Projection

Ping Yu

School of Economics and Finance
The University of Hong Kong

Overview

- Whenever we discuss projection, there must be an underlying Hilbert space since we must define "orthogonality".
- We explain projection in two Hilbert spaces (L^2 and $\mathbb{R}^n$) and integrate many estimators in one framework.
- Projection in the L^2 space: linear projection and regression (linear regression is a special case)
- Projection in $\mathbb{R}^n$: Ordinary Least Squares (OLS) and Generalized Least Squares (GLS)
- One main topic of this course is the (ordinary) least squares estimator (LSE).
- Although the LSE has many interpretations, e.g., as a MLE or a MoM estimator, the most intuitive interpretation is that it is a projection estimator.

Hilbert Space and Projection Theorem

Hilbert Space

Definition (Hilbert Space)

A complete inner product space is called a **Hilbert space**.[a] An inner product is a bilinear operator $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$, where H is a real vector space, satisfying for any $x, y, z \in H$ and $\alpha \in \mathbb{R}$,

(i) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$;

(ii) $\langle \alpha x, z \rangle = \alpha \langle x, z \rangle$;

(iii) $\langle x, z \rangle = \langle z, x \rangle$;

(iv) $\langle x, x \rangle \geq 0$ with equal if and only if $x = 0$.

We denote this Hilbert space as $(H, \langle \cdot, \cdot \rangle)$.

[a] A metric space (H, d) is **complete** if every Cauchy sequence in H converges in H, where d is a metric on H. A sequence $\{x_n\}$ in a metric space is called a **Cauchy sequence** if for any $\varepsilon > 0$, there is a positive integer N such that for all natural numbers $m, n > N$, $d(x_m, x_n) < \varepsilon$.

Angle and Orthogonality

- An important inequality in the inner product space is the Cauchy–Schwarz inequality:

$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|,$$

where $\|\cdot\| \equiv \sqrt{\langle \cdot, \cdot \rangle}$ is the norm induced by $\langle \cdot, \cdot \rangle$.

- Due to this inequality, we can define

$$\text{angle}(x, y) = \arccos \frac{\langle x, y\rangle}{\|x\| \cdot \|y\|}.$$

- We assume the value of the angle is chosen to be in the interval $[0, \pi]$.

[Figure Here]

- If $\langle x, y\rangle = 0$, $\text{angle}(x, y) = \frac{\pi}{2}$; we call x is **orthogonal** to y and denote it as $x \perp y$.

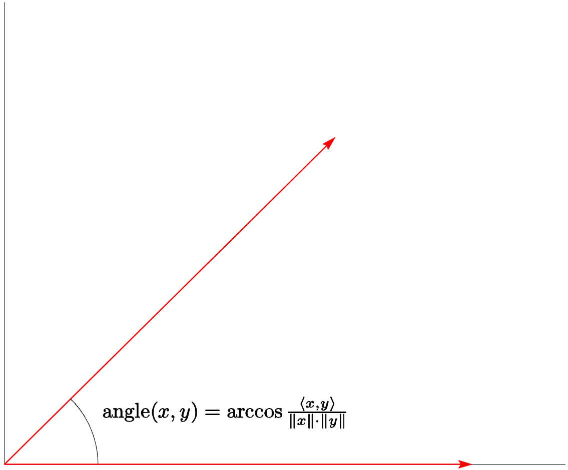

Figure: Angle in Two-dimensional Euclidean Space

Projection and Projector

- The ingredients of a projection are $\{y, M, (H, \langle\cdot,\cdot\rangle)\}$, where M is a subspace of H.
- Note that the same H endowed with different inner products are different Hilbert spaces, so the Hilbert space is denoted as $(H, \langle\cdot,\cdot\rangle)$ rather than H.
- Our objective is to find some $\Pi(y) \in M$ such that

$$\Pi(y) = \arg\min_{h \in M} \|y - h\|^2 . \tag{1}$$

- $\Pi(\cdot)$: $H \to M$ is called a **projector**, and $\Pi(y)$ is called a **projection** of y.

Direct Sum, Orthogonal Space and Orthogonal Projector

Definition

Let M_1 and M_2 be two disjoint subspaces of H so that $M_1 \cap M_2 = \{0\}$. The space

$$V = \{h \in H | h = h_1 + h_2, h_1 \in M_1, h_2 \in M_2\}$$

is called the **direct sum** of M_1 and M_2 and it is denoted by $V = M_1 \oplus M_2$.

