General linear model

Pierre Legendre, Université de Montréal
August 2009

1 - Introduction

The *general linear model* (abbreviated GLM) is a statistical linear model of the form:

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

where **Y** is a matrix with *n* multivariate measurements, **X** may be a design matrix or a matrix containing *m* explanatory variables, **B** is a matrix of parameters that are to be estimated, and **E** is a matrix containing errors. R function: *lm()*.

The residual matrix $\mathbf{U} = \mathbf{Y} - \hat{\mathbf{Y}}$ is usually assumed to contain normal error. If the residuals are not multivariate normal, *generalized linear models* (also abbreviated GLM; R function *glm()*; not to be confused with the *general linear model*, R function *lm()*) may be used to relax assumptions about **Y** and **E**.

t-test, ANOVA, MANOVA and (multiple) linear regression are special cases of the general linear model.

2 - Comparison of sample means by regression

In the study of the comparison of two sample means, Section 3, an alternative computation method was mentioned: a two-group *t*-test is equivalent to a test of the linear regression coefficient between the

response variable and a dummy (binary) variable representing the two groups. This was a first application of the principle of equivalence between the *t*-test and the test in linear regression, in the framework of the general linear model. An exercise was proposed in Section 5 of the *Practicals*.

We will develop that idea and generalize it to all forms of mean comparisons.

Dummy variable coding for factors

The one-way analysis-of-variance problem is equivalent to the problem of explaining the variation in a response variable using a series of dummy variables coding for the groups. Exactly (k – 1) coding variables are necessary to represent the membership of the observations into k groups.

The naive solution is to represent the membership into k groups by k dummy variables. It turns out, however, that when k dummy variables are used, the sum-of-squares-and-cross-products (SSCP) matrix of these variables is collinear and cannot be inverted by ordinary inversion, (R function *solve()*) whereas inverting that matrix is necessary to compute the regression coefficients; see the lecture on multiple regression.

Consider the following example where k = 3 groups of objects are coded for regression using 3 dummy (binary, 0-1) variables, plus the column of '1' that are used to estimate the intercept:

	Code intercept	**Binary Gr. 1**	**Binary Gr. 2**	**Binary Gr. 3**
Group1, replicate1	1	1	0	0
Group1, replicate2	1	1	0	0
Group1, replicate3	1	1	0	0

	Code intercept	Binary Gr. 1	Binary Gr. 2	Binary Gr. 3
Group1, replicate4	1	1	0	0
Group1, replicate5	1	1	0	0
Group2, replicate1	1	0	1	0
Group2, replicate2	1	0	1	0
Group2, replicate3	1	0	1	0
Group2, replicate4	1	0	1	0
Group2, replicate5	1	0	1	0
Group3, replicate1	1	0	0	1
Group3, replicate2	1	0	0	1
Group3, replicate3	1	0	0	1
Group3, replicate4	1	0	0	1
Group3, replicate5	1	0	0	1

That coding, which uses 3 dummy variables, fully and unambiguously describes the group membership of the observations, but the corresponding SSCP matrix cannot be inverted (see *Practicals*). So it cannot be used in regression analysis.

The following coding also fully and unambiguously describes the group membership of the observations. It offers the advantage that the corresponding SSCP matrix can be inverted (see *Practicals*):

	Code intercept	Binary Gr. 1	Binary Gr. 2
Group1, replicate1	1	1	0
Group1, replicate2	1	1	0
Group1, replicate3	1	1	0
Group1, replicate4	1	1	0
Group1, replicate5	1	1	0

	Code intercept	Binary Gr. 1	Binary Gr. 2
Group2, replicate1	1	0	1
Group2, replicate2	1	0	1
Group2, replicate3	1	0	1
Group2, replicate4	1	0	1
Group2, replicate5	1	0	1
Group3, replicate1	1	0	0
Group3, replicate2	1	0	0
Group3, replicate3	1	0	0
Group3, replicate4	1	0	0
Group3, replicate5	1	0	0

Any one of the three “Group” columns could have been removed.

