The General Linear Model

Thomas D. Wickens
Department of Psychology
University of California, Berkeley

Mathematics in Brain Imaging
Graduate Summer School Program
Institute for Pure and Applied Mathematics
University of California, Los Angeles
July, 2004

When I begin to plan this talk, then looked over the other offerings at this conference, I realized that in a sense I am here to present a null hypothesis. The other talks will describe methods that go far beyond the general linear model. Like many null hypotheses, the general linear model is simple, elegant, even beautiful. Unlike many null hypotheses, it is also both very useful and widely used. It is a component of, or the origin of, the greater part of the work to come. Thus, it is essential to understand it clearly from the outset.

1 Statistical models

In choosing a statistical model, there is always a tension between simplicity and completeness. Simple models tend to be easier to understand, computationally more tractable, but they are frequently at odds with data. On the other hand, complicated models tend to fit the data better and to capture richer conceptual pictures, but they can be computationally awkward or intractable. When they are too complicated, they are hard to replicate. For practical reasons, the classical procedures such as the general linear model have had to be computationally manageable. However, computational power is no longer the limit it once was. Many things that were impossible before—iterative algorithms, Monte Carlo methods, repeated tests, the whole range of Bayesian approaches—now can be routine (or nearly so). Moreover, imaging data have properties that make them differ substantially from the traditional data structures. Nevertheless, the classical method are still best for many questions, and as they are the basis for the newer approaches, and we must keep them in mind.

2 Data for General Linear Model

The data for the models I will discuss have the general structure of observations by variables:

		Variables							
		Conditions				Responses			
		$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_p$	$\mathbf{y}_1$	$\mathbf{y}_2$	...	$\mathbf{y}_q$
Observations	1	–	–		–	–	–		–
	2	–	–		–	–	–		–
	⋮								
	n	–	–		–	–	–		–

The rows—the observations—may refer to observations at different times, to different subjects, or to different brain locations, depending on the particular analysis. The columns are divided into two sets. One set, denoted by $\mathbf{x}_j$ are variables describing the condition in which the observation was made. They could be dummy variables describing groups, subjects, or the particular response condition, or they could be other covariates. The second set, denoted by $\mathbf{y}_k$ are the observed measurements, for example activations measured different locations. Much of the analysis I will describe focuses on a single observation $\mathbf{y}$, although in imaging

t	Stim	x_0	x_1	x_2	y_1	y_2	y_3	y_4
1	0	1	0	0	9.45	13.25	11.23	16.48
2	0	1	0	0	9.86	10.26	11.13	13.62
3	0	1	0	0	10.17	13.90	11.74	15.13
4	1	1	1	0	12.97	11.76	10.97	16.63
5	1	1	1	0	11.31	13.83	10.65	16.42
6	1	1	1	0	12.70	10.96	10.12	17.85
7	0	1	0	0	11.38	12.95	11.15	13.65
8	0	1	0	0	10.29	12.12	11.56	15.96
9	0	1	0	0	11.82	10.29	12.73	14.27
10	2	1	0	1	10.27	12.45	14.15	19.39
11	2	1	0	1	11.54	13.25	14.33	18.49
12	2	1	0	1	8.93	8.93	14.32	16.73
13	0	1	0	0	11.01	11.69	10.40	17.31
14	0	1	0	0	8.92	11.52	10.87	14.62
15	0	1	0	0	11.04	12.85	11.09	14.00
16	2	1	0	1	9.45	11.65	15.50	17.54
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table 1: A miniature set of illustrative data. Two stimuli are presented at different times, and responses are recorded from four locations.

studies a large number of responses are typically recorded. The basic goal of the analysis is to find a way to describe the $\mathbf{y}_k$ as functions of the $\mathbf{x}$'s.

Table 1 shows a simple (synthetic) example of the type of data involved. A response is measured in four locations (y_1 to y_4) at a sequence of time points, of which 16 are shown. One of two stimuli are presented during some of these periods—stimulus 1 during intervals 4, 5, and 6, stimulus 2 during intervals 10, 11, and 12, then again starting in interval 16. From the stimulus events, three dummy variables are created: x_0 always equals 1 and corresponds to background activity, x_1 is zero except in the presence of stimulus 1, and x_2 is zero except in the presence of stimulus 2. Properly these variables would be convolved with a hemodynamic response function, but for simplicity I have not done that here.

