Linear Models

by Kiranmoy Chatterjee.

An important question that one often tries to answer through statistical models is the following: How can an observed quantity y be explained by a number of other quantities, $x_1, x_2, \ldots, x_{p-1}$? Perhaps, the simplest model that is used to answer this question is the linear model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon,$$

where $\beta_0, \beta_1, \ldots, \beta_p$ are constants and ϵ is an error term that accounts for uncertainties. For a set of n observations on y and x's the explicit form of the above equation is $y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi} + \epsilon_i$, $i = 1(1)n$.

The general linear model is assumed to be the following matrix form:

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \qquad \ldots\ldots\ldots \quad (1)$$

where

$$\underline{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \underline{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \underline{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Here $\underline{y}$ is the vector of n observations, hence known, on response variable,

X is the $n \times p$ order matrix, called as design matrix, hence known,

$\underline{\beta}$ is the vector of $p+1$ unknown model parameters and

$\underline{\epsilon}$ is the vector of ~~random~~ errors representing the irregular components of the observed values $\underline{y}$.

Characteristically, $\underline{\beta}$ and design matrix X are fixed since they together constitute the non-random regular components of the observed values $\underline{y}$.

Obviously, there are discrepancies between the observed $\underline{y}$ and the regular fixed component $X\underline{\beta}$. $\underline{\epsilon}$ refers the vector of such discrepancies i.e. $\underline{\epsilon} = \underline{y} - X\underline{\beta}$ and termed as 'error'.

Now, a primary objective is to make inference – either point or interval estimation or statistical hypothesis testing, on the model parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

To make statistical inference, we need to consider the observed $\underline{y}$ as the realization (on random response variable Y) of a random sample of n individuals from our population of interest. Therefore, we assume $\underline{\epsilon}$ to be random vector with $E(\underline{\epsilon}) = \underline{0}$ and

$$V(\underline{\epsilon}) = \sigma^2 I_n = \text{Diag}(\underbrace{\sigma^2, \sigma^2, \ldots, \sigma^2}_{n \text{ terms}}).$$

Since $\underset{n\times 1}{\underline{y}}$ is assumed to be the realizations of ~~independent~~ random variables, then probabilistically the model (1) can be reconsidered with same notation as

$$\underline{y} = X\underline{\beta} + \underline{\epsilon}, \quad \text{----------} \quad (2)$$

where $\underline{y}$ is a random vector of n components and the randomness of $\underline{y}$ is modelled only through $\underline{\epsilon}$ as $X\underline{\beta}$ is ~~left~~ left as fixed part. If the random variables $(y_1, \ldots, y_n)'$ are assumed to be independent to each other then $\text{Cov}(y_i, y_j) = \text{Cov}(\epsilon_i, \epsilon_j) = 0$, then $V(\underline{\epsilon})$ will be a diagonal matrix, otherwise the off-diagonal elements of $V(\underline{\epsilon})$ will be non-zero.

Further, if the random variables $(y_1, y_2, \ldots, y_n)'$ are assumed to be homoscadastic; then $V(y_i) = V(\epsilon_i)$ will be identical for all $i = 1(1)n$. We assume that identical value as σ^2.

Hence for iid random variables $(y_1, y_2, \ldots, y_n)$ on interest variable Y, the general linear model is the model (2) with the following assumptions

(i) $E(\underline{\epsilon}) = \underline{0}$, (ii) $V(\underline{\epsilon}) = \sigma^2 I_n$.

Apart from making inference on $\beta_1, \beta_2, \ldots, \beta_p$ primarily, we may also be interested to make estimation or testing of some linear combination of $\beta_1, \beta_2, \ldots, \beta_p$. We may also wish to estimate or test regarding the variance parameter σ^2.

Usually, $n > p+1$ or $n >> p+1$, the no. of unknown model parameters, but we are not assuming this. Now, it is obvious that $\text{Rank}(X) \leq \text{Min}(n, p+1)$

If Rank $(X) = p+1 \leq n$, then the model (2) is said to be a 'Full Rank Model', otherwise it is described as a 'Non-full Rank Model'.

Now, in order to be estimate $\underline{\beta}$, we need to determine a vector $\hat{\underline{\beta}}$ such that $\underline{y}$ will be 'close' to its model part $X\underline{\beta}$. This closeness can simply be measured by the residuals, $\underline{e} = \underline{y} - X\hat{\underline{\beta}}$. Thus, the scalar quantity $\underline{e}'\underline{e}$ will be the least value among all values of $\underline{\epsilon}'\underline{\epsilon} = (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta})$, for different values of $\underline{\beta}$, i.e.

$$\underline{e}'\underline{e} = \min_{\underline{\beta}} \underline{\epsilon}'\underline{\epsilon} = \min_{\underline{\beta}} (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}).$$

$$\text{and} \quad \hat{\underline{\beta}} = \arg\min_{\underline{\beta}} (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}).$$

This method of estimating $\underline{\beta}$ is called 'Least square method'. To obtain $\hat{\underline{\beta}}$, the least square estimate of $\underline{\beta}$, we need to find the solution of $\underline{\beta}$ satisfying

$$\frac{d}{d\underline{\beta}} (\underline{\epsilon}'\underline{\epsilon}) = \frac{d}{d\underline{\beta}} (\underline{y}-X\underline{\beta})'(\underline{y}-X\underline{\beta}) = \underline{0} \quad \text{----(3)}$$

From (3), we have

$$\frac{d}{d\underline{\beta}} \left(\underline{y}'\underline{y} - 2\underline{\beta}'X'\underline{y} + \underline{\beta}'X'X\underline{\beta} \right) = \underline{0}$$

$$\Rightarrow \quad -2X'\underline{y} + 2(X'X)\underline{\beta} = \underline{0}$$

$$\Rightarrow \quad (X'X)\underline{\beta} = X'\underline{y} \quad \text{----------(4)}.$$

The (1+p) no. of equations in (4) are called 'Normal Equations'.

Now two basic questions are appeared at this stage of obtaining the solution for $\underline{\beta}$ and they are

(a) Does there exist any solution of equation (4)? If so, what is the necessary and sufficient condition (n.s.c) for a solution to exist?

(b) Does the condition in (a) always provide us the unique solution? If not so, then what is the n.s.c for unique solution?

To answer the question (a), theory of linear equations says a necessary and sufficient condition for a solution of equation (4) to exist is the 'consistency' condition i.e

$$\text{Rank}(X'X \mid X'\underline{y}) = \text{Rank}(X'X).$$

The above condition can be rewritten ~~simply~~ generally as

$$\text{Rank}(A) = \text{Rank}(A \mid \underline{u})$$

for any system of linear equations $\underset{n\times 1}{\underline{x}}$ $\overset{m\times n}{A}\underline{x} = \underline{u}$ for unknown vector $\underline{x}$ for which solution is to be found.

Any $n\times m$ matrix A^- satisfying the relation $AA^-A = A$ is called a 'Generalized Inverse' of the $m\times n$ matrix A and it has the property that $A^-\underline{u}$ is a solution of the equation $A\underline{x} = \underline{u}$, for every vector $\underline{u}$ satisfying $\text{Rank}(A) = \text{Rank}(A \mid \underline{u})$. In fact, one can obtain an infinite number of generalized inverses or g-inverses and hence, infinite number of solutions of the equations $A\underline{x} = \underline{u}$ involving unknown $\underline{x}$. Note that g-inverse exists for any matrix whether it is square or rectangular. In particular, here, $A = X'X$ is a square matrix and $\underline{u} = X'\underline{y}$. So, in terms of any g-inverse of $X'X$, denoted as $(X'X)^-$, the solution of (4) in $\underline{\beta}$ is not unique and that will be $\tilde{\underline{\beta}} = (X'X)^-X'\underline{y}$ for different $(X'X)^-$.

To answer the question (b), we require more strong condition on $X'X$ or X itself. Only if $X'X$ (or X) satisfies that condition then only we will have unique solution to (4). If $X'X$ is ~~of full rank~~ non-singular, then only inverse of $X'X$ exists and hence,

$$\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y} \qquad \text{---------- (5)}$$

becomes the unique solution to (4).

From matrix theory, we know, $\text{Rank}(A'A) = \text{Rank}(A)$ and further, if $\text{Rank}(X) = p+1 (\leq n)$, then $X'X$ is positive

definite (p.d.) and hence, it must be non-singular.

Thus, if X is of full-rank, $X'X$ is non-singular and the solution to (4) is given by (5).

Therefore, $\underline{e} = \underline{y} - X\hat{\underline{\beta}} = \underline{y} - \hat{\underline{y}}$ is the vector of residuals, $e_1 = y_1 - \hat{y}_1$, $e_2 = y_2 - \hat{y}_2$, ----, $e_n = y_n - \hat{y}_n$. The residual vector $\underline{e}$ estimates the vector of error terms ϵ in the model (2). This $\underline{e}$ can be used to check the validity of the fitted model $X\hat{\beta}$ and the attendant assumptions.

Example: [See Freund and Minton (1979, p.p. 36-39)] Consider the multivariate data presented in the following:

Sl. NO.	y	x_1	x_2
1	2	0	2
2	3	2	6
3	2	2	7
4	7	2	5
5	6	4	9
6	8	4	8
7	10	4	7
8	7	6	10
9	8	6	11
10	12	6	9
11	11	8	15
12	14	8	13

Now if we are to fit a linear model $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$, $i = 1(1)12$, to the data on response variable Y based on two explanatory variables x_1 and x_2, then the associated design matrix X in the linear model (2) can be written as

$$X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 2 & 2 & 2 & 4 & 4 & 4 & 6 & 6 & 6 & 8 & 8 \\ 2 & 6 & 7 & 5 & 9 & 8 & 7 & 10 & 11 & 9 & 15 & 13 \end{pmatrix}^T$$

and $\underline{\beta} = (\beta_0, \beta_1, \beta_2)^T$

So, $X'X = \begin{pmatrix} 12 & 52 & 102 \\ 52 & 395 & 536 \\ 102 & 536 & 1004 \end{pmatrix}$ and Rank$(X'X) = 3$, full rank

and $X'\underline{y} = \begin{pmatrix} 90 \\ 482 \\ 872 \end{pmatrix}$

Therefore, inverse of $X'X$ is computed as

$$(X'X)^{-1} = \begin{pmatrix} 0.97476 & 0.24290 & -0.22871 \\ 0.24290 & 0.16207 & -0.11120 \\ -0.22871 & -0.11120 & 0.08360 \end{pmatrix}$$

So, finally using (5) we obtain,

$$\hat{\underline{\beta}} = (X'X)^{-1} X'\underline{y} = \begin{pmatrix} 5.3754 \\ 3.0118 \\ -1.2855 \end{pmatrix}$$

Hence, the fitted ~~linear model~~ or predicted is value of $\underline{y}$ is $\hat{y} = X\hat{\underline{\beta}}$ and the fitted linear model is

$$y_i = 5.3754 + 3.0118 x_{1i} - 1.2855 x_{2i}, \quad i = 1(1)12. \blacksquare$$

Properties of the Least-Squares Estimator $\hat{\underline{\beta}}$:

The least-squares estimator $\hat{\underline{\beta}}$ in (5) was obtained without using the assumptions $E(\underline{y}) = X\underline{\beta}$ and $Var(\underline{y}) = \sigma^2 I_n$ which are equivalent to the assumptions already considered in model (2). So, it is clear that least-square estimation does not require any probabilistic assumptions on the response variable Y. However, from the following Theorem we will see that the assumptions made under the model (2), i.e. $E(\underline{\epsilon}) = 0$ (or equivalently, $E(\underline{y}) = X\underline{\beta}$) and $Var(\underline{\epsilon}) = \sigma^2 I_n$ (or equivalently, $Var(\underline{y}) = \sigma^2 I_n$), are needed actually to justify the least-square estimator $\hat{\underline{\beta}}$ as an estimator with good properties.

Theorem

In connection to the linear model $\underline{y} = X\underline{\beta} + \underline{\epsilon}$,

(i) if $E(\underline{y}) = X\underline{\beta}$, or equivalently, $E(\underline{\epsilon}) = \underline{0}$, then $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$ is an unbiased estimator for $\underline{\beta}$,

(ii) if $Cov(\underline{y}) = \sigma^2 I_n$, or equivalently, $Var(\underline{\epsilon}) = \sigma^2 I_n$, then $Var(\hat{\underline{\beta}}) = \sigma^2(X'X)^{-1}$, if the design matrix X has full rank.

Proof: $E(\hat{\underline{\beta}}) = E\left((X'X)^{-1}X'\underline{y}\right) = (X'X)^{-1}X'E(\underline{y}) = (X'X)^{-1}X'X\underline{\beta} = \underline{\beta}$.

$$Var(\hat{\underline{\beta}}) = Var\left\{(X'X)^{-1}X'\underline{y}\right\} = (X'X)^{-1}X'\,Var(\underline{y})\left[(X'X)^{-1}X'\right]'$$
$$= \sigma^2(X'X)^{-1}X'X(X'X)^{-1}$$
$$= \sigma^2(X'X)^{-1}.$$ ∎

Thus, if $E(\underline{y}) \neq X\underline{\beta}$, the model $\underline{y} = X\underline{\beta} + \underline{\epsilon}$ could still be fitted to the data, but in that case $\underline{\beta}$ will have poor properties such as 'lack of unbiasedness'. Further, if $Var(\underline{y}) \neq \sigma^2 I_n$, there may be some adverse effect on the estimator $\hat{\underline{\beta}}$. One can further check whether the estimator $\hat{\underline{\beta}}$ for $\underline{\beta}$ has the minimum variance among all unbiased estimators of $\underline{\beta}$, i.e. MVUE (minimum variance unbiased estimator) or $\hat{\underline{\beta}}$ may possess some other good properties involving its variance. We will consider this study later.

