General Linear Model

K. Gibert [1]

[1] *Department of Statistics and Operation Research*

Knowledge Engineering and Machine Learning group at
Intelligent Data Science and Artificial Intelligence research center
Universitat Politècnica de Catalunya, Barcelona

Statistical Modelling

Data= Fit+Error

- Fit:
  - Structural
  - Law governing the phenomenon
  - Analytic Function

- Error:
  - Random
  - Variability arround Fit (null expectation)
  - Probabilistic model

©K. Gibert

Statistical models

- Determine the family of fits:
  - Linear
  - Quadratic
  - Exponential
  - .....

- Determine the law of error:
  - Normal
  - Poisson
  - Binomial....

General Linear Model

- Family of models based on :
  - Response variable: Continuous
  - Explanatory variables: Continous or Categorical
  - Model: Linear
  - Estimation method: Mean least squares

- Particular cases of General Linear Model Family
  - t-Student hypothesis test
  - Linear simple regression
  - Linear mutiple regression
  - ANOVA
  - ANCOVA

General Linear Model

- *Formalization:*

  I=i:n observations

  Y: Response variable

  $X_1 \ldots\ldots X_K$ *: ExplanatoryVariables*

  Find $\beta_{0\ldots}\ \beta_K$ such that

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K + \varepsilon$$

- *Assumptions:*
  - *Linearity:* $E(Y \mid X = x) = \mu_{y|X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K$; $E[\varepsilon] = 0$

    Population regression line
  - *Normality:* $\varepsilon_1, \ldots, \varepsilon_n \sim \mathcal{N}(0, \sigma_i)$, i=1:n
  - *Homokedasticity:* $Var[\varepsilon_i] = \sigma^2$ forall i
  - *Independence:* $Cov(\varepsilon_i, \varepsilon_j) = 0$ forall i,,j

Two Sample t Student test

A particular case of General linear model

- *T-Test*

Two samples $X_A \sim \mathcal{N}(\boldsymbol{\mu_A}, \boldsymbol{\sigma_A})$, $X_B \sim \mathcal{N}(\boldsymbol{\mu_B}, \boldsymbol{\sigma_B})$ *of size* n_A *and* n_B $\boldsymbol{\sigma_A = \sigma_B}$

$$\boldsymbol{t = \frac{(\overline{X_A} - \overline{X_B}) - (\mu_A - \mu_B)}{\sqrt{\frac{S_A^2}{n_A} - \frac{S_B^2}{n_B}}} \sim t_{(n_A + n_B - 2)}}$$

- *Equivalent linear model*

n= $n_A + n_B$ *observations;* *Y=* (X_A, X_B)*;* X =*(0 (*n_A *times) , 1 (*n_B *times))*

$$Y = \beta_0 + \beta_1 X + \mathcal{E}, \qquad \varepsilon_n \sim \mathcal{N}(0,\sigma)$$

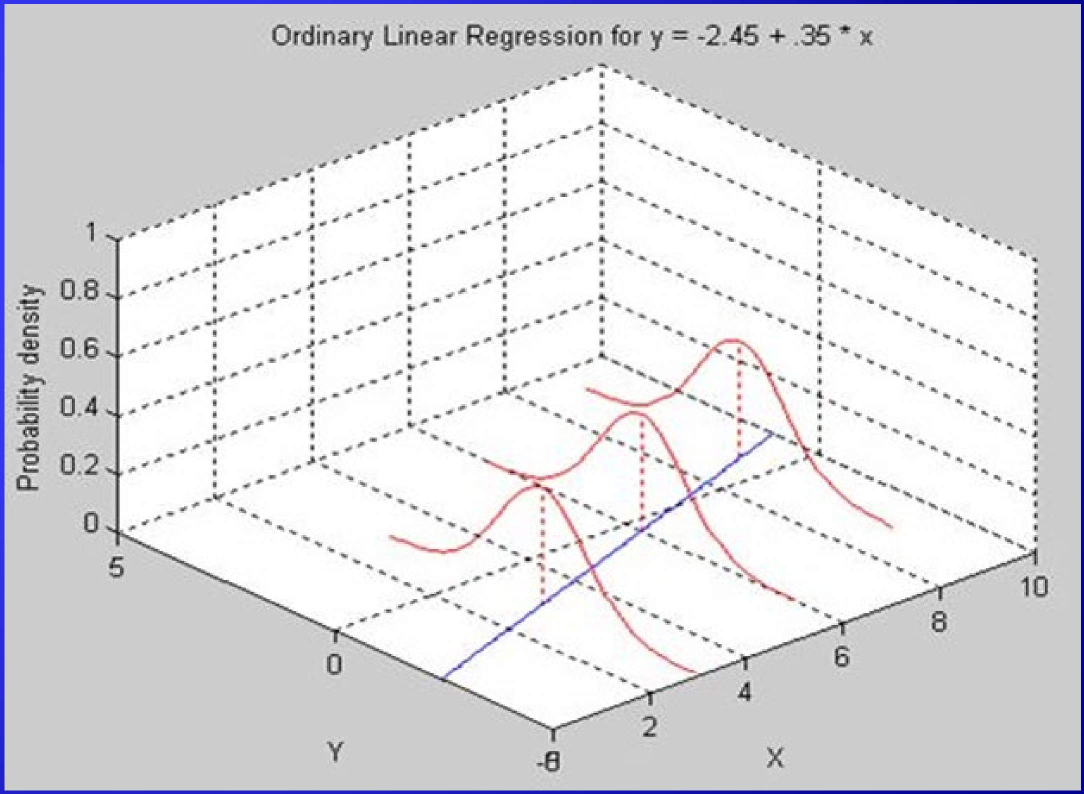
$$\hat{\beta}_1 = \boldsymbol{\overline{X_A} - \overline{X_B}} \qquad t_1 = \frac{\hat{\beta}_1}{S_{\hat{\beta}_1}} \sim t_{n-2}$$

