# A GxE analysis of Triticale in Spain - 2: modelling the VCOV

In the previous exercise we have investigated modelling the mean. Here we will investigate modelling the (co)variances, where genetic effects are treated as random. (The idea is that the genotypes are a sample from a large population of potential genotypes).

1. **Raw data**
   First look at the results of question 1 of the previous exercise, particularly at the variances per environment (site), and correlations between environments. This should already give an indication of the kind of random-effects model you need. Based on the results of question 1 of the previous exercise, what do you think should be important properties of the random-effects model?

2. **The simplest model: a random genotype effect (CS)**
   The simplest random effects model contains a random genotype effect.
   $$\underline{y}_{ij} = \mu + E_j + \underline{G}_i + \underline{\varepsilon}_{ij}$$
   This is also called a compound symmetry model (CS). In the context of GxE, animal breeders would simply call it a single trait model. Fit this model using the file Triticale_CS.as.
   a. Observe the likelihood and calculate the AIC (Akaike's information criterion; in simple terms, "the lower the AIC, the better the model". AIC serves for model comparison, its absolute value is not important), AIC = 2k – 2ln(L), ln(L) being the log-likelihood and *k* the number of parameters fitted. You can use *k* = 1 here, *i.e.*, count only the random effect, since all models in this assignment will have the same fixed effects.
   b. What is the genetic and phenotypic correlation between environments in this model? From the phenotypic correlation, does the performance in one environment predict performance in another environment accurately?

3. **Some improvement: heterogeneous variances (CSH)**
   The model can be improved by allowing for heterogeneous residual variances (CSH),
   $$\underline{y}_{ij} = \mu + E_j + \underline{G}_i + \underline{\varepsilon}_{ij}$$
   where $\text{var}(\underline{\varepsilon}_{ij}) = \text{var}(\underline{\varepsilon}_j)$; in other words, there is a separate residual variance for each environment. (Model 1 had $\text{var}(\underline{\varepsilon}_{ij}) = \text{var}(\underline{\varepsilon})$). Fit this model using the file Triticale_CSH.as.
   a. Observe the likelihood and calculate the AIC.
   b. Compare the previous and this model; are there indications that the variance is heterogeneous?
   c. What are the genetic and phenotypic correlations between Coruna and Cordoba in this model? How do you judge the similarity of the phenotypes of a genotype between Coruna and Cordoba?

4. **A homogeneous diagonal model (DIAG)**
   An alternative (simpler) model has no random genetic effect, and a homogeneous diagonal residual variance structure (DIAG),
   $$\underline{y}_{ij} = \mu + E_j + \underline{\varepsilon}_{ij}$$
   where $\text{var}(\underline{\varepsilon}_{ij}) = \text{var}(\underline{\varepsilon})$ and $\text{cov}(\underline{\varepsilon}, \underline{\varepsilon}) = 0$; in other words, there is single residual variance, and residuals are independent. Fit this model using the file Triticale_DIAG.as.

a. Observe the likelihood and calculate the AIC.
b. What is the assumption about the genetic and phenotypic correlation of genotypes between environments?

**5. A heterogeneous diagonal model (DIAGH)**

An more flexible DIAG model allows for heterogeneous variances between environments (DIAGH),

$$\underline{y}_{ij} = \mu + E_j + \underline{\varepsilon}_{ij}$$

where $var(\underline{\varepsilon}_{ij}) = var(\underline{\varepsilon}_j)$ and $cov(\underline{\varepsilon}, \underline{\varepsilon}) = 0$; in other words, there is a separate variance for each environment. Fit this model using the file Triticale_DIAGH.as.

a. Observe the likelihood and calculate the AIC.
b. Compare the model to DIAG and CS and CSH. What is more important, fitting a correlation between environments, or fitting the variance correctly?

**6. A full model (FULL)**

An much more flexible model is a full model, that allows for a separate variance in each environment, and for any covariance between environments,

$$\underline{y}_{ij} = \mu + E_j + \underline{\varepsilon}_{ij}$$

where $var(\underline{\varepsilon}_{ij}) = var(\underline{\varepsilon}_j)$ and $cov(\underline{\varepsilon}_{ij}, \underline{\varepsilon}_{ik}) = cov_{jk}$. In other words, there is a separate variance for each environment, and a separate covariance between each pair of environments.

For the animal breeders: This is the "multitrait model", yield in each environment is treated as a different trait. This is fitted in the residual structure of the model (i.e. there is no random genotypic effect in the model), because there is only a single observation for each genotype-environment combination. So $\underline{\varepsilon}_{ij}$ actually represents $GE_{ij} + e_{ij}$, which is treated as a separate trait for each ij.

There are two as-files that you can use to fit this model. The file Triticale_FULL.as is the typical multitrait implementation as used in animal breeding (notice it uses another format of the data set). The file Triticale_FULLb.as fits the same model by defining a residual variance structure. (The file Triticale_FULLb.as has a list of starting value very close to the ultimate estimates, to facilitate convergence. The Triticale_FULL.as converges without any starting values, hence is more stable).

a. Compare both .as files and see whether you understand them.
b. Before fitting the model, think about the number of parameter of this model.
c. Observe the likelihood and calculate the AIC.
d. How do you judge the fit of this model, any reason to look further?
e. What does this model suggest with respect to GxE-interaction?

**7. A Factor analytical model (FA)**

The full model has very many parameters. A better fit may be obtained by giving a structure to the GE-matrix. This can be done with a factor-analytical model (see ppt). The file Triticale_FA.as implements a first-order factor analytical model.

a. Before fitting the model, think about the number of parameter of this model. Note: there are 10 environments.
b. Observe the likelihood and AIC. How does this model compare to the full model?

c. In the asr-file, you find 20 parameter estimates. The first 10, indicated FAC_L, are the $\lambda$, and the second 10, indicated FAC_V are the $\psi$. From those estimates, calculate the variance in Coruna (environment 3), and the covariance between Coruna and Cordoba (environment 4). Also calculate the estimated genetic correlation between Coruna and Cordoba. Check whether you get the right answers, by comparing with the values given in the (co)variance/correlation matrix.
d. Compare the genetic correlations from the full model to those of the factor analytical model. From the asr-file, you can copy the correlation estimates into e.g. Excel, then subtract both correlation matrices and observe the difference. Is the FA-model a good approximation to the full model in terms of GxE-fitting?