The Gauss-Markov Linear Model

\[ y = X\beta + \epsilon \]

- \( y \) is an \( n \times 1 \) random vector of responses.
- \( X \) is an \( n \times p \) matrix of constants with columns corresponding to explanatory variables. \( X \) is sometimes referred to as the design matrix.
- \( \beta \) is an unknown parameter vector in \( \mathbb{R}^p \).
- \( \epsilon \) is an \( n \times 1 \) random vector of errors.
- \( E(\epsilon) = 0 \) and \( \text{Var}(\epsilon) = \sigma^2 I \), where \( \sigma^2 \) is an unknown parameter in \( \mathbb{R}^+ \).

The Column Space of the Design Matrix

- \( X\beta \) is a linear combination of the columns of \( X \):

\[
X\beta = [x_1, \ldots, x_p] \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} = \beta_1 x_1 + \cdots + \beta_p x_p.
\]

- The set of all possible linear combinations of the columns of \( X \) is called the column space of \( X \) and is denoted by

\[ C(X) = \{Xa : a \in \mathbb{R}^p\}. \]

- The Gauss-Markov linear model says \( y \) is a random vector whose mean is in the column space of \( X \) and whose variance is \( \sigma^2 I \) for some positive real number \( \sigma^2 \), i.e.,

\[ E(y) \in C(X) \text{ and } \text{Var}(y) = \sigma^2 I, \sigma^2 \in \mathbb{R}^+. \]

An Example Column Space

\[
X = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies C(X) = \{Xa : a \in \mathbb{R}^p\} = \bigg\{ \begin{bmatrix} 1 \\ 1 \\ a_1 \end{bmatrix} : a_1 \in \mathbb{R} \bigg\} = \bigg\{ a_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} : a_1 \in \mathbb{R} \bigg\} = \bigg\{ \begin{bmatrix} a_1 \\ a_1 \end{bmatrix} : a_1 \in \mathbb{R} \bigg\}
\]

Another Example Column Space

\[
X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \implies C(X) = \bigg\{ \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \end{bmatrix} : a \in \mathbb{R}^2 \bigg\} = \bigg\{ a_1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} + a_2 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} : a_1, a_2 \in \mathbb{R} \bigg\} = \bigg\{ \begin{bmatrix} a_1 \\ a_1 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ a_2 \\ a_2 \end{bmatrix} : a_1, a_2 \in \mathbb{R} \bigg\} = \bigg\{ \begin{bmatrix} a_1 \\ a_1 \\ a_2 \\ a_2 \end{bmatrix} : a_1, a_2 \in \mathbb{R} \bigg\}
\]
Another Column Space Example

\[
X_1 = \begin{bmatrix}
1 & 0 \\
1 & 0 \\
0 & 1 \\
0 & 1
\end{bmatrix} \quad X_2 = \begin{bmatrix}
1 & 1 & 0 \\
1 & 1 & 0 \\
0 & 1 & 1 \\
0 & 1 & 1
\end{bmatrix}
\]

\[x \in C(X_1) \implies x = X_1 a \text{ for some } a \in \mathbb{R}^2 \]
\[\implies x = X_2 \begin{bmatrix}
0 \\
a
\end{bmatrix} \text{ for some } a \in \mathbb{R}^2 \]
\[\implies x = X_2 b \text{ for some } b \in \mathbb{R}^3 \]
\[\implies x \in C(X_2) \]

Thus, \( C(X_1) \subseteq C(X_2) \).

Another Column Space Example (continued)

\[x \in C(X_1) \implies x = X_1 a \text{ for some } a \in \mathbb{R}^2 \]
\[\implies x = X_2 \begin{bmatrix}
0 \\
a
\end{bmatrix} \text{ for some } a \in \mathbb{R}^3 \]
\[\implies x = X_2 a \text{ for some } a \in \mathbb{IR}^3 \]
\[\implies x = a_1 \begin{bmatrix}
1 \\
1 \\
1
\end{bmatrix} + a_2 \begin{bmatrix}
1 \\
0 \\
0
\end{bmatrix} + a_3 \begin{bmatrix}
0 \\
0 \\
1
\end{bmatrix} \text{ for some } a \in \mathbb{IR}^3 \]
\[\implies x = \begin{bmatrix}
ad_1 + ad_2 \\
ad_1 + ad_3 \\
ad_1 + ad_3
\end{bmatrix} \text{ for some } a, d_1, d_2, d_3 \in \mathbb{IR} \]
\[\implies x = X_1 \begin{bmatrix}
ad_1 + ad_2 \\
ad_1 + ad_3 \\
ad_1 + ad_3
\end{bmatrix} \text{ for some } a, d_1, d_2, d_3 \in \mathbb{IR} \]

Thus, \( C(X_2) \subseteq C(X_1) \).

