

General Linear Model

Introduction, Classification and Estimation

An aim of scientific enquiry:

- To describe or to discover relationships among events (variables) in the controlled (laboratory) conditions or in the real world.

The underlying purpose may be to

- Prediction of future events (outcome),
- Control the outcome of future events,
- Develop understanding of the underlying phenomena
- Test some specified hypotheses.

- We must build a MODEL (a mathematical equation) involving deterministic (controlled) variables, stochastic variables, as well as unknown parameters.
 - Note that the assumptions about the probability distribution of the random variables are considered part of the model.

- The unknown model parameters can be estimated (learned) from the available training data.

- **Proposition:** Suppose that y is a quantity in the real world (response) about which we want to learn. We assume that *there exists*
 - a finite, although possibly large, collection of variables $\{x_1, x_2, \dots, x_r\}$,
 - a function g , such that $y = g(x_1, x_2, \dots, x_r)$.

Thus, y and $\{x_1, x_2, \dots, x_r\}$ are assumed to be functionally related.

- We are not implying that these variables are all observable, and/or the function g is known.
- When this relationship is not exact, it may be a good and useful *approximation* to the exact model. This is called a statistical model.

General Linear Models: A popular class of linear models is known as:

- Signal plus noise model - These variables are known and observable, but g may be unknown and one is willing to assume that

$$g(x_1, x_2, \dots, x_k) = \mu(x_1, x_2, \dots, x_k) + \varepsilon,$$

where, the signal $\mu(x_1, x_2, \dots, x_k)$ is known up to a set of unknown parameters and the additive error ε acts as a random (uncontrollable or unexplainable) noise.

- Sometimes not all the x_i 's, that determine the response y , may be known. One can assume that $g(x_1, x_2, \dots, x_k) = \mu(x_1, x_2, \dots, x_p) + \eta(x_{p+1}, x_{p+2}, \dots, x_k)$, and for conditional on the values of key explanatory variables x_1, x_2, \dots, x_p , the quantities $x_{p+1}, x_{p+2}, \dots, x_k$ change so that $\eta(x_{p+1}, x_{p+2}, \dots, x_k)$ behaves like a random error ε . This error is called the equation error or the specification error.

- In certain other situations, the underlying response, y^* , itself may not be observable exactly. Instead, we measure $y = y^* + \varepsilon$, where ε denote the measurement error.

- For simplicity, one can write the model for the measurement (observable quantity) Y as

$$Y = \mu(x_1, x_2, \dots, x_r) + \varepsilon, \quad (\text{P})$$

So that, the random noise ε could be either the specification error, or the measurement error, or a mixture of both these errors.

- If the errors are not additive, sometimes these models are called Generalized linear models (GLIM), e.g., logistic regression models, etc.

General Linear Model (GLM):

- The Population Model (P) where,
 - Y and ε are random variables,
 - x_1, x_2, \dots, x_r are deterministic variables, and
 - **The mean response function is linear in unknown parameters** $\{\beta_1, \beta_2, \dots, \beta_r\}$, i.e., for all (x_1, x_2, \dots, x_r) ,

$$\mu(\mathbf{x}) = \sum_{j=1}^p \beta_j f_j(x_1, x_2, \dots, x_r).$$

Here, the *features* f_i 's are assumed to be completely *known* functions of x_1, x_2, \dots, x_r . In engineering, one talks about feature extraction process, which searches for appropriate features to describe the response. In statistical science, this is called model selection.

- The variable Y is called
 - a **response** variable, or
 - an *endogenous* variable.
- The variables x_1, x_2, \dots, x_r [or the features f_i 's] are called
 - the *independent variables*,
 - the *explanatory variables*,
 - the *predictor variables*, or
 - the *exogenous variables*.

Broad Classification of General Linear Models:

Linear Regression Models: $(Y, X_1, X_2, \dots, X_r)$ are a set of jointly distributed random variables, such that $E[Y | X_1, X_2, \dots, X_r] = \mu(\mathbf{x}) = \sum_{j=1}^p \beta_j f_j(x_1, x_2, \dots, x_r)$, and expression (1) above holds. For analysis purposes, we treat the regression models as particular case of the GLM. Here, we are segmenting (stratifying) the whole population based on the values of the variables (X_1, X_2, \dots, X_r) and studying the conditional expectation of the response variable as a function of these variables. [Why expectation? Why not some quantile?]

Experimental Design Models: If the x_1, x_2, \dots, x_r in the GLM are qualitative, or categorical levels of certain factors or traits under study. These levels are represented as $\{0,1\}$ or $\{-1,1\}$ indicating presence or absence of traits, and the GLM is called an experimental design or ANOVA model.

In the past, these experiments were analyzed separately, because the analysis can be simplified quite a bit due the underlying structure in the set of explanatory variables. These days research effort is devoted to finding optimal designs that allow optimal estimation of a specified set of parametric functions, based on some optimality criterion.

These models are called the fixed effects models. Examples include,

- One-Way, Two-Way, Cross-classified multifactor experiments or nested designs.
- When the experiment includes some design variables, as well as some continuous explanatory variables, these models are called ANCOVA models.

Variance Component Models (Random Effect Models): In many experiments, the levels of a factor are assumed to be randomly drawn from a population of levels. The effects due to this factor are (unobserved) random variables, following a distribution with mean zero and unknown variance. These are called random effect models.

Mixed Effects Models: The mixed effects models have some factors with fixed effects and some that have random effects.

Remark: Functionally related variables, when all the variables are subject to measurement errors, are called Error-in-variables models. These should not be treated as a particular case of the GLM.