Definition

Let M be a subspace of H. The space

$$M^{\perp} \equiv \{h \in H | \langle h, M \rangle = 0\}$$

is called the **orthogonal space** or **orthogonal complement** of M, where $\langle h, M \rangle = 0$ means h is orthogonal to every element in M.

Definition

Suppose $H = M_1 \oplus M_2$. Let $h \in H$ so that $h = h_1 + h_2$ for unique $h_i \in M_i$, $i = 1, 2$. Then P is a **projector** onto M_1 along M_2 if $Ph = h_1$ for all h. In other words, $PM_1 = M_1$ and $PM_2 = 0$. When $M_2 = M_1^{\perp}$, we call P as an **orthogonal projector**.

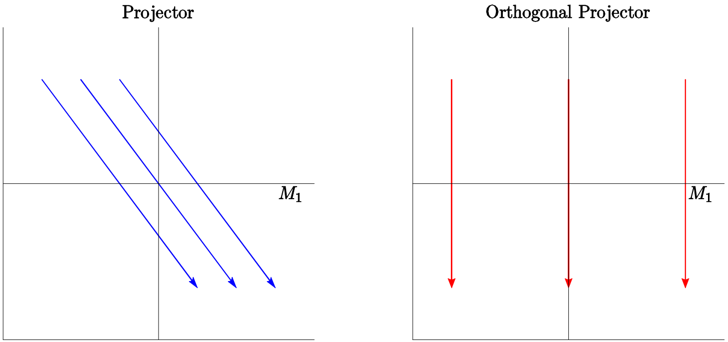

Figure: Projector and Orthogonal Projector

- What is M_2?

[Back to Lemma 9]

Hilbert Projection Theorem

Theorem (Hilbert Projection Theorem)

*If M is a **closed** subspace of a Hilbert space H, then for each $y \in H$, there exists a **unique** point $x \in M$ for which $\|y - x\|$ is minimized over M. Moreover, x is the closest element in M to y **if and only if** $\langle y - x, M \rangle = 0$.*

- The first part of the theorem states the existence and uniqueness of the projector.
- The second part of the theorem states something related to the first order conditions (FOCs) of (1) or, simply, orthogonal conditions.
- From the theorem, given any closed subspace M of H, $H = M \oplus M^{\perp}$.
- Also, the closest element in M to y is determined by M itself, not the vectors generating M since there may be some redundancy in these vectors.

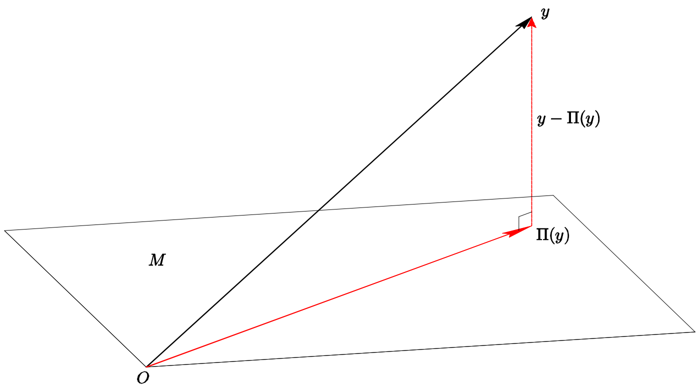

Figure: Projection

Sequential Projection

Theorem (Law of Iterated Projections or LIP)

If M_1 and M_2 are closed subspaces of a Hilbert space H, and $M_1 \subset M_2$, then $\Pi_1(y) = \Pi_1(\Pi_2(y))$, where $\Pi_j(\cdot)$, $j = 1, 2$, is the orthogonal projector of y onto M_j.

Proof.

Write $y = \Pi_2(y) + \Pi_2^{\perp}(y)$. Then

$$\Pi_1(y) = \Pi_1(\Pi_2(y) + \Pi_2^{\perp}(y)) = \Pi_1(\Pi_2(y)) + \Pi_1(\Pi_2^{\perp}(y)) = \Pi_1(\Pi_2(y)),$$

where the last equality is because $\langle \Pi_2^{\perp}(y), x \rangle = 0$ for any $x \in M_2$ and $M_1 \subset M_2$. □

- We first project y onto a larger space M_2, and then project the projection of y (in the first step) onto a smaller space M_1.
- The theorem shows that such a sequential procedure is equivalent to projecting y onto M_1 directly.
- We will see some applications of this theorem below.