The columns of the data table are variables, and algebraically they are column vectors. It is useful, both for understanding the methods and for interpreting them, to think of these vectors geometrically. Each variable, then, is represented by a spatial vector. The length of the vector characterizes the variability of the variable and the angle between two vectors describes the association between the variables. Specifically, the squared length of the vector is the sum of squares associated with the variable,

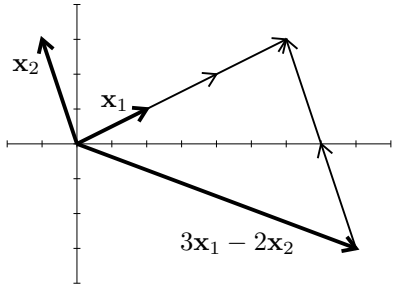
$$|\mathbf{y}|^2 = SS_{\mathrm{y}},$$

and the cosine of the angle equals the correlation between the variables:

$$r_{jk} = \mathrm{corr}(\mathbf{x}_j, \mathbf{x}_k) = \cos \angle(\mathbf{x}_j, \mathbf{x}_k).$$

In particular, two variables that are simple multiples of each other, hence that convey equivalent information, have vectors that point in the same (or exactly the opposite) direction, and so are collinear. Two variables that are linearly unrelated to each other have vectors lying at right angles and are orthogonal. It takes a little practice to interpret these relationships when more than three variables are involved, but I have found that with some practice I have obtained useful information about five- or six-dimensional relationships by projecting them back into three dimensional space.

A linear combination of two variables is represented by their vector sum, which lies in the space spanned by its components. For example, the sum of $3\mathbf{x}_1 - 2\mathbf{x}_2$ is represented by the vector sum[1]



The important thing about this representation is that the vector operations correctly represent distributional properties of the new variable, such as its variability and its correlation with other variables.

Note that these vector representation makes use of the fact that the vectors are expressed in ordinary Euclidean space, a point that is important when we consider the distribution of errors.

3 The General Linear Model

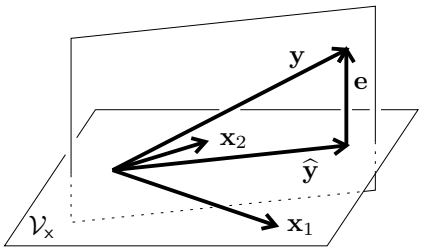
The general linear model is a procedure by which a single target variable $\mathbf{y}$ is represented by a combination of variables $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_p$. The combination in question is linear:

$$\widehat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_p x_{ip},$$

or in matrix notation,

$$\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}.$$

Geometrically, this representation is obtained by projecting $\mathbf{y}$ into the into the space $\mathcal{V}_{\mathbf{x}}$ of linear combinations of $\mathbf{x}_1$, $\mathbf{x}_2$, ..., $\mathbf{x}_p$:



Algebraically, the coefficient vector $\mathbf{b}$ for this projection is determined by multiplying $\mathbf{y}$ by the matrix $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}$:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

This equation will be seen frequently. As the diagram above shows, it decomposes $\mathbf{y}$ into two orthogonal vectors, the prediction $\widehat{\mathbf{y}}$ and the error $\mathbf{e}$. These vectors are chosen so that $|\widehat{\mathbf{y}}|$ is maximal, and consequently $|\mathbf{e}|$ is minimal.

For the example in Table 1, applying this formula to each of the four $\mathbf{y}_k$ fits the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}$ to each set of activations, where $\mathbf{X}$ is the 3×16 matrix given in the three $\mathbf{x}$ columns of the table and $\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}$ (of course, ordinarily the data would include more than the 16 observations shown in Table 1). The result is the set of four estimates

[1] Diagrams in this talk are taken from my book *The geometry of multivariate statistics* (Erbaum, 1995).

	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$	$\mathbf{y}_4$
b_0	10.44	12.09	11.32	15.00
b_1	1.89	0.09	−0.74	1.96
b_2	−0.39	−0.52	3.25	3.04

Examining these estimates shows that $\mathbf{y}_1$ has a positive relationship to signal one, $\mathbf{y}_2$ has little relationship to either signal, $\mathbf{y}_3$ has a positive relationship to signal 2, and $\mathbf{y}_4$ has a positive relationship to both signals. The primary use of the general linear model with imaging data is to detect relationships such as these.

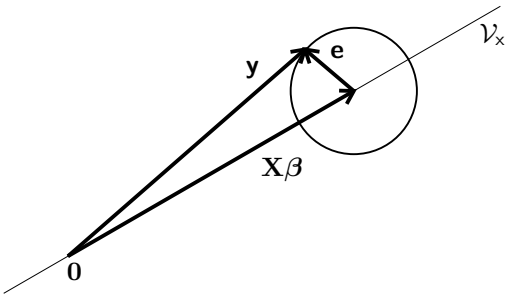
The model for $\widehat{\mathbf{y}}$ given above is not a statistical model for the scores. The complete general linear model includes an error component:

$$\mathsf{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{p2} + \mathsf{e}_i,$$

or in matrix terms,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}.$$

I have indicated the random variables typographically here with a sans-serif font. The letter β is used instead of b to indicate that it refers to the "true," if unobserved, values of the parameters. The errors e_i in this model are independent and identically distributed, having a Gaussian (i.e., normal) distribution with mean zero and variance σ^2. This assumption means that the actual observation $\mathbf{y}$ has a spherically symmetrical distribution about the true model value $\mathbf{X}\boldsymbol{\beta}$:



The isotropy of the error keeps $\mathbf{X}\boldsymbol{\beta}$ uncorrelated with the error $\mathbf{e}$ and allows $\mathbf{b}$ to be estimated by projection. It is an important consequence of the assumption of the Gaussian distribution.

