Linear Modeling

- Probability Theory
  - Axioms
  - Basic Principles for probability modeling and computation
  - Law of Total Probability & Bayesian Theorem
  - Data Summaries and EDA
  - Distributions (http://www.socr.ucla.edu/htmls/SOCR_Distributions.html)
  - Experiments & Demos (http://www.socr.ucla.edu/htmls/SOCR_Experiments.html)
- Statistical Inference
  - Hypothesis Testing & Confidence intervals
  - Parameter Estimation
  - Parametric vs. Non-parametric inference (http://www.socr.ucla.edu/htmls/SOCR_Analyses.html)
  - CLT & LLN
- **Linear modeling**
  - Simple linear regression, Multiple linear regression
  - ANOVA & GLM

Fitted Value and Residual

The fitted value of $\mathbf{y}$, denoted $\hat{\mathbf{y}}$, is :

$$\underset{n\times 1}{\hat{\mathbf{y}}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

and the residual terms :

$$\underset{n\times 1}{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$$

Since population ε is unknonw, we estimate σ^2 from sample :

$$\mathbf{s}^2(e) = MSE$$

Multiple Regression in Matrix Form

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \beta_0 \begin{pmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} + \beta_1 \begin{pmatrix} x_{11} \\ x_{12} \\ x_{13} \\ \vdots \\ x_{1n} \end{pmatrix} + \beta_2 \begin{pmatrix} x_{21} \\ x_{22} \\ x_{23} \\ \vdots \\ x_{2n} \end{pmatrix} + \beta_3 \begin{pmatrix} x_{31} \\ x_{32} \\ x_{33} \\ \vdots \\ x_{3n} \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$= \begin{pmatrix} 1 & x_{11} & x_{21} & x_{31} \\ 1 & x_{12} & x_{22} & x_{32} \\ 1 & x_{13} & x_{23} & x_{33} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & x_{3n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Interpreting Multiple Regression Model

For a multiple regression model :

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$$

β_1 should be interpreted as change in y when 1 unit change is observed in x_1 and x_2 is kept constant. This statement is not very clear when x_1 and x_2 are not independent.

- Misunderstanding: β_i always measures the effect of x_i on E(y), independent of other x variables.
- Misunderstanding: a statistically significant β value establishes a cause and effect relationship between x and y.

Multiple Regression and LSE

The general multiple regression model is :

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{where} \quad V(\boldsymbol{\varepsilon}) = \sigma^2 \quad \mathrm{V}(\mathbf{y}) = \mathrm{V}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \sigma^2$$

Is based on model assumptions. Why?

$$\mathbf{X} = (X_1, X_2, \cdots X_p)$$

$$X_i = (x_{1i}, x_{2i}, \cdots x_{ni})'$$

The LSE solution for $\boldsymbol{\beta}$ will be :

$$Min\ SSE = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_{p-1} x_{p-1})^2$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

In matrix notation :

$$\mathbf{X}'\mathbf{y} = \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \Rightarrow \hat{\boldsymbol{\beta}} = \underset{p\times p}{(\mathbf{X}'\mathbf{X})^{-1}} \underset{p\times 1}{(\mathbf{X}'\mathbf{y})}$$

$$X'\mathbf{y} = \begin{pmatrix} 1'\mathbf{y} \\ X_1'\mathbf{y} \\ X_2'\mathbf{y} \\ X_3'\mathbf{y} \end{pmatrix} \qquad X'X = \mathrm{SSCP} = \begin{pmatrix} 1'1 & 1'X_1 & \cdots & 1'X_p \\ X_1'1 & X_1'X_1 & \cdots & X_1'X_p \\ X_2'1 & X_2'X_1 & \cdots & X_2'X_p \\ \vdots & \vdots & \ddots & \vdots \\ X_p'1 & X_p'X_1 & \cdots & X_p'X_p \end{pmatrix} \qquad (X'X)^{-1} = \begin{pmatrix} c_{00} & c_{01} & \cdots & c_{op} \\ c_{10} & c_{11} & \cdots & c_{1p} \\ \vdots & \vdots & \ddots & \vdots \\ c_{p0} & c_{p1} & \cdots & c_{pp} \end{pmatrix}$$

Properties of Coefficient Estimate

- It can be shown that:

$$\underset{\mathbf{p}\times\mathbf{p}}{\boldsymbol{\sigma}^2(\hat{\boldsymbol{\beta}})} = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$$

$$E(\hat{\beta}_i) = \beta_i$$

$$V(\hat{\beta}_i) = c_{ii}\sigma^2$$

$$\sigma_{\hat{\beta}_i} = \sigma\sqrt{c_{ii}}$$

$$Cov(\hat{\beta}_i, \hat{\beta}_j) = c_{ij}\sigma^2$$

- In the simplest case when there is only one x,

$$\sigma_{\hat{\beta}_1} = \sigma / \sqrt{SS_{xx}}$$

Properties of Coefficient Estimate

- Proof

$$Y_{n\times1} = X_{n\times k}\beta_{k\times1} + \varepsilon_{n\times1}; \qquad \varepsilon_{n\times1} \sim N(0, \sigma^2 I_{n\times n})$$