In addition to the above, one may also be interested on the distributional form of the estimator $\hat{\underline{\beta}}$. Note that $\hat{\underline{\beta}}$ is based on the random variable $\underline{y}$ only since X is non-random and fixed. Thus, distribution of $\hat{\underline{\beta}}$ can be obtained from distribution of $\underline{y}$ only and the distribution of $\underline{y}$ comes from the assumption of distributional form of $\underline{\epsilon}$. We will also study this later.

Another issue is that, when the design matrix X, hence $X'X$, does not have full rank. In that case all possible solutions to the (4) are given by $\tilde{\underline{\beta}} = (X'X)^{-}X'\underline{y}$ using all possible values of $(X'X)^{-}$, the g-inverse of $X'X$ (as discussed earlier). Since the solutions are infinite in number, none of the $\tilde{\underline{\beta}}$ values themselves have

any meaning. However, interestingly, the resulting prediction of $\underline{y}$, $\hat{\underline{y}} = X\tilde{\underline{\beta}}$ is unique, since $X(X'X)^{-}X'$ remains the same for any value of $(X'X)^{-}$, and $X(X'X)^{-}X'$ is symmetric. Therefore, in that case, we are ~~partial~~ specially concerned about any linear combination of the vector $X\hat{\underline{\beta}}$ that can be used as a good estimator for any given linear combination of the parameters $(\beta_1, \beta_2, \ldots, \beta_p)' = \underline{\beta}$. This kind of problem can also bring interest when the X has full rank. We will take up this problem in the next section.

Remark: From the discussion on General Linear Model made so far, ~~one~~ we state some remarks to understand this model in more better sense.

(1) The Linear Model (in (1)) is such a model that is linear in its parameters, not in its other quantities (i.e. x's). In that sense, for examples

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}^2 + \beta_3 x_{3i}^4 + \epsilon_i,$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3 + \beta_4 x_{1i}^4 + \epsilon_i$$

are ~~not at all~~ also linear models, but

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_1^2 x_{2i} + \beta_3^2 x_{3i} + \beta_3^3 x_{4i} + \epsilon_i$$

is not a linear model for $i = 1(1)n$.

(2) General classes of linear models:

- Least square Model: $\underline{y} = X\underline{\beta} + \underline{\epsilon}$. This model makes no assumption on the random nature of $\underline{\epsilon}$. The parameter space is $\Theta = \{\underline{\beta} : \underline{\beta} \in \mathbb{R}^{p+1}\}$.
- Gauss-Markov Model: $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = \underline{0}$ and $\text{Var}(\underline{\epsilon}) = \sigma^2 I_n$. Here, $\Theta = \{(\underline{\beta}, \sigma^2) : (\underline{\beta}, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}^{+}\}$.
- Aitken Model: $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = 0$, $\text{Var}(\underline{\epsilon}) = \sigma^2 V$, V is known. Thus, $\Theta = \{(\underline{\beta}, \sigma^2) : (\underline{\beta}, \sigma^2) \in \mathbb{R}^{p+1} \times \mathbb{R}^{+}\}$.

Geometrical Interpretation of least Square Estimation — Projection

Let us recall the least-square estimation of $\underset{\sim}{\beta}$ based on data $\underset{n\times 1}{\underset{\sim}{y}}$ and given design matrix $\underset{n\times p+1}{X}$, where $\underset{\sim}{y}$, the observed responses is assumed to be expressed in the form of a linear model in (2). The resulting normal equations generated from the least-square method is

$$(X'X)\underset{\sim}{\beta} = X'\underset{\sim}{y} \qquad \text{(presented in equation (4))}$$

Now, if $X'X$ is found to be non-singular, i.e. $(X'X)^{-1}$ exists, then only solution of the above equation is found to be unique and in that case solution is $\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y}$. Otherwise, one will be found a solution $\tilde{\underset{\sim}{\beta}} = (X'X)^{-}X'\underset{\sim}{y}$ for a choice of the g-inverse of $X'X$, that is $(X'X)^{-}$. As $(X'X)^{-}$ varies, the corresponding solution $\tilde{\underset{\sim}{\beta}}$ also varies. Thus $\tilde{\underset{\sim}{\beta}}$ is not unique unlike $\hat{\underset{\sim}{\beta}}$, if exists. However, irrespective of the existence of inverse of $(X'X)$, one property between $\hat{\underset{\sim}{\beta}}$ and $\tilde{\underset{\sim}{\beta}}$ is common and that comes from the 'Least Square' method'. This is

$$X\hat{\underset{\sim}{\beta}} = M\underset{\sim}{y} = X\tilde{\underset{\sim}{\beta}} \quad (\text{if } \hat{\underset{\sim}{\beta}} \text{ exists})$$

As in this ~~scop~~ course, g-inverse concept is not much needed (actually beyond the scope of the course), hence we assume here that $\hat{\beta}$ exists. Thus, we have $X\hat{\underset{\sim}{\beta}} = M\underset{\sim}{y}$. Since,

$$\hat{\underset{\sim}{\beta}} = (X'X)^{-1}X'\underset{\sim}{y} \Rightarrow X\hat{\underset{\sim}{\beta}} = X(X'X)^{-1}X'\underset{\sim}{y},$$

so, $M = X(X'X)^{-1}X'$ is a $n\times n$ matrix. We have denoted earlier that predicted value of $\underset{\sim}{y}$ by least-square method is $\hat{\underset{\sim}{y}}$, so

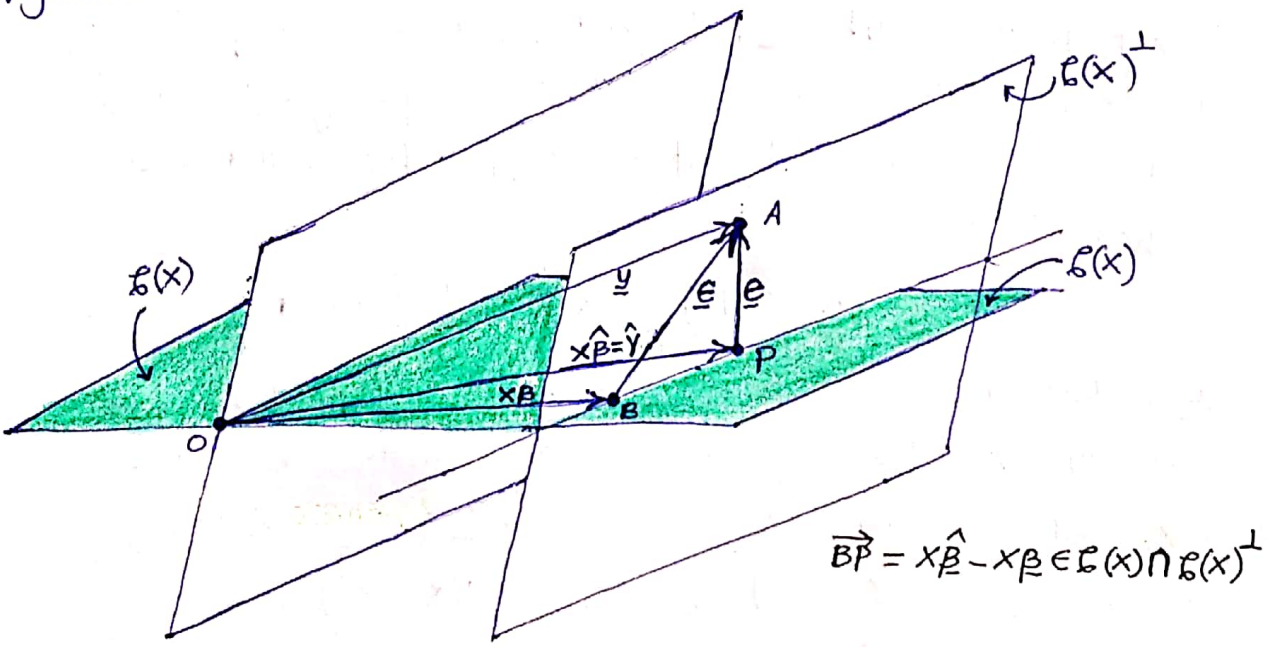
$$\hat{\underset{\sim}{y}} = X\hat{\underset{\sim}{\beta}} = M\underset{\sim}{y}.$$

Thus, predicted $\underset{\sim}{y}$, i.e. $\hat{\underset{\sim}{y}}$ is nothing but a projection of the observed data $\underset{\sim}{y}$ by the projection matrix (operator) M. In

this context, a the projection matrix is commonly denoted as P_x, therefore $P_x = x(x'x)^{-1}x'$.

Now, a simple query should be generated our mind that-In which space (i.e. vector space) the projection matrix P_x projects the $\underline{y}$ to?

To answer this, we have $\hat{\underline{Y}} = P_x \underline{y}$. Since, P_x is post-multiplied by $\underline{y}$, so, $\hat{\underline{Y}} = P_x \underline{y} \in \mathcal{C}(P_x)$, column space of P_x. Again, in P_x, x is post-multiplied by $(x'x)^{-1}x'$, so $\underline{u} \in \mathcal{C}(x')$, column space of x, ~~Hence, $P_x \in \mathcal{C}(x)$ shows~~ for every columnvector $\underline{u}$ of $x(x'x)^{-1}x'$. Here, for every such $\underline{u}$, $\underline{u}$ is linear combination of ~~rows of x~~ columns of X. Thus, $\hat{\underline{Y}} = P_x \underline{y} \in \mathcal{C}(X)$, i.e, P_x projects $\underline{y}$ onto $\mathcal{C}(X)$. Let us see the geometric view of least square estimate $\hat{\underline{Y}}$ of $\underline{y}$ through the concept of (orthogonal) projection.



The space (coloured in 'green') refers to $\mathcal{C}(X)$. Figure shows $\hat{\underline{Y}} \in \mathcal{C}(X)$, $\underline{e}, \underline{e} \in \mathcal{C}(X)^{\perp}$, an orthogonal space to $\mathcal{C}(X)$. Here, $\mathcal{C}(X)^{\perp} = \mathcal{C}(I_n - P_x)$.

The projection matrix P_x has some properties such as P_x is (i) idempotent, (ii) Symmetric, (iii) Rank (P_x) = Rank (X), (iv) $\mathcal{C}(X) = \mathcal{C}(P_x)$. The proofs of the above results are simple, hence left for readers. To know more about projection, see 'Sengupta & Jammalamadaka (2003)'.

Orthogonal Complement space to $\mathcal{C}(X)$ & ANOVA Table*

From the last discussion we have seen $\text{Rank}(P_X) = \dim(\mathcal{C}(P_X)) = \dim(\mathcal{C}(X)) = \text{Rank}(X)$. Let us consider the following result.

Result: Suppose X is $n \times (p+1)$ with rank $r \leq (p+1)$, and let P_X be the ~~perpendicular~~ orthogonal projection matrix onto $\mathcal{C}(X)$. Then $\text{Rank}(I_n - P_X) = n - r$.

Proof: We know $(I_n - P_X)$ is the ~~perpendicular~~ orthogonal projection matrix onto $\mathcal{C}(X)^{\perp}$, so it is idempotent.

$$(I_n - P_X)(I_n - P_X) = I_n - P_X - P_X + P_X^2 = I_n - P_X$$

(since P_X is idempotent).

So, $\text{Rank}(I_n - P_X) = \text{tr}(I_n - P_X) = \text{tr}(I_n) - \text{tr}(P_X) = n - r$. ∎

Recall the least square estimate $\hat{\underline{\beta}}$ or $\underline{\beta}$ in the context of linear model (2) i.e. GM model. We have learned that if $X'X$ is not non-singular (i.e. singular), then $\hat{\underline{\beta}}$ is not unique but

$$P_X \cdot \underline{y} = X\hat{\underline{\beta}} = \hat{\underline{y}}$$

is always unique irrespective of the situation $X'X$ is non-singular or not. We have seen that $\hat{\underline{y}}$, the vector of the fitted values, belongs to $\mathcal{C}(X)$ that is closest to $\underline{y}$. As P_X is the orthogonal projection matrix onto $\mathcal{C}(X)$. Similarly, $(I_n - P_X)$ is the orthogonal projection matrix onto $\mathcal{C}(X)^{\perp}$.

Result: $\mathcal{C}(X)^{\perp} = \mathcal{N}(X')$.