Another form of coding, called Helmert contrasts, can be used. For the example, the coding variables are the following:

	Code intercept	Helmert 1	Helmert 2
Group1, replicate1	1	2	0
Group1, replicate2	1	2	0
Group1, replicate3	1	2	0
Group1, replicate4	1	2	0
Group1, replicate5	1	2	0
Group2, replicate1	1	–1	1
Group2, replicate2	1	–1	1
Group2, replicate3	1	–1	1
Group2, replicate4	1	–1	1
Group2, replicate5	1	–1	1
Group3, replicate1	1	–1	–1

	Code intercept	Helmert 1	Helmert 2
Group3, replicate2	1	–1	–1
Group3, replicate3	1	–1	–1
Group3, replicate4	1	–1	–1
Group3, replicate5	1	–1	–1

The number of Helmert contrasts necessary to correctly and unambiguously describes the group membership of the observations is 2; it is the same as with dummy variables. Helmert contrasts are orthogonal to each other. They are also centred on mean 0: the column sums are 0.

Regression using Helmert contrasts produces the same R^2 as regression with dummy variables. In crossed ANOVA, however, the use of Helmert contrasts will be necessarily to correctly code for the interaction.

The coding rule for Helmert contrasts is illustrated by the following examples:

2 groups: 1 variable

$$\begin{bmatrix} +1 \\ -1 \end{bmatrix}$$

3 groups: 2 variables

$$\begin{bmatrix} +2 & 0 \\ -1 & +1 \\ -1 & -1 \end{bmatrix}$$

4 groups: 3 variables

$$\begin{bmatrix} +3 & 0 & 0 \\ -1 & +2 & 0 \\ -1 & -1 & +1 \\ -1 & -1 & -1 \end{bmatrix}$$

5 groups: 4 variables etc.

$$\begin{bmatrix} +4 & 0 & 0 & 0 \\ -1 & +3 & 0 & 0 \\ -1 & -1 & +2 & 0 \\ -1 & -1 & -1 & +1 \\ -1 & -1 & -1 & -1 \end{bmatrix}$$

Note: the number of Helmert or binary variables used in the regression equation is equal to the number of degrees of freedom of the among-group variation in ANOVA, which is $(k - 1)$.

Function to construct Helmert contrasts in R: *contr.helmert()*. Other types of contrasts are available: see *?contr.helmert*.

Applications

In Section 4 of the document on the comparison of two sample means, a paragraph mentioned that a two-group *t*-test for related samples is equivalent to a *t*-test of the linear regression coefficient between the response variable and a dummy variable representing the two groups, in the presence of binary or Helmert covariables representing the related observations.

Two exercises are proposed in the *Practicals*, one using binary variables to code for the related observations, the other Helmert-coded variables.

Interaction

In crossed ANOVA, a significant interaction between two factors indicates that the effects of one of the factors are not consistent across the levels of the other factor.

With more than one factor, one must represent the main factors using Helmert contrasts. The interaction dummy variables are the direct products of the Helmert variables coding for the main factors. Example:

y	Factor A		Factor B	Interaction AxB	
y_{111}	2	0	1	2	0
y_{112}	2	0	1	2	0
y_{121}	2	0	–1	–2	0
y_{122}	2	0	–1	–2	0
y_{211}	–1	1	1	–1	1
y_{212}	–1	1	1	–1	1
y_{221}	–1	1	–1	1	–1
y_{222}	–1	1	–1	1	–1
y_{311}	–1	–1	1	–1	–1
y_{312}	–1	–1	1	–1	–1
y_{321}	–1	–1	–1	1	1
y_{322}	–1	–1	–1	1	1

The number of variables representing the interaction is equal to the number of variables coding for A times the number of variables coding for B. The Helmert variables representing the main factors A and B are orthogonal by design of the experiment, since the main factors are crossed and the design is balanced. The variables representing the

interaction are orthogonal among themselves. By construction, they are also orthogonal to the variables representing the main factors A and B.

Application

We can use this form of coding to revisit the rat feeding example, where 12 rats, males and females, were fed fresh and rancid lard (pig fat). The response variable recorded how much fat (in g) each rat had eaten. The data can be coded using Helmert contrasts for the main factors, and from them, the interaction variables can be constructed:

Consumption	Sex	Lard	Interaction
709	+1	+1	+1
679	+1	+1	+1
699	+1	+1	+1
592	+1	–1	–1
538	+1	–1	–1
476	+1	–1	–1
657	–1	+1	–1
594	–1	+1	–1
677	–1	+1	–1
508	–1	–1	+1
505	–1	–1	+1
539	–1	–1	+1

These data will be analysed in the *Practicals* and the results compared to the regular ‘aov’ results.