Under the general linear model, the estimate $\mathbf{b}$ of $\boldsymbol{\beta}$ has two somewhat different interpretations. One is as a least-squares fit of $\widehat{\mathbf{y}}$ to $\mathbf{y}$. The estimate $\mathbf{b}$ gives the linear combination $\widehat{\mathbf{y}}$ that minimizes the length of the error $\mathbf{e}$. Specifically, it gives the smallest possible value to sum of squared differences between $\mathbf{y}$ and $\widehat{\mathbf{y}}$:

$$|\mathbf{y} - \widehat{\mathbf{y}}|^2 = \sum (y_i - \widehat{y}_i)^2 = \sum (y_i - \mathbf{x}_i'\mathbf{b})^2.$$

The second interpretation is that of maximum likelihood and is based on the probability distribution of y_i. Because this interpretation is so important to later developments, I will briefly describe it.

1. Under the general linear model, the probability of observation y_i depends on the parameters $\boldsymbol{\beta}$ and σ^2, and can be written $P(y_i; \boldsymbol{\beta}, \sigma^2)$.

2. Because the observations are independent, the probability of the complete set of data $\mathbf{y}$ is the product of the probabilities of its individual observations:

$$P(\mathbf{y}; \boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{N} P(y_i; \boldsymbol{\beta}, \sigma^2).$$

3. In an estimation problem, a set of data has been obtained, hence it can no longer be treated as a random quantity. Instead, treat the probability above as a function of the parameters and call in the likelihood. For several reasons it is more convenient to work with the logarithm of the likelihood than the likelihood itself, so define:
$$\mathcal{L}(\boldsymbol{\beta}, \sigma^2) = \log P(\mathbf{y}; \boldsymbol{\beta}, \sigma^2).$$

4. To estimate the parameters $\boldsymbol{\beta}$ and σ^2, find the values $\mathbf{b}$ and s^2 that maximize $\mathcal{L}(\mathbf{b}, s^2)$. These are the maximum-likelihood estimates.

The least-squares and maximum-likelihood interpretations of the estimates are equivalent under the assumptions that observations are independent, identically distributed, and Gaussian. When they are not, the two procedures will give different estimates.

4 Goodness of fit tests

The variability parameter σ^2 in general linear model is estimated by (to give one among several formulas)

$$\widehat{\sigma}^2 = \frac{\mathbf{y}'(\mathbf{y} - \mathbf{X}\mathbf{b})}{n - p}.$$

From it, the covariance matrix of the estimates is obtained:

$$\mathbf{S}_\mathbf{b} = \widehat{\sigma}^2 (\mathbf{X}'\mathbf{X})^{-1}.$$

The goodness of fit of the model can be measured by the proportion of the variability that the model can explain, that is by the relative length of the original and estimated vector:

$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{|\widehat{\mathbf{y}}|^2}{|\mathbf{y}|^2}.$$

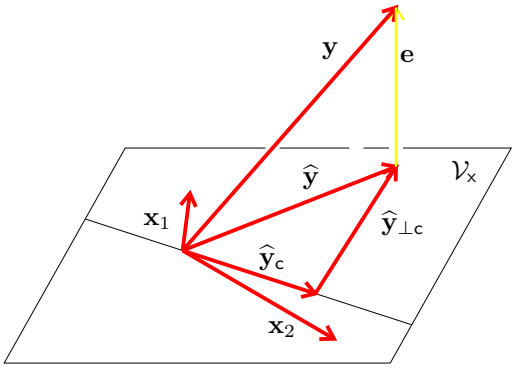
This quantity is known as the correlation ratio, and, in the analysis of variance context often denoted by η^2.

A test of the null hypothesis that $\boldsymbol{\beta} = 0$ is obtained by comparing the per-dimension squared lengths of the two components of $\mathbf{y}$ under the general linear model. With p explanatory variables and n observations,

$$F = \frac{|\widehat{\mathbf{y}}|^2 / p}{|\widehat{\mathbf{e}}|^2 / (n - p)} = \frac{MS_\text{model}}{MS_\text{error}}.$$

In this form, the test is of limited value—the assertion that $\boldsymbol{\beta} = 0$ reduces the model to simply $y_i = \mathsf{e}_i$, which is not very informative. More useful are tests of hypotheses about particular values of the parameters.

Most tests of the parameters of the general linear model are made by imposing linear constraints on $\boldsymbol{\beta}$. These hypotheses have the form $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, for some constraint matrix $\mathbf{C}$. Using the example in Table 1, a test that $\mathbf{y}$ is unaffected by the second stimulus is given by the matrix $\mathbf{C} = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$, which holds the coefficient β_2 of $\mathbf{x}_2$ to zero. A test of whether the responses to the two stimuli are the same is given by the matrix $\mathbf{C} = \begin{bmatrix} 0 & 1 & -1 \end{bmatrix}$, which causes β_1 and β_2 to be equal. Compound hypotheses are possible: a test of the model that sets both signal parameters to zero is obtained from $\mathbf{C} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$. Geometrically any such linear hypothesis is tested by splitting $\widehat{\mathbf{y}}$ into two orthogonal parts, a vector $\widehat{\mathbf{y}}_\mathbf{c}$ that is consistent with the hypothesis and a vector $\widehat{\mathbf{y}}_{\perp\mathbf{c}}$ that violates it:



If $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$, then the component $\widehat{\mathbf{y}}_{\perp\mathsf{c}}$ is entirely due to error, and its size should be comparable to that of $\mathbf{e}$. If it is false, then the per-dimension length of $\widehat{\mathbf{y}}_{\perp\mathsf{c}}$ is much greater than that of $\mathbf{e}$.