Proof: Since $\hat{\underline{y}} = X\hat{\underline{\beta}} \in \mathcal{C}(P_X) = \mathcal{C}(X)$, $\underline{y} - \hat{\underline{y}} = \underline{e} \in \mathcal{C}(X)^{\perp}$ because, $\underline{e}'(X\hat{\underline{\beta}} - X\underline{\beta})$ $\quad [\because X\hat{\underline{\beta}} - X\underline{\beta} \in \mathcal{C}(X)]$

$$= (\underline{y} - P\underline{y})'(P\underline{y} - X\underline{\beta})$$
$$= \underline{y}'P\underline{y} - \underline{y}'X\underline{\beta} - \underline{y}'P'P\underline{y} + \underline{y}'P'X\underline{\beta} \quad [\because P' = P \,\&\, P^2 = P]$$
$$= \underline{y}'P'X\underline{\beta} - \underline{y}'X\underline{\beta}$$
$$= 0 \quad [\because P'X = PX = X(X'X)^{-1}X'X = X]$$

So, $\mathcal{E}(X)^{\perp}$ is the orthocomplement space to $\mathcal{E}(X)$. Consider any $\underline{z} \in \mathcal{E}(X)^{\perp}$ and we know $\mathcal{E}(X)^{\perp} = \mathcal{E}(I_n - P_X)$, then for some $\underline{g}$, $\underline{z} = (I_n - P_X)\underline{g}$ ~~$= X\hat{\beta}) = X'(\underline{y} - P_X\underline{y}) = X'\underline{y} - X'P_X\underline{y}$~~ $[\because P'X = X]$

then $X'\underline{z} = X'(I_n - P_X)\underline{g} = X'\underline{g} - X'P_X\underline{g} = X'\underline{g} - X'\underline{g}$

$= \underline{0}$

Hence, according to the definition 'null space'

$\underline{z} \in \mathcal{N}(X')$,

and therefore $\mathcal{E}(X) = \mathcal{E}(I_n - P_X) = \mathcal{N}(X')$. ■

Thus, the vector of residuals, $\underline{e} \in \mathcal{N}(X')$, where $\underline{y} = \hat{\underline{y}} + \underline{e}$. As $\hat{\underline{y}} \in \mathcal{E}(X)$, $\mathcal{N}(X')$ and $\mathcal{E}(X)$ are orthogonal complements, $\hat{y}$ & $\underline{e}$ are orthogonal vectors. Note that

$$\underline{y}'\underline{y} = \underline{y}'I_n\underline{y} = \underline{y}'(P_X + I_n - P_X)\underline{y}$$

$$= \underline{y}'P_X\underline{y} + \underline{y}'(I_n - P_X)\underline{y}$$

$$= \underline{y}'P_X'P_X\underline{y} + \underline{y}(I_n - P_X)'(I_n - P_X)\underline{y}$$

$[\because P_X$ & $(I_n - P_n)$ are idempotent & symmetric$]$

$$= \hat{\underline{y}}'\hat{\underline{y}} + \underline{e}'\underline{e}' \quad [\because P_X\underline{y} = \hat{\underline{y}},\ (I_n - P_X)\underline{y} = \underline{e}]$$

This orthogonal decomposition of $\underline{y}'\underline{y}$ (i.e. data sum of squares as a measure of data 'variability') is often given in a tabular form display, called an analysis of variance (ANOVA) table. In this context the ANOVA table looks like

Source of variation	d.f.	SS (uncorrected)
Model	r	$\hat{\underline{y}}'\hat{\underline{y}}$ $(= \underline{y}'P_X\underline{y})$
Residual	$n-r$	$\underline{e}'\underline{e}$ $(= \underline{y}'(I_n - P_X)\underline{y})$
Total	n	$\underline{y}'\underline{y}$ $(= \underline{y}'I_n\underline{y})$

The term 'SS' stands for 'sum of squares'. The 'd.f.' (i.e. degrees of freedom) column catalogues the ranks of associated quadratic form matrices i.e. $\text{Rank}(P_X) = r$, $\text{Rank}(I_n - P_X) = n - r$ and $\text{Rank}(I_n) = n$. When 'Model SS' is sufficiently larger than 'Residual SS', one can say that linear model $\underline{y} = X\underline{\beta} + \underline{\epsilon}$ fitted the data $\underline{y}$ well.

Examples of some of the special cases involving relevant linear models:

From our discussions so far, we have obtained the estimate of the estimate of $\underline{\beta}$ in the model (1). which is nothing but the 'Least Square Model' which has no probabilistic assump-tion on $\underline{\epsilon}$. However, Later, we have understood the requirement of these assumptions. Thus, one can assume the 'Gauss Markov Model'. In the remaining course of this paper, we specifically discuss several issues related to the estimation involving the model parameters $\underline{\beta}$ and σ^2 for different variants of Gauss Markov (GM) model. We now highlight some special cases of this model.

Example 1. Modeling of response variable with common mean.

Let $y_1, y_2, \ldots, y_n$ be the n observed data on the response variable Y with common mean $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2 > 0 \quad \forall\ i = 1(1)n$. Therefore the GM model is

$$\underset{n\times 1}{\underline{y}} = X\underset{1\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{\epsilon}},$$

where,

$$\underset{n\times 1}{\underline{y}} = (y_1, y_2, \ldots, y_n)',$$

$$\underset{1\times 1}{\underline{\beta}} = \mu,$$

$$\underset{n\times 1}{X} = (1, 1, \ldots, 1)' \text{ and}$$

$$\underset{n\times 1}{\underline{\epsilon}} = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)' \text{ with } E(\underline{\epsilon}) = \underline{0},\ V(\underline{\epsilon}) = \sigma^2 I_n.$$

Example 2. Linear regression model.

Suppose that a response variable Y is linearly related to several independent (explanatory) variables, say $X_1, X_2, \ldots, X_p$. Hence, each observed data on Y can be written through the following model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \epsilon_i,$$

for $i = 1(1)n$. Here, the GM model has the exact form of the original GM model but X, the design matrix here is a $n \times (p+1)$ matrix with

entries $((x_{ij}))_{i=1(1)n, j=1(1)\overline{p+1}}$, where x_{ij} denotes the observed quantity (or value) of the $(j-1)^{th}$ explanatory variable corresponding to the i^{th} individual for $j = 2(1)\overline{p+1}$, $i = 1(1)n$ and $x_{i1} = 1$ for all $i = 1(1)n$, i.e.

$$\underset{n\times\overline{p+1}}{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$$

Hence, this model is called 'multiple linear regression model'. In particular, when $p=1$ i.e. we have only one explanatory variable (say, x_1), the reduced GM model with $\underset{n\times 2}{X}$ as design matrix and $\underline{\beta} = (\beta_0, \beta_1)'$, is called 'simple linear regression'. In both the cases, $E(\underline{\epsilon}) = \underline{0}$ and $Var(\underline{\epsilon}) = \sigma^2 I_n$.

Example 3. One-way ANOVA model.

Consider an experiment that is performed to compare $p(\geq 2)$ different levels of a treatment/factor. For the i^{th} treatment level, suppose that n_i experimental units are selected at random and assigned them to the i^{th} level. Therefore, we can consider the model for the observed responses

$$y_{ij} = \mu + \alpha_i + \epsilon_{ij},$$

for $i = 1, 2, \ldots, p$, $j = 1(1)n_i$, $\sum_{i=1}^{p} n_i = n$, where $E(\epsilon_{ij}) = 0$ and $Var(\epsilon_{ij}) = \sigma^2$ $\forall$ i,j. If all the levels $\alpha_1, \alpha_2, \ldots, \alpha_p$ are considered as fixed constant then this model is a special case of the GM model where

$$\underset{n\times 1}{\underline{y}} = \begin{pmatrix} y_{11} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ \vdots \\ y_{2n_2} \\ \vdots \\ y_{pn_p} \end{pmatrix}, \quad \underset{n\times\overline{p+1}}{X} = \begin{pmatrix} \underline{1}_{n_1} & \underline{1}_{n_1} & \underline{0}_{n_1} & \cdots & \underline{0}_{n_1} \\ \underline{1}_{n_2} & \underline{0}_{n_2} & \underline{1}_{n_2} & \cdots & \underline{0}_{n_2} \\ \vdots & & & & \\ \underline{1}_{n_p} & \underline{0}_{n_p} & \underline{0}_{n_p} & \cdots & \underline{1}_{n_p} \end{pmatrix} \quad \text{and}$$

$\underset{(p+1)\times 1}{\underline{\beta}} = (\mu, \alpha_1, \alpha_2, \ldots\ldots, \alpha_p)'$,

where $\underset{n\times 1}{\underline{\epsilon}} = (\epsilon_{11}, \epsilon_{12}, \ldots, \epsilon_{1n_1}, \epsilon_{21}, \ldots, \epsilon_{2n_2}, \ldots\ldots, \epsilon_{pn_p})'$ and $\underline{1}_{n_i}$ denotes an $n_i \times 1$ vector of 1's and $\underline{0}_{n_i}$ denotes an $n_i \times 1$ vector of 0s with $E(\underline{\epsilon}) = \underline{0}$ and $V(\underline{\epsilon}) = \sigma^2 I_n$.

However, the first column of X is the sum of the last p columns that means there is a linear dependence in the columns of X here. Thus, X is not of full rank matrix, infact the Rank $(X) = p < p+1$. This is the common character-istic of ANOVA model. On the other hand, (linear) regression models (Example. 2) are the models having GM form with the $\underset{n\times(p+1)}{X}$ having full rank (column) i.e. $(p+1)$.

Example 4. Two-way ANOVA model

If we extend the previous model by incorporating another one factor which has q (say) no. of levels then the resulting model would be

$$y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij},$$

where β_j's are fixed constant as like α_i for $i = 1(1)p$, $j = 1(1)q$. Hence, the sample size, $n = p \times q$. There are some variants of this two-way model where atleast one of the set from $(\alpha_1, \alpha_2, \ldots, \alpha_p)$ or $(\beta_1, \beta_2, \ldots, \beta_q)$ can be considered as random in nature. If α's and β's are random then the above model is called two-way random effect model, whereas if either α's or β's are assumed to be random ~~and~~ then the associated ~~this~~ model is called two-way mixed effect model. However, in ~~all the~~ all such cases the corresponding X matrix is

$$\underset{n\times(p+q+1)}{X} = \begin{pmatrix} \underline{1}_q & \underline{1}_q & \underline{0}_q & \cdots & \underline{0}_q & q\underline{e}_1 & q\underline{e}_2 & \cdots & q\underline{e}_q \\ \underline{1}_q & \underline{0}_q & \underline{1}_q & \cdots & \underline{0}_q & q\underline{e}_1 & q\underline{e}_2 & \cdots & q\underline{e}_q \\ \vdots & & & & & & & & \\ \underline{1}_q & \underline{0}_q & \underline{0}_q & \cdots & \underline{1}_q & q\underline{e}_1 & q\underline{e}_2 & \cdots & q\underline{e}_q \end{pmatrix}$$

(p-terms: $\underline{1}_q \ \underline{0}_q \ \underline{0}_q \cdots \underline{1}_q$; q-terms: $q\underline{e}_1 \ q\underline{e}_2 \cdots q\underline{e}_q$)

and $\underset{p+q+1\times 1}{\underline{\beta}} = (\mu, \alpha_1, \alpha_2, \ldots, \alpha_p, \beta_1, \beta_2, \ldots, \beta_q)'$,

where $\underset{q\times 1}{\underline{e}}_k$ denotes the vector of order $(q\times 1)$ for $k = 1(1)q$. However, as par GM model is concerned, the two-way model with only fixed effects satisfies the definition of GM model when ϵ_{ij} are assumed to be independent random variables with $E(\epsilon_{ij}) = 0$ and $V(\epsilon_{ij}) = \sigma^2 > 0$, $\forall\, i,j$. The X matrix is not of full column rank because if we add up the last 'q' columns of X we get the first column.

Example 5. Analysis of covariance model.

Consider the same experiment of comparing p (≥ 2) different levels of a treatment/factor but after adjusting for the effects of a covariate, say Z. Therefore the modified model, which is called the analysis of covariance model with one concomitant variable (here Z) is given by

$$y_{ij} = \mu + \alpha_i + \beta_i x_{ij} + \epsilon_{ij},$$

for $i = 1(1)p$ and $j = 1(1)n_i$, $\sum_{1}^{p} n_i = n$. Here β_i denotes the slope of the line that relates $\underline{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})$ to $\underline{x}_i = (x_{i1}, x_{i2}, \ldots, x_{in_i})$ for the i^{th} treatment (or i^{th} level of the factor), Here x_{ij}'s are assumed to be fixed. This model is another special case of GM model since all other quantities of the above model carry same meaning and assumptions (if any) of the model describing in Example 3. for example, with $p = 3$ and $n_1 = n_2 = n_3 = 3$, we have

$$\underline{y} = (y_{11}, y_{12}, y_{13}, y_{21}, y_{22}, y_{23}, y_{31}, y_{32}, y_{33})'$$

$$\underline{\beta} = (\mu, \alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3)' \quad \text{and}$$

$$X_{9\times 7} = \begin{pmatrix} 1 & 1 & 0 & 0 & x_{11} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{12} & 0 & 0 \\ 1 & 1 & 0 & 0 & x_{13} & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{21} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{22} & 0 \\ 1 & 0 & 1 & 0 & 0 & x_{23} & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{31} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{32} \\ 1 & 0 & 0 & 1 & 0 & 0 & x_{33} \end{pmatrix}$$

Here the matrix X is not of full column rank. If the last three columns are linearly independent, the $\text{Rank}(X) = 6 < 7$, the number of columns of X.

Remark 1. The random effects or mixed effects ANOVA model, which are precisely discussed in Example 4, are not at all the GM model since in those cases $\text{Var}(\underline{\epsilon}) \neq \sigma^2 I_n$.

Remark 2. Time Series Model.