Two-way anova with interaction computed in R

Two-way anova with replication, Model I (two fixed factors)

Rats data

```
Consumption = c(709, 679, 699, 592, 538, 476, 657, 594, 677, 508, 505, 539)
Sex  = gl(2,6)
Lard = gl(2, 3, length=12)

Sex
# [1] 1 1 1 1 1 1 2 2 2 2 2 2
# Levels: 1 2
Lard
# [1] 1 1 1 2 2 2 1 1 1 2 2 2
# Levels: 1 2
```

1. Two-way anova using the aov() function of the stats package (parametric tests)

```
aov.res4 = aov(Consumption ~ Sex*Lard)

summary(aov.res4)
#             Df Sum Sq Mean Sq F value    Pr(>F)
# Sex          1   3781    3781  2.5925 0.1460358
# Lard         1  61204   61204 41.9685 0.0001925 ***
# Sex:Lard     1    919     919  0.6300 0.4502546
# Residuals    8  11667    1458
```

2. Two-way anova by RDA using the rda() function of the vegan package (permutation tests)

2.1. Construct the matrix of Helmert contrasts. Remove the first column (intercept).

```
helmert = model.matrix(~ Sex*Lard, contrasts=list(Sex="contr.helmert",
Lard="contr.helmert"))[,-1]
```

2.2. Test the interaction by RDA. Factors Sex and Lard form the matrix of covariables.

```
interaction.rda = rda(Consumption, helmert[,3], helmert[,1:2])
anova(interaction.rda, step=10000, perm.max=10000)

# Permutation test for rda under reduced model
# Model: rda(X = Consumption, Y = helmert[, 3], Z = helmert[, 1:2])
#          Df     Var    F N.Perm Pr(>F)
# Model     1   83.52 0.63   9999 0.4404
# Residual  8 1060.61
```

F = 0.6300, p = 0.4404

2.3. Test the main factor Sex by RDA. Lard and the interaction form the matrix of covariables.

```
Sex.rda = rda(Consumption, helmert[,1], helmert[,2:3])
anova(Sex.rda, step=10000, perm.max=10000)
```

```
# Model: rda(X = Consumption, Y = helmert[, 1], Z = helmert[, 2:3])
#          Df    Var      F N.Perm Pr(>F)
# Model     1  343.7 2.5925   9999 0.1435
# Residual  8 1060.6
```

F = 2.5925, p = 0.1435

2.4. Test the main factor Lard by RDA. Sex and the interaction form the matrix of covariables.

```
Lard.rda = rda(Consumption, helmert[,2], helmert[,c(1,3)])
anova(Lard.rda, step=10000, perm.max=10000)
```

```
# Model: rda(X = Consumption, Y = helmert[, 2], Z = helmert[, c(1, 3)])
#          Df    Var      F N.Perm Pr(>F)
# Model     1 5564.0 41.968    999  0.001 ***
# Residual  8 1060.6
```

F = 41.968, p = 0.0016

2.5. Compute the R-square corresponding to the amount of variation of Consumption explained by each of the main factors

```
Sex.rda          # Summary of the RDA results for factor Sex
```

```
# Call: rda(X = Consumption, Y = helmert[, 1], Z = helmert[, 2:3])
#                 Inertia Proportion Rank
# Total         7.052e+03  1.000e+00
# Conditional   5.648e+03  8.009e-01    2
# Constrained   3.437e+02  4.874e-02    1  <= R2 = 0.04874
# Unconstrained 1.061e+03  1.504e-01    1
# Inertia is variance
```

```
Lard.rda rda       # Summary of the RDA results for factor Lard
```

```
# Call: rda(X = Consumption, Y = helmert[, 2], Z = helmert[, c(1, 3)])
#                 Inertia Proportion Rank
# Total         7.052e+03  1.000e+00
# Conditional   4.272e+02  6.058e-02    2
# Constrained   5.564e+03  7.890e-01    1  <= R2 = 0.7890
# Unconstrained 1.061e+03  1.504e-01    1
# Inertia is variance
```

3 - Regression using quantitative and ‘factor’ explanatory variables

Regression analysis (R function l*m()*) can use combinations of quantitative and qualitative (‘factor’) explanatory variables. The analysis can contain several quantitative and several ‘factor’ explanatory variables. If an interaction between factors is included in the analysis, make sure it is represented by products of Helmert contrasts.

The analysis of covariance (ANCOVA) is a mixed analysis of this type, involving quantitative and qualitative (‘factor’) explanatory variables. In R, it can be computed by *lm()*, which accepts combinations of quantitative and ‘factor’ variables. In ANCOVA, one is particularly interested in determining the interplay between the quantitative explanatory variable(s) and the factor(s). By testing the difference between nested models, one can check if the response can be modelled

• by a single model of the quantitative variable(s),

• by models that differ in their intercepts, depending on the level of the factor(s), but have the same slope,

• by models that differ in their slopes, depending on the level of the factor(s), but have the same intercept or centroid,

• or by entirely different linear models.