We previously showed that \( C(X_1) \subseteq C(X_2) \).

Thus, it follows that \( C(X_1) = C(X_2) \).

Another Column Space Example (continued)

Estimation of \( E(y) \)

- A fundamental goal of linear model analysis is to estimate \( E(y) \).
- We could, of course, use \( y \) to estimate \( E(y) \).
- \( y \) is obviously an unbiased estimator of \( E(y) \), but it is often not a very sensible estimator.
- For example, suppose

\[
\begin{bmatrix}
y_1 \\
y_2
\end{bmatrix} = \begin{bmatrix}
1 \\
1
\end{bmatrix} \mu + \begin{bmatrix}
e_1 \\
e_2
\end{bmatrix}, \text{ and we observe } y = \begin{bmatrix}
6.1 \\
2.3
\end{bmatrix}.
\]

Should we estimate \( E(y) = \begin{bmatrix}
\mu \\
\mu
\end{bmatrix} \text{ by } y = \begin{bmatrix}
6.1 \\
2.3
\end{bmatrix} \)?
Estimation of $E(y)$

- The Gauss-Markov linear models says that $E(y) \in C(X)$, so we should use that information when estimating $E(y)$.
- Consider estimating $E(y)$ by the point in $C(X)$ that is closest to $y$ (as measured by the usual Euclidean distance).
- This unique point is called the *orthogonal projection* of $y$ onto $C(X)$ and denoted by $\hat{y}$ (although it could be argued that $\hat{E}(y)$ might be better notation).
- By definition, $||y - \hat{y}|| = \min_{z \in C(X)} ||y - z||$, where $||a|| \equiv \sqrt{\sum_{i=1}^{n} a_i^2}$.

Orthogonal Projection Matrices

It can be shown that...

- $\forall y \in \mathbb{R}^n$, $\hat{y} = P_X y$, where $P_X$ is a unique $n \times n$ matrix known as an *orthogonal projection matrix*.
- $P_X$ is idempotent: $P_X P_X = P_X$.
- $P_X$ is symmetric: $P_X = P_X'$.
- $P_X X = X$ and $X' P_X = X'$.
- $P_X = X (X'X)^{-} X'$, where $(X'X)^{-}$ is any generalized inverse of $X'X$.

Why Does $P_X X = X$?

\[
P_X X = P_X [x_1, \ldots, x_p] \\
= [P_X x_1, \ldots, P_X x_p] \\
= [x_1, \ldots, x_p] \\
= X.
\]

Generalized Inverses

- $G$ is a *generalized inverse* of a matrix $A$ if $AGA = A$.
- We usually denote a generalized inverse of $A$ by $A^-$.
- If $A$ is nonsingular, i.e., if $A^{-1}$ exists, then $A^{-1}$ is the one and only generalized inverse of $A$.
  \[
  AA^{-1} A = AI = IA = A
  \]
- If $A$ is singular, i.e., if $A^{-1}$ does not exist, then there are infinitely many generalized inverses of $A$. 
Invariance of $P_X = X(X'X)^{-1}X'$ to Choice of $(X'X)^{-1}$

- If $X'X$ is nonsingular, then $P_X = X(X'X)^{-1}X'$ because the only generalized inverse of $X'X$ is $(X'X)^{-1}$.
- If $X'X$ is singular, then $P_X = X(X'X)^{-1}X'$ and the choice of the generalized inverse $(X'X)^{-1}$ does not matter because $P_X = X(X'X)^{-1}X'$ will turn out to be the same matrix no matter which generalized inverse of $X'X$ is used.
- Suppose $(X'X)_1^{-1}$ and $(X'X)_2^{-1}$ are any two generalized inverses of $X'X$. Then
  
  $$X(X'X)_1^{-1}X' = X(X'X)_2^{-1}X'(X'X)_1^{-1}X' = X(X'X)_2^{-1}X'.$$

An Example Orthogonal Projection

Thus, the orthogonal projection of $y = \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}$ onto the column space of $X = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ is $P_Xy = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix} = \begin{bmatrix} 4.2 \\ 4.2 \end{bmatrix}$.