There are three ways to get a test statistic for these hypotheses, all of which are comparable under the general linear model with a Gaussian error distribution.

1. By comparing the mean squared lengths of the vectors $\widehat{\mathbf{y}}_{\perp\mathsf{c}}$ and $\mathbf{e}$ using an F statistic as described for the overall test. If there are r linear restrictions (rows of $\mathbf{C}$), then this statistic is

$$F = \frac{|\widehat{\mathbf{y}}_{\perp\mathsf{c}}|^2/r}{|\widehat{\mathbf{e}}|^2/(N-p)} = \frac{MS_{\text{restriction}}}{MS_{\text{error}}}.$$

2. When the hypothesis is unidimensional—it contains only one restriction—then it can be tested by calculating the observed size of its violation and the standard error of that quantity:

$$V = \mathbf{C}\mathbf{b} \qquad \text{and} \qquad s_\mathsf{v}^2 = \mathbf{C}\mathbf{S}_b\mathbf{C}'.$$

   The hypothesis is tested with a t statistic, $t = V/s_\mathsf{v}$, which is the square root of the F above.

3. By fitting two models, one that includes the tested parameters and one that does not. If the hypothesis holds, then the error of these two models will be about the same size, although the unrestricted model, for which the hypothesis is not imposed, with have a slightly smaller error. The amount by which imposing the restriction increases the error is compared to the overall error by the statistic

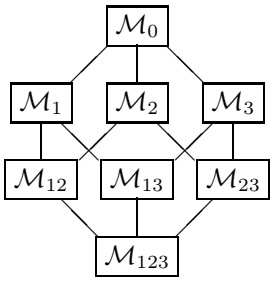
$$F = \frac{(|\mathbf{e}_{\text{restricted}}|^2 - |\mathbf{e}_{\text{unrestricted}}|^2)/r}{|\mathbf{e}_{\text{unrestricted}}|^2/(N-p)}.$$

5 Model selection and comparison

The issue of selection of a representation or model for data is a very important one, and I will digress briefly from the general linear model to discuss it in general. I find it very useful to think of the problem graphically by constructing a lattice of possible models. For example, with three predictor variables, $\mathbf{x}_1$, $\mathbf{x}_2$ and $\mathbf{x}_3$ we can constrict eight models:

$$\begin{array}{ll}
\mathcal{M}_0\colon & y = b_0 \\
\mathcal{M}_1\colon & y = b_0 + b_1x_1 \\
\mathcal{M}_2\colon & y = b_0 + \quad b_2x_2 \\
\mathcal{M}_3\colon & y = b_0 + \quad\quad b_3x_3 \\
\mathcal{M}_{12}\colon & y = b_0 + b_1x_1 + b_2x_2 \\
\mathcal{M}_{13}\colon & y = b_0 + b_1x_1 + \quad b_3x_3 \\
\mathcal{M}_{23}\colon & y = b_0 + \quad b_2x_2 + b_3x_3 \\
\mathcal{M}_{123}\colon & y = b_0 + b_1x_1 + b_2x_2 + b_3x_3
\end{array}$$

These representations fit into a hierarchical lattice with the simplest model $\mathcal{M}_0$ at the top and the most complex one $\mathcal{M}_{123}$ at the bottom:



Models that are restrictions of each other are connected in this lattice by a line. The lattice could be expanded to include many other models, such as those that equate the coefficient of some of the explanatory variables (e.g., $y = b + 0 + b_{12}x_1 + b_{12}x_2$) or those that include powers of their value (e.g., $y = b_0 + b_{11}x_1 + b_{12}x_1^2$).

The models toward the bottom of this lattice include more parameters than those toward the top, and consequently they typically fit better. Where the models are connected by a line (or sequence of ascending or descending lines) they have a hierarchical relationship to each other and the upper one is strictly included in the lower. For example, $\mathcal{M}_1 \subset \mathcal{M}_{13} \subset \mathcal{M}_{123}$. The lower models in this sequence will fit at least as well as the upper models. Where the models are not connected—the relationship is nonhierarchical—the lower model will not necessarily fit better. For example although $\mathcal{M}_{12}$has three parameters and $\mathcal{M}_3$ has two, it will not necessarily fit a given set of data better. Whether the relationship between two models is hierarchal or not makes a difference in how they can be compared.