$$\textbf{\textit{Leasts quares solution is}}: \ \hat{\beta} = (X'X)^{-} X'Y =: A'_{k\times n} Y_{n\times1}$$

$$\underline{E(\hat{\beta})} = E((X'X)^{-} X'Y) = (X'X)^{-} X'E(Y) = (X'X)^{-} X'X\beta = \beta$$

$$\textbf{\textit{General Property of Matrices}}: Var(A'_{k\times n}Y_{n\times1}) = A'_{k\times n}Var(Y_{n\times1})A_{n\times k}$$

$$\underline{Var(\hat{\beta})} = Var(A'_{k\times n}Y_{n\times1}) = A'_{k\times n}Var(Y_{n\times1})A_{n\times k} = A'_{k\times n}Var(\varepsilon_{n\times1})A_{n\times k} =$$

$$A'_{k\times n}\sigma^2 I_{n\times n}A_{n\times k} = \sigma^2 A'_{k\times n}A_{n\times k} = \sigma^2((X'X)^{-}X')(X(X'X)^{-}) \Rightarrow$$

$$\underline{Var(\hat{\beta})} = \sigma^2 (X'X)^{-} X'X (X'X)^{-} = \sigma^2 (X'X)^{-} (X'X(X'X)^{-}) = \sigma^2 (X'X)^{-}$$

Properties of Coefficient Estimate

- Example:

$$(X'X)^{-} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.714 & 0.225 & 0.322 & -0.313 & -0.414 & -0.137 \\ 0 & 0 & 0.225 & 0.793 & 0.194 & -0.339 & -0.167 & -0.247 \\ 0 & 0 & 0.322 & 0.194 & 0.67 & -0.172 & -0.396 & -0.216 \\ 0 & 0 & -0.313 & -0.339 & -0.172 & 0.551 & 0.194 & 0.128 \\ 0 & 0 & -0.414 & -0.167 & -0.396 & 0.194 & 0.524 & 0.0141 \\ 0 & 0 & -0.137 & -0.247 & -0.216 & 0.128 & 0.0141 & 0.0366 \end{bmatrix}_{8\times8}, rank = 6$$

$$\hat{\beta} = (\beta_0, \beta_{11}, \beta_{12}, \beta_{13}, \beta_{14}, \beta_{21}, \beta_{22}, \beta_{23})' = (0.0, 0.0, 0.52, 0.82, 1.47, 1.05, 2.04, 4.3)'$$

Let the contrast $k = (0\ \text{-}1\ 1\ 0\ 0\ 0\ 0\ 0)'$, *then* $k'\times b = (0\ \text{-}1\ 1\ 0\ 0\ 0\ 0\ 0)\times b \Rightarrow$

$k'\times b = (0\ \text{-}1\ 1\ 0\ 0\ 0\ 0\ 0)\times b = b_{12} - b_{11} = 0.052.$

Note that k ***could have been*** $k = (0\ 0\ 1\ 0\ 0\ 0\ 0\ 0)'$

$\Rightarrow Var(k'b) = k'Var(b)k = k'Var(b)k = \sigma^2 k'(X'X)^{-}k = \mathbf{0.714}\ \sigma^2$

Estimate $\sigma^2 \approx s^2(e) = MSE = \dfrac{SSE}{n-r}$, ***where r*** $= rank(X) = 6$, $\boldsymbol{n} = 12$

and SSE $= Y'Y - b'X'Y = \mathbf{0.0031} \Rightarrow s^2(e) = \mathbf{0.0005.}$

Properties of Coefficient Estimate

- Example: Suppose that

$$Y = \begin{bmatrix} 2.1 \\ 2.15 \\ 2.2 \\ 2.35 \\ 2.33 \\ 2.22 \\ 2.25 \\ 2.27 \\ 2.5 \\ 2.6 \\ 2.4 \\ 2.45 \end{bmatrix}_{12\times1}$$

$$X = \begin{bmatrix} 1\,1\,0\,0\,0\,1\,0\,0 \\ 1\,0\,1\,0\,0\,1\,0\,0 \\ 1\,0\,0\,1\,0\,1\,0\,0 \\ 1\,0\,0\,0\,1\,0\,1\,0 \\ 1\,0\,0\,0\,1\,0\,1\,0 \\ 1\,1\,0\,0\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,0\,1\,0 \\ 1\,0\,1\,0\,0\,0\,1\,0 \\ 1\,0\,0\,1\,0\,0\,0\,1 \\ 1\,0\,0\,0\,1\,0\,0\,1 \\ 1\,1\,0\,0\,0\,0\,0\,1 \\ 1\,1\,0\,0\,0\,0\,0\,1 \end{bmatrix}_{12\times8} \Rightarrow (X'X)^{-} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0.714 & 0.225 & 0.322 & -0.313 & -0.414 & -0.137 \\ 0 & 0 & 0.225 & 0.793 & 0.194 & -0.339 & -0.167 & -0.247 \\ 0 & 0 & 0.322 & 0.194 & 0.67 & -0.172 & -0.396 & -0.216 \\ 0 & 0 & -0.313 & -0.339 & -0.172 & 0.551 & 0.194 & 0.128 \\ 0 & 0 & -0.414 & -0.167 & -0.396 & 0.194 & 0.524 & 0.0141 \\ 0 & 0 & -0.137 & -0.247 & -0.216 & 0.128 & 0.0141 & 0.0366 \end{bmatrix}_{8\times8}, rank = 6$$