An Example Orthogonal Projection Matrix

Suppose $\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \end{bmatrix}$, and we observe $y = \begin{bmatrix} 6.1 \\ 2.3 \end{bmatrix}$.

$$X(X'X)^{-1}X' = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}' \begin{bmatrix} 1 \\ 1 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}'$$

$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\right)^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \begin{bmatrix} 1 \\ 1 \end{bmatrix} 2^{-1} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \begin{bmatrix} 1/2 \\ 1/2 \end{bmatrix}$$

$$= \begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}.$$
Why is \( P_X \) called an orthogonal projection matrix?

Suppose \( X = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \) and \( y = \begin{bmatrix} 2 \\ 3/4 \end{bmatrix} \).
Why is $P_X$ called an orthogonal projection matrix?

- The angle between $\hat{y}$ and $y - \hat{y}$ is $90^\circ$.
- The vectors $\hat{y}$ and $y - \hat{y}$ are orthogonal.

$$\hat{y}'(y - \hat{y}) = \hat{y}'(y - P_Xy) = \hat{y}'(I - P_X)y$$

$$= (P_Xy)'(I - P_X)y = y'(I - P_X)y$$

$$= y'(P_X'I - P_X)y = y'(P_X - P_XP_X)y$$

$$= y'(P_X - P_X)y = 0.$$ 

Optimality of $\hat{y}$ as an Estimator of $E(y)$

- $\hat{y}$ is an unbiased estimator of $E(y)$:

$$E(\hat{y}) = E(P_Xy) = P_XE(y) = P_XX\beta = X\beta = E(y).$$

- It can be shown that $\hat{y} = P_Xy$ is the best estimator of $E(y)$ in the class of linear unbiased estimators, i.e., estimators of the form $My$ for $M$ satisfying

$$E(My) = E(y) \forall \beta \in \mathbb{R}^p \iff MX \beta = X\beta \forall \beta \in \mathbb{R}^p \iff MX = X.$$

- Under the Normal Theory Gauss-Markov Linear Model, $\hat{y} = P_Xy$ is best among all unbiased estimators of $E(y)$.

Ordinary Least Squares (OLS) Estimation of $E(y)$

- OLS: Find a vector $b^* \in \mathbb{R}^p$ such that

$$Q(b^*) \leq Q(b) \ \forall \ b \in \mathbb{R}^p, \ \text{where} \ Q(b) = \sum_{i=1}^{n} (y_i - x_i'b)^2.$$

- Note that

$$Q(b) = \sum_{i=1}^{n} (y_i - x_i'b)^2 = (y - Xb)'(y - Xb) = ||y - Xb||^2.$$

- To minimize this sum of squares, we need to choose $b^* \in \mathbb{R}^p$ such that $Xb^*$ will be the point in $C(X)$ that is closest to $y$.

- In other words, we need to choose $b^*$ such that $Xb^* = P_Xy = X(X'X)^{-1}X'y$.

- Clearly, choosing $b^* = (X'X)^{-1}X'y$ will work.

Ordinary Least Squares and the Normal Equations

- Often calculus is used to show that $Q(b^*) \leq Q(b) \ \forall b \in \mathbb{R}^p$ if and only if $b^*$ is a solution to the normal equations:

$$X'Xb = X'y.$$ 

- If $X'X$ is nonsingular, multiplying both sides of the normal equations by $(X'X)^{-1}$ shows that the only solution to the normal equations is $b^* = (X'X)^{-1}X'y$.

- If $X'X$ is singular, there are infinitely many solutions that include $(X'X)^{-1}X'y$ for all choices of generalized inverse of $X'X$.

$$X'X[(X'X)^{-1}X'y] = X'[X(X'X)^{-1}X']y = X'P_Xy = X'y.$$ 

- Henceforth, we will use $\hat{\beta}$ to denote any solution to the normal equations.
Ordinary Least Squares Estimator of $E(y) = X\beta$

- We call $X\hat{\beta} = P_XX\hat{\beta} = X(X'X)^{-1}X'y = X(X'X)^{-1}X'y = P_Xy = \hat{y}$ the OLS estimator of $E(y) = X\beta$.

- It might be more appropriate to use $\hat{X}\beta$ rather than $X\hat{\beta}$ to denote our estimator because we are estimating $X\beta$ rather than pre-multiplying an estimator of $\beta$ by $X$.

- As we shall soon see, it does not make sense to estimate $\beta$ when $X'X$ is singular.

- However, it does make sense to estimate $E(y) = X\beta$ whether $X'X$ is singular or nonsingular.