For a comparison of hierarchical models, there are usually good tests available. A test statistic (such as the F statistic above) can be formed by looking at the difference between the fit of the unrestricted model (that lower in the lattice) and the restricted model above it

$$\text{Test statistic} = \text{Fit}(\mathcal{M}_{\text{unrestricted}}) - \text{Fit}(\mathcal{M}_{\text{restricted}})$$

The magnitude of this statistic is evaluated against a distribution determined by the number of restrictions that separate the models. When the log likelihood can be calculated, the difference can be referred to the chi-square distribution: for a test of r restrictions,

$$2[\mathcal{L}(\mathcal{M}_{\text{unrestricted}}) - \mathcal{L}(M_{\text{restricted}})] \sim \chi^2_r.$$

Comparisons of nonhierarchical models are more problematic. There is no assurance that a model with more parameters will fit better than one with fewer parameters, and the distribution of the difference in fit is not generally available. It is reasonable to construct a measure of the successfulness of a model that combines a measure of fit with a penalty for the number of parameters used:

$$\text{Successfulness}(\mathcal{M}) = \text{Fit}(\mathcal{M}) - f[\text{Parameters}(\mathcal{M})].$$

In a comparison of several models, the most successful one can be chosen. The difficulty here is to determine the appropriate penalty function. Many of these have been suggested. In general they depend on the likelihood to measure fit and use a penalty function $f(n, p)$ that depends on the number of observations and the number of parameters. The two most familiar of these are the Akaike information criterion,

$$\text{AIC} = 2\mathcal{L}(\mathcal{M}) - 2p$$

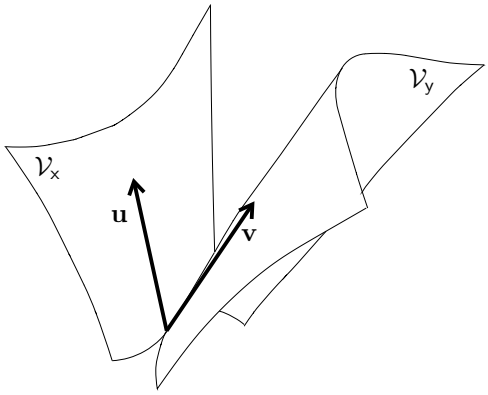
and the Bayesian information criterion,

$$\text{BIC} = 2\mathcal{L}(\mathcal{M}) - \log(n)p.$$

Because the penalties differ, these measures can order a set of models differently. They also have no sampling theory attached to them. Thus, in my opinion, while these measures give useful and necessary information for model selection, they must be treated more heuristically than the tests that compare hierarchical model.

6 Multivariate generalization

An important property of the general linear model is that it generalizes to multivariate observations readily, simply by replacing the univariate Gaussian distribution of error with a multivariate distribution. The problem comparable to that of the general linear model is to relate two sets of variables, $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p\}$ and $\{\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_q\}$. This is done by constructing new variables that are linear combinations of the variables in the two sets, $\mathbf{u} = \mathbf{Xb}$ from the space $\mathcal{V}_\mathsf{x}$ generated by the $\mathbf{x}$ vectors and $\mathbf{v} = \mathbf{Yc}$ from the space $\mathcal{V}_\mathsf{y}$ generated by the $\mathbf{y}$ vectors:[2]



The coefficients $\mathbf{b}$ and $\mathbf{c}$ are chosen so that the correlation between $\mathbf{u}$ and $\mathbf{v}$ is as large as possible, that is, so that the angle between them is minimal.

Because the spaces $\mathcal{V}_\mathsf{x}$ and $\mathcal{V}_\mathsf{y}$ are both multivariate, additional vectors can be chosen. Variables within each set are orthogonal to those already chosen, and between sets they are chosen to maximize their correlation. The result is a series of linked spaces of successively greater dimension:

$$\begin{aligned}
\mathbf{u}_1 &\longleftrightarrow \mathbf{v}_1, \\
\{\mathbf{u}_1, \mathbf{u}_2\} &\longleftrightarrow \{\mathbf{v}_1, \mathbf{v}_2\}, \\
\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3\} &\longleftrightarrow \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}, \\
\{\mathbf{u}_1, \mathbf{u}_2, \mathbf{u}_3, \mathbf{u}_4\} &\longleftrightarrow \{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4\}, \text{ etc.}
\end{aligned}$$

This procedure has various names, depending on the interpretation of the variables involved (multivariate analysis of variance, multivariate multiple regression, canonical correlation, etc.)

7 The importance of the General Linear Model

As a summary of the general linear model, note the features that make is attractive.

1. It is comprehensible, both easy to describe and to interpret. In particular the linear model $\mathbf{y} = \mathbf{Xb}$ is both simple and very powerful, and the error distribution is cleanly separated from the effects.

2. The random structure is simple. The Gaussian distribution is completely described by its first two moments, μ and σ^2 (or, in the multivariate case, by the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$). These moments are stochastically independent, making them easier to estimate. The geometric space induced by the error is Euclidean and isotropic.