When measurements on interest variable as well as some other auxiliary variables are taken over 'time', a linear model of the form $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = \underline{0}$ and $V(\underline{\epsilon}) = \sigma^2 V$ is appropriate. The form of V is chosen to model the correlation of the observed responses. for example,

$$y_t = \beta_0 + \beta_1 t + \epsilon_t,$$

for $t = 1(1)n$, where $\epsilon_t = \rho\epsilon_{t-1} + a_t$, $a_t \overset{iid}{\sim} N(0, \sigma^2)$ and $|\rho| < 1$ (this is a 'stationarity' condition on $\{y_t\}$). This model is called a 'linear trend model' where the error process $\{\epsilon_t : t = 1(1)n\}$ follows the AR(1) model. It is easy to show that $E(\epsilon_t) = 0$ $\forall t$ and $\text{Cov}(\epsilon_t, \epsilon_s) = \sigma^2 \rho^{|t-s|}$ $\forall\ t, s = 1(1)n$. So, the present model is not at all the GM model since $V(\underline{\epsilon}) \neq \sigma^2 I_n$.

Linear Estimation

Estimability is one of the most important concept in linear models. Consider the general linear model

$$\underset{n\times 1}{\underline{y}} = \underset{n\times(p+1)}{X}\underset{(p+1)\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{\epsilon}}, \qquad \text{.....(5)}$$

where $E(\underline{\epsilon}) = \underline{0}$. The further assumption of 'homoscadasticity' of model i.e. $V(\underline{\epsilon}) = \sigma^2 I_n$. is not needed as per as estimability of $\underline{\beta}$ is concerned. Suppose, $\text{Rank}(X) = r \le p+1 (\le n)$. If $r = p+1$ (as in regression models, see Example 2), estimability concern vanishes as $\underline{\beta}$ is estimated by $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$. If $r < p+1$, a common characteristic of ANOVA models, ANCOVA models (see Examples 3, 4 & 5), then $\underline{\beta}$ cannot be estimated uniquely. However, when $\underline{\beta}$ is not estimable, certain functions of $\underline{\beta}$ may be estimable. Following is an example of situation where one may be interested on the estimability of certain function of $\underline{\beta}$.

In some situation/study, some of the parameters of a linear model may be redundant. If one has 10 observations, all of them measuring the combined weight of an apple and an orange, one cannot hope to estimate weight of the orange alone from these measurements. In general, only some functions of the $\underline{\beta}$ and not all, can be estimated from the data. We discuss this issue in this section. Further, if it is possible to estimate the complete $\hat{\underline{\beta}}$ or certain functions of it, the next question is how to estimate it in an 'optimal' manner. We shall discuss this in the next section.

Some basic facts

Irrespective of the nature of the associated design matrix $\underset{n\times(p+1)}{X}$, whether it has full rank or not, classical inference problems related to the linear model (5), often concern a 'linear parametric function' of $\underline{\beta}$, say $\underline{\lambda}'\underline{\beta}$, where $\underline{\lambda}' = (\lambda_1, \lambda_2, \ldots, \lambda_{p+1})$ is a set

of p known scalar quantity such that $\lambda_i \in \mathbb{R}\ \forall i = 1(1)\overline{p+1}$. In that case, one can think of the estimation of $\underline{\lambda}'\underline{\beta}$ by a linear function of the response, say $\underline{l}'\underline{y}$, where $\underline{l}' = (l_1, l_2, \ldots, l_n)$ is a set of real unknowns to be found. The logic of considering $\underline{l}'\underline{y}$ as an estimator of $\underline{\lambda}'\underline{\beta}$ is the following. Since $\underline{y}$ itself is modelled as a linear function of the parameter $\underline{\beta}$ plus error (i.e. model (5)), it is reasonable to expect that one may be able to estimate $\underline{\beta}$ by some kind of a linear transformation in the reverse direction. This is why we try to estimate any linear parametric function by 'linear estimator', that is, as linear functions of $\underline{y} = (y_1, y_2, \ldots, y_n)'$.

Note that least square estimate $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y} = A\underline{y}$, if $r = \overline{p+1}$ where A is an unique $\overline{p+1} \times n$ matrix. Hence, least square estimate $\hat{\beta}_i = \underline{a}_i'\underline{y}$ which is indeed a linear function of $\underline{y}$ for any i, $i = 0(1)p$, where $\underline{a}_i'$ denotes the i^{th} row vector of the matrix A. One point is further noted that for some i, the scalar parameter ~~is also~~ β_i can also be written in the form of the linear parametric function $\underline{\lambda}'\underline{\beta}$ and in that case $\underline{\lambda}' = (0, 0, \ldots, \underset{\uparrow}{1}, 0, \ldots 0)$ (i^{th} position). Thus, $\underline{\lambda}'\underline{\beta}$, being a certain function $\underline{\beta}$ for known $\underline{\lambda}$, is the most general form of a parameter (since any parametric function is itself a parameter) on which one may be interested irrespective of the situation whether $\text{Rank}(X) = r = \overline{p+1}$ or $< \overline{p+1}$.

Estimability of linear parametric function (LPF)

For accurate estimation of the LPF $\underline{\lambda}'\underline{\beta}$, it is desirable that the estimator is not systematically away from the 'true' value of the parameter i.e. LPF $\underline{\lambda}'\underline{\beta}$. Hence, the concept of 'estimability' is invoked which assures the closeness of the estimator to its true value in expectation. This is none other than the 'unbiasedness' property. To have a linear estimator, this unbiasedness should come from any linear estimator in $\underline{y}$, i.e. $\underline{l}'\underline{y}$.

Definition : Linear Unbiased Estimator (LUE)

A statistic of the form $\underline{l}'\underline{y} = \sum_{j=1}^{n} l_j y_j$ is said to be a linear unbiased estimator of $\underline{\lambda}'\underline{\beta}$ if $E(\underline{l}'\underline{y}) = \underline{\lambda}'\underline{\beta}$ for all possible values of $\underline{\beta}$.

Definition : Estimability.

A linear parametric function (say $\underline{\lambda}'\underline{\beta}$) is said to be estimable if and only if there exists a linear unbiased estimator for it.

~~Since~~ The existence of 'linear unbiased estimator' is the primary concern here and 'unbiasedness' property refers the term 'estimable'. Hence, a more appropriate name for such a function should be 'linearly estimable'.

Result

Under the model assumption $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = \underline{0}$, a linear parametric function $\underline{\lambda}'\underline{\beta}$ is estimable (linearly) iff there exists a vector $\underset{n\times 1}{\underline{l}}$ such that $\underline{\lambda}' = \underline{l}'X$ that is $\underline{\lambda}' \in R(X)$, the row space of X.

Proof: **If Part:** Suppose that $\underline{\lambda}' \in R(X)$ i.e. there exists a vector $\underset{n\times 1}{\underline{l}}$ such that $\underline{\lambda}' = \underline{l}'X$ (pre-multiplication of X by $\underline{l}'$). Now, if we consider $\underline{l}'\underline{y}$ as an linear estimator, then $E(\underline{l}'\underline{y}) = \underline{l}'E(\underline{y}) = \underline{l}'X\underline{\beta} = \underline{\lambda}'\underline{\beta}$ $\forall \underline{\beta}$.

Therefore $\underline{\lambda}'\underline{\beta}$ is estimable.

Only if part: Suppose that $\underline{\lambda}'\underline{\beta}$ is estimable. Then there must exist a linear unbiased estimator for $\underline{\lambda}'\underline{\beta}$. Assume $\underline{l}'\underline{y}$ is the linear estimator which possesses the unbiased property, that is

$$E(\underline{l}'\underline{y}) = \underline{\lambda}'\underline{\beta}$$

Again we have $E(\underline{l}'\underline{y}) = \underline{l}'E(\underline{y}) = \underline{l}'X\underline{\beta}$. So,

$$\underline{\lambda}'\underline{\beta} = \underline{l}'X\underline{\beta} \Leftrightarrow \underline{\lambda}' = \underline{l}'X \Leftrightarrow \underline{\lambda}' \in R(X) \quad \forall \underline{\beta} \neq \underline{0} \quad \blacksquare$$

Remark 1. • For a matrix $\underset{m\times n}{A} = (\underline{a}_1\ \underline{a}_2 \dots \underline{a}_n)$, where $\underline{a}_j$ is $m\times 1$ vector for $j = 1(1)n$, the column space of A is defined as

$$C(A) = \left\{ \underline{\omega} \in \mathbb{R}^m : \underline{\omega} = \sum_{j=1}^{n} c_j \underline{a}_j \ ;\ c_j \in \mathbb{R},\ j = 1(1)n \right\}$$
$$= \left\{ \underline{\omega} \in \mathbb{R}^m : \underline{\omega} = A\underline{c} \ ;\ \underline{c} \in \mathbb{R}^n \right\}.$$

· Thus column space of a matrix, say $\underset{m\times n}{A}$, is the set of all $m\times 1$ vectors spanned by the columns of A and it is denoted by $C(A)$. On the other way, $C(A)$ is the set of all possible linear combinations of the columns of A. The dimension of $C(A)$ is the column rank of A.

• For the matrix $\underset{m\times n}{A} = \begin{pmatrix} \underline{a}_1^{*\prime} \\ \underline{a}_2^{*\prime} \\ \vdots \\ \underline{a}_m^{*\prime} \end{pmatrix}$, where $\underline{a}_i^{*\prime}$ is $1\times n$ vector or $\underline{a}_i^{*}$ is $n\times 1$ vector for $i = 1(1)m$, the row space of A is defined as

$$R(A) = \left\{ \underline{u}' \in \mathbb{R}^n : \underline{u}' = \sum_{i=1}^{m} d_i \underline{a}_i^{*\prime} \ ;\ d_i \in \mathbb{R},\ i = 1(1)m \right\}$$
$$= \left\{ \underline{u}' \in \mathbb{R}^n : \underline{u}' = \underline{d}'A \ ;\ \underline{d} \in \mathbb{R}^m \right\}.$$

Thus row space of a matrix, say $\underset{m\times n}{A}$, is the set of all $1\times n$ vectors spanned by the rows of A and it is denoted by $R(A)$. On the other words, $R(A)$ is the set of all possible linear combinations of the rows of A. The dimension of $R(A)$ is the row rank of A.

• Dimension of $R(A)$ (or $C(A)$) equals to no. of independent vectors belong to the $R(A)$ (or $C(A)$).

Remark 2. The necessary and sufficient condition $\underset{1\times(p+1)}{\lambda'} \in \mathcal{R}(X)$ for estimability of $\underline{\lambda}'\underline{\beta}$, mentioned in the last result, can also be equivalently written as $\underline{\lambda} \in \mathcal{C}(X')$, the 'column space' of X'. The condition $\underline{\lambda} \in \mathcal{C}(X')$ means there exist a vector $\underline{\ell}$ such that $\underline{\lambda} = X'\underline{\ell}$ (i.e. post-multiplication of X' by $\underline{\ell}$).

Remark 3. The condition $\underset{1\times(p+1)}{\underline{\lambda}'} \in \mathcal{R}(X)$ refers to the $\underline{\lambda}' = \underset{n\times 1}{\underline{\ell}'}X$ for some $\underline{\ell} \in \mathbb{R}^n$, which says $\underline{\lambda}'$ is a linear combination of rows of the X, i.e.

$$\underset{1\times(p+1)}{\underline{\lambda}'} = \underline{\ell}'X = \underline{\ell}'\begin{pmatrix} \underline{x}_{(1)}' \\ \underline{x}_{(2)}' \\ \vdots \\ \underline{x}_{(n)}' \end{pmatrix} = \sum_{j=1}^{n} \ell_j \cdot \underline{x}_{(j)}',$$

where $\underset{1\times(p+1)}{\underline{x}_{(j)}}$ denotes the j^{th} row of matrix $\underset{n\times(p+1)}{X}$.

Similarly, the equivalent condition $\underline{\lambda} \in \mathcal{C}(X')$ refers that $\underline{\lambda} = X'\underline{\ell}$ for some $\underset{n\times 1}{\underline{\ell}} \in \mathbb{R}^n$. This means $\underline{\lambda}$ is a linear combination of the columns of the X' matrix of order $(p+1)\times n$, i.e.

$$\underset{(p+1)\times 1}{\underline{\lambda}} = X'\underline{\ell} = \left(\underline{x}_{(1)}, \underline{x}_{(2)}, \ldots, \underline{x}_{(n)}\right)\begin{pmatrix} \ell_1 \\ \ell_2 \\ \vdots \\ \ell_n \end{pmatrix}$$

$$= \sum_{j=1}^{n} \ell_j \underline{x}_{(j)},$$

where $\underset{(p+1)\times 1}{\underline{x}_{(j)}}$ denotes the j^{th} column of matrix $\underset{(p+1)\times n}{X'}$.

Example 6. Consider the one-way fixed effects ANOVA model stated in Example 3. Take $p=3$ and $n_i=2 \ \forall \ i=1(1)p$.

Hence,

$$\underset{6\times 1}{\underline{y}} = (y_{11}, y_{12}, y_{21}, y_{22}, y_{31}, y_{32})'$$

$$\underset{6\times 3+1}{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}, \text{ and } \underset{3+1\times 1}{\underline{\beta}} = \begin{pmatrix} \mu \\ \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{pmatrix}.$$

Note that $\text{Rank}(X)=3$, so X is not of full rank, therefore $\underline{\beta}$ is not uniquely estimable as $(X'X)^{-1}$ does not exist. In this situation, one may be interested in estimation of linear parametric function. Let us consider some linear parametric functions $\underline{\lambda}'\underline{\beta}$ and check their estimability.

