

General Linear Model (GLM):

- Vector differentiation:

Vector of Partial derivatives of a linear form $\mathbf{l}'\mathbf{u} = \sum_{i=1}^p l_i u_i$, and quadratic form $\mathbf{u}'\mathbf{A}\mathbf{u} = \sum_{i=1}^p \sum_{j=1}^p a_{ij} u_i u_j$,

where \mathbf{A} is a symmetric matrix.

$$\frac{\partial}{\partial \mathbf{u}} (\mathbf{l}'\mathbf{u}) = \begin{pmatrix} \frac{\partial}{\partial u_1} (\mathbf{l}'\mathbf{u}) \\ \frac{\partial}{\partial u_2} (\mathbf{l}'\mathbf{u}) \\ \vdots \\ \frac{\partial}{\partial u_p} (\mathbf{l}'\mathbf{u}) \end{pmatrix} = \begin{pmatrix} l_1 \\ l_2 \\ \vdots \\ l_p \end{pmatrix} = \mathbf{l}; \quad \frac{\partial}{\partial \mathbf{u}} (\mathbf{u}'\mathbf{A}\mathbf{u}) = \begin{pmatrix} \frac{\partial}{\partial u_1} (\mathbf{u}'\mathbf{A}\mathbf{u}) \\ \frac{\partial}{\partial u_2} (\mathbf{u}'\mathbf{A}\mathbf{u}) \\ \vdots \\ \frac{\partial}{\partial u_p} (\mathbf{u}'\mathbf{A}\mathbf{u}) \end{pmatrix} = \begin{pmatrix} 2a_{11}u_1 + 2\sum_{j \neq 1} a_{1j}u_j \\ 2a_{22}u_2 + 2\sum_{j \neq 2} a_{2j}u_j \\ \vdots \\ 2a_{nn}u_n + 2\sum_{j \neq n} a_{nj}u_j \end{pmatrix} = 2 \begin{pmatrix} \mathbf{a}'_1 \mathbf{u} \\ \mathbf{a}'_2 \mathbf{u} \\ \vdots \\ \mathbf{a}'_n \mathbf{u} \end{pmatrix} = 2\mathbf{A}\mathbf{u}.$$

When \mathbf{A} is not symmetric, $\mathbf{u}'\mathbf{A}\mathbf{u} = \mathbf{u}'\{(\mathbf{A} + \mathbf{A}')/2\}\mathbf{u}$, with $\{(\mathbf{A} + \mathbf{A}')/2\}$ symmetric. Therefore,

$$\frac{\partial}{\partial \mathbf{u}} (\mathbf{u}'\mathbf{A}\mathbf{u}) = (\mathbf{A} + \mathbf{A}')\mathbf{u}.$$

- 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 .

- 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

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

$$E[Y_i | \mathbf{x}_i] = \sum_{j=1}^p \beta_j x_{ij}. \quad (3)$$

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

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

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

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, E[\boldsymbol{\varepsilon}] = 0, \text{Cov}[\boldsymbol{\varepsilon}] = ((\text{cov}(\varepsilon_i, \varepsilon_j))) = \sigma^2 \mathbf{I}. \quad (4)$$

- For this model, $E[\mathbf{Y}] = \boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}, \text{Cov}(\mathbf{Y}) = \sigma^2 \mathbf{I}.$ (5)

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

$$\min_{\tilde{\boldsymbol{\beta}} \in R^p} l_2(\tilde{\boldsymbol{\beta}}) = \min_{\tilde{\boldsymbol{\beta}} \in R^p} \sum_{i=1}^n (Y_i - \sum_{j=1}^p \tilde{\beta}_j x_{ij})^2, \text{ or, in matrix notation,}$$

$$\bullet \min_{\tilde{\boldsymbol{\beta}} \in R^p} \mathbf{e}'\mathbf{e} = \min_{\tilde{\boldsymbol{\beta}} \in R^p} (\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \quad (6)$$

- Expand $S(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}.$
- On setting the partial derivatives of $S(\boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$ equal to zero, we get the **normal equations**

$$\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{Y}. \quad (7)$$

- Any solution of these equations is an optimal solution to the least square problem.
- **Full rank case:** If the matrix $\mathbf{X}'\mathbf{X}$ is non-singular (the design matrix \mathbf{X} is of full column rank p), inverse of $\mathbf{X}'\mathbf{X}$ exists, and the unique optimal least square solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y}. \quad (8)$$

- Since, $\text{Cov}(\mathbf{T}\mathbf{Y}) = \mathbf{T}\text{Cov}(\mathbf{Y})\mathbf{T}'$, the covariance matrix of the LS estimator $\hat{\boldsymbol{\beta}}$ is given by $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}.$

