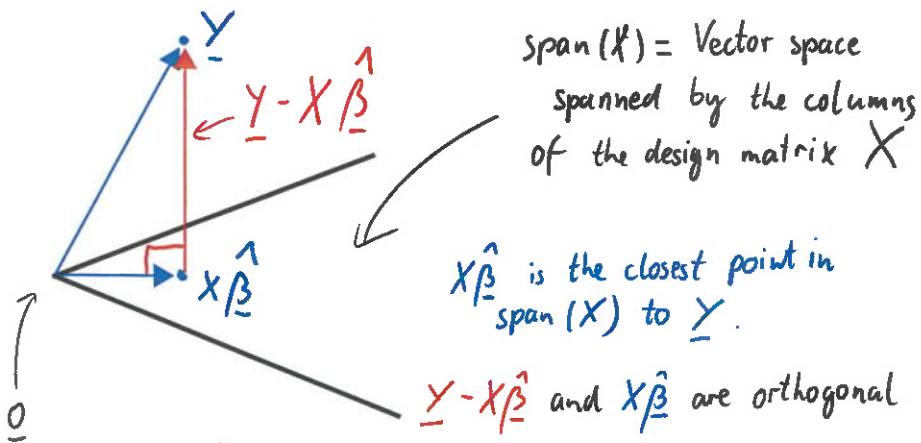


Geometrical Interpretation



To see that $\underline{Y} - \underline{X}\hat{\beta}$ is orthogonal to $\underline{X}\hat{\beta}$:

$$(\underline{X}\hat{\beta})^T(\underline{Y} - \underline{X}\hat{\beta}) = \hat{\beta}^T \underline{X}^T (\underline{Y} - \underline{X}\hat{\beta}) = \hat{\beta}^T \underbrace{(\underline{X}^T \underline{Y} - \underline{X}^T \underline{X}\hat{\beta})}_{=0 \text{ by LSE}} = 0$$

9.7 Properties of Least Squares Estimation

In this section: assume (FR) and (SOA)
Then

$$\hat{\beta} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y}$$

- $\hat{\beta}$ is linear in \underline{Y} .

More precisely: the function $\hat{\beta} : \mathbb{R}^n \rightarrow \mathbb{R}^p$, $\mathbf{y} \mapsto (\underline{X}^T \underline{X})^{-1} \underline{X}^T \mathbf{y}$ is a linear mapping.

- $\hat{\beta}$ is unbiased for β .

Indeed, for all β :

$$E[\underline{Y}] = \underline{X}\beta$$

$$E[\hat{\beta}] = (\underline{X}^T \underline{X})^{-1} \underline{X}^T E[\underline{Y}] = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{X}\beta = \underline{\beta}$$

- $\text{cov}(\hat{\beta}) = \sigma^2 (\underline{X}^T \underline{X})^{-1}$.

$$\underline{Y} = \underline{X}\beta + \varepsilon$$

$$\text{cov}(\varepsilon) = \sigma^2 \mathbf{I} \Rightarrow \text{cov}(\underline{Y}) = \sigma^2 \mathbf{I}$$

Indeed, letting $A = (X^T X)^{-1} X^T$ we have

$$\begin{aligned}\text{cov}(\hat{\beta}) &\equiv \text{cov}(A\mathbf{Y}) = A \text{cov}(\mathbf{Y}) A^T = \sigma^2 A A^T \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} = \sigma^2 (X^T X)^{-1}\end{aligned}$$

$$\text{cov}(\gamma) = \sigma^2 I$$

We now compute the least squares estimator explicitly in one simple situation.

Example 53 (Simple linear regression)

Let

$$Y_i = \beta_1 + \beta_2 a_i + \epsilon_i, \quad E \epsilon_i = 0 \quad i = 1, \dots, n,$$

where a_1, \dots, a_n are known deterministic constants. Assume that $n \geq 2$

$$\mathbf{Y}^T = (Y_1, \dots, Y_n), \quad \beta^T = (\beta_1, \beta_2) \text{ and } X = \begin{pmatrix} 1 & a_1 \\ \vdots & \vdots \\ 1 & a_n \end{pmatrix}. \quad X^T = \begin{pmatrix} 1 & \dots & 1 \\ a_1 & \dots & a_n \end{pmatrix}$$