- Then:

Confidence Intervals and Tests of Hypotheses for β's

One - tailed test	***Two - tailed test***
$H_0: \beta_i = 0$	$H_0: \beta_i = 0$
$H_a: \beta_i > 0$ ***or*** $(\beta_i < 0)$	$H_a: \beta_i \neq 0$

test statistic : $t = \dfrac{\hat{\beta}_i}{s\sqrt{c_{ii}}}$, ***where*** s ***= sample SD***

Rejection region :

$t > t_\alpha$ ***(or*** $t < t_\alpha$***)***	$\lvert t\rvert > t_{\alpha/2}$

$t_{\alpha/2}$ ***is based on [n - (p + 1)]df, p is number of independent variables in the model***

Properties of Coefficient Estimate

$$(X'X)_{8\times8} \times b_{8\times1} = (X'Y)_{8\times1} \quad \Leftrightarrow$$

$$\begin{bmatrix} 12 & 4 & 3 & 2 & 3 & 3 & 5 & 4 \\ 4 & 4 & 0 & 0 & 0 & 1 & 1 & 2 \\ 3 & 0 & 3 & 0 & 0 & 1 & 2 & 0 \\ 2 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 3 & 0 & 0 & 0 & 3 & 0 & 2 & 1 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 5 & 1 & 2 & 0 & 2 & 0 & 5 & 0 \\ 4 & 2 & 0 & 1 & 1 & 0 & 0 & 4 \end{bmatrix} \begin{bmatrix} b_0 \\ b_{11} \\ b_{12} \\ b_{13} \\ b_{14} \\ b_{21} \\ b_{22} \\ b_{23} \end{bmatrix} = \begin{bmatrix} 27.82 \\ 9.17 \\ 6.67 \\ 4.70 \\ 7.28 \\ 6.45 \\ 11.42 \\ 9.95 \end{bmatrix}$$

Two-way ANOVA

- Two treatment factors, with **g** and **b** levels
- There are *1 ≤ l≤* ***g*** levels of factor 1
- *1 ≤ k≤* ***b*** levels of factor 2
- **gb** combinations of levels *(l,k)*
- **N** independent observations

Univariate Analysis of Variance
Two-way Fixed Effects Model with Interaction

The ANOVA model (Linear Model) can be written as:

$$y_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + e_{lkr}$$

μ is the grand mean
τ is the fixed effect for factor 1, $1 \le l \le g$ levels of factor 1
β is fixed effect of factor 2, $1 \le k \le b$ levels of factor 2
γ is the interaction
r replicates

Hypotheses tested by ANOVA:

1) Does the effect of one factor on the response variable(s) depend on level of the other factor?

H_0: There is no interaction between Factor 1 and Factor 2

$$\mu_{lk} - \mu_{l'k} - \mu_{lk'} + \mu_{l'k'} = 0$$

2) Do the levels of Factor 1 differ in the effects on the response variable(s)

H_0: There is no main effect of Factor 1 on the response

$$\mu_{1.} = \mu_{2.} = \cdots = \mu_{p.}$$

3) Do the levels of Factor 2 differ in their effects on the response variable(s)

$$\mu_{.1} = \mu_{.2} = \cdots = \mu_{.p}$$

The Expected Response

$$\sum_{l=1}^{g} \tau_l = \sum_{k=1}^{b} \beta_k = \sum_{l=1}^{g} \gamma_{lk} = \sum_{k=1}^{b} \gamma_{lk} = 0$$

Noise: e_{lkr} are independent $N(0, \sigma^2)$

ANOVA Table & Variance Decomposition

Source of Variation	Sum of Squares (SS)	Degrees of Freedom	F-ratios
Factor 1	$= \sum_{l=1}^{g} bn(\bar{x}_{l.} - \bar{x})^2$	g - 1	
Factor 2	$= \sum_{k=1}^{b} gn(\bar{x}_{.k} - \bar{x})^2$	b - 1	
Interaction	$= \sum_{l=1}^{g} \sum_{k=1}^{b} n(\bar{x}_{lk} - \bar{x}_{l.} - \bar{x}_{.k} + \bar{x})^2$	(g-1)(b–1)	
Residual (error)	$\sum_{l=1}^{g} \sum_{k=1}^{b} \sum_{r=1}^{n} (\bar{x}_{lkr} - \bar{x}_{lk})^2$	gb(n – 1)	
Total (Corrected)	$\sum_{l=1}^{g} \sum_{k=1}^{b} \sum_{r=1}^{n} (\bar{x}_{lkr} - \bar{x})^2$	gbn(n - 1)	