The analysis of covariance will not be developed further in this Workshop. Please consult advanced textbooks on ANCOVA.

4 - Generalized linear model

The *generalized linear model* (GLM; Nedler and Wedderbum 1972[*]) is a flexible generalization of ordinary least squares regression. The generalization is obtained by allowing the linear model to be related to the response variable via a link function and by allowing the magnitude of the variance of each measurement to be a function of its predicted value.

R function: *glm()*.

The general linear model, logistic regression, and Poisson regression are special cases of the generalized linear model.

[*] Nelder, J. and R. Wedderburn. 1972. Generalized linear models. J*ournal of the Royal Statistical Society, Series A (General)* 135: 370–384.

Analysis of variance by canonical redundancy analysis (RDA)

Examples of complex analyses of variance for species data

Hooper, E., R. Condit and P. Legendre. 2002. Responses of 20 native tree species to reforestation strategies for abandoned farmland in Panama. *Ecological Applications* 12: 1626-1641. [PDF available on the Web page http://www.bio.umontreal.ca/legendre/reprints/]

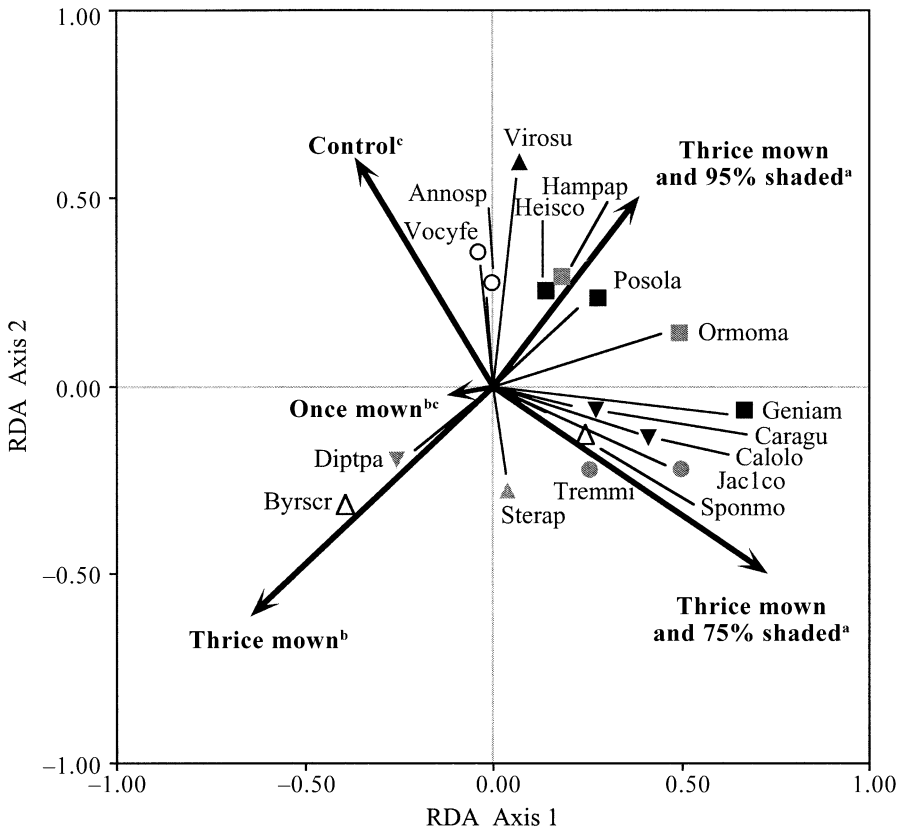

FIG. 6. Ordination biplot illustrating the significant ($P < 0.001$) effect of *Saccharum* treatment on the relative height growth (RHT) of 16 tree species in the wet season, between May and July 1997. Same-letter superscripts indicate no significant difference ($P < 0.05$) based on post hoc analysis. *Saccharum* treatment explained 19.8% of the variance between species, with the first and second axes explaining 11.2% and 4.5%, respectively. Arrows indicate treatments and lines indicate species vectors. Codes associated with each species are listed in Table 1, and the symbols match Fig. 2.

relatively high performance in the *Saccharum* control (Table 4). We conclude that the *Saccharum* did not completely limit their regeneration. A reforestation effort starting with seed and minimal pre-sowing treatment is likely to succeed with these large-seeded, shade-tolerant species.