Assume SOA and that not all a_i 's are equal (to ensure FR). Then

$$\begin{aligned}X^T X &= \begin{pmatrix} n & n\bar{a} \\ n\bar{a} & \sum a_i^2 \end{pmatrix} & \bar{a} &= \frac{1}{n} \sum_{i=1}^n a_i \\ (X^T X)^{-1} &= \frac{1}{n \sum a_i^2 - n^2 \bar{a}^2} \begin{pmatrix} \sum a_i^2 & -n\bar{a} \\ -n\bar{a} & n \end{pmatrix} \\ X^T \mathbf{Y} &= \begin{pmatrix} n\bar{Y} \\ \sum a_i Y_i \end{pmatrix}.\end{aligned}$$

Now we can find $\hat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$, hence

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \frac{1}{\sum a_i^2 - n\bar{a}^2} \begin{pmatrix} \bar{Y} \sum a_i^2 - \bar{a} \sum a_i Y_i \\ \sum a_i Y_i - n\bar{a} \bar{Y} \end{pmatrix}.$$

$$\hat{\beta}_2 = \frac{\sum (a_i - \bar{a})(Y_i - \bar{Y})}{\sum (a_i - \bar{a})^2}$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{a}.$$

$$\text{cov}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} = \frac{\sigma^2}{n \sum (a_i - \bar{a})^2} \begin{pmatrix} \sum a_i^2 & -n\bar{a} \\ -n\bar{a} & n \end{pmatrix}.$$

$$Y_i = \gamma_1 + b_i \gamma_2 + \varepsilon_i, \quad b_i = a_i - \bar{a}, \quad b_i = 0$$

If $\bar{a} = 0$ everything becomes easier: the covariance matrix is diagonal and $\hat{\beta}_1 = \bar{Y}$.

To get to this situation, we can re-parametrise the model by letting $\gamma_1 = \beta_1 + \bar{a}\beta_2$ and $\gamma_2 = \beta_2$. Then

ADD AND SUBTRACT $\bar{a}\beta_2$

$$E(\mathbf{Y}) = X\beta = \begin{pmatrix} \beta_1 + a_1\beta_2 \\ \vdots \\ \beta_1 + a_n\beta_2 \end{pmatrix} \stackrel{\uparrow}{=} \begin{pmatrix} \gamma_1 + (a_1 - \bar{a})\gamma_2 \\ \vdots \\ \gamma_1 + (a_n - \bar{a})\gamma_2 \end{pmatrix} = \begin{pmatrix} 1 & (a_1 - \bar{a}) \\ \vdots & \vdots \\ 1 & (a_n - \bar{a}) \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \end{pmatrix}$$

Then $\sum_{i=1}^n (a_i - \bar{a}) = 0$, $\hat{\gamma}_1 = \bar{Y}$, $\hat{\gamma}_2 = \hat{\beta}_2$.

If the main interest lies in β_2 then one can work with the simpler transformed model and get the same estimates via γ_2 .

$$\text{Var}(\hat{\gamma}) = \frac{\sigma^2}{n \sum b_i^2} \begin{pmatrix} \sum b_i^2 & 0 \\ 0 & n \end{pmatrix}$$

The following theorem justifies the use of the least squares estimator - it can be used to construct a best linear unbiased estimator (*BLUE*).

An estimator $\hat{\gamma}$ is called linear if there exists $\mathbf{L} \in \mathbb{R}^{n \times p}$ such that $\hat{\gamma} = \mathbf{L}^T \mathbf{Y}$.

Theorem 7 (The Gauss-Markov Theorem for full-rank linear models)

Assume (FR), (SOA). Let $\mathbf{c} \in \mathbb{R}^p$ and let $\hat{\beta}$ be a least squares estimator of β in a linear model. Then the following holds: The estimator $\mathbf{c}^T \hat{\beta}$ has the smallest variance among all linear unbiased estimators for $\mathbf{c}^T \beta$.

i-th

Remark For $i \in \{1, \dots, n\}$, let $\mathbf{e}_i = (0, \dots, 0, 1, 0, \dots, 0)^T$ with the 1 being in the i th component. Choosing $\mathbf{c} = \mathbf{e}_i$ we have $\mathbf{c}^T \beta = \beta_i$ and $\mathbf{c}^T \hat{\beta} = \hat{\beta}_i$. Thus $\hat{\beta}_i$ has the smallest variance among all linear unbiased estimators of β_i .