Parameter (Parametric function)	$\underline{\lambda}'$	Estimable? (Yes, iff $\underline{\lambda}' \in \mathcal{R}(X)$)
$\underline{\lambda}_1'\underline{\beta} = \mu$	$\underline{\lambda}_1' = (1,0,0,0)$	No
$\underline{\lambda}_2'\underline{\beta} = \alpha_1$	$\underline{\lambda}_2' = (0,1,0,0)$	No
$\underline{\lambda}_3'\underline{\beta} = \mu + \alpha_1$	$\underline{\lambda}_3' = (1,1,0,0)$	Yes
$\underline{\lambda}_4'\underline{\beta} = \alpha_1 - \alpha_2$	$\underline{\lambda}_4' = (0,1,-1,0)$	Yes.
$\underline{\lambda}_5'\underline{\beta} = \mu + \alpha_2 + \alpha_3$	$\underline{\lambda}_5' = (1,0,1,1)$	No
$\underline{\lambda}_6'\underline{\beta} = \alpha_1 - \frac{\alpha_2+\alpha_3}{2}$	$\underline{\lambda}_6' = (0,1,-\frac{1}{2},-\frac{1}{2})$	Yes.

Since $\underline{\lambda}_3'\underline{\beta}$, $\underline{\lambda}_4'\underline{\beta}$ and $\underline{\lambda}_6'\underline{\beta}$ are linearly estimable, there must exist linear unbiased estimators for them. Note that

- $E(y_{11}+y_{12}) = (\mu+\alpha_1) + (\mu+\alpha_1)$

  $\Rightarrow E\left[\frac{1}{2}(y_{11}+y_{12})\right] = \mu+\alpha_1 = \underline{\lambda}_3'\underline{\beta}$. Here, $\underline{\ell}_3' = \left(\frac{1}{2}, \frac{1}{2}, 0, 0, 0, 0\right)$.

- $E(y_{11} - y_{21}) = (\mu+\alpha_1) - (\mu+\alpha_2) = \alpha_1 - \alpha_2 = \underline{\lambda}_4'\underline{\beta}$

  Here, $\underline{\ell}_4' = (1, 0, -1, 0, 0, 0)$.

- $E(y_{2+}) = E(y_{21}+y_{22}) = 2(\mu+\alpha_2)$

  $\Rightarrow E\left(\frac{y_{21}+y_{22}}{2}\right) = E(\bar{y}_{2+}) = \mu+\alpha_2$

  Similarly, $E(\bar{y}_{3+}) = \mu+\alpha_3$

Then, $E\left[\bar{y}_{1+} - \left(\frac{\bar{y}_{2+}+\bar{y}_{3+}}{2}\right)\right]$

$= (\mu+\alpha_1) - \left(\frac{2\mu+\alpha_2+\alpha_3}{2}\right)$

$= \alpha_1 - \frac{\alpha_2+\alpha_3}{2}$

Here, $\underline{\ell}_6' = \left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}, -\frac{1}{4}\right)$.

All the ~~estimators~~ linear estimators, i.e. linear function of the observations ~~of the form~~ $\underline{\ell}'\underline{y}$, obtained above are expressed in the form of $\underline{\ell}'\underline{y}$, where $\underline{\ell}'$ for the respective estimators are given above.

Definition: Linearly Independent Parametric functions

Linear parametric functions $\underline{\lambda}_1'\underline{\beta}$, $\underline{\lambda}_2'\underline{\beta}$, ..., $\underline{\lambda}_k'\underline{\beta}$ are said to be linearly independent if $\underline{\lambda}_1, \underline{\lambda}_2, \ldots, \underline{\lambda}_k$ comprise a set of linearly independent vectors, i.e. $\Lambda = (\underline{\lambda}_1, \underline{\lambda}_2, \ldots, \underline{\lambda}_k)$ as a matrix of order $\overline{p+1} \times k$, has rank k.

In this context, a natural question arises that how many such 'linearly independent parametric functions' may exist. To answer this let us state and prove the following Result.

Result

Under the model assumptions $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = \underline{0}$, we can always find $r = \text{Rank}(X)$ linearly independent estimable functions. Moreover, no collection of linear estimable functions can have more than r linearly independent functions.

Proof. Since $\text{Rank}(X) = r$, there exists r independent rows of the matrix X. Let the independent rows are $\underline{x}_1', \underline{x}_2', \ldots, \underline{x}_r'$. Therefore, $\underline{x}_1'\underline{\beta}, \underline{x}_2'\underline{\beta}, \ldots, \underline{x}_r'\underline{\beta}$ are said to be linearly independent estimable functions.

Let $\underset{k\times(p+1)}{\Lambda'}\underline{\beta} = (\underline{\lambda}_1'\underline{\beta}, \underline{\lambda}_2'\underline{\beta}, \ldots, \underline{\lambda}_k'\underline{\beta})'$ be any collection of k estimable functions. Then, $\underline{\lambda}_i' \in Q(X)$ $\forall\ i = 1(1)k$ and hence,

$$\underline{\lambda}_i' = \sum_{j=1}^{n} d_j \underline{x}_j^{*\prime}, \text{ where } \underline{x}_j^{*\prime} \text{ denotes the } j^{th} \text{ row vector of}$$

the matrix X and d_j's are scalars such that not all d_j's are zero. Hence, there exists a matrix D of order k by n such that $\underset{k\times(p+1)}{\Lambda'} = (\underline{\lambda}_1', \underline{\lambda}_2', \ldots, \underline{\lambda}_k')' = \begin{pmatrix} \underline{\lambda}_1' \\ \underline{\lambda}_2' \\ \vdots \\ \underline{\lambda}_k' \end{pmatrix} = DX$.

Now, $\text{Rank}(\Lambda') = \text{Rank}(DX) \leq \text{Rank}(X) = r$. Hence, there can be almost r linearly independent estimable functions. ∎

Result

Under the model assumptions $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = \underline{0}$, the least squares estimator $\underline{\lambda}'\hat{\underline{\beta}}$ of an estimable linear parametric function $\underline{\lambda}'\underline{\beta}$ is a linear unbiased estimator of $\underline{\lambda}'\underline{\beta}$.

Proof. The least square estimator of $\underline{\beta}$ is $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$, assuming full rank of X. Then, $\underline{\lambda}'\hat{\underline{\beta}} = \underline{\lambda}'(X'X)^{-1}X'\underline{y} = \underline{l}'\underline{y}$,

, where $\underline{l}' = \underline{\lambda}'(X'X)^{-1}X'$ or $\underline{l} = \left(\underline{\lambda}'(X'X)^{-1}X'\right)' = X(X'X)^{-1}\underline{\lambda}$ is a $n\times 1$ vector as X and λ have $n\times\overline{p+1}$ and $\overline{p+1}\times 1$ order respectively.

Hence, $\underline{\lambda}'\hat{\underline{\beta}}$ is a linear estimator in $\underline{y}$.

Next, to show that $\underline{\lambda}'\hat{\underline{\beta}}$ is unbiased, we at first know

$\underline{\lambda}'\underline{\beta}$ is estimable $\Leftrightarrow$ $\underline{\lambda}' \in \mathcal{R}(X)$, row space of X

$\Leftrightarrow$ $\underline{\lambda}' = \underline{a}'X$, for some $\underline{a}$.

Thus, $E(\underline{\lambda}'\hat{\underline{\beta}}) = E\left(\underline{\lambda}'(X'X)^{-1}X'\underline{y}\right)$

$= \underline{\lambda}'(X'X)^{-1}X'E(\underline{y})$

$= \underline{\lambda}'(X'X)^{-1}X'X\underline{\beta}$ $\quad\left[\because \underline{y} = X\underline{\beta} + \underline{\epsilon}\right.$ and $\left.E(\underline{\epsilon}) = \underline{0}\right]$

$= \underline{\lambda}'\underline{\beta}$

Hence the proof. ∎

Equivalent necessary and sufficient conditions for estimability of $\underline{\lambda}'\underline{\beta}$:

From the discussion made so far we know $\underline{\lambda}' \in \mathcal{R}(X)$ or $\underline{\lambda} \in \mathcal{C}(X')$ is the necessary and sufficient condition of $\underline{\lambda}'\underline{\beta}$ to be estimable. By considering the concept of 'null space' of a matrix, one can eshtablish an equivalent necessary and sufficient condition for the same.

Let $\underset{n\times\overline{p+1}}{\text{Rank}(X)} = r$. Therefore, one can find r linearly independent columns of X and express the remaining $s = \overline{p+1} - r$ columns as linear combinations of r linearly independent columns of X. So, those s columns are linearly dependent and hence, those exist some $\underset{\overline{p+1}\times 1}{\underline{c}_1}, \underset{\overline{p+1}\times 1}{\underline{c}_2}, \ldots, \underset{\overline{p+1}\times 1}{\underline{c}_s}$ such that

$$\underset{n\times\overline{p+1}}{X}\underline{c}_i = \underset{n\times 1}{\underline{0}}$$

for $i = 1(1)s$, that is, $\underline{c}_i \in \mathcal{N}(X)$ for $i = 1(1)s$. Here $\mathcal{N}(X)$ denotes the null space of X.

Definition : 'Null space' of a matrix.

The set $\mathcal{N}(A) = \{\underset{nx1}{\underline{c}} : \underset{mxn}{A}\underline{c} = \underline{0}, \underline{c} \neq \underline{0}\}$ is called the 'null space' of A, denoted by $\mathcal{N}(A)$.

Thus, $\underline{c} \in \mathcal{N}(A)$ is such a vector scalar coefficients such that linear combination of the column vectors of A w.r.t. the coefficients $\underline{c} = (c_1, c_2, \ldots, c_n)'$ is zero, that is $\sum_{i=1}^{n} c_i . \underline{a}_i = \underline{0}$, where $\underset{mxn}{A} = (\underline{a}_1, \underline{a}_2, \ldots, \underline{a}_n)$. So, $\underline{c}\ (\neq \underline{0})$ is such a vector consisting the coefficients which makes the columns of A linearly dependent. Note that dimension of $\mathcal{N}(A) = n - \text{Rank}(A)$ is called the 'nullity' of A.

In continuation of the discussion, before the above definition of null space, the $\{\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_s\}$ forms a basis for $\mathcal{N}(X)$, i.e., $\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_s$ are linearly independent. Therefore, the necessary and sufficient condition

$$\underline{\lambda}' \in \mathcal{R}(X) \Leftrightarrow \underline{\lambda}' = \underset{nx1}{\underline{\ell}'}X, \text{ for some } \underline{\ell} \neq \underline{0}$$

$$\Leftrightarrow \underline{\lambda}'\underline{c}_i = \underline{\ell}' X \underline{c}_i \quad [\because \underline{c}_i \neq \underline{0}\ \forall i = 1(1)s]$$

$$\Leftrightarrow \underline{\lambda}'\underline{c}_i = \underline{0} \quad [\text{by definition of } \mathcal{N}(X),\ X\underline{c}_i = \underline{0}\ \forall i]$$

Thus $\underline{\lambda}'\underline{c}_i = \underline{0}$ for $i = 1(1)s$ are the equivalent necessary and sufficient conditions for $\underline{\lambda}'\underline{\beta}$ to be estimable.

There are two spaces of interest : $\mathcal{C}(X') = \mathcal{R}(X)$ and $\mathcal{N}(X)$. If the design matrix X is of order $n \times \overline{p+1}$ has rank $r\ (< \overline{p+1} \leq n)$, then $\dim\{\mathcal{C}(X')\} = r = \dim\{\mathcal{R}(X)\}$ and $\dim\{\mathcal{N}(X)\} = \overline{p+1} - r = s$. If $\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_s$ are linearly independent, then $\{\underline{c}_1, \underline{c}_2, \ldots, \underline{c}_s\}$ must be a basis for $\mathcal{N}(X)$.

$\underline{\lambda}'\underline{\beta}$ estimable $\Leftrightarrow \underline{\lambda} \in \mathcal{C}(X')$

$\Leftrightarrow \underline{\lambda}$ is orthogonal to every vector in $\mathcal{N}(X)$

$\Leftrightarrow \underline{\lambda} \perp \underline{c}_i$ for all $i = 1(1)s$, since $\{\underline{c}_1, \ldots, \underline{c}_s\}$ is basis.

$\Leftrightarrow \underline{\lambda}'\underline{c}_i = \underline{0} \quad \forall\ i = 1(1)s.$

Example 7. (Two-way crossed ANOVA with no interaction)

Reconsider the linear model stated in Example 4 (Page 15) but here the model captures replications unlike to the previous Example. i.e. Example 4 (Page 15). Hence the two-way fixed effects ANOVA model is written as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},$$

for $i = 1(1)p$, $j = 1(1)q$, $k = 1(1)n_{ij}$. In the previous Example 4, we have shown that X is not of full rank for $n_{ij} = 1 \ \forall \ i,j$. In this present replicated model Rank (X) will be unchanged as in replicated model all the newly added rows of X are nothing but the repeatation of the existing rows. One can check that Rank$(X) = p+q-1$. So, dimension $\mathcal{N}(X) = \overline{p+q+1} - \overline{p+q-1} = 2$. Taking

$$\underline{c}_1 = \begin{pmatrix} 1 \\ -\underline{1}_p \\ \underline{0}_q \end{pmatrix} \quad \text{and} \quad \underline{c}_2 = \begin{pmatrix} 1 \\ \underline{0}_p \\ -\underline{1}_q \end{pmatrix}$$

produces $X\underline{c}_1 = \underline{0}$ and $X\underline{c}_2 = \underline{0}$ respectively. Further, $\underline{c}_1$ and $\underline{c}_2$ are linearly independent. Hence, $\{\underline{c}_1, \underline{c}_2\}$ is a basis for $\mathcal{N}(X)$. Thus necessary and sufficient conditions for $\underline{\lambda}'\underline{\beta}$ to be estimable are

$$\underline{\lambda}'\underline{c}_1 = 0 \Rightarrow \lambda_0 = \sum_{i=1}^{p} \lambda_i.$$

$$\underline{\lambda}'\underline{c}_2 = 0 \Rightarrow \lambda_0 = \sum_{j=1}^{q} \lambda_j.$$

In this context, some estimable functions includes

(i) $\mu + \alpha_i + \beta_j$, (ii) $\alpha_i - \alpha_k$, (iii) $\beta_j - \beta_k$,

(iv) any contrast in α's, i.e. $\sum_{i=1}^{p} \delta_i \alpha_i$ where $\sum_{i=1}^{p} \delta_i = 0$.