In other words

$E(y_{lkr})$	=	μ	+	τ_l	+	β_k	+	γ_{lk}
Mean Response	=	Overall level	+	Effect of Factor 1	+	Effect of Factor 1	+	Interaction

ANOVA in Matrix Notation

- Regardless of the complexity of the ANOVA model, we can express it in matrix notation

$$y = X\beta + \varepsilon$$

- **X** is a matrix of 0's and 1s that follows the experimental plan and its' linear model

The General Linear Model

$$y = Xb + e$$

y is the column vector of responses for N individuals	X is the (N × r) "design matrix"
b is a vector of parameters	e is a vector of residuals

GML vs. Multiple Regression

- The multiple regression limitations:
- It can be used to analyze only a single dependent variable
- It cannot provide a solution for the regression coefficients when the *X* variables are not approx linearly independent (the inverse of X'X therefore does not exist).
- These restrictions can be overcome by transforming the multiple regression model into the general linear model.

GML vs. Multiple Regression

- The general purpose of **multiple regression** is to quantify the relationship between several independent (or predictor) variables (X) and one dependent (or response) variable (Y).

$$Y = b_0 + b_1X_1 + b_2X_2 + \ldots + b_kX_k$$

- There are *k* predictors (X) and the regression coefficients ($b_1 \ldots b_k$) represent the *independent* contributions of each independent variable to the prediction of the dependent variable, i.e., *X1* is (partially) correlated with the *Y* variable, after controlling for all other independent variables.
- Example: we can find a significant positive correlation between brain volume and height in the population (i.e., short people have smaller brains). Let's add the variable Gender into the multiple regression equation, this correlation would probably disappear. This is because women, on the average, have smaller head-size than men; they are also shorter on the average than men. Thus, after we remove this gender difference by entering Gender into the equation, the relationship between Brain Volume and height may disappear, as brain volume may *not* make any unique contribution to the prediction of height, above and beyond what it shares in the prediction with variable Gender. I.e., controlling for the variable Gender, the *partial correlation* between brain volume and height is zero.

GML

- The general linear model differs from the multiple regression model is in terms of the number of dependent variables that can be analyzed. The ***Y*** vector of *n* observations of a single *Y* variable can be replaced by a ***Y*** matrix of *n* observations of *m* different *Y* variables (in fact replaced with linear combinations of responses).
- Similarly, the ***b*** vector of regression coefficients for a single ***Y*** variable can be replaced by a ***b*** matrix of regression coefficients, with one vector of ***b*** coefficients for each of the *m* dependent variables.
- These substitutions yield what is sometimes called the multivariate regression model - the matrix formulations of the multiple and multivariate regression models are identical, except for the number of columns in the ***Y*** and ***b*** matrices.
- The method for solving for the ***b*** coefficients is also identical, that is, *m* different sets of regression coefficients are separately found for the *m* different dependent variables in the multivariate regression model.

GML - Multiple Regression

- The multiple regression model in matrix notation then can be expressed as

  $$Y = Xb + e$$

  b is a column vector of 1 (for the intercept) + *k* unknown regression coefficients. Recall that the goal of multiple regression is to minimize the sum of the squared residuals. Regression coefficients that satisfy this criterion are found by solving the set of normal equations

  $$X'Xb = X'Y$$

- If the *X* variables are linearly independent (i.e., they are nonredundant, yielding an ***X'X*** matrix which is of full rank) there is a unique solution to the normal equations.
- Premultiplying both sides of the matrix formula for the normal equations by the inverse of ***X'X*** gives

  $$(X'X)^{-1}X'Xb = (X'X)^{-1}X'Y \Rightarrow b = (X'X)^{-1}X'Y$$

- 3 basic matrix operations
  - matrix transposition, exchange the rows and columns of a matrix
  - matrix multiplication, sum of the products of the elements for each row and column combination of two conformable
  - matrix inversion, which involves finding the matrix equivalent of a numeric reciprocal, that is, the matrix that satisfies

GML

- The general linear model also differs from the multiple regression model in its ability to provide a solution for the normal equations when the *X* **variables are not linearly independent** and the inverse of ***X'X*** does not exist. Redundancy of the *X* variables may be incidental (e.g., two predictor variables are perfectly correlated), accidental (e.g., two copies of the same variable) or designed (e.g., indicator variables with exactly opposite values might be used in the analysis, as when both *Male* and *Female* predictor variables are used in representing *Gender*).
- Finding the regular inverse of a non-full-rank matrix is analogous to finding the reciprocal of 0 in ordinary arithmetic. No such inverse or reciprocal exists because division by 0 is not permitted. This problem is solved in the general linear model by using a generalized inverse of the ***X'X*** matrix in solving the normal equations. A generalized inverse (A^-) is any matrix A that satisfies

$$AA^-A = A$$

GML

- A generalized inverse is unique and coincides with the regular inverse if the matrix ***A*** is full rank.
- A generalized inverse for a non-full-rank matrix can be computed by zeroing the elements in redundant rows and columns of the matrix.
- Suppose that an ***X'X*** matrix with *r* non-redundant columns is partitioned as

$$X'X = \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix}$$

- where A_{11} is an *r* by *r* matrix of rank *r*. Then the regular inverse of ***A11*** exists and a generalized inverse of ***X'X*** is

$$(X'X)^{-} = \begin{bmatrix} A_{11} & O_{21} \\ O_{12} & O_{22} \end{bmatrix}$$

- where each ***0*** (null) matrix is a matrix of 0's (zeroes) and has the same dimensions as the corresponding ***A*** matrix.