Proof $\mathbf{c}^T \hat{\beta}$ is linear and unbiased

Let $\hat{\gamma} = \mathbf{L}^T \mathbf{Y}$ be any other linear unbiased estimator of $\mathbf{c}^T \beta$.

$$\text{Var}(\hat{\gamma}) \geq \text{Var}(\mathbf{c}^T \hat{\beta}) \xrightarrow{\text{ADD AND SUBTRACT } \mathbf{c}^T \hat{\beta}}$$

$$\begin{aligned} \text{Var}(\hat{\gamma}) &= \text{Var}(\mathbf{L}^T \mathbf{Y}) = \text{Var}(\mathbf{c}^T \hat{\beta} + (\mathbf{L}^T - \mathbf{c}^T (X^T X)^{-1} X^T) \mathbf{Y}) \\ &\quad =: \mathbf{D}^T \\ &= \text{cov}(\mathbf{c}^T \hat{\beta} + \mathbf{D}^T \mathbf{Y}, \mathbf{c}^T \hat{\beta} + \mathbf{D}^T \mathbf{Y}) \\ &= \text{Var}(\mathbf{c}^T \hat{\beta}) + \text{Var}(\mathbf{D}^T \mathbf{Y}) + 2 \text{cov}(\mathbf{c}^T \hat{\beta}, \mathbf{D}^T \mathbf{Y}) \end{aligned}$$

USING DEF OF $\hat{\beta}$ AND PROPERTIES OF COV.

$$\text{cov}(\mathbf{c}^T \hat{\beta}, \mathbf{D}^T \mathbf{Y}) = \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \underbrace{\text{cov}(\mathbf{Y})}_{= \sigma^2 I_n} \mathbf{D} = \mathbf{c}^T (\mathbf{X}^T \mathbf{X})^{-1} (\underbrace{\mathbf{D}^T \mathbf{X}}_{(*) \mathbf{0}^T})^T \sigma^2 = 0.$$

To see (*): both est unbiased $\Rightarrow 0 = E(\hat{\gamma}) - E(\mathbf{c}^T \hat{\beta}) = E(\mathbf{D}^T \mathbf{Y}) = \mathbf{D}^T \mathbf{X} \beta$ for all β . Hence, $\mathbf{D}^T \mathbf{X} = \mathbf{0}^T$. Thus,

$$\text{Var}(\hat{\gamma}) = \text{Var}(\mathbf{c}^T \hat{\beta}) + \text{Var}(\mathbf{D}^T \mathbf{Y}) \geq \text{Var}(\mathbf{c}^T \hat{\beta}).$$

≥ 0

9.8 Projection Matrices

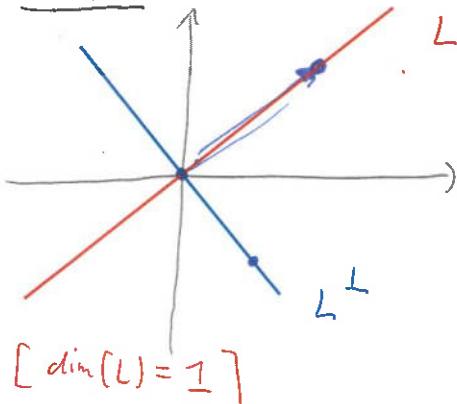
Let L be a linear subspace of \mathbb{R}^n , $\dim L = r \leq n$.

Definition 19

$P \in \mathbb{R}^{n \times n}$ is a projection matrix onto L , if

1. $P\mathbf{x} = \mathbf{x} \quad \forall \mathbf{x} \in L$
2. $P\mathbf{x} = \mathbf{0} \quad \forall \mathbf{x} \in L^\perp = \{\mathbf{z} \in \mathbb{R}^n : \mathbf{z}^T \mathbf{y} = 0 \forall \mathbf{y} \in L\}$

Example: $n=2$



L^\perp is called orthogonal complement

By definition, rank $P = \dim L = r$.

$\dim L^\perp = n-r$ of L

Remark Projection matrices will be very useful when proving results for linear models. We will often use $L = \text{span}(X)$, the space spanned by the columns of the design matrix X , or $L = \text{span}(X)^\perp$.