(v) any contrast in β's, i.e. $\sum_{j=1}^{q} \gamma_j \beta_j$ where $\sum_{j=1}^{q} \gamma_j = 0$.

Readers are directed to check the estimability of the above functions : A set of linearly independent estimable functions is

$\{\mu+\alpha_1+\beta_1,\ \alpha_1-\alpha_2,\ \ldots,\ \alpha_1-\alpha_p,\ \beta_1-\beta_2,\ \ldots,\ \beta_1-\beta_q\}$.

Please also verify this statement.

It is important to note that when replication occurs; i.e., when $n_{ij} > 1\ \forall (i,j)$, our estimability findings remain unchanged as replication does not change $Q(X)$. When some $n_{ij} = 0$, i.e., there are missing cells, estimability may be affected due to those missing cells.

Best linear unbiased estimation & Gauss-Markov Theorem

The definition of 'estimability' of a linear parametric function guarantees only the existence of at least one unbiased estimate of an estimable linear parametric function. It does not explicitly give a method of obtaining it, nor does it say that it is the "best" estimate.

Let us consider the Gauss-Markov (linear) model and $\underline{\lambda}'\underline{\beta}$ be an estimable linear parametric function. Let us denote the linear unbiased estimator of $\underline{\lambda}'\underline{\beta}$ by $\underline{\ell}'\underline{y}$ for some $\underline{\ell} \neq \underline{0}$. Now, given the LUE $\underline{\ell}'\underline{y}$, one can always construct another LUE by simply adding another linear combination of the response data, say $\underline{a}'\underline{y}$, such that $E(\underline{a}'\underline{y}) = 0\ \forall \underline{\beta}$. This type of linear functions of response data are called linear zero function (LZF). One may construct infinitely many LZFs or atleast a large number of LZFs. Thus, there is a large class of LUEs of any

of the form $\underline{u}'\underline{y} = \underline{l}'\underline{y} + a'\underline{y} = (\underline{l}' + \underline{a}')\underline{y}$

given estimable LPF $\lambda'\beta$. Now the question is how to choose the 'best' LUE from this large class of LUEs to a given estimable LPF. Equivalently, which one is the 'best' LUE from this large class? By 'best' LUE of the LPF, say $\underline{\lambda}'\underline{\beta}$, we mean the LUE which has smallest variance among all the LUE's for $\underline{\lambda}'\underline{\beta}$. So, this 'best' LUE is clearly a linear function of response data that is unbiased for $\underline{\lambda}'\underline{\beta}$ and has smallest variance among all such unbiased linear estimator. We define this formally below:

Definition (Best Linear Unbiased Estimator)

A linear function $\underline{a}'\underline{y}$ of the observations $\underline{y}$ in the model (2) is said to be the Best Linear Unbiased Estimator (BLUE) of a parametric function $\underline{\lambda}'\underline{\beta}$, if it satisfies the following:

(i) $E(\underline{a}'\underline{y}) = \underline{\lambda}'\underline{\beta}$,

(ii) $\text{Var}(\underline{a}'\underline{y}) \leq \text{Var}(\underline{a}^{*\prime}\underline{y})$

for any $\underline{a}^* \in \mathbb{R}^n$ satisfying (i) for all non-zero $\underline{\beta}$.

In the following section we shall deal with the problem of obtaining the BLUE of an estimable LPF $\underline{\lambda}'\underline{\beta}$.

Gauss-Markov Theorem

The following theorem, which is known as the Gauss-Markov theorem is extreamly important and helpful in the theory of general linear model, as it provides an easy method of obtaining the best linear unbiased estimator (BLUE) of any estimable linear parametric function (LPF) $\underline{\lambda}'\underline{\beta}$, in the model (2) i.e. Gauss-Markov model.

Theorem (Gauss Markov Theorem)

For the model, $\underset{n\times 1}{\underline{y}} = X\underline{\beta} + \underline{\epsilon}$, $E(\underline{\epsilon}) = \underline{0}$, $V(\underline{\epsilon}) = \sigma^2 I_n$, where $\underline{y}$ is observed, X is known and $\underline{\beta}, \sigma^2$ are unknown, the BLUE of an estimable linear parametric function $\underline{\lambda}'\underline{\beta}$ (where $\underline{\lambda}$ is known) is $\underline{\lambda}'\hat{\underline{\beta}}$, where $\hat{\underline{\beta}}$ being any solution of the normal equations $X'X\underline{\beta} = X'\underline{y}$ which are obtained by minimizing the quantity $\underline{\epsilon}'\underline{\epsilon} = (\underline{y} - X\underline{\beta})'(\underline{y} - X\underline{\beta})$ with respect to the unknown vector $\underline{\beta}$.

Proof. $E(\underline{\lambda}'\hat{\underline{\beta}}) = \underline{\lambda}' E(\hat{\underline{\beta}})$ [as $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$ assuming $X'X$ non-singular]

$= \underline{\lambda}'(X'X)^{-1}X'E(\underline{y})$

$= \underline{\lambda}'(X'X)^{-1}X'X\underline{\beta}$ [since $E(\underline{y}) = X\underline{\beta}$]

$= \underline{\lambda}'\underline{\beta}$

It remains to prove that the variance of $\underline{\lambda}'\underline{\beta}$ is not larger than that of any other unbiased estimator of $\underline{\lambda}'\underline{\beta}$.

Let $\underline{u}'\underline{y}$ be any other unbiased estimator of $\underline{\lambda}'\underline{\beta}$. Then

$\underline{\lambda}'\underline{\beta} = E(\underline{u}'\underline{y}) = \underline{u}'E(\underline{y}) = \underline{u}'X\underline{\beta}$ which implies

$\underline{\lambda}' = \underline{u}'X$. [$\because \underline{\beta} \neq \underline{0}$]

Let us split $\underline{u}'\underline{y}$ as $\underline{u}'\underline{y} = (\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}) + \underline{\lambda}'\hat{\underline{\beta}}$.

Therefore, $V(\underline{u}'\underline{y}) = V(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}) + V(\underline{\lambda}'\hat{\underline{\beta}}) + 2\,Cov(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}, \underline{\lambda}'\hat{\underline{\beta}})$

Now, $Cov(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}, \underline{\lambda}'\hat{\underline{\beta}})$

$= Cov(\underline{u}'\underline{y} - \underline{\lambda}'(X'X)^{-1}X'\underline{y}, \underline{\lambda}'(X'X)^{-1}X'\underline{y})$

$= Cov[(\underline{u}' - \underline{\lambda}'(X'X)^{-1}X')\underline{y}, \underline{\lambda}'(X'X)^{-1}X'\underline{y}]$

$= Cov[\underline{s}'\underline{y}, \underline{t}'\underline{y}]$, where $\underline{s}' = \underline{u}' - \underline{\lambda}'(X'X)^{-1}X'$, $\underline{t}' = \underline{\lambda}'(X'X)^{-1}X'$

$= \underline{s}'\,Var(\underline{y})\,\underline{t}$

$= \underline{s}'\underline{t}\cdot\sigma^2$ [$\because Var(\underline{y}) = \sigma^2 I_n$]

$= (\underline{u}' - \underline{\lambda}'(X'X)^{-1}X')[\underline{\lambda}'(X'X)^{-1}X']'\sigma^2$

$$= [\underline{u}' - \underline{\lambda}'(X'X)^{-1}X'][X(X'X)^{-1}\underline{\lambda}]\sigma^2$$

$$= [\underline{u}'X - \underline{\lambda}'(X'X)^{-1}X'X](X'X)^{-1}\underline{\lambda}.\sigma^2$$

$$= [\underline{\lambda}' - \underline{\lambda}'](X'X)^{-1}.\underline{\lambda}\sigma^2 \quad (\because \underline{u}'X = \underline{\lambda}').$$

$$= \underline{0}.$$

Thus, $V(\underline{u}'\underline{y}) = V(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}) + V(\underline{\lambda}'\hat{\underline{\beta}})$

$$\geq V(\underline{\lambda}'\hat{\underline{\beta}}), \qquad \text{---------- (6)}$$

since variance of any variable is certainly non-negative.

Hence the proof. ∎

Note that the equality sign holds only when $V(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}) = 0$.

But, $E(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}}) = E(\underline{u}'\underline{y}) - E(\underline{\lambda}'\hat{\underline{\beta}}) = \underline{\lambda}'\underline{\beta} - \underline{\lambda}'\underline{\beta} = 0$.

Thus, if the equality holds in (6), the difference $(\underline{u}'\underline{y} - \underline{\lambda}'\hat{\underline{\beta}})$ has both mean and variance equal to 0 that means $\underline{u}'\underline{y}$ and $\underline{\lambda}'\hat{\underline{\beta}}$ are identical with probability 1. Thus, we can say that the BLUE of an estimable parametric function is estimable.

Thus the Gauss-Markov theorem provides a very convenient method of obtaining the BLUE of an estimable parametric function $\underline{\lambda}'\underline{\beta}$. Obtain a solution $\hat{\underline{\beta}}$ of the normal equations (4) and substitute $\hat{\underline{\beta}}$ for $\underline{\beta}$ in the $\underline{\lambda}'\underline{\beta}$ to get its BLUE $\underline{\lambda}'\hat{\underline{\beta}}$.

Variance and Covariance of BLUEs

We have proved that BLUE of LPF $\underline{\lambda}'\underline{\beta}$ is $\underline{\lambda}'\hat{\underline{\beta}}$, where $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{y}$, assuming that $X'X$ or equivalently X is non-singular. Now, as $\underline{\lambda}'\hat{\underline{\beta}}$ is an estimator, so standard error (s.e.) of $\underline{\lambda}'\hat{\underline{\beta}}$ has a common interest to practitioner.

$$\text{Var}(\underline{\lambda}'\hat{\underline{\beta}}) = \text{Var}(\underline{\lambda}'(X'X)^{-1}X'\underline{y})$$

$$= \underline{\lambda}'(X'X)^{-1}X'(\sigma^2 I_n)[\underline{\lambda}'(X'X)^{-1}X']'$$

$$[\because \text{Var}(A\underline{z}) = A\,\text{Var}(\underline{z}).A']$$

$$= \underline{\lambda}'(X'X)^{-1}X'X(X'X)^{-1}\underline{\lambda}.\sigma^2 = \underline{\lambda}'(X'X)^{-1}\underline{\lambda}\sigma^2.$$

Therefore, square root of the $Var(\underline{\lambda}'\hat{\underline{\beta}})$ is taken in order to obtain the s.e. of $\underline{\lambda}'\hat{\underline{\beta}}$.

If we consider two BLUEs, say $\underline{\lambda}_{(1)}'\hat{\underline{\beta}}$ and $\underline{\lambda}_{(2)}'\hat{\underline{\beta}}$ of two estimable LPFs $\underline{\lambda}_{(1)}'\underline{\beta}$ and $\underline{\lambda}_{(2)}'\underline{\beta}$ respectively, their covariance is given by

$$\begin{aligned} Cov\left(\underline{\lambda}_{(1)}'\hat{\underline{\beta}}, \underline{\lambda}_{(2)}'\hat{\underline{\beta}}\right) &= \underline{\lambda}_{(1)}' Var(\hat{\underline{\beta}}).\underline{\lambda}_{(2)} \quad [\because Cov(A\underline{x}, B\underline{z}) \\ &\qquad = A\,Cov(\underline{x},\underline{z}).B'] \\ &= \underline{\lambda}_{(1)}'(X'X)^{-1}X'(\sigma^2 I_n)\left[(X'X)^{-1}X'\right]'\underline{\lambda}_{(2)} \\ &= \underline{\lambda}_{(1)}'(X'X)^{-1}\underline{\lambda}_{(2)}.\sigma^2 \end{aligned}$$

Example 8. A trivial example on 'measurement error model'

Suppose that an orange and an apple with unknown weights α_1 and α_2, respectively, are weighted separately with a crude scale. Each measurement is followed by a 'dummy' measurement with nothing on the scale, in order to get an idea about typical 'measurement errors'. Let us assume that the measurements satisfy the following linear model

$$\underset{4\times 1}{\underline{y}} = X\underline{\alpha} + \underline{\epsilon},$$

where $\underline{y} = (y_1, y_2, y_3, y_4)'$, $\underline{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4)'$

$$\underset{4\times 2}{X} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}, \quad \underset{2\times 1}{\underline{\alpha}} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix}, \quad E(\epsilon_i) = 0,\ V(\epsilon_i) = \sigma^2 \text{ for } i = 1(1)4.$$

So, clearly,

$$\begin{aligned} y_1 &= \alpha_1 + \epsilon_1 \\ y_2 &= \epsilon_2 \\ y_3 &= \alpha_2 + \epsilon_3 \\ y_4 &= \epsilon_4 \end{aligned}$$

Thus, y_2 and y_4, being direct measurement of error, may be used to estimate the σ^2. The other two observations, y_1 and y_3,

carry information about the two parameters α_1 and α_2, respectively. There are several unbiased estimators of α_1, such as y_1, y_1+y_2, y_1+y_4. The 'natural estimator' y_1 is the BLUE of α_1. One can check this by finding BLUE $\underline{\lambda}'\hat{\underline{\alpha}}$ for $\underline{\lambda}'\underline{\alpha}$, where $\underline{\lambda}' = (1,0)$.