GML

- Overparameterized model of categorical predictors.
- The second basic method for recoding categorical predictors is the indicator variable approach. In this method a separate predictor variable is coded for each group identified by a categorical predictor variable. Example, females might be assigned a value of 1 and males a value of 0 on a first predictor variable identifying membership in the female *Gender* group. Males would then be assigned a value of 1 and females a value of 0 on a second predictor variable identifying membership in the male *Gender* group.
- This method of recoding categorical predictor variables will almost always lead to ***X'X*** matrices with redundant columns, and thus require a generalized inverse for solving the normal equations. As such, this method is often called the overparameterized model for representing categorical predictor variables, because it results in more columns in the ***X'X*** than are necessary for determining the relationships of categorical predictor variables to responses on the dependent variables.
- The general linear model can be used to perform analyses with categorical predictor variables which are coded using either Standard of Overparameterized models.

GML

- There are infinitely many generalized inverses of a non-full-rank ***X'X*** matrix. Thus, infinitely many solutions to the normal equations. So, the regression coefficients can change depending on the particular generalized inverse chosen for solving the normal equations. However, many results obtained using the general linear model have invariance properties (e.g., correlation is linearly invariant).
- Example: If both *Male* and *Female* predictor variables with exactly opposite values are used in an analysis to represent *Gender*, it is essentially arbitrary as to which predictor variable is considered to be redundant (e.g., *Male* can be considered to be redundant with *Female*, or vice versa).
- The predicted values and the corresponding residuals for males and females will be unchanged -- no matter which predictor variable is considered to be redundant, no matter which corresponding generalized inverse is used in solving the normal equations, and no matter which resulting regression equation is used for computing predicted values on the dependent variables. Using the general linear model, finding a particular arbitrary solution to the normal equations is primarily a means to accounting for responses effects on the dependent variables.

GML - Calculations

- The general linear model can be expressed as

YM = Xb + e

Example: Y1=Systolic
Y2=Diastolic Pressure
MAP=(Y1+2*Y2)/3
Mean Arterial Pressure

- Here ***Y***, ***X***, ***b***, and ***e*** are multivariate response, Desing matrix, parameter matrix, residual matrix and ***M*** is an *m* x *s* matrix of coefficients defining *s* linear transformation of the dependent variables. The normal equations are

X'Xb = X'YM

- and a solution for the normal equations is given by **b = (X'X)⁻X'YM**
- The inverse of ***X'X*** is a generalized inverse if ***X'X*** contains redundant columns
- Allows for analyzing linear combinations of multiple dependent variables, add a method for dealing with redundant predictor variables and recoded categorical predictor variables, and the major limitations of multiple regression are overcome by the general linear model.

GML

- In multiple regression model, the *X* variables are continuous. The general linear model is frequently applied to analyze
  - ANOVA or MANOVA design with categorical predictor variables
  - ANCOVA or MANCOVA design with both categorical and continuous predictor variables
  - Multiple or multivariate regression design with continuous predictor variables.
- Example: *Gender* is clearly a nominal level variable. There are two basic methods by which *Gender* can be coded into one or more (non-offensive) predictor variables, and analyzed using the general linear model.
- Standard model of categorical predictors. Males and females can be assigned any two distinct values on a single predictor variable. Typically, the values corresponding to group membership are chosen to facilitate interpretation of the regression coefficient associated with the predictor variable. For example, the two groups are assigned values of 1 and -1 on the predictor variable, so that if the regression coefficient for the variable is positive, the group coded as 1 on the predictor variable will have a higher predicted value (i.e., a higher group mean) on the dependent variable, and if the regression coefficient is negative, the group coded as -1 on the predictor variable will have a higher predicted value on the dependent variable. An advantage is that since each group is coded with a value one away from zero - helps in interpreting the magnitude of differences in predicted values between groups, because regression coefficients reflect the units of change in the dependent variable for each unit change in the predictor variable.

GML - Calculations

$$[Y]_{n\times m}[M]_{m\times s} = [X]_{n\times k}[b]_{k\times 1} + [\varepsilon]_{k\times 1}$$

GML – ANOVA example

- A design with a single categorical predictor variable is called a one-way ANOVA design. For example, a study of 4 different populations (NC, MCI, AD-1, AD-2), with four levels for the factor *disease*.
- In general, consider a single categorical predictor variable *A* with 1 case in each of its 4 categories. Using the Standard model coding of A into 3 quantitative contrast variables, the matrix ***X*** defining the between design is

$$X = \begin{array}{c} A_1 \\ A_2 \\ A_3 \\ A_4 \end{array} \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 \\ 1 & -1 & -1 & -1 \end{bmatrix}}_{X_0 \quad X_1 \quad X_2 \quad X_3}$$

- That is, cases in groups *A1*, *A2*, *A3* and A4 are all assigned values of 1 on *X0* (the intercept), the case in group *A1* is assigned a value of 1 on *X1* and a value 0 on other *X's*, the case in group *A2* is assigned a value of 1 on *X2* and a value 0 on other *X's*, and the case in group *A3* is assigned a value of -1 on *X1* and *X2*.

