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
- Parametric vs. Non-Parametric inference
- Simple linear regression, Multiple linear regression
- ANOVA & GLM

Fitted Value and Residual

The fitted value of y, denoted \( \hat{y} \), is:

\[
\hat{y} = \mathbf{X} \hat{\beta}
\]

and the residual terms:

\[
e_i = y - \hat{y} = y - \mathbf{X} \hat{\beta}
\]

Since population \( \sigma \) is unknown, we estimate \( \sigma^2 \) from sample:

\[
s^2 = \frac{\sum (y_i - \hat{y})^2}{n - p - 1}
\]

Multiple Regression in Matrix Form

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i
\]

Interpreting Multiple Regression Model

For a multiple regression model:

\[
y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i
\]

\( \hat{\beta} \) should be interpreted as change in y when 1 unit change is observed in \( x_i \), and \( x_j \) is kept constant. This statement is not very clear when \( x_i \) and \( x_j \) are not independent.

- Misunderstanding: \( \hat{\beta} \) always measures the effect of \( x_i \) on \( y \) independent of other \( x \) variables.
- Misunderstanding: a statistically significant \( \hat{\beta} \) value establishes a cause and effect relationship between \( x \) and \( y \).
Example: Suppose that

\[ Y_{eit} = X_{eit} \beta_{et} + e_{eit}, \quad e_{eit} \sim N(0, \sigma^2 I_{n_{eit}}) \]

Least squares solution is:

\[ \hat{\beta} = (X'X)^{-1}X'Y \]

General Property of Matrices:

\[ \text{Var}(\hat{\beta}) = \sigma^2 (X'X)^{-1} \]

**Properties of Coefficient Estimate**

- **Example:** Suppose that

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) vector, \( X \) is an \( n \times p \) matrix of independent variables, \( \beta \) is a \( p \times 1 \) vector of parameters, and \( \epsilon \) is an \( n \times 1 \) vector of error terms.

Properties of Coefficient Estimate

- Example: Suppose that

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) vector, \( X \) is an \( n \times p \) matrix of independent variables, \( \beta \) is a \( p \times 1 \) vector of parameters, and \( \epsilon \) is an \( n \times 1 \) vector of error terms.

Properties of Coefficient Estimate

- Example: Suppose that

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) vector, \( X \) is an \( n \times p \) matrix of independent variables, \( \beta \) is a \( p \times 1 \) vector of parameters, and \( \epsilon \) is an \( n \times 1 \) vector of error terms.

Properties of Coefficient Estimate

- Example: Suppose that

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) vector, \( X \) is an \( n \times p \) matrix of independent variables, \( \beta \) is a \( p \times 1 \) vector of parameters, and \( \epsilon \) is an \( n \times 1 \) vector of error terms.

Confidence Intervals and Tests of Hypotheses for \( \beta \)’s

- Example: Suppose that

\[ Y = X\beta + \epsilon \]

where \( Y \) is an \( n \times 1 \) vector, \( X \) is an \( n \times p \) matrix of independent variables, \( \beta \) is a \( p \times 1 \) vector of parameters, and \( \epsilon \) is an \( n \times 1 \) vector of error terms.

Two-way ANOVA

- Two treatment factors, with \( g \) and \( b \) levels.

- There are \( l \leq g \) levels of factor 1.

- \( l \leq b \) levels of factor 2.

- \( gb \) combinations of levels \( (l,b) \).

- \( N \) independent observations.
The ANOVA model (Linear Model) can be written as:

\[ y_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lkr} + e_{lkr} \]

- \( \mu \) is the grand mean
- \( \tau \) is the fixed effect for factor 1, \( 1 \leq l \leq g \) levels of factor 1
- \( \beta \) is fixed effect of factor 2, \( 1 \leq k \leq b \) levels of factor 2
- \( \gamma \) is the interaction
- \( r \) replicates

**The Expected Response**

\[ \sum_{l=1}^{g} \tau_l = \sum_{k=1}^{b} \beta_k = \sum_{l=1}^{g} \gamma_{l} = \sum_{k=1}^{b} \gamma_{l} = 0 \]

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

**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
     - \( \mu_1 - \mu_2 - \cdots - \mu_p = 0 \)

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)
   - \( \mu_1 = \mu_2 = \cdots = \mu_p \)