Lemma 11

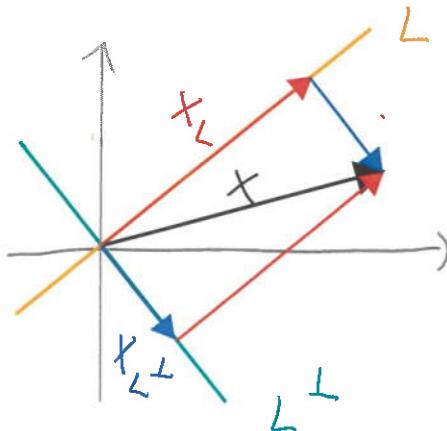
P is a projection matrix $\iff \underbrace{P^T = P}_{P \text{ symmetric}}$ and $\underbrace{P^2 = P}_{P \text{ idempotent}}$.

Proof \Rightarrow :

Recall that any $\mathbf{x} \in \mathbb{R}^n$ can be uniquely written as $\mathbf{x} = \mathbf{x}_L + \mathbf{x}_{L^\perp}$, where $\mathbf{x}_L \in L$ and $\mathbf{x}_{L^\perp} \in L^\perp$.

$$\begin{aligned} P\mathbf{x}_L &= \mathbf{x}_L \\ P\mathbf{x}_{L^\perp} &= \mathbf{0} \end{aligned} \Rightarrow P\mathbf{x} = P(\mathbf{x}_L + \mathbf{x}_{L^\perp}) = P\mathbf{x}_L = \mathbf{x}_L$$

Let $\mathbf{x} \in \mathbb{R}^n$. Then $P^2\mathbf{x} = P(P\mathbf{x}) = P\mathbf{x}_L = P\mathbf{x}$. Hence, $\underline{P^2 = P}$.



$$\cdot Y = Y_L + Y_{L^\perp} \quad (\text{VECTOR})$$

- IF WE MULTIPLY AN ELEMENT OF L WITH AN $\cancel{\text{ELEMENT}}$ OF L^\perp WE GET 0
For all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$:

$$\begin{aligned} \mathbf{x}^T P^T \mathbf{y} &= (\underbrace{Px}_{\in L})^T \mathbf{y} = (\underbrace{Px}_{\in L})^T \mathbf{y}_L = \mathbf{x}_L^T P \mathbf{y} = \mathbf{x}^T P \mathbf{y} \\ \text{Hence, } P^T &= P \quad \Rightarrow \quad \mathbf{x}^T P \mathbf{y} = (\mathbf{x}_L + \mathbf{x}_{L^\perp})^T P \mathbf{y} = \mathbf{x}_L^T P \mathbf{y} \\ &\quad \cancel{+ (Px)^T Y_{L^\perp} = 0} \quad \text{BECAUSE } \mathbf{x}_{L^\perp}^T P \mathbf{y} = 0 \end{aligned}$$

$\Leftarrow:$

Let L be the space spanned by the columns of P .

$$L = \text{Span}(P) = \{Pz : z \in \mathbb{R}^m\}$$

- Let $\mathbf{x} \in L$. Then $\exists \mathbf{z} \in \mathbb{R}^n : \mathbf{x} = P\mathbf{z}$. Hence, $\underline{Px = P^2z} \stackrel{\text{idempot}}{=} \underline{Pz = x}$.
- Let $\underline{\mathbf{x} \in L^\perp}$. Then for all $\mathbf{y} \in \mathbb{R}^n$: $(Px)^T \mathbf{y} = \mathbf{x}^T P^T \mathbf{y} \stackrel{\text{symm}}{=} \underbrace{\mathbf{x}^T P \mathbf{y}}_{\in L} = \mathbf{0}$. Hence $\underline{Px = 0}$.

The projection matrix is unique. Indeed, for each i , the vector \mathbf{e}_i can be uniquely written as $\mathbf{e}_i = \mathbf{x} + \mathbf{y}$, where $\mathbf{x} \in L$ and $\mathbf{y} \in L^\perp$. Then the i th column of P is $P\mathbf{e}_i = \mathbf{x}$.

If $\mathbf{x}_1, \dots, \mathbf{x}_r$ are a basis of L then the projection onto L is given by

$$P = X(X^T X)^{-1} X^T,$$

where $X = (\mathbf{x}_1, \dots, \mathbf{x}_r)$. [prove this directly via the definition of the projection matrix or check $P^2 = P$, $P^T = P$, $\underbrace{\text{span}(P)}_{\text{space spanned by the columns of } P} = L$ or .]

If $\mathbf{x}_1, \dots, \mathbf{x}_r$ are an *orthonormal* basis then $P = XX^T$.

$I_n - P$ is the projection matrix onto L^\perp