- Learning from Data: We wish to learn (make inference: estimate, test hypotheses) about the unknown parameters $\beta_1, \beta_2, \dots, \beta_p$ based on a Training Set (Sample).
- Start with the Sample model for the observations in the training set. But, it is assumed that the *population model* is valid, to enable us to relate the response y to x_1, x_2, \dots, x_r for unobserved or out-of-sample units in the population.
- Training Sample Model: Given n observations $[(Y_i, \mathbf{x}_i), \mathbf{x}_i = (x_{i1}, \dots, x_{ir})]$, $i = 1, 2, \dots, n$, the sample model can be expressed as

$$Y_i = \mu(x_{i1}, x_{i2}, \dots, x_{ir}) + \varepsilon_i, i = 1, 2, \dots, n, \quad (1)$$

where, $\varepsilon_i, i = 1, 2, \dots, n$, denote the noise (random errors), each with mean zero and variance σ^2 . Clearly,

$$E[Y_i | \mathbf{x}_i] = \mu(x_{i1}, x_{i2}, \dots, x_{ir}), i = 1, 2, \dots, n. \quad (2)$$

- From now on, we denote the features $f_j, j = 1, 2, \dots, p$, themselves as coded predictor variables x_1, x_2, \dots, x_p . In the simplest setting, the random errors are also assumed to be uncorrelated. Thus the sample GLM can be expressed as

$$Y_i = \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i, E[\varepsilon_i] = 0, Var[\varepsilon_i] = \sigma^2, Cov(\varepsilon_i, \varepsilon_k) = 0, i \neq k. \quad (3)$$

$$E[Y_i | \mathbf{x}_i] = \sum_{j=1}^p \beta_j x_{ij}. \quad (4)$$

- Examples:
 - Simple linear regression model
 - Multiple linear regression model,
 - Polynomial regression model,
 - One-way fixed-effect ANOVA model,
 - One-way random-effects ANOVA model.
- Vector/matrix notation for the response, predictor variables, error terms and the unknown coefficients:

$$\mathbf{Y}_{n \times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X}_{n \times p} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_p], \text{ where } \mathbf{x}_j = \begin{bmatrix} x_{1j} \\ x_{2j} \\ \vdots \\ x_{nj} \end{bmatrix}, \boldsymbol{\beta}_{p \times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} \text{ and } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

- Given the response vector \mathbf{Y} , and the design matrix \mathbf{X} , the sample GLM can be written in matrix notation as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, E[\boldsymbol{\varepsilon}] = 0, Cov[\boldsymbol{\varepsilon}] = ((cov(\varepsilon_i, \varepsilon_j))) = \sigma^2 \mathbf{I}. \quad (5)$$

- For this model, $E[\mathbf{Y}] = \mathbf{X}\boldsymbol{\beta}, Cov(Y) = \sigma^2 I.$ (6)

- In order to use this model for prediction of future response given a set of predictor values, the unknown parameter needs to be estimated (learned) from the training sample.
- For any reasonable estimate $\tilde{\boldsymbol{\beta}}$ of the vector $\boldsymbol{\beta}$, estimated errors (residuals) $e_i = (Y_i - \sum_{j=1}^p \tilde{\beta}_j x_{ij})$ must be as small as possible. In general, the choice of $\tilde{\boldsymbol{\beta}}$ is based on solving an optimization problem: Minimize a loss function $l(\tilde{\boldsymbol{\beta}})$, an implicit function of the estimated errors, $\{e_1, e_2, \dots, e_n\}$, that tends to keep the errors as small as possible. For example,

- $l_1(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n |e_i|$, or $l_2(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n e_i^2$, $l_w(\tilde{\boldsymbol{\beta}}) = \sum_{i=1}^n w_i e_i^2$, where

- l_1 : The absolute error (L^1 loss) criterion
- l_2 : The ordinary least squared error (L^2 loss) criterion, and
- l_w : The weighted least squares error criterion.

- Ordinary Least Square (OLS) Criterion: Find the estimated coefficient vector $\hat{\beta} = \arg \min_{\tilde{\beta} \in R^p} l_2(\tilde{\beta})$, that minimizes the sum of squared errors, i.e.,

$$\min_{\tilde{\beta} \in R^p} l_2(\tilde{\beta}) = \min_{\tilde{\beta} \in R^p} \sum_{i=1}^n (Y_i - \sum_{j=1}^p \tilde{\beta}_j x_{ij})^2.$$

- Historically, the LS criterion has been popular, since one could find its solution analytically. Therefore, its statistical properties can be studied easily.
 - The absolute error loss criterion required solving a linear programming problem, so it is difficult to derive its statistical properties.
- Nowadays, regularized versions (minimization subject to some upper bound on the size of the vector β) of both these criteria are popular in data mining applications. For example,
 - Ridge regression: Minimize the squared error loss subject to an upper bound on the L^2 -norm of the coefficient vector,
 - LASSO: Minimize the squared error loss subject to an upper bound on the L^1 -norm of the coefficient vector.
- Note that the OLS criterion is equivalent to minimizing the residual sum of squares, i.e.,

- $$\min_{\tilde{\beta} \in R^p} \mathbf{e}'\mathbf{e} = \min_{\tilde{\beta} \in R^p} (\mathbf{Y} - \mathbf{X}\tilde{\beta})'(\mathbf{Y} - \mathbf{X}\tilde{\beta}).$$

- Expand $S(\beta) = (\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\beta - \beta'\mathbf{X}'\mathbf{Y} + \beta'\mathbf{X}'\mathbf{X}\beta$.
- In order to be able to write these models in a compact notation, we need to have some background in linear algebra. In the next few lectures, we will review some of these tools.