[2]The spaces $\mathcal{V}_\mathsf{x}$ and $\mathcal{V}_\mathsf{y}$ in this diagram are actually linear (flat) spaces; the diagram is my attempt to project a four-dimensional configuration onto a two-dimensional page and show the relationship of $\mathbf{u}$ and $\mathbf{v}$.

3. The model is computationally tractable. Closed-form estimates of the parameters are available, and the mean and variance can be estimated separately. The estimated linear parameters are obtained by projection, $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, and are linear combinations of the original data $\mathbf{y}$. The power of linear algebra can be harnessed for this and other calculations. The computations can even be done by hand!

4. The different interpretation of the estimates (least squares and maximum likelihood) and the different forms of test yield identical results.

5. The model generalizes readily from univariate to multivariate form. Both the conditional and marginal distributions of the multivariate Gaussian distribution are also Gaussian, so tests on these distributions have the same form as those on the full multivariate model.

When the assumptions of the general linear model are relaxed, some of these characteristics go with it. Nevertheless, in some situations the model is too restrictive. I will discuss briefly at some of its extensions, particularly those that relax the assumptions that

- All observations have the same weight
- The error distribution is Gaussian
- The observations are independent

8 Unequal variability

In some sets of data, the variance of the y_i differ:

$$\text{var}(\mathsf{y}_i) = \sigma_i^2 \neq \text{var}(\mathsf{y}_k) = \sigma_k^2, \text{ for some } i \neq k.$$

These variabilities, obviously, cannot be measured on a single observation, but they may be available when the scores derive from a source that includes measurement error or when the variances are theoretically motivated.

Differences in the variance make the observations unequally informative. This inequality is accommodated in the analysis by weighting the observations inversely by their variance. For example, the weighted mean is calculated as

$$\overline{y} = \frac{\sum w_i y_i}{\sum w_i}, \qquad \text{where} \qquad w_i = \frac{1}{\sigma_i^2}.$$

Estimates of the parameters of a linear model $\widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}$ are obtained by weighting the least squares estimation equation:

$$\mathbf{b} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{W}\mathbf{y})$$

where the weight matrix $\mathbf{W}$ contains the inverse of the variances:

$$\mathbf{W} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}^{-1}.$$

9 Non-Gaussian distributions and the Generalized Linear Model

Some data generation processes do not yield observations with Gaussian distributions. Positing a non-Gaussian distribution has several consequences, which are readily seen by considering binary observations distributed according to a Bernoulli distribution.

1. The mean and the variance are related. For example, with a Bernoulli random variable y for which $P(\mathsf{y}{=}1) = \pi$ and $P(\mathsf{y}{=}0 = 1{-}\pi$, the mean and variance are given by

$$\mathsf{E}(\mathsf{y}) = \pi \qquad \text{and} \qquad \text{var}(\mathsf{y}) = \pi(1{-}\pi).$$

   Any model that does not predict a constant mean, therefore also predicts a changing variance. With distributions other than the Bernoulli (for example, Poisson counts), the distributions may be nearly Gaussian when the mean is large, but the dependence of the variance on the mean remains.

2. A linear model may not be an appropriate way to model the data. For example, with Bernoulli data, a linear model $\widehat{\boldsymbol{\pi}} = \mathbf{Xb}$ can give values of $\widehat{\pi}$ that lie outside the interval $[0, 1]$ so cannot be probabilities.

3. The changing variance makes least squares estimates (i.e., simple projection) inappropriate as an estimation technique.

A procedure known as the generalized linear model extends the general linear model to accommodate these properties. This generalization has three characteristics.

1. The distribution of the observations is drawn from a larger set of possibilities known as the exponential family. The density of these distributions has the form

$$f(y) = \exp\left[\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi)\right],$$

   where θ is a parameter that determines the location of the distribution, ϕ is a parameter that scales the variability, and $a(\cdot)$, $b(\cdot)$ and $c(\cdot, \cdot)$ are functions. This family includes many (but not all) useful distributions:

   - Among continuous distributions, the Gaussian, log-normal, inverse-Gaussian, exponential, gamma (chi-square), Pareto, Gompers, Weibull, etc.
   - Among discrete distributions, the Bernoulli, binomial, Poisson, geometric, negative binomial, etc.

2. The variance of observations is naturally linked to mean. For example, with Poisson distributed data, $\text{var}(\mathsf{y}) = \mathsf{E}(\mathsf{y})$, and with Gamma distributed data, $\text{var}(\mathsf{y}) = [\mathsf{E}(\mathsf{y})]^2$. The quantity $a(\phi)$ (usually just ϕ) scales the variablity up or down without changing this relationship, thus allowing the model to accommodate overdispersed or (less often) underdispersed observations.

3. The mean μ is related to a linear model via a location parameter η and a nonlinear link function:

$$g(\mu) = \eta = \mathbf{x}'\boldsymbol{\beta}.$$

   Although any function can be used for the link, the model is simplest when the location parameter of the link function is the same as the location parameter θ of the distribution. The link function then has a canonical form that is determined by the distribution itself. For example, with Bernoulli data, the canonical link function is logistic (making them model that of logistic regression), and for Gamma data, it is the reciprocal.