Here $X'X = I_2$. So, $\hat{\underline{\alpha}} = (X'X)^{-1}X'\underline{y} = \underline{X}'\underline{y} = \begin{pmatrix} y_1 \\ y_3 \end{pmatrix}$.

$$\text{Var}(\underline{\lambda}'\hat{\underline{\alpha}}) = \text{Var}(y_1) = V(\epsilon_1) = \sigma^2 = \underline{\lambda}'(X'X)^{-1}\underline{\lambda}\sigma^2$$

Another ~~estim~~ unbiased estimator (y_1+y_2) for α_1 has variance $\text{Var}(y_1+y_2) = \text{Var}(\epsilon_1+\epsilon_2) = 2\sigma^2$. Similarly, $\text{Var}(y_1+y_4) = 2\sigma^2$. Here, the another unbiased estimator of α_1, y_1+y_2, has the additional baggage of the y_2, which inflate the variance. This example shows the 'bestness' of the BLUE over all the unbiased estimator of a linear parametric function $\underline{\lambda}'\underline{\beta}$ (here, $\underline{\lambda}'\underline{\alpha}$).

Estimation Space & Error Space

Let us consider the following important theorem on characteristics of BLUE of a linear parametric function.

Theorem

If every BLUE (of linear parametric function) is expressed in terms of the observations $\underline{y}$ as $\underline{\ell}'\underline{y}$, the coefficient vector $\underline{\ell}$ is a linear combination of the columns of X. and conversely, every linear function of the observations $\underline{\ell}'\underline{y}$, such that the coefficient vector $\underline{\ell}$ is a linear combination of the columns of X, is the BLUE of its expected value.

Proof. Let $\underline{\lambda}'\underline{\beta}$ be an estimable parametric function. Then its BLUE is $\underline{\lambda}'\hat{\underline{\beta}}$, where $\hat{\underline{\beta}}$ satisfies the normal equation

$$(X'X)\hat{\underline{\beta}} = X'\underline{y}$$

i.e. $\hat{\beta} = (X'X)^{-1}X'\underline{y}$, assuming $X'X$ is non-singular.

So, $\underline{\lambda}'\hat{\underline{\beta}} = \underline{\lambda}'(X'X)^{-1}X'\underline{y} = \underline{l}'\underline{y}$, where $\underline{l}' = \underline{\lambda}'(X'X)^{-1}X'$

or, $\underline{l} = (\underline{\lambda}'(X'X)^{-1}X')' = X(X'X)^{-1}\underline{\lambda} = X\underline{\omega}$, with $\underline{\omega} = (X'X)^{-1}\underline{\lambda}$

Thus, $\underline{l}$ is the column vector obtained by post multiplication of X by $\underline{\omega}$. So, $\underline{l}$ is a linear combination of the columns of X.

Conversely, if $\underline{l}$ is a linear combination of the columns of X, therefore, $\underline{l}$ can be written as

$$\underline{l} = X\underline{u} \text{ for some column vector } \underline{u}.$$

Then $E(\underline{l}'\underline{y}) = \underline{l}'X\beta = \underline{u}'X'X\beta$ (since $\underline{l} = X\underline{u}$)

Now, by the Gauss-Markov Theorem, BLUE of the estimable parametric function $\underline{u}'X'X\beta$ is

$$\underline{u}'X'X\hat{\underline{\beta}} = \underline{u}'X'\underline{y}, \text{ since } \hat{\underline{\beta}} \text{ satisfies normal equation } (X'X)\hat{\underline{\beta}} = X'\underline{y}$$

$$= \underline{l}'\underline{y}$$

as $\underline{l} = X\underline{u}$ is given. Thus, $\underline{l}'\underline{y}$, for every $\underline{l}$, is the BLUE of its expected value.

Hence the proof. ∎

The above theorem shows that the coefficient vectors of all BLUE's are linear combinations of columns of design matrix X and conversely. The space spanned by the column vectors of X is therefore called the 'Estimation Space'. Since Rank$(X) = r$, it is obvious that, there can atmost be 'r' linearly independent estimable linear parametric functions (LPFs) and hence their (unique) BLUEs.

Thus it can be said that class of all vectors $\underline{l}$ such that $\underline{l}'\underline{y}$ is a BLUE (of any estimable LPF) is referred to as the 'estimation space'. Let us define the 'estimation space' in this context of linear model, i.e. Gauss-Markov Model.

Definition: (Estimation Space)

The 'estimation space' of the Gauss Markov linear model is the class of all coefficient vectors such that linear combinations of the observations with respect to any of such coefficient vector is a BLUE of its expected value, i.e. estimation space is defined as

$$S_{es} = \{\underline{l} : \underline{l}'\underline{y} \text{ is the BLUE of } \underline{\lambda}'\underline{\beta};\ \underline{\lambda} \in \mathcal{R}(X),\ \underline{\lambda} \in \mathbb{R}^{p+1},\ \underline{\lambda} \neq \underline{0}\}.$$

Remark From the last theorem, it is clear that estimation space of the Gauss-Markov linear model $(\underline{y}, X\underline{\beta}, \sigma^2 I)$ is $\mathcal{C}(X)$, column space of X.

Let us discuss about another space which also has an important role in Linear Model. We have introduced the concept of Linear zero function (LZF) in Page 29. Let us collect all such vector $\underline{a}'$ such that $E(\underline{a}'\underline{y}) = 0\ \forall\ \underline{\beta}$. This collection constitutes a 'space'. This space contains a special vectors in the context of Linear Model. The term 'e' refers the error in the fitted linear model. 'e' is popularly known as 'residual' vector. One can check that $E(\underline{e}'\underline{y}) = 0\ \forall\ \underline{\beta}$. We have $X'\underline{e} = X'(\underline{y} - X\hat{\underline{\beta}}) = X'\underline{y} - X'X\hat{\underline{\beta}} = X'\underline{y} - X'\underline{y} = \underline{0}$.

So, $E(\underline{e}'\underline{y}) = \underline{e}'X\underline{\beta} = \underline{0}'\underline{\beta} = 0$. Thus, having a special member $\underline{e}$, this space is termed as 'Error Space'. Let us define this space formally.

Definition: (Error Space)

The 'error space' of the Gauss Markov linear model is the class of all such vectors $\underline{a}$ such that $E(\underline{a}'\underline{y}) = 0$ for any value of $\underline{\beta}$, i.e.

$$S_{er} = \{\underline{a} : E(\underline{a}'\underline{y}) = 0, \text{ for all possible values of } \underline{\beta}\}.$$

Let us assume $\underline{b} \in \mathcal{E}r$. Therefore, $E(\underline{b}'\underline{y}) = \underline{b}'X\underline{\beta} = \underline{0}$ for all values of $\underline{\beta}$. It implies and implied by

$$\underline{b}'X = \underline{0}' \text{ or } X'\underline{b} = \underline{0}$$

$$\Leftrightarrow \underline{b}'(\underline{X}_{(1)}, \underline{X}_{(2)}, \ldots, \underline{X}_{(p+1)}) = \underline{0}'$$

where $\underline{X}_{(i)}$ is the i^{th} column of X

$$\Leftrightarrow \underline{b}'\underline{X}_{(i)} = 0 \quad \forall\ i = 1(1)\overline{p+1}.$$

$\Leftrightarrow$ $\underline{b}$ is orthogonal to the columns of X

i.e. $\underline{b} \perp \underline{X}_{(i)} \quad \forall\ i = 1(1)\overline{p+1}$.

Thus we have the following result.

Result: A linear function of the observations $\underline{y}$ for which the coefficient vector belongs to the error space if and only if the coefficient vector is orthogonal to the columns of X.

Remark: From the last theorem, we know $\underline{\ell} \in \mathcal{E}(X)$. further from the above result we have $\underline{b} \perp \mathcal{E}(X)$. Thus, we can say such $\underline{b} \in \mathcal{E}(X)^{\perp}$, the orthogonal space to the column space of X.

Thus we have another result.

Result: The coefficient vector of any BLUE (when expressed in terms of linear combination of the observations) is orthogonal to the coefficient vector, of any linear combination of the observations, belonging to the error space.

Proposition: (i) Every BLUE is linear function of the vector of the fitted values, and (ii) every linear zero function (LZF) is a linear function of the residuals.

Proof. (i) Let $\underline{\lambda}'\underline{\beta}$ be an estimable linear parametric function (LPF). So, $\underline{\lambda}' \in \mathcal{R}(X)$ or $\underline{\lambda} \in \mathcal{E}(X')$. Then, according to the

Gauss-Markov Theorem, the BLUE of $\underline{\lambda}'\underline{\beta}$ is

$$\underline{\lambda}'\hat{\underline{\beta}} = \underline{\lambda}'(X'X)^{-1}X'\underline{y} = \underline{\lambda}'X^{-1}X\hat{\underline{\beta}} \text{ (assuming } X^{-1} \text{ exists)}$$
$$= \underline{\lambda}'X^{-1}\underline{y}$$

Thus, BLUE is linear function of the fitted values $\hat{Y}_1, \hat{Y}_2, \ldots, \hat{Y}_n$.

(ii) If $\underline{a}'\underline{y}$ is an LZF i.e. $E(\underline{a}'\underline{y}) = 0 \Leftrightarrow \underline{a}'E(\underline{y}) = 0$

$$\Leftrightarrow \underline{a}'X\underline{\beta} = 0$$
$$\Leftrightarrow \underline{a}'X = \underline{0}' \quad \forall \underline{\beta} \neq 0.$$
$$\Leftrightarrow \underline{a}'\underline{X}_{(i)} = 0 \quad \forall i = 1(1)\overline{p+1},$$

where $X = (\underline{X}_{(1)}, \underline{X}_{(2)}, \ldots, \underline{X}_{(p+1)})$ and $\underline{X}_{(i)}$ denotes the i^{th} column of X.

So, $\underline{a} \perp \mathcal{C}(X) \Rightarrow \underline{a} \perp \mathcal{C}(P_X)$.

Now if $\underline{z} \in \mathcal{C}(P_X)$, $\underline{z} = P_X\underline{m}$, for some $\underline{m} \neq \underline{0}$.

Therefore, $\underline{a}'\underline{z} = 0$ (∵ they belong to two orthogonal spaces)

$$\Rightarrow \underline{a}'P_X\underline{m} = 0$$

This will only happen if $\underset{1\times n}{\underline{a}'} = \underline{d}'(I - P_X)$ for some $\underset{n\times 1}{\underline{d}} \neq \underline{0}$

Then only $\underline{a}'P_X\underline{m} = \underline{d}'(I-P_X)P_X\underline{m}$

$$= \underline{d}'(P_X - P_X^2)\underline{m}$$
$$= 0 \quad (\because P_X = P_X^2 \text{ due to idempotent})$$

Thus, LZF is of the form $\underline{a}'\underline{y} = \underline{d}'(I-P_X)\underline{y}$

$$= \underline{d}'(\underline{y} - P_X\underline{y})$$
$$= \underline{d}'\underline{e} \quad (\because P_X\underline{y} = \hat{\underline{Y}}) \quad \blacksquare$$

Remark: With continuation to the last remark, now we can say that if $\underline{b}'\underline{y}$ is an LZF or $\underline{b}$ belongs to the 'error space', $\underline{b} \in \mathcal{C}(X)^{\perp}$ i.e. $\underline{b} \in \mathcal{C}(I-P_X)$., since $\underline{b}$ can be written as $\underline{b} = (I-P_X)\underline{d}$ or $\underline{b}' = \underline{d}'(I-P_X)$ for some $\underline{d}$.

Estimation of the error variance σ^2 in the GM model

Consider the Gauss-Markov model $\underline{y} = X\underline{\beta} + \underline{\epsilon}$, where $E(\underline{\epsilon}) = \underline{0}$ and $Var(\underline{\epsilon}) = \sigma^2 I_n$. Here $X\underline{\beta}$ is estimable and therefore, BLUE of $X\underline{\beta}$ is

$$X\hat{\underline{\beta}} = X(X'X)^{-1}X'\underline{y} = P_X\underline{y} = \hat{\underline{y}} \text{ (say)}.$$

Here $\hat{\underline{y}}$ is the orthogonal projection of $\underline{y}$ onto $\mathcal{C}(X)$. The resulting residuals are given by

$$\underline{e} = \underline{y} - \hat{\underline{y}} = \underline{y} - P_X\underline{y} = (I_n - P_X)\underline{y},$$

the orthogonal projection of $\underline{y}$ onto $\mathcal{C}(I_n - P_X)$ or $\mathcal{N}(X')$.