Least Squares Estimates of b

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

GML – ANOVA example

- If there were 1 case in group *A1*, 2 cases in group *A2*, 1 case in group *A3*, and 3 cases in A4, the ***X*** matrix would be

$$X = \begin{array}{c} A_{11} \\ A_{21} \\ A_{22} \\ A_{31} \\ A_{41} \\ A_{41} \\ A_{43} \end{array} \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \end{bmatrix}}_{X_0 \quad X_1 \quad X_2 \quad X_3}$$

- The first subscript for *A identifies the group and the second* gives the replicate number. Usually replicates are not shown when describing ANOVA designs
- One-way designs with an equal number of cases in each group, **Standard Model coding** yields *X1* … *Xk* variables all of which have means of 0.

Elaboration of Matrix Elements

The transpose of the parameter vector is (r×1):

$$b' = [\tau_1 \cdots \tau_g, \beta_1 \cdots \beta_b, \gamma_{11} \cdots \gamma_{1b} \cdots \gamma_{g1} \cdots \gamma_{gb}, \mu]$$

$$y_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + e_{lkr}$$

GML – ANOVA example

- Using the Underparameterized model to represent A, the ***X*** matrix defining the between design is just

$$X = \begin{array}{c} A_1 \\ A_2 \\ A_3 \\ A_4 \end{array} \underbrace{\begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{bmatrix}}_{X_0 \quad X_1 \quad X_2 \quad X_3 \quad X_4}$$

- The ***X*** matrix serves two purposes:
  - Specifies the coding for the levels of the original predictor variables on the *X* variables used in the analysis
  - Shows the between variable design.

Design Matrix

$$A = \begin{bmatrix} 1\cdots0 & 1\cdots0 & 1\cdots0\cdots0\cdots0 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 1\cdots0 & 0\cdots1 & 0\cdots1\cdots0\cdots0 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 0\cdots1 & 1\cdots0 & 0\cdots0\cdots1\cdots0 & 1 \\ \cdots & \cdots & \cdots & \cdots \\ 0\cdots1 & 0\cdots1 & 0\cdots0\cdots0\cdots1 & 1 \end{bmatrix}_{Nxr} \qquad \overset{b}{\begin{bmatrix} \tau_1 \\ \cdots \\ \tau_g \\ \beta_1 \\ \cdots \\ \beta_b \\ \gamma_{11} \\ \cdots \\ \cdots \\ \gamma_{bg} \\ \mu \end{bmatrix}}_{r\times1}$$

Each column of the design matrix corresponds with the appropriate element of the parameter vector.

Assumptions of ANOVA

- Normal distribution
- Independence of residuals
- Homoscedasticity of Variances
  - Variances are ≈ Equal

Full Model

$$Y_i = E(Y/X_i) + \varepsilon_i$$

ε_i is referred to as an: Error or Residual

Regression Analysis

- Most widely applied technique for assessing relationships among variables
- Used to investigate relationship between a **response** (dependent) variable and one or more **predictor** (independent) variables.
- Regression analysis is concerned with estimating and predicting the population mean value of the response variable Y on the basis of known (fixed) values of one or more predictor (or explanatory) variable(s)

Properties of Population Model

- Postulates the condition means are linear functions of the X_i.
- The β's are known as regression coefficients.
- The intercept gives E(Y|X=0)
- The slope describes the change in Y for a fixed unit change in X

The Population-based Regression Model

$$E(Y/X_i) = \beta_0 + \beta_1 X_i$$

β_0, β_1 are unknown, but fixed parameters

β_0, - intercept

β_1 – slope

$$Y_i = E(Y/X_i) + \varepsilon_i$$

Assumptions of Regression Analysis

- Y's are normally distributed
- X's are fixed,
- Residuals (e_i) are normal, independent random variables.

Sample-based Regression Model

$$E(Y_i / X_i) = b_0 + b_1 X$$

or

$$Y_i = b_0 + b_1 X_i + e_i$$

Matrix Notation for Linear Regression

$$Y = X\beta + \varepsilon$$

We can estimate the regression parameters using the simple expression:

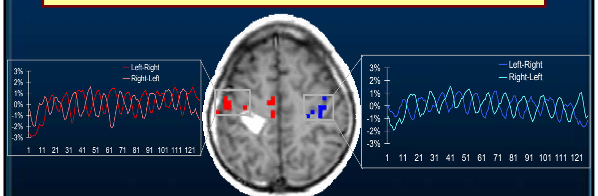
$$\hat{\beta} = [X'X]^{-1} X'y$$

How to estimate b_0 and b_1.