The generalized linear model does not have closed-form estimation equations. However, there are many iterative algorithms by which it can be fitted. For example, maximum likelihood fits can be obtained by repeatedly applying the weighted least squares equation given above after generating the variances from the model. Software to make these estimated is readily available and appears in the major statistical packages.

The fit of the generalized linear model is measured by a quantity known as the deviance, which is just twice the negative log likelihood:

$$\text{deviance} = -2 \log \mathcal{L}.$$

When the model allows for overdispersion or underdispersion by estimating the parameter ϕ, the deviance cannot be interpreted as measuring goodness of fit—there is no independent estimate of the error as there is for the general linear model. However, tests for hierarchical models are expressed by difference in deviance

$$\text{deviance}(\mathcal{M}_{\text{restricted}}) - \text{deviance}(\mathcal{M}_{\text{unrestricted}}) \sim \chi^2_r.$$

Although it shares some features with the general linear model, the generalized linear model cannot be interpreted quite as easily. Unless a Gaussian distribution is specified, the space of observations of $\mathbf{y}$ is nonisotropic and non-Euclidean. Intuitions based on this representation must be used with caution. The non-Euclidean error space also makes tests based on variability and tests based on model comparison give different results. A given hypothesis can often be tested in three or more ways with slightly different results (likelihood-ration tests, scores test, Wald test, etc.)

10 Dependent observations

In both the general linear model and the generalized linear model, the observations $y_1, \ldots, y_n$ are assumed to be independent. However, there are many causes for dependencies among the y_i, many of which arise with image data. A prime cause of dependence is in observations that take place over time, where the value at one time is influenced by the values at earlier times. The most common representation of this association is with an autoregressive model in which an observation depends linearly on the value at one or more earlier trial:

$$\begin{aligned}
&\text{AR}(1)\colon y_t = \mathbf{x}_t'\mathbf{b} + a_1 y_{t-1},\\
&\text{AR}(2)\colon y_t = \mathbf{x}_t'\mathbf{b} + a_1 y_{t-1} + a_2 y_{t-2},\\
&\text{AR}(3)\colon y_t = \mathbf{x}_t'\mathbf{b} + a_1 y_{t-1} + a_2 y_{t-2} + a_3 y_{t-3}, \text{ etc.}
\end{aligned}$$

The autoregressive models induce constant correlations between observations a fixed number of time intervals apart.

Another source of dependence is the presence of clustered observations. Common instances of clusters arise when several observations are made from each subject, when several observations are collected during each of several sessions, or when subjects can be groups, say into families or litters. Each type of dependence leads to a characteristic covariance structure within the cluster. Typically, the clusters are assumed to be independent. More complex sources of dependence arise when several sources of dependence are present. For example, observations of activation may be made in several related (or unrelated) brain regions at the same time (as if the 16 observations of four activations in Figure 1 were unstrung into a single column of 64 single observations. The spatial structure will induce one correlation structure, the time series another.

Consider the clustered observations in more detail. The presence of clustering give a hierarchical structure to the data. Often this structure can be represented by two (or more) sampling stages. A simple version of this hierarchical, random-effects, structure extends the general linear model by treating the parameter vectors as sampled

- Stage One: The scores within each cluster are determine by a linear model with independent errors and a cluster-specific parameter vector

$$\mathbf{y}_1 = \begin{bmatrix} y_{11} \\ \vdots \\ y_{1c} \end{bmatrix} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{e}_1, \qquad \mathbf{y}_2 = \begin{bmatrix} y_{21} \\ \vdots \\ y_{2c} \end{bmatrix} = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{e}_2, \qquad \text{etc.},$$

  where $\mathbf{e}_j \sim \mathcal{N}(0, \boldsymbol{\Sigma}_\mathbf{e})$, and $\boldsymbol{\Sigma}_\mathbf{e}$ is diagonal.
- Stage Two: The model parameters are sampled from a Gaussian distribution: $\beta_j \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$.

The model, therefore, depends on the mean $\boldsymbol{\mu}_\beta$ of the distribution of $\boldsymbol{\beta}$ and on two covariance matrices, $\boldsymbol{\Sigma}_\mathbf{e}$ and $\boldsymbol{\Sigma}_\beta$.

There are a variety of interpretations and ways that this model can be fitted, including as a random-effects model, a hierarchical linear model, an empirical Bayes model, or through the use of what are known as generalized estimating equations. These approaches differ somewhat in their simplicity or applicability from case to case, although they are often equivalent and usually give comparable results.

11 Simultaneous testing

Finally, I want to consider the problems that arise when multiple statistical tests are made, an issue of considerable importance to image data.[3] I will talk only about the conventional solutions.

In the standard hypothesis testing framework, a null hypothesis H_0 that no relationship is present is contrasted with its negation H_a that some unknown relationship is present. The properties of any scheme for deciding between these alternatives is measured by the two conditional error rates

- Type I: $\alpha = P(\text{decide } H_0 \text{ false} | H_0 \text{ true})$,
- Type II: $\beta = P(\text{decide } H_0 \text{ false} | H_a \text{ true})$.