So, residual sum of squares is

$$\underline{e}'\underline{e} = (\underline{y} - \hat{\underline{y}})'(\underline{y} - \hat{\underline{y}}) = (\underline{y} - X\hat{\underline{\beta}})'(\underline{y} - X\hat{\underline{\beta}})$$
$$= \underline{y}'(I_n - P_X)(I_n - P_X)\underline{y}$$
$$= \underline{y}'(I_n - P_X)\underline{y} \quad \text{------(*)}$$

Now we consider a result.

Result: Suppose $\underline{z}$ is a random vector with $E(\underline{z}) = \mu$ and $Var(\underline{z}) = \Sigma$. Then, for a non-random matrix A,

$$E(\underline{z}'A\underline{z}) = \mu'A\mu + \text{trace}(A\Sigma).$$

Applying the above result we have from (*),

$$E(\underline{e}'\underline{e}) = E(\underline{y}'(I_n - P_X)\underline{y}) = (X\underline{\beta})'(I_n - P_X)X\underline{\beta} + \text{trace}((I_n - P_X)\sigma^2 I_n)$$
$$= \sigma^2(\text{tr}(I_n) - \text{tr}(P_X))$$
$$\left(\text{since } X'(I_n - P_X)X = X'X - X'X(X'X)^{-1}X'X = 0\right)$$
$$= \sigma^2(n - r), \text{ where } \text{tr}(P_X) = \text{Rank}(P_X) = \text{Rank}(X) = r$$

So, $\hat{\sigma}^2 = \frac{1}{n-r}\left[\underline{y}'(I_n - P_X)\underline{y}\right]$ is the unbiased estimate of σ^2.

$$= \frac{1}{n-r}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \text{ where } \hat{y}_i = X_{(i)}^{*\prime}\hat{\underline{\beta}},\ X_{(i)}^{*\prime} = i^{th} \text{ row of } X.$$

There are different ways to look at the comparision of models. So far we have looked at comparing/selecting models based on R^2 or adjusted R^2 values, model significance test, t-tests for variable(s) added last. These are good way to look at, but they are ineffective in cases when

- ✓ explanatory variables work together in a group
- ✓ we want to test some hypotheses for some $\beta_i = b_i$, $i = 1(1)k \leq p$. for example, we may want to test $H_0: \beta_1 = 3, \beta_4 = 5$ against the alternative hypothesis that atleast one of those is false.

These kind of problem involving more than one parameter is commonly formulated in the form of

$$H_0: C\underset{\sim}{\beta} = \underset{\sim}{d}$$

against H_1: not H_0,

where $\underset{\sim}{\beta}$ is the $\overline{p+1} \times 1$ order coefficient vector in the linear model $\left(\underset{n\times 1}{\underset{\sim}{y}}, X\underset{\sim}{\beta}, \sigma^2 I_n\right)$ and $\underset{k\times \overline{p+1}}{C}$ and $\underset{k\times 1}{\underset{\sim}{d}}$ are known. Here, H_0 comprises k number of hypothesis simultaneously.

To solve this testing problem, General Linear Hypotheses look at the difference between unrestricted (i.e. full) model and restricted (i.e. under H_0) model in terms of 'residual sum of squares (SS)'.

Let us denote the unrestricted residual SS by SSE or S_1^2. further, restricted residual SS is denoted by SS_H or S_2^2. Note that $S_2^2 \geq S_1^2$ always, since 'restricted model' has lesser number of unknown

parameters than that of 'unrestricted' or full model. So, 'restricted model' contains lesser number of estimated parameters which means it has lesser variability in the fitted model part, that is,

$$Var(\underline{\hat{Y}}_H) < Var(\underline{\hat{Y}})$$

$$\Leftrightarrow Var(\underline{e}_H) > Var(\underline{e}) \quad [\because V(\underline{\hat{Y}}_H)+V(\underline{e}_H) = V(\underline{\hat{Y}})+V(\underline{e}) = V(\underline{y})]$$

where $\underline{\hat{Y}}_H$ and $\underline{\hat{e}}_H$ denote the predicted value of $\underline{y}$ and associated residual vector, respectively, under H_0.

'Sum of squares (SS)' measures the variability in the data. Therefore, we will get an F-test that compares the two models — unrestricted and restricted.

For example, let us consider a linear model involving X_1, X_2, X_3, X_4 and X_5 as explanatory variable. If any one like to compare this model with another model involving X_1, X_2, X_3 only, then certainly, he/she has to test

$H_0: \beta_4 = \beta_5 = 0$

against H_1: β_4 and β_5 are not both 0.

The General Linear hypothesis test statistic is

$$F = \frac{(S_2^2 - S_1^2) / (df(S_2^2) - df(S_1^2))}{S_1^2 / df(S_1^2)}$$

$$\overset{H_0}{\sim} F_{n_1, n_2},$$

where $n_1 = df(S_2^2) - df(S_1^2)$ = no. of extra variables (Here it is 2).

$n_2 = df(S_1^2)$

We reject H_0 if p-value $\leq \alpha$, for given $\alpha \in (0,1)$.

If H_0 is rejected, one can conclude that atleast one of X_4 and X_5 is useful for predicting $\underline{y}$ that already contains

the variables x_1, x_2 and x_3.

Example Suppose $n = 100$ and we are testing x_1, x_2, x_3, x_4, x_5 (full/Unrestricted) vs. x_1, x_2, x_3 (reduced/restricted). Our hypotheses are

$$H_0 : \beta_4 = \beta_5 = 0$$

against H_1 : β_4 and β_5 are not both 0.

Since there are 4 ~~regression~~ model parameters in the reduced model (due to removal of 2 variables x_4 and x_5), the numerator df is $6-4=2$. The denominator df is $n-6 = 94$ (since there are 6 parameters including the drift in the full model). Suppose for a given data the p-value $= Pr.(F > F(\text{observed}) | H_0)$ is obtained as 0.032 or the $F(\text{observed}) = 4.25$. Then p-value is found to be lesser than 0.05 or. $F(\text{observed})$ is greater than the tabulated $F_{2,94;0.05} = 3.105$. Thus, we reject the H_0 and in that case we conclude that either x_4 or x_5 or both contain additional information that is useful for predicting $\underline{y}$ in a linear model that also includes x_1, x_2 and x_3.

Example Let us consider the same problem mentioned in the previous Example. If we additionally wish to test whether x_2 and x_3 are equally ~~effective~~ i.e. $H_0' : \beta_2 = \beta_3$. Note that in H_0, there are two hypotheses

$$\beta_4 = 0$$

and $\beta_5 = 0$.

Then considering $H_0' : \beta_2 - \beta_3 = 0$, total 3 hypotheses are to be tested simultaneously. Note that the 3 hypotheses state 3 values for 3 linear parametric function of $\underline{\beta}$. So,

the hypotheses to be tested in this problem can be written as

$$H_0: \quad C\underline{\beta} = \underline{d},$$

where C is an 3×6 order matrix ~~such~~ given by

$$C = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{pmatrix} \quad \text{and} \quad \underset{6\times 1}{\underline{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{pmatrix}, \ \underset{3\times 1}{\underline{d}} = \underset{3\times 1}{\underline{0}}.$$

Clearly, one can see $C\underline{\beta} = \underline{d} \Rightarrow \begin{cases} \beta_4 = 0 \\ \beta_5 = 0 \\ \beta_2 - \beta_3 = 0 \end{cases}$

Now, before proceeding further, as $C\underline{\beta}$ is a collection of LPFs, so to draw any inference (here, hypothesis testing) on these LPFs, one should assure about the ability of the LPFs in order to make statistical inference in terms of hypothesis testing. We should check whether the LPFs i.e. elements of $C\underline{\beta}$ are testable or not. Let us define the testable hypothesis.

Definition: Testable Hypotheses

A linear hypothesis $C\underline{\beta}$ is testable if the elements of $C\underline{\beta}$ are all estimable.

Now come to the problem. Here the matrix C is assumed to be full column rank matrix. Now, in this problem

Full/Unrestricted model: $E(\underline{y}) = X\underline{\beta}$.

or $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_5 x_5$

Here X is of $n\times 6$ matrix and $\underline{\beta} = (\beta_0, \beta_1, \cdots, \beta_5)'$.

Now, under H_0, that is, assuming H_0 is true, the Reduced/Restricted Model : $E(\underline{y}) = Z\underline{\alpha}$

$$\text{or, } E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_2 + x_3)$$
$$= \alpha_0 + \alpha_1 z_1 + \alpha_2 z_2 .$$

Therefore, we can compute s_1^2 and s_2^2, defined earlier and thereafter, compute the test statistic

$$F = \frac{(s_2^2 - s_1^2)/n_1}{s_1^2/n_2} ,$$

where $n_1 = (n-3) - (n-6) = 3 = d.f.(s_2^2) - d.f.(s_1^2)$.

$n_2 = n-6 = d.f.(s_1^2)$

and take the decision about the H_0 based on p-value or tabular value of F-random variable for given n_1 and n_2.

Remark So, it is clear how to carry out a 'General Linear Hypothesis' testing of null hypothesis of the form

$$H_0 : \underset{k \times p+1}{C}\underline{\beta} = \underline{b}$$

in connection to the linear model $(\underline{y}, X\underline{\beta}, \sigma^2 I_n)$. We just need to compute s_1^2 and s_2^2 and their corresponding d.f.s. to obtain the value of the F-statistic. Now, let us formulate s_1^2 and s_2^2.

$s_1^2 = SSE$ = unrestricted residual SS $= \underline{e}'\underline{e} = \underline{y}'(I_n - P_x)\underline{y}$,

$s_2^2 = SS_H$ = restricted residual SS $= \underline{e}_H'\underline{e}_H = \underline{y}'(I_n - P_z)\underline{y}$,

where projection matrix in unrestricted/full model, ~~$P_x = X(X'X)^{-1}X'$~~

$\underline{y} = X\underline{\beta} + \underline{\epsilon}$, is $P_x = X(X'X)^{-1}X'$ and

projection matrix in restricted/reduced model, $\underline{y} = Z\underline{\alpha} + \underline{\epsilon}_H$,

is $P_z = Z(Z'Z)^{-1}Z'$, where Z has $n \times (p+1-k)$ order.

$\underline{e}$ and $\underline{e}_H$ denote the residual vectors from unrestricted & restricted models, respectively. So, $\underline{e} = \underline{y} - X\hat{\underline{\beta}}$ and $\underline{e}_H = \underline{y} - Z\hat{\underline{\alpha}}$,

Model, Gauss-Markov Linear Model & its classification

To summarize the concept of linear model and its classification, let us throw some basic question on modelling. These are

(1) What does a 'model' mean?

(2) Why do we require 'modelling' of an observable variable?

To answer the first question, let us consider $y_1, y_2, \ldots, y_n$ be the n observations on the observable continuous random variable Y. In all cases, we can assume the observed value to be composed of two parts

$$y_i = \mu_i + \epsilon_i,$$

where μ_i is the 'true value satisfying a specific pattern governed by a mathematical function' of some explanatory (stochastic and/or non-stochastic) variables that may have some impact on Y. and ϵ_i is the error which is the difference (or discrepency) between real observed quantity y_i and its theoretical value satifying the function, i.e. $\epsilon_i = y_i - \mu_i$ $\forall$ $i = 1(1)n$. The true value μ_i is the part which is obtained based on 'assignable causes', whereas the error ϵ_i is due to various 'chance causes' that have impact on Y but not identified or included in μ_i. Certainly, the error term ϵ_i is assumed to be random for all i.

To answer the second question, let us pose a question: 'Are a person's brain size, height and weight predictive of his or her intelligence?' To answer this kind of question, we require 'modelling' of the interest variable (here, intelligence) based on some other given variable(s) (here, brain size, height, weight). One should opt such model (i.e. function or relation) of interest variable which best fit (or describe) the 'true relation'

between the interest variable (denoted by Y) and the set of chosen predictor or explanatory variables denoted by $(x_1, x_2, \ldots, x_p)$ for some $p \geq 1$. That optimum or best fitted model is therefore used to predict the intelligence level of a new person (not belongs to the used sample but must belong to the population from which the sample has been drawn) based on his/her measurements on brain size, height and weight. Thus, precisely one can say that a model gives us a regular or systematic pattern of the values of response variable in terms of the chosen explanatory variable(s), say predictor(s). The 'best' model consumes best match with the observed relationship between Y and $(x_1, x_2, \ldots, x_p)$.

Now we come to the point on 'Linear Model'. When the μ_i is assumed to be of the form

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_p x_{pi}$$

in (8), where x_{ji}'s, for $i = 1(1)n$, $j = 1(1)p$, are known, then we call this model (8) for each y_i as 'linear model' since here the μ_i is a linear function of $(p+1)$ unknown quantities, $\beta_0, \beta_1, \ldots, \beta_p$, called effects. These unknown quantities are the parameters of the linear model. In addition, if we assume the error term ϵ_i has the random nature with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$, $Cov(\epsilon_i, \epsilon_{i'}) = 0 \ \forall \ i \neq i' = 1(1)n$, then the linear model is called Gauss-Markov (Linear) Model. Now, the Gauss-Markov (GM) model can be classified further into three models depending on the nature of known x_{ji}'s.

I. when x_{ji}'s are continuous, GM model $\Rightarrow$ Regression model.

II. when $x_{ji} = 0$ or $1 \ \forall \ i, j$., GM model $\Rightarrow$ ANOVA model

III. when x_{ji}'s are continuous for atleast one j and $x_{j'i} = 0$ or 1 for atleast one j', $j, j' = 1(1)p \ni j + j' = p$, $i = 1(1)n$, GM model $\Rightarrow$ ANCOVA model.