- Use Ordinary Least Squares approach.
  - i.e., minimize error sum of squares.

$$\text{minimize} \sum_{i=1}^{n} \hat{e}^2{}_i$$

Example of am fMRI Study

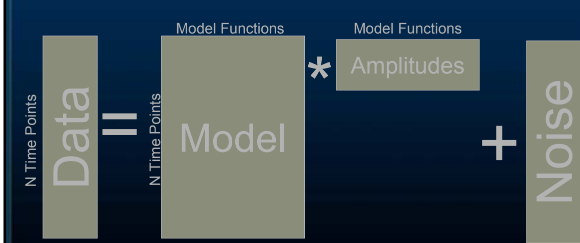


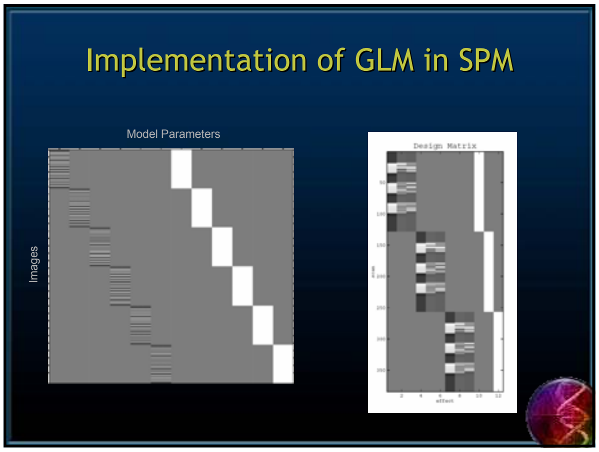
•Correlation - Special case of General Linear Model
  •Blocked t-test is equivalent to correlation with square wave function
•Correlation coefficient describes match between observation and expectation, $-1 \le R \le 1$. R is "almost" linearly invariant!
•Problems with using the correlation:
  •Limited by choice of HDR
  •Poorly chosen HDR can significantly impair power
  •Assume random variation around HDR
  •Does not model variability contributing to noise (e.g., scanner drift) - such variability is usually removed in preprocessing steps
  •Does not model interactions between successive events

ANOVA Table for Regression

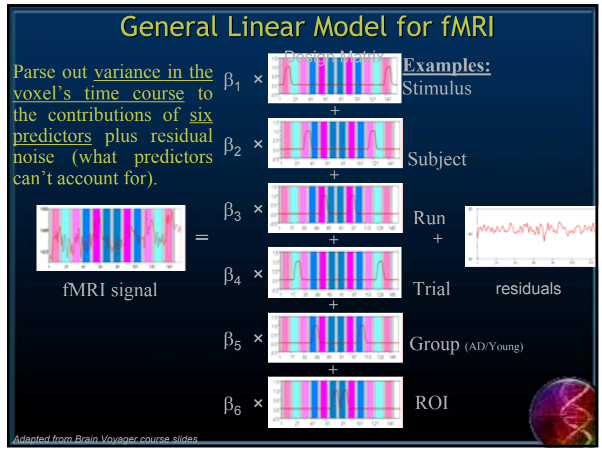
Source of Variation		Sum of Squares	DF	Mean Square
Total	$[Y_i - \bar{Y}]$	$\sum y^2$	n –1	
Linear Regression	$[\hat{Y}_i - \bar{Y}]$	$\frac{(\sum xy)^2}{\sum x^2}$	1	Regression SS/ Regression df
Residual	$[Y_i - \hat{Y}]$	Total SS- Regression SS	n - 2	Residual SS/ Residual df

Form of the GLM

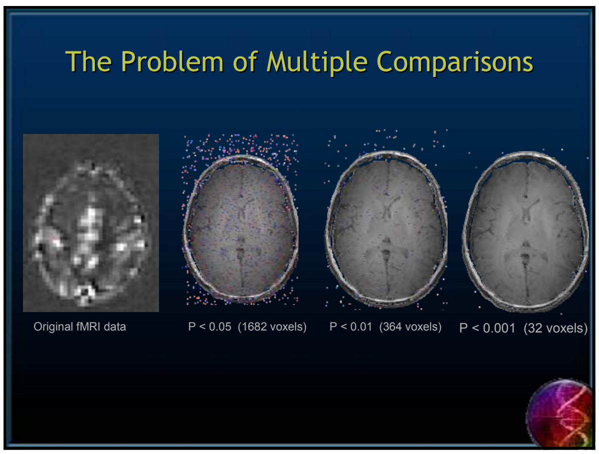




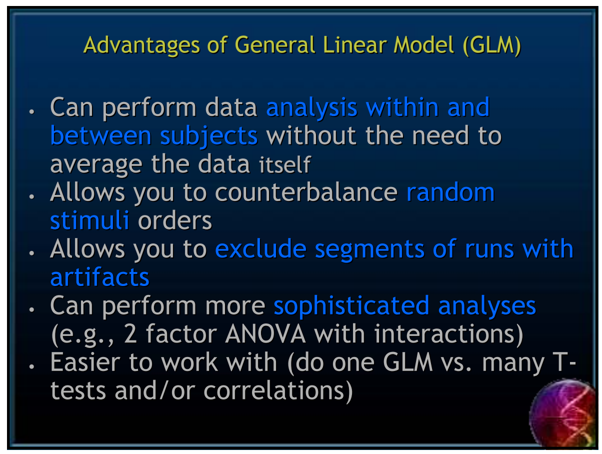
Implementation of GLM in SPM
Model Parameters
Images
Design Matrix



