be removed by including residual variance components for each genotype ($\sigma_{E(i)}^2$). This model can be written:

$$Y_{ij} = \mu + \alpha_i + B_j + \sum_{h=1}^{H} \rho_{ih} z_{ijh} + E_{ij}$$

The expectation is not modified but the variance becomes:

$$V(Y_{ij}) = \sigma_B^2 + \sigma_{E(i)}^2$$

Estimation of variance components and fixed parameters were done using REML (REstricted Maximum Likelihood) through Genstat (version 5, release 4.1).

Reduced-rank regression

Reduced-rank regression (van Eeuwijk *et al*, 1996) allows the incorporation of several covariates consuming few degrees of freedom:

$$\mathrm{E}(Y_{ij}) = \mu + \alpha_i + \beta_j + \rho_i \left(\sum_{h=1}^{H} \delta_{1h} z_{jh} \right)$$

The information given by the covariates is summed-up into a synthetic covariate,

$$\zeta_j = \sum_{h=1}^{H} \delta_{1h} z_{jh},$$

which is the most explanatory linear combination of the original covariates. Reduced-rank regression belongs to biadditive regression models described by (Denis, 1998). Because the data set contains missing values, we used the EM (Expectation-Maximisation) algorithm to fit the model.

Optimum model construction

Environmental covariates were selected according to their mean squares computed in fixed models. An estimate of the experimental error was obtained by pooling the errors estimated at the plot level in each trial. This pooled error was used for all F-tests. Three disease pressure covariates, lodging (LODG), *phomopsis* (PHOMO) and *sclerotinia* on capitulum (SCC), which were each significant for explaining the interaction, were positively correlated between themselves. So the effect of each one could not be separated. We computed a synthetic covariate (called SYN) using reduced-rank regression. LODG and PHOMO highly contributed to the synthetic covariate, while the contribution of SCC to SYN was very small. The other covariates related to diseases pressure were not significantly correlated neither between themselves nor with LODG, PHOMO or SCC; so they were used individually. Environmental covariates selection was performed by progressively adding to the additive model the best covariates. The first two were used in the heteroscedastic mixed factorial regression model.
Once environmental covariates had been chosen for the fixed part of the model, the best and significant genotypic covariates were introduced in the random part of the model. Alternatively, individual residual variances were included. Wald tests for fixed effects were calculated after heteroscedasticity had been modelled. These statistics have an asymptotic χ^2 distribution with the degrees of freedom equal to those of the fixed term.

Results

The fixed part of the model explains the interaction by estimating differential responses between genotypes to environmental characteristics. The more explanatory environmental covariates were the synthetic covariate (SYN), traducing disease pressure (mainly lodging and *phompsis*) and the covariate quantifying the occurrence of high temperatures during flowering and maturating stages (HT45). Both covariates were significant, with P-value below 10^{-6} and 5.10^{-3} respectively.

Table 1: comparison of the models

	Model						
	Additive model	Env. Cov.	Geno. Cov.	Stability Var.	df	Deviance	σ_E^2
1	x				464	1639	4.86
2	x	x			456	1562	3.67
3	x	x	x		455	1553	3.00
4	x	x		x	453	1534	

Moisture content at harvest (MC) was the only significant genotypic covariate. Its introduction in the random part of the model 3 (Table1) decreased the deviance by 9 consuming 1 degree of freedom (P-value<10^{-3}). Alternatively, the addition of genotypic individual variances (model 4) led to a strong decrease of the deviance (29 for 3 degrees of freedom, which corresponds to a P-value below 10^{-5}).

The residual part of heteroscedastic mixed factorial regression models comprises not only the experimental error, but also the interaction not accounted for by the covariates or the stability variances. In the additive model (model 1), the estimated residual variance (4.85) is much larger than the pooled error (1.30) because it comprises all the interaction. In model 2, the estimated residual variance has dropped down to 3.67, indicating that genotypic regressions on the two environmental covariates explained some of the interaction. In model 4, including stability variances, the smallest genotypic error variance is 1.74. This indicates that a part of the interaction remained unexplained. This part of the interaction was not due to heteroscedasticity (as genotypic stability variances accounts for all of it), but to some important environmental covariates not taken into account in the fixed part of the model.

