Analysis of variance and the general linear model or performing t-tests with matrix algebra

Christophe Pallier[*]

April 13, 2003

Abstract

This note is the beginning of an attempt to try and explain how to use contrasts to test hypothesis in the general linear model framework. I have only addressed the two-samples T-test for the moment... Comments are welcome to improve, simplify and/or extend this note.

Contents

[*] christophe AT pallier.org. Many thanks to J.-B. Poline for discussions about this topic. My main reference is the book *Plane answers to complex questions* by R. Christensen (Springer, 1996). Any comment to improve this note are welcome.

1 Introduction

Many statistical questions can be expressed as “how do some variables influence another variable?” Here are some examples:

1. What is the influence of parents’ income on the IQ of their children?

   Income and IQ are two continous numeric variables which can be analysed with a ’simple regression’ procedure where income is treated as the predictor and IQ as the dependent variable..

2. What are the influences of parents’ income and of their own IQ on their children’s IQ?

   This is similar to above except that there are now two continuous predictors: parents’ income and IQ. This can analysed with ’multiple regression’.

3. What is the influence of the sex (male vs. female) on IQ?

   Here, the predictor variable ’sex’ is a categorical variable which can take two values. This problem would classicaly be statistically analysed with Student T test or one way analysis of variance. However, it can be analysed with the same technique as problems 1 and 2.

The general linear model (GLM) subsumes regression (where the predictors are continous) and analysis of variance (where the predictors are categorical). That means that the same formalism can be applied to the two types of problems. Analysis of covariance models, which mixes categorical and continuous predictors, are also quite naturally expressed in this framework.

Most modern statistical packages use the general linear model approach because of its flexibility. This is also the case of the brain imaging data analysis software “SPM”.

2 Formalism

In situations relevant for the use of the GLM, there are predictor variables (continous or discrete) and one “dependent” (continous) variable. It is assumed that these variables are measured on n statistical cases. Continous variables can each be represented by a column vectors with n elements. Categorical variables can be represented by a set of as many vectors as there are categories, where each vector contains 0s and 1s; 1s indicating the cases that belong to the corresponding category (different types of coding for categorical variable are also possible but let us ignore this for the moment).

The column vectors corresponding to the predicors can be packed into a matrix X. Then, the general linear model takes the form:

$$Y = X\beta + e$$

Where:

- X is the design matrix, that is a the matrix whose column-vectors correspond to the predictor variables (several columns may be associated to one discrete predictor).

- Y is the dependent variable.

- β are parameters which represent 'weights' associated to each predictor variable. They are apriori unknown and will be estimated by the algorithm explained below.

- e is a vector of random errors; For the hypothesis tests to work, these errors are assumed to be independent and distributed normally with a variance σ^2. Like the β, these errors are also unknown, and must be estimated from the data.

This equation expresses that Y is the sum of a linear combination of the column vectors of X and of random errors. This assertion may be true or false, useful or useless. Statistics methods can only help judge whether this model is better than other models.

One algorithm used to "solve" the above equation is called the "ordinary least squares" (OLS). The methods searches for β and e that give the best fit, that is minimize the "square error": $\|e\|^2 = \|Y - X\beta\|^2$.

This is achieved by finding a $\hat{\beta}$ such that $X\hat{\beta}$ is as near as possible to Y, that is, a $\hat{\beta}$ such that $X\hat{\beta}$ is the orthogonal projection of Y onto $C(X)$, the vector space spanned by the columns of X. Therefore, if M is the orthogonal projection matrix onto $C(X)$, $\hat{\beta}$ must verify:

$$MY = X\hat{\beta}$$

This equation is most interesting:

First it expresses the relationships between the parameters $\hat{\beta}$ and the data. Understanding what the parameters mean is obviously crucial to be able to manipulate contrasts. In one-way anova situations, MY provides the means of the various groups, and the above equations gives the relationship between the $\hat{\beta}$ and the means.

Second, it is very important to realise that because M depends only on the space spanned by X columns (C(X)), two models with different design matrices X_1 and X_2 such that $C(X_1) = C(X_2)$ are actually quite similar: They explain the same amount of data ($X.\beta$), have the same error components, and each contrast in one model will have a counterpart in the other, which will lead to the same statistical conclusions.

