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

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

- A MODEL (a mathematical equation) involving deterministic (controlled) variables, stochastic variables, as well as unknown parameters, will be useful in performing this task.

- Note: Assumptions about the probability distribution of the stochastic variable may be made. These are considered part of the model.

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

**Proposition:** Suppose that \( y \) denotes a quantity in the real world (response) about which we want to learn. We assume that there exists

- a finite, possibly large, collection of variables \( \{x_1, x_2, \ldots, x_r\} \), called explanatory variables, and
- a function \( g \), such that \( y = g(x_1, x_2, \ldots, x_r) \).

Thus, \( y \) and \( \{x_1, x_2, \ldots, x_r\} \) are assumed to be functionally related.

- We are not implying that these variables are known and 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.
Signal plus noise model: A popular class of linear models

- The explanatory variables are known and observable, but \( g \) may be unknown and one is willing to assume that
  \[
  g(x_1, x_2, \cdots, x_k) = \mu(x_1, x_2, \cdots, x_k) + \varepsilon,
  \]
  such that the signal \( \mu(x_1, x_2, \cdots, x_k) \) is known up to a set of unknown parameters and the additive error \( \varepsilon \) acts as a random (uncontrollable or unexplainable) noise.

- Sometimes not all the \( x_i \)'s, that determine the response \( y \), may be known. However, one can assume that
  \[
  g(x_1, x_2, \cdots, x_k) = \mu(x_1, x_2, \cdots, x_p) + \eta(x_{p+1}, x_{p+2}, \cdots, x_k),
  \]
  where, conditional on the values of key explanatory variables \( x_1, x_2, \cdots, x_p \), the quantities \( x_{p+1}, x_{p+2}, \cdots, x_k \) change so that \( \eta(x_{p+1}, x_{p+2}, \cdots, x_k) \) behaves like a random error \( \varepsilon \). 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 \( \varepsilon \) denote the measurement error.

- For simplicity, one can write the model for the measurement (observable quantity) \( Y \) as
  \[
  Y = \mu(x_1, x_2, \cdots, x_r) + \varepsilon, \tag{P}
  \]
  Thus, the random noise \( \varepsilon \) 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 \( \varepsilon \) are random variables,
  - \( x_1, x_2, \ldots, x_r \) are deterministic variables, and
  - The mean response function \( E\{Y \mid x\} = \mu(x) \), is linear in unknown parameters \( \{\beta_1, \beta_2, \ldots, \beta_r\} \), i.e., for all \( (x_1, x_2, \ldots, x_r) \in X \),
    \[
    \mu(x) = \sum_{j=1}^{p} \beta_j f_j(x_1, x_2, \ldots, x_r).
    \]

Here, the features \( f_i \)'s are assumed to be completely known functions of \( x_1, x_2, \ldots, x_r \). 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 \( x_1, x_2, \ldots, x_r \) [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: \((Y, X_1, X_2, \ldots, X_r)\) are a set of jointly distributed random variables, such that \( E[Y \mid X_1, X_2, \ldots, X_r] = \mu(x) = \sum_{j=1}^{p} \beta_j f_j(x_1, x_2, \ldots, x_r) \), and expression (P) above holds, e.g., Simple linear regression, or Multiple linear regression. 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 \((X_1, X_2, \ldots, X_r)\) and studying the conditional expectation of the response variable as a function of these variables. [Why do we choose conditional mean, not some quantile?]
Experimental Design Models: Each explanatory variable in the GLM is quantitative or categorical levels of certain factors or traits under study. The categorical 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. Basically, we are interested in understanding major causes of variability in the response variable.

Classically, the analysis of such experiments could 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 Analysis of Covariance (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 of 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 \( \beta_1, \beta_2, \ldots, \beta_p \) 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 \( x_1, x_2, \ldots, x_r \) for unobserved or out-of-sample units in the population.
Training Sample Model: Given n observations \([(Y_i, x_i) \text{, } x_i = (x_{i1}, \cdots, x_{ip})]\), \(i = 1, 2, \cdots, n\), the sample model can be expressed as
\[
Y_i = \mu(x_{i1}, x_{i2}, \cdots, x_{ip}) + \varepsilon_i, \quad i = 1, 2, \cdots, n,
\]
where \(\varepsilon_i, i = 1, 2, \cdots, n\), denote the noise (random errors), each with mean zero and equal variance \(\sigma^2\). Clearly,
\[
E[Y_i | x_i] = \mu(x_{i1}, x_{i2}, \cdots, x_{ip}), \quad i = 1, 2, \cdots, n.
\]

From now on, we denote the features \(f_j, j = 1, 2, \cdots, p\), themselves as coded predictor variables \(x_1, x_2, \cdots, x_p\). In the simplest setting, the random errors are also assumed to be uncorrelated. Thus the sample GLM can be expressed as
\[
Y_i = \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad E[\varepsilon_i] = 0, Var[\varepsilon_i] = \sigma^2, Cov(\varepsilon_i, \varepsilon_k) = 0, i \neq k.
\]

Examples:
- Simple linear regression model
- Multiple linear regression model,
- Polynomial regression model,
- One-way fixed-effect ANOVA model,
- One-way random-effects ANOVA model.

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 \(\hat{\beta}\) of the vector \(\beta\), estimated errors (residuals) in the response \(e_i = (Y_i - \sum_{j=1}^{p} \tilde{\beta}_j x_{ij})\) should be as small as possible. The choice of \(\hat{\beta}\) is based on solving an optimization problem:

Minimize a loss function \(l(\hat{\beta})\), an implicit function of the estimated errors, \(\{e_1, e_2, \cdots, e_n\}\), that tends to keep the errors as small as possible. For example,

\[
\sum_{i=1}^{n} |e_i|, \quad \sum_{i=1}^{n} e_i^2, \quad \sum_{i=1}^{n} w_i e_i^2, \quad \text{where}
\]
\( l_1 \): 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 \( \hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} l_2(\bar{\beta}) \), that minimizes the sum of squared errors, i.e.,

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

- Historically, the LS criterion has been popular, since one could find its solution analytically as well as geometrically. Therefore, its statistical properties can be studied easily.

  - The absolute error loss criterion required solving a linear programing problem, thus it was difficult to derive its statistical properties analytically.

- Nowadays, regularized versions (minimization subject to some upper bound on the size of the vector \( \beta \) ) 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 without a concise notation, it is tedious to express these quantities.

- **Vector/matrix notation for the response, predictor variables, error terms and the unknown coefficients:**

\[
Y_{nx1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X}_{nxp} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_p], \quad \text{where} \quad \mathbf{x}_j = \begin{bmatrix} x_{1j} \\ \vdots \\ x_{nj} \end{bmatrix}, \quad \beta_{p \times 1} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{and} \quad \mathbf{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.
\]
- Given the response vector $Y$, and the design matrix $X$, the sample GLM can be written in matrix notation as

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

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

- Thus the OLS criterion is equivalent to minimizing the residual sum of squares, i.e.,

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

- Expand $S(\beta) = (Y - X\beta)'(Y - X\beta) = Y'Y - Y'X\beta - \beta'X'Y + \beta'X'X\beta.$

- 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.