**ANOVA Table & Variance Decomposition**

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares (SS)</th>
<th>Degrees of Freedom</th>
<th>Variance Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 1</td>
<td>( \sum_{l=1}^{g}(\mu_l - \bar{\mu})^2 )</td>
<td>( g - 1 )</td>
<td>( g(bn-1) )</td>
</tr>
<tr>
<td>Factor 2</td>
<td>( \sum_{k=1}^{b}(\mu_k - \bar{\mu})^2 )</td>
<td>( b - 1 )</td>
<td>( gb(n-1) )</td>
</tr>
<tr>
<td>Interaction</td>
<td>( \sum_{l=1}^{g} \sum_{k=1}^{b}(\mu_{lkr} - \bar{\mu})^2 )</td>
<td>( (g-1)(b-1) )</td>
<td>( gb(n-1) )</td>
</tr>
<tr>
<td>Residual (error)</td>
<td>( \sum_{l=1}^{g} \sum_{k=1}^{b} \sum_{r=1}^{n}(y_{lkr} - \mu_{lkr})^2 )</td>
<td>( gb(n-1) )</td>
<td>( gb(n-1) )</td>
</tr>
<tr>
<td>Total (Corrected)</td>
<td>( \sum_{l=1}^{g} \sum_{k=1}^{b} \sum_{r=1}^{n}(y_{lkr} - \mu_{lkr})^2 )</td>
<td>( gb(n-1) )</td>
<td>( gb(n-1) )</td>
</tr>
</tbody>
</table>

**ANOVA in Matrix Notation**

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

\[ y = X\beta + \epsilon \]

- \( X \) is a matrix of 0’s and 1’s that follows the experimental plan and its’ linear model

**In other words**

<table>
<thead>
<tr>
<th>( E[y_{lkr}] )</th>
<th>( \mu )</th>
<th>( \tau )</th>
<th>( \beta )</th>
<th>( \gamma )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Response</td>
<td>Overall level</td>
<td>Effect of Factor 1</td>
<td>Effect of Factor 2</td>
<td>Interaction</td>
</tr>
</tbody>
</table>
The General Linear Model

\[ y = Xb + e \]

- \( y \) is the column vector of responses for \( N \) individuals
- \( X \) is the \((N \times p)\) “design matrix”
- \( b \) is a vector of parameters
- \( e \) is a vector of residuals

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., \( X \) 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 average, are smaller than men (on average); thus they are also shorter than the average 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 - Multiple Regression

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

\[ Y = Xb + e \]

\( b \) is a column vector 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

\[ X'Xb = X'Y \rightarrow b = (X'X)^{-1}X'Y \]

- 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 matrices
  - matrix inversion, which involves finding the inverse equivalent of a numeric reciprocal, that is, the matrix that satisfies

\[ AA = A \]
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

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

where $A_{11}$ is an $r \times r$ matrix of rank $r$. Then the regular inverse of $A_{11}$ exists and a generalized inverse of $X'X$ is

$$(X'X)^{-1} = A_{11}^{-1} - A_{11}^{-1}A_{12}(A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1}A_{21}A_{11}^{-1}$$

where each $\Theta$ (null) matrix is a matrix of 0’s (zeros) and has the same dimensions as the corresponding $A$ matrix.

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, as 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 column of the normal equations is primarily a matter of accounting for responses effects on the dependent variables.

The general linear model can be expressed as

$$Y = Xb + \varepsilon$$

where $Y$ is an $n \times 1$ response vector, $X$ is an $n \times k$ matrix of covariates, $b$ is a $k \times 1$ vector of regression coefficients, and $\varepsilon$ is a $n \times 1$ vector of error terms. The normal equations are

$$X'Xb = X'Y$$

and a solution for the normal equations is given by

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

Overparameterized models. The second basic method for recoding categorical predictor variables is the indicator variable approach. In this method a separate predictor variable is coded for each group identified by the categorical predictor variable. Example: females might be assigned a value of 1 and males a value of 0 in a first predictor variable identifying membership in a female gender group. Males would then be assigned a value of 1 and females a value of 0 in 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 $X'X$ matrix 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 or Overparameterized models.