Finally, it should be clear that if X is not full rank, there are infinitely many β which fullfill the equation. The choice of $\hat{\beta}$ is partly arbitrary. This as an important application for contrasts. Only some of them will be meaningfull. They are called *estimable* contrasts: A contrast λ is *estimable* iff $\lambda.\beta$ is invariant when β spans the solutions of $MY = X\beta$. In other words, the choice of a particular $\hat{\beta}$ solution to $MY = X\beta$, should not change thex conclusion reached by examining $\lambda.\beta$. The next section provides an example of estimable contrasts.

It is sometimes useful to have the general formula for the orthogonal projection matrix onto $C(X)$:

$$M = X(X'X)^{-}X'$$

In the special cases of the one-way anova design matrices (with one columns per category, dummy coding for that category), computations show that

$$M = BlkDiag[1/n_i * J_{n_i}]$$

where J_p is a square matrix $p \times p$ filled with 1s. That is, given the projection matrix transforms a vector Y into a vector where all values are replace by the mean of the group they belong to. For example, for the following two-samples t-test design matrix:

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

then

$$MY = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix} .Y = \begin{bmatrix} \bar{y_1} \\ \bar{y_1} \\ \bar{y_1} \\ \bar{y_2} \\ \bar{y_2} \\ \bar{y_2} \end{bmatrix}$$

where $\bar{y_i}$ is the means of observations in group i.

3 Three models for the two-samples t-test

Matrix	Parameters	Contrasts
$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\{ \begin{array}{c} \hat{\beta}_1 = \bar{y_1} \\ \hat{\beta}_2 = \bar{y_2} \end{array}$	$(1,0).\hat{\beta} = \bar{y_1}$ $(0,1).\hat{\beta} = \bar{y_2}$ $(1,-1).\hat{\beta} = \bar{y_1} - \bar{y_2}$ $(.5,.5).\hat{\beta} = \text{mean}(\bar{y_1}, \bar{y_2})$
$X = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$	$\{ \begin{array}{c} \hat{\beta}_1 + \hat{\beta}_2 = \bar{y_1} \\ \hat{\beta}_2 = \bar{y_2} \end{array}$	$(1,1).\hat{\beta} = \bar{y_1}$ $(0,1).\hat{\beta} = \bar{y_2}$ $(1,0).\hat{\beta} = \bar{y_1} - \bar{y_2}$ $(.5,1).\hat{\beta} = \text{mean}(\bar{y_1}, \bar{y_2})$
$X = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$	$\{ \begin{array}{c} \hat{\beta}_1 + \hat{\beta}_3 = \bar{y_1} \\ \hat{\beta}_2 + \hat{\beta}_3 = \bar{y_2} \end{array}$	$(1,0,1).\hat{\beta} = \bar{y_1}$ $(0,1,1).\hat{\beta} = \bar{y_2}$ $(1,-1,0).\hat{\beta} = \bar{y_1} - \bar{y_2}$ $(.5,.5,1).\hat{\beta} = \text{mean}(\bar{y_1}, \bar{y_2})$

The space spanned by these three design matrices is the same, so the projection matrix M is the same in all three cases:

$$M = \begin{bmatrix} 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \end{bmatrix}$$

Errors will be the same and the statistical tests will all lead to the same conclusions.

- The first matrix is the simplest: there is one column for each group, containing the corresponding indicator variables, with 1 for each case belonging to the group. As the parameters correspond to the means of the groups, the contrasts are straigthforward.

  Note that the constant variable (that is a column full of 1s) is absent from this model.

- In the second matrix, the second category is not represented, yet the constant is. The contrast $(1,0).\hat{\beta}$, equal to $\hat{\beta}_1$, estimates the difference between the first and the second condition.

- The third matrix has both indicator variables for each group and also include the global mean. It has the interesting property of not being full-rank (that it has its columns vectors are not linearly independent: the last is the sum of the first two, and together, these vectors span a space of dimension 2).

  This implies that there are an infinite number of $\hat{\beta}$ solutions to the $MY = X\beta$ equation (which can also be seen from the equations linking $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$ to $\bar{y_1}$ and $\bar{y_2}$: there are 3 unknowns for only two equations.)