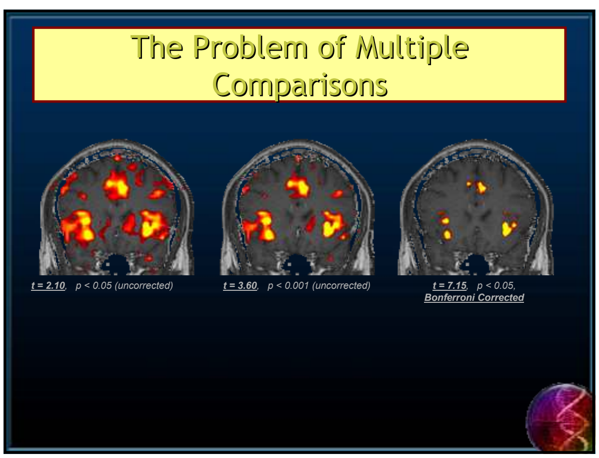
General Linear Model for fMRI
Parse out variance in the voxel's time course to the contributions of six predictors plus residual noise (what predictors can't account for).
fMRI signal
=
β1 ×
β2 ×
β3 ×
β4 ×
β5 ×
β6 ×
+
Examples:
Stimulus
Subject
Run
Trial
Group (AD/Young)
ROI
residuals
Adapted from Brain Voyager course slides



The Problem of Multiple Comparisons
Original fMRI data
P < 0.05 (1682 voxels)
P < 0.01 (364 voxels)
P < 0.001 (32 voxels)



Advantages of General Linear Model (GLM)
• Can perform data analysis within and between subjects without the need to average the data itself
• Allows you to counterbalance random stimuli orders
• Allows you to exclude segments of runs with artifacts
• Can perform more sophisticated analyses (e.g., 2 factor ANOVA with interactions)
• Easier to work with (do one GLM vs. many T-tests and/or correlations)



The Problem of Multiple Comparisons
t = 2.10, p < 0.05 (uncorrected)
t = 3.60, p < 0.001 (uncorrected)
t = 7.15, p < 0.05, Bonferroni Corrected

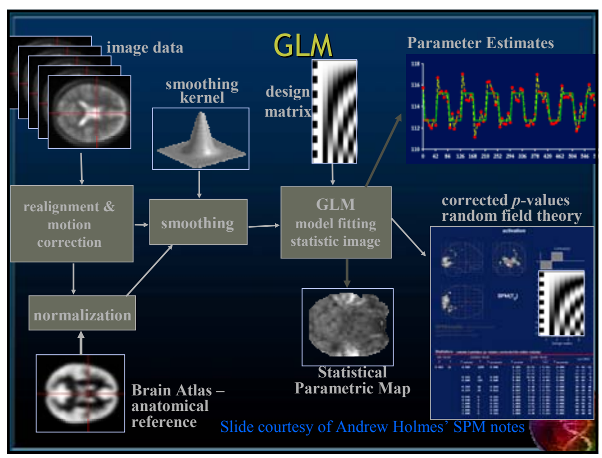
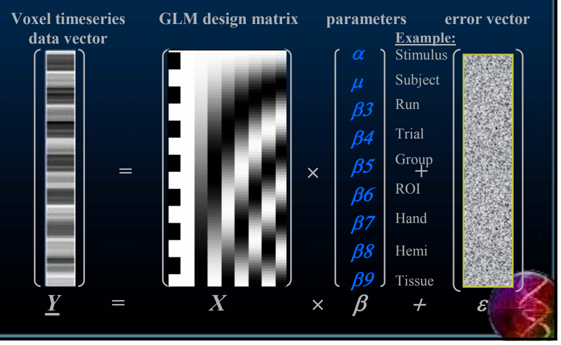


image data
GLM
Parameter Estimates
smoothing kernel
design matrix
realignment & motion correction
smoothing
GLM model fitting statistic image
corrected p-values random field theory
normalization
Statistical Parametric Map
Brain Atlas – anatomical reference
Slide courtesy of Andrew Holmes' SPM notes

General Linear Model Approach



Options for Multiple Comparisons

- Statistical Correction
  - Gaussian Field Theory (Worsley, et al.)
  - False discovery rate (Taylor, et al.)
  - Bonferroni (Dinov, et al.)
  - Tukey (Mills, et al.)
- Cluster Analyses (Müller, et al.)
- ROI Approaches (e.g., CCB Probabilistic Atlas; Mega, et al.)

Why Use Nonparametric Statistics?

- Parametric tests are based upon assumptions that may include the following:
  - The data have the **same variance**, regardless of the treatments or conditions in the experiment.
  - The data are **normally distributed** for each of the treatments or conditions in the experiment.
- What happens when we are not sure that these assumptions have been satisfied?