- **Example** – Simple Linear Regression
- **Not full-rank case:** Example
- There are infinitely many solutions to the normal equations (8). All of these can be expressed $\beta^0 = (\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{Y}$, where $(\mathbf{X}'\mathbf{X})^-$ is some generalized inverse of $\mathbf{X}'\mathbf{X}$.
 - All these can't be regarded as the optimal estimator of the vector β . Why?
 - Minimum norm least square estimator: $\beta^+ = (\mathbf{X}'\mathbf{X})^+ \mathbf{X}'\mathbf{Y}$, where $(\mathbf{X}'\mathbf{X})^+$ is the Moore-Penrose generalized inverse (Pseudo-inverse) of $\mathbf{X}'\mathbf{X}$.
- **Coordinate Free Approach:** The model $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$ can be interpreted as $\boldsymbol{\mu} \in C[\mathbf{X}]$. However, note that $\hat{\boldsymbol{\mu}} = \mathbf{X}\boldsymbol{\beta}^0 = \mathbf{P}\mathbf{Y} = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{Y}$. The symmetric matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'$ is the orthogonal projection matrix onto $C[\mathbf{X}]$, the space spanned by the columns of \mathbf{X} .
- The projection $\hat{\boldsymbol{\mu}} = \mathbf{P}\mathbf{Y}$ (also called $\hat{\mathbf{Y}}$) is uniquely defined, even though there are infinitely many solutions to the normal equations in non-full rank case. Thus the matrix \mathbf{P} does not change with the choice of a generalized inverse of $\mathbf{X}'\mathbf{X}$.
- For a vector $\mathbf{u} \in C[\mathbf{X}]$, i.e., $\mathbf{u} = \mathbf{X}\mathbf{b}$, for some vector \mathbf{b} ,

$$\mathbf{P}\mathbf{u} = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{u} = \mathbf{X}(\mathbf{X}'\mathbf{X})^- \mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}\mathbf{b} = \mathbf{u}.$$
 - Thus the projection of a vector $\mathbf{u} \in C[\mathbf{X}]$ onto $C[\mathbf{X}]$ is \mathbf{u} itself.
- Furthermore, for an arbitrary vector $\mathbf{Y} \in V_n$, $\mathbf{P}\mathbf{Y} \in C[\mathbf{X}]$, therefore, $\mathbf{P}(\mathbf{P}\mathbf{Y}) = \mathbf{P}\mathbf{Y}$ holds true for all $\mathbf{Y} \in \mathbb{R}^n$. Hence, $(\mathbf{P}^2 - \mathbf{P}) = \mathbf{0} \Leftrightarrow \mathbf{P}(\mathbf{I} - \mathbf{P}) = \mathbf{0}$. Thus, $\mathbf{P}^2 = \mathbf{P}$, (i.e., \mathbf{P} is an idempotent matrix). It can also be shown that \mathbf{P} is symmetric.
 - In fact, every symmetric idempotent matrix is an orthogonal projection matrix onto the space spanned by its columns.
- Since $\mathbf{P}(\mathbf{I} - \mathbf{P}) = \mathbf{0}$, rows (columns) of \mathbf{P} are orthogonal to the columns(rows) of $(\mathbf{I} - \mathbf{P})$, i.e., $\mathbf{P}\mathbf{y}$ and $(\mathbf{I} - \mathbf{P})\mathbf{y}$ are orthogonal. Note that since,

$\mathbf{X}\beta^0 = \mathbf{P}\mathbf{Y} = \hat{\mathbf{Y}}$, $\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{P})\mathbf{Y} = \mathbf{e}$, the vector of residuals, it follows that the vectors $\hat{\mathbf{Y}}$ and \mathbf{e} are orthogonal, *i.e.*, $\hat{\mathbf{Y}}'\mathbf{e} = \sum_{i=1}^n \hat{y}_i e_i = \mathbf{0}$.

- Example: Details of Simple linear regression model.
- In the Not-full rank case, there are infinitely many least squares solutions β^0 , even though $\hat{\mathbf{Y}} = \mathbf{X}\beta^0$ is unique. Not all can be optimal.
 - **The Key Question is:**
 - What is the class of linear functions $\mathbf{c}'\beta = \sum_{j=1}^p c_j \beta_j$ that can be estimated uniquely through the least squares solutions?
- **Estimable functions:** A linear function $\mathbf{c}'\beta$ is said to be estimable, if there exist at least one unbiased estimator. Why Estimable functions? We will discuss its connection with the concept of *Identifiability*.
- Note that, Y_i is an unbiased estimator of $\mathbf{x}'_{[i]}\beta$, where $\mathbf{x}'_{[i]}$ is the i^{th} row of \mathbf{X} . Thus $\mathbf{x}'_{[i]}\beta$ is estimable for each row of the matrix \mathbf{X} . Hence, $\mathbf{c}'\beta$ where the vector \mathbf{c}' is some linear combination of rows of \mathbf{X} is also estimable. In fact, these are the only linear functions of β_j 's that are estimable. (Prove it.)
- **Gauss-Markov Theorem** - $\mathbf{c}'\beta^0$ Best Linear Unbiased Estimator (B.L.U.E.) of an estimable function $\mathbf{c}'\beta$
- **Generalized Least Squares:** $\text{Var}(\varepsilon) = \sigma^2\mathbf{V}$, where \mathbf{V} is a known p.d. matrix. Reduce this problem to the OLS problem by a non-singular transformation.