**GML - Calculations**

The general linear model can be expressed as

$$Y = Xb + \varepsilon$$

where $Y$ is an $n \times 1$ response vector, $X$ is an $n \times k$ matrix of covariates, $b$ is a $k \times 1$ vector of regression coefficients, and $\varepsilon$ is a $n \times 1$ vector of error terms. The normal equations are

$$X'Xb = X'Y$$

and a solution for the normal equations is given by

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

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 responses, categorical predictor variables, and the major limitations of multiple regression are overcome by the general linear model.

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

\[
X = \begin{bmatrix}
A_1 & A_2 & A_3 & A_4 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
\end{bmatrix}
\]

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

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:

\[
X = \begin{bmatrix}
A_1 & A_2 & A_3 & A_4 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & -1 \\
\end{bmatrix}
\]

The first subscript for A identifies the group and the second subscript 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 XI. . . Xₖ variables all of which have means of 0.

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

\[
X = \begin{bmatrix}
A_1 & A_2 & A_3 & A_4 \\
1 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 \\
\end{bmatrix}
\]

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.

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

\[
b' = (\tau_1 \ldots \tau_k, \beta_1, \ldots \beta_k, \gamma_{11}, \ldots \gamma_{41} \ldots \gamma_{4k}, \mu)
\]

\[
y'_{lkr} = \mu + \tau_l + \beta_k + \gamma_{lk} + e_{lkr}
\]

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 \( \approx \) Equal

Regressions 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)

The Population-based Regression Model

\[
E(Y/X_i) = \beta_0 + \beta_1 X_i \\
\beta_0, \beta_1 \text{ are unknown, but fixed parameters} \\
\beta_0 \text{ - intercept} \\
\beta_1 \text{ - slope} \\
Y_i = E(Y/X_i) + \epsilon_i
\]

Full Model

\[
Y_i = E(Y/X_i) + \epsilon_i
\]

\( \epsilon \) is referred to as an Error or Residual

Properties of Population Model

- Postulates the condition means are linear functions of the \( X \)
- The \( \beta \)'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 \( X \)

Assumptions of Regression Analysis

- \( Y \)'s are normally distributed
- \( X \)'s are fixed,
- Residuals (\( \epsilon \)) are normal, independent random variables.
Sample-based Regression Model

\[ E(Y_i/X_i) = b_0 + b_1X \]

or

\[ Y_i = b_0 + b_1X_i + e_i \]

How to estimate \( b_0 \) and \( b_1 \).

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

\[ \text{minimize} \sum_{i=1}^{n} e_i^2 \]

Matrix Notation for Linear Regression

\[ Y = X\beta + \varepsilon \]

We can estimate the regression parameters using the simple expression:

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

ANOVA Table for Regression

<table>
<thead>
<tr>
<th>Source of Variation</th>
<th>Sum of Squares</th>
<th>df</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>( Y_i - \bar{Y} ) \sum y^2</td>
<td>n - 1</td>
<td></td>
</tr>
<tr>
<td>Linear Regression</td>
<td>( \hat{Y}_i - \bar{Y} ) ( \sum \hat{y}y ) / ( \sum \hat{y}^2 )</td>
<td>1</td>
<td>Regression SS / Regression df</td>
</tr>
<tr>
<td>Residual</td>
<td>( Y_i - \hat{Y}_i ) Residual SS / Residual df</td>
<td>n - 2</td>
<td></td>
</tr>
</tbody>
</table>

Form of the GLM

Data \( \times \) Model Functions \( \rightarrow \) Amplitudes + Noise

Example of an fMRI Study

- Correlation - Special case of General Linear Model
- Block-design is equivalent to correlation with square wave function
- Correlation coefficient describes match between observation and expectation, \(-1 \leq R \leq 1\)
- 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
Implementation of GLM in SPM

The Problem of Multiple Comparisons

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).

Advantages of General Linear Model (GLM)
**General Linear Model Approach**

\[ Y = \mathbf{X} \times \mathbf{\beta} + \mathbf{\varepsilon} \]

- Voxel timeseries data vector
- GLM design matrix parameters error vector

**Example:**
- Stimulus
- Subject
- Run
- Trial
- Group
- ROI
- Hand
- Hemi
- Tissue

**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?