Multiple Linear Regression

A particular case of General Linear Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0,\sigma), \; X \text{ cont}$$

- All properties of simple linear regression hold, except one
  - Determination coefficient (R^2) still measures goodness of fit
  - But R^2 is not equal to the squared correlation coefficient in multiple regression
- Analysis of the residuals is done for every regressor X_k



ANOVA test

A particular case of General linear model

- *ANOVA-Test*

  K samples $X_k \sim \mathcal{N}(\boldsymbol{\mu_k}, \boldsymbol{\sigma_k})$ *size* n_k, $\boldsymbol{\sigma_k} = \boldsymbol{\sigma}$ *forall k in 1:K*

$$\boldsymbol{F} = \frac{\frac{s_B{}^2}{K-1}}{\frac{s_W{}^2}{n-K}} \sim \boldsymbol{F_{K-1,n-K}}, \qquad S_B^2 = \sum_{k=1}^{q} n_k (\overline{x}_k - \overline{x})^2 \qquad S_W^2 = \sum_{k=1}^{q} \sum_{i=1}^{n_k} (x_{ki} - \overline{x}_k)^2$$

- *Equivalent linear model*

  n= $n_A + n_B$ *observations;* $Y = (X_1, X_2 \ldots\ldots X_K)$*;* X_{ki} *=(1 if i= Group k, else 0)*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_{K-1} + \varepsilon$$

Test F is the same as ANOVA test and indicates significance of model

The β_0 *represents effect of* X_K

ANCOVA test

A particular case of General linear model

- *Equivalent linear model*

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_K X_K + \varepsilon$$

Where X_k can be

- Numerical
- A Dummy of a category of a qualitative variable

All categories excepte one of a given qualitative variable enters into the model

Matricial formulation

Regression fit criterion: $\min_r E\left[\left(y_i - r(x_{i1},\cdots,x_{ip})\right)^2\right]$

$$r(x_{i1},\cdots,x_{ip}) = E\left[y_i \middle| x_{i1},\cdots,x_{ip}\right]$$

$$E\left[y_i \middle| x_{i1},\cdots,x_{ip}\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

Estimation of coefficients

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + e_i$$

$$\mathrm{var}(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & . & 0 \\ 0 & \sigma^2 & . & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & . & \sigma^2 \end{pmatrix}$$

In matrix notation

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1\, x_{11} \cdots x_{1p} \\ \vdots \quad \vdots \quad \ddots \quad \vdots \\ 1\, x_{n1} \cdots x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix} \equiv y = Xb + e = \hat{y} + e$$

Mean least squares solution $\hat{\beta} = (X^T X)^{-1} X^T Y$

Validation

- Technical Assumptions
  - normality, linearity, independence, homokedasticity

  - Tools
    - Graphical residuals analysis
    - Influence-point indicators (hi)

- Quality:
  - R2 (determination coeficient): goodness, reliability
  - s-2: noise, precision
  - Both guarantee generalizability (only interpolation)