Discussion

Genotypic responses to environmental characteristics

Table 2: parameters estimates with their standard error (s.e.)

			Regressions on environmental covariates				Genotypic variances			
Genotypes	mean	s.e.	SYN	s.e.	(x1000) HT45	s.e.	$\sigma_T^2 x_i^2$	s.e.	$\sigma_{E(i)}^2$	s.e.
ALBENA	29.26	0.52	-10.24	7.16	-0.26	6.50	0.02	0.01	1.74	0.41
EUROSOL	28.71	0.54	-29.15	7.27	-3.22	6.65	0.02	0.01	3.05	0.56
SANTAFE	27.67	0.58	-12.99	8.05	6.03	7.16	0.68	0.26	7.38	1.22
VIKI	28.18	0.52	-36.01	7.17	-7.99	6.51	1.23	0.46	3.17	0.54

The comparison of estimated regression coefficients on the SYN covariate shows that EUROSOL and above all VIKI were the genotypes most penalised by high pressure of lodging and *phomopsis* (Table 2). This is in accordance with the known sensitivities to *phomopsis* of the four varieties. One can think that sensitivity to lodging can be related to earliness at harvest and plant height, late and tall genotypes being more affected than early and small ones. This hypothesis would allow to interpret the difference in regression

coefficient between EUROSOL and VIKI: both have the same sensitivity to *phomopsis*, but VIKI was more penalised by lodging because it is latest than EUROSOL. Both varieties have similar height.

High temperatures during and after flowering (HT45) decrease oil content and total grain yield (Merrien, 1992). VIKI was the most penalised by high temperatures, maybe because it is a high oil genotype. All coefficient are expected to be negative. Though, SANTAFE had a positive regression coefficient on HT45. This may not traduce a kind of tolerance of this variety to high temperatures, but could result from a compensation phenomenon when estimating regression coefficients on HT45 once grain yield had been adjusted for SYN, though these two covariates were not significantly correlated. Hence, SANTAFE positive regression coefficient on HT45 must be cautiously interpreted.

Variance modelling using either genotypic covariates or stability variances

The random term involving environmental regressions on genotypic covariate MC results in heteroscedasticity related to differences in earliness between genotypes. Hence, the earliest variety (SANTAFE) and, above all, the latest one (VIKI), had the greatest earliness-related variances (Table 2). Stability variances were much larger than earliness-related genotypic variances, indicating that differences between genotypes in other characteristics (than earliness) contributed to heteroscedasticity. Comparing stability variances, SANTAFE appeared to be the less stable variety whereas ALBENA was the more stable one. VIKI, though having the highest earliness-related variance, is no more unstable than EUROSOL.

Conclusion

The heteroscedastic mixed factorial regression model succeeded in removing a large part of GE interaction. It provided good parameters for characterising the behaviour of the four check varieties in the large set of environments resulting from the 1993-1996 French official trials.

References

Denis J-B, Piepho H-P, van Eeuwijk FA (1997) Modelling expectation and variance for genotype by environment data. Heredity 79:162-171

Denis J-B (1998) BIAREG: Splus functions to perform Biadditive Regressions. Report INRA-Versailles

van Eeuwijk FA, Denis J-B and Kang MS (1996) Incorporating additional information on genotypes and environments in models for two-way genotype by environment tables In Kang, M.S. et Gauch, H.G., eds. Genotype-By-Environment Interaction, CRC-Press, Boca Raton, Fl, USA pp15-49

Genstat 5 Committee (1993) Genstat 5, Release 3, Reference Manual. Clarendon Press, Oxford

Merrien A (1992) Physiologie du tournesol. Les points techniques du CETIOM, Ed CETIOM, 65 p