  Another way to put it, is to notice that $X.[1,1,-1] = 0$, and therefore that $\forall c : X.\hat{\beta} = X.(\hat{\beta} + c.[1,1,-1])$. So a meaningful, estimable, contrast $\lambda = [\lambda_1, \lambda_2, \lambda_3]$ must be orthogonal to $[1,1,-1]$, that is:

$$\lambda \quad \text{is estimable iff} \quad \lambda_1 + \lambda_2 - \lambda_3 = 0$$

4 Relevance to SPM

4.1 Two-samples T-test

To compare two groups in SPM, the two-samples T-test basic model will produce a design matrix like the third in the previous section. That is, SPM creates a column for each group and add the constant term. Therefore, the following contrasts can be used:

$$\begin{aligned} &(1,0,1).\hat{\beta} = \bar{y_1} \\ &(0,1,1).\hat{\beta} = \bar{y_2} \\ &(1,-1,0).\hat{\beta} = \bar{y_1} - \bar{y_2} \\ &(.5,.5,1).\hat{\beta} = \text{mean}(\bar{y_1}, \bar{y_2}) \end{aligned}$$

As one can see, the same model can be used to test the main effect of each group (vs. 0), or even the global main effect (vs. 0). This can replace the use of simple-T tests for each group. The appendix A provide the code for generating these contrasts with the SPM99 batch system.

This approach is valid only if using a pooled error term for both group is acceptable, that is if the variances in each group do not differ widely. Note also that relative to simple t-tests, the degrees of freedom double as the variance is estimated on twice as much data (supposing the two groups are balanced).

In any case, using these contrasts to assess the main effects of group 1 and group 2, and compare them, is a good idea inasmuch as any difference that may be revealed will have to be due to the differences in the size of effects rather than to variations in within-group variances.

4.2 Modelling or not the null event

People often wonder whether they should or not model the baseline, or the 'null' event in event-related studies.

The comparison between model 2 and model 3 above may clarify the situation: even if the null event is not modeled (like in model 2), it is automatically taken into account by the fact that there is the constant term in the model. Whether or not to include it in the model is a matter of taste, but then, of course, the contrasts change according to the model chosen.

4.3 About unbalanced designs

Sometimes, the numbers of cases differ between groups, that is, the number of subject is not balanced accross groups. This is not at all a problem for a anova with a single between subject factor. The projection matrix Mtakes into account the number of cases per group. Therefore the contrasts remain the same as in the balanced design.

Unbalanced groups can only be problematic for factorial designs with at least two between-subject factors. When subgroups have different sizes, there are possibly two different main effects which may be relevant for each factor. When estimating the effect of factor 1, one can take into account or not the structure imposed by factor 2. The default solution provided (I believe but this should be checked) by SPM, does not take into account factor 2 and corresponds to the so-called 'weigthed' solution. The second, so-called 'unweigthed' solution (which is often the more meaningful one), the sizes of subgroups are ignored, and each subgroup has the same weight in computing the means.

Fortunately, when subgroups are "nearly balanced", the main effects assessed by the weigthed and unweighted solutions do not differ much, and the choice is irrelevant.

Note also that the 'difficulty' of unbalancedness does not apply to interactions.

A Code for generating contrasts with SPM99 batch system

Here are two files, `batch_contrasts.m` and `cc1.m`, that I use to automatically generate the contrasts of interest for the “two samples t-test” case using the spm99 batch system (using the command “`spm_bch('batch_contrasts')`”).

```
%%%%%%%%%%%%%%%%%%%% file batch_contrasts.m
analyses = struct( ...
      'type',          [ 2 ],...
      'index',          [ 1 ],...
      'work_dir',       [ 1 ],...
      'mfile',          [ 1 ]...
);

type = {'model','contrasts','defaults_edit','headers',...
        'means','realignment','normalisation','smooth'};

work_dir = {  '.' };

mfile = { 'cc1.m' };

%%%%%%%%%%%%%%%% file cc1.m
% Generate the contrasts for the SPM 'two way t test' design
% to compare two groups: G1 vs. G2

% trick to erase all previous contrasts ((C) Christophe ;-))
load xCon
xCon(2:end)=[]
save xCon

c1=[1 0 1]; % main effect of G1>0
c2=-c1;     % main effect of G1<0

c3=[0 1 1]; % main effect of G2>0
c4=-c3;     % main effect of G2<0

c5=[.5 .5 1]; % global main effect (G1+G2)>0
c6=-c5;     % global main effect (G1+G2)<0

c7=[1 -1 0]; % G1 > G2
c8=[-1 1 0]; % G2 > G1

contrasts(1) = struct( ...
 'names', {{'G1>0','G1<0','G2>0','G2<0', 'G1+G2>0', 'G1+G2<0',...
   'G1>G2', 'G2>G1'}},...
 'types', {{'T','T','T','T','T','T','T','T'}}, ...
 'values', {{c1,c2,c3,c4,c5,c6,c7,c8}}...
);
```