These probabilities relate to the possible true state and the decision according to the familiar table

	Decision made	
True state	Effect present	Effect absent
Effect present	$1-\beta$	β
Effect absent	α	$1-\alpha$

There is a tradeoff between these error rates: unless more information can somehow be extracted from the data, a procedure that increases one will decrease the other. Although H_0 is a specific hypothesis, H_a is a composite of many particular alternatives. Thus, while α is usually readily calculated, β is not. For this reason, if for no other, the usual approach to hypothesis testing controls only the type I error.

When several tests are made, the aggregate, or familywise, errors rates (that is, the chance of making one or more errors) increase. Unless some form of adjustment is made, this increase affects both error types. The goal of the simultaneous-testing procedures is to manage these error rates in some rational way. Specifically, control is to be exerted over the familywise Type I rate without concomitantly allowing the individual Type II error rate to fall to the point where interesting discoveries are lost.

Many procedures have been developed to accomplish this task, each with its advantages and limitations. They can roughly be divided into three, occasionally overlapping, classes.

1. Bonferroni tests. A fundamental probability relationship, known as the Bonferroni inequality, states that for any set of k tests the relationship between the familywise and individual error rates satisfy the inequalities

   $$\alpha_{\text{familywise}} \leq 1 - (1-\alpha_{\text{individual}})^k < k\alpha_{\text{individual}}.$$

   The familywise error rate is always less than k times the individual rate, so it can be limited by setting the individual error rate to a fraction of the desired value:

   $$\alpha_{\text{individual}} = \frac{\text{Desired } \alpha_{\text{familywise}}}{k}.$$

   Because the Bonferroni inequality is an inequality, the actual familywise error is less than its nominal value.

2. The union-intersection principle. In many simultaneous testing situations, an omnibus test exist for the composite null hypothesis H_{C0} that all the individual null hypotheses hold. For example, the overall test of the hypothesis $\boldsymbol{\beta} = \mathbf{0}$ in the general linear model combines the tests that each individual $\beta_j = 0$. The compound null hypothesis, therefore, is the intersection of these simple hypotheses, while

[3] Because of time limitations, I did not cover this material during my talk at the meeting.

its alternative is the union of their negation. For example, if $\boldsymbol{\beta}$ has three components, the composite hypothesis and its alternative are

$$H_{\mathsf{C}0}\colon (\beta_1{=}0) \cap (\beta_2{=}0) \cap (\beta_3{=}0),$$
$$H_{\mathsf{C}a}\colon (\beta_1{\neq}0) \cup (\beta_2{\neq}0) \cup (\beta_3{\neq}0).,$$

In general, writing the individual hypotheses as H_{01}, H_{02}, etc.,

$$H_{\mathsf{C}0}\colon H_{01} \cap H_{02} \cap \ldots,$$
$$H_{\mathsf{C}a}\colon H_{a1} \cup H_{a2} \cup \ldots,$$

To control the familywise error rate in this situation, it is sufficient to set the level of evidence required for any of the component tests H_{0j} to be great enough that it will also cause the omnibus test of $H_{\mathsf{C}0}$ to be rejected at the desired $\alpha_{\text{familywise}}$. Scheffé's test in analysis of variance uses this principle.

3. Sequential testing. These approaches adopt a multistage testing procedure. An omnibus test (or series of tests) is applied first to limit the overall error rate to $\alpha_{\text{familywise}}$. Only when this "gatekeeper" test has been passed do further tests take place. Because the initial test has controlled the overall Type I error rate, the subsequent tests can adopt less stringent standards without compromising the overall error rate. Familiar instances of sequential tests are Fisher and Newmann-Keuls procedures in the analysis of variance.

All these methods emphasize control of Type I errors, and, by adopting more stringent criteria, they increase the Type II error rate.

Each of these strategies has its strengths and weaknesses. In brief, the Bonferroni procedure becomes increasingly conservative as the number of tests gets large, particularly when they are not independent. Tests based on the union-intersection relationship also tend to be quite conservative, and they can only be used when a satisfactory omnibus statistic exists. The sequential procedures are often more power than the others, but they get that power from a more complex relationship to their hypotheses. They are particularly unsatisfactory when the null hypothesis is partially false because (in their simpler forms) their protection is applied only at the level of $H_{\mathsf{C}0}$. One false subhypothesis can disable the gatekeeper and drop control over the rest of the subhypotheses.

Much of imaging data seems to me to fall outside the structure that has given us these approaches. With such data, the number of potential subhypotheses (say, activation of a given voxel), the subhypotheses are highly dependent, and the compound null hypothesis is partially (or even completely) false. Under such circumstances, it may be better to think of the simultaneous testing problem as one of separating the important effects that are worth following up from those that small and uninteresting. Instead of the table above, we have

	Decision made	
Magnitude of effect	Connection identified	Connection not identified
Interesting	Useful discovery	Missed opportunity
Uninteresting	False lead	Correct avoidance

Approaches that emphasize this structure, such as the false discovery rate, have been developed, but I will leave them to the other speakers.