Fires burn yearly in the dry season in these *Saccharum*-dominated grasslands, and our data show that these wildfires are also a major barrier to tree regeneration. Fire killed most species and significantly lowered the germination of all except *Trema* and *Byrsonima* (but those two cannot compete with established *Saccharum*). Resprouting from cut stems or stumps is a very common mechanism for reestablishment following disturbance (Aide et al. 1995), and we found that seedlings of several large-seeded species (*Carapa*, *Dipteryx*, *Virola*, *Ormosia*, and *Calophyllum*) could resprout following fire. Recurring fires, as a result of grass invasion following pasture abandonment, arrest natural tree regeneration in abandoned pastures at other Neotropical sites as well (Janzen 1988, Nepstad et al. 1990, Aide and Cavelier 1994).

Management suggestions

Fire is a major barrier to tree regeneration at our sites, limiting both establishment and species diversity. We therefore recommend the establishment of firebreaks, which have also been an integral part of reforestation strategies in Costa Rica (Janzen 1988) and the Amazon (Nepstad et al. 1990). The breaks must be large for effective fire protection because the flame height of *Saccharum* wildfires can reach >15 m.

Many alternatives have been suggested for forest restoration throughout the wet tropics. These range, in order of increasing cost, from simply allowing natural regeneration to proceed, to planting seeds or seedlings to assist natural regeneration, through establishing tree plantations and allowing recruitment of tree seedlings below them (Brown and Lugo 1994, Guariguata et al. 1995, Kuusipalo et al. 1995). Our goal was to find a low-cost strategy for extensive forest restoration in abandoned Panamanian farmland, and our results suggest that even with the removal of fire, natural tree regeneration will not proceed unassisted because the *Saccharum* poses a formidable barrier to the small-

APPENDIX A

Summary of repeated-measures ANOVA on tree seedling germination, abundance, survival, height, and relative height growth (RHT).

Source	Germination†		Abundance†		Survival†		Height‡		RHT‡	
	df	*F*	df	*F*	df	*F*	df	*F*	df	*F*
Between subjects										
Distance	2	0.52	2	3.39	2	1.11	2	3.52	2	4.77
Site × distance	4§									
Treatment	4	11.68**	4	13.04**	4	2.79*	4	3.22**	4	4.86**
Treatment × distance	8	0.69	8	0.45	8	0.43	8	0.22	8	0.58
Site × treatment + Site × treatment × distance	24‖									
Within subjects										
Time	2	36.31**	3	80.73**	2	1.71	2	5.68**	2	0.86
Time × distance	4	0.59	6	1.01	4	0.15	4	2.17	4	0.13
Time × site × distance	8§									
Time × treatment	8	1.84	12	2.82**	8	1.09	8	1.08	8	0.26
Time × treatment × distance	16	0.92	24	0.80	16	0.74	16	0.36	16	1.08
Time × site × treatment + Time × site × treatment × distance	48‖									

Notes: The analysis followed a repeated-measures, split-plot design. Sources of variation included distance from the forest as the main plot factor (distance), shading and mowing treatments of the *Saccharum* as the subplot factor (treatment), and their interactions. Site was included as a blocking factor.

* $P < 0.05$; ** $P < 0.01$.

† Sample size: $n = 45$ subjects (i.e., 45 unburned subplots).

‡ Sample size: $n = 371$ subjects (i.e., 371 seedlings were present over all three time periods in the 45 unburned subplots).

§ Main plot error.

‖ Subplot error.

APPENDIX B

Summary of RDA in a MANCOVA-like design on matrices of germination per species per subplot per time period (germination) or relative growth rates for height (RHT) per species per subplot per time period.

Source	Germination		RHT	
	df†	*F*	df‡	*F*
Site	2		2	
Distance	2	1.03	2	1.41
Treatment	4	2.89**	4	4.67**
Treatment × distance	8	0.70	8	1.25
Time	3	31.51**	1	12.08**
Time × distance	6	1.01	2	2.46*
Time × treatment	12	1.81**	4	2.85**
Time × treatment × distance	24	0.72	8	1.36

Notes: Sources of variation included distance from the forest (distance), shading and mowing treatments of the *Saccharum* (treatment), time, and their interactions. Site and the interaction of site with all main factors and interactions were used as covariables.

* $P < 0.05$; ** $P < 0.01$ (determined using permutation testing).

† For all factors and interactions, denominator df = 60.

‡ For all factors and interactions, denominator df = 16.