Projection in the L^2 Space

Linear Projection

- A random variable $x \in L^2(P)$ if $E\left[x^2\right] < \infty$.
- $L^2(P)$ endowed with some inner product is a Hilbert space.
- $y \in L^2(P)$, $x_1, \cdots, x_k \in L^2(P)$, $M = span\left(x_1, \cdots, x_k\right) \equiv span(\mathbf{x})$,[1] $H = L^2(P)$ with $\langle \cdot, \cdot \rangle$ defined as $\langle x, y \rangle = E\left[xy\right]$.

$$\begin{aligned} \Pi(y) &= \arg\min_{h \in M} E\left[(y-h)^2\right] \\ &= \mathbf{x}' \cdot \arg\min_{\beta \in \mathbb{R}^k} E\left[\left(y - \mathbf{x}'\beta\right)^2\right] \end{aligned} \tag{2}$$

is called the **best linear predictor** (BLP) of y given $\mathbf{x}$, or the linear projection of y onto $\mathbf{x}$.

[1] $span(\mathbf{x}) = \left\{z \in L^2(P) | z = \mathbf{x}'\alpha, \alpha \in \mathbb{R}^k\right\}$.

continue...

- Since this is a concave programming problem, FOCs are sufficient[2]:

$$-2E\left[\mathbf{x}\left(y-\mathbf{x}'\beta_0\right)\right]=\mathbf{0}\Rightarrow E\left[\mathbf{x}u\right]=\mathbf{0} \tag{3}$$

  where $u=y-\Pi(y)$ is the error, and $\beta_0=\arg\min_{\beta\in\mathbb{R}^k}E\left[\left(y-\mathbf{x}'\beta\right)^2\right]$.

- $\Pi(y)$ always exists and is unique, but β_0 needn't be unique unless $x_1,\cdots,x_k$ are linearly independent, that is, there is no nonzero vector $\mathbf{a}\in\mathbb{R}^k$ such that $\mathbf{a}'\mathbf{x}=0$ almost surely (a.s.).
- Why? If $\forall\ \mathbf{a}\neq 0$, $\mathbf{a}'\mathbf{x}\neq 0$, then $E\left[\left(\mathbf{a}'\mathbf{x}\right)^2\right]>0$ and $\mathbf{a}'E\left[\mathbf{x}\mathbf{x}'\right]\mathbf{a}>0$, thus $E\left[\mathbf{x}\mathbf{x}'\right]>0$. So from (3),

$$\beta_0=\left(E\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1}E\left[\mathbf{x}y\right]\ \text{(why?)} \tag{4}$$

  and $\Pi(y)=\mathbf{x}'\left(E\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1}E\left[\mathbf{x}y\right]$.

- In the literature, β with a subscript 0 usually represents the true value of β.

[2] $\frac{\partial}{\partial\mathbf{x}}\left(\mathbf{a}'\mathbf{x}\right)=\frac{\partial}{\partial\mathbf{x}}\left(\mathbf{x}'\mathbf{a}\right)=\mathbf{a}$

Regression

- The setup is the same as in linear projection except that $M = L^2(P, \sigma(\mathbf{x}))$, where $L^2(P, \sigma(\mathbf{x}))$ is the space spanned by any function of $\mathbf{x}$ (not only the linear function of $\mathbf{x}$) as long as it is in $L^2(P)$.

$$\Pi(y) = \arg\min_{h \in M} E\left[(y-h)^2\right] \tag{5}$$

- Note that

$$\begin{aligned}
& E\left[(y-h)^2\right] \\
&= E\left[(y - E[y|\mathbf{x}] + E[y|\mathbf{x}] - h)^2\right] \\
&= E\left[(y - E[y|\mathbf{x}])^2\right] + 2E\left[(y - E[y|\mathbf{x}])(E[y|\mathbf{x}] - h)\right] + E\left[(E[y|\mathbf{x}] - h)^2\right] \\
&\overset{?}{=} E\left[(y - E[y|\mathbf{x}])^2\right] + E\left[(E[y|\mathbf{x}] - h)^2\right] \geq E\left[(y - E[y|\mathbf{x}])^2\right] \equiv E[u^2],
\end{aligned}$$

so $\Pi(y) = E[y|\mathbf{x}]$, which is called the **population regression function** (PRF), where the error u satisfies $E[u|\mathbf{x}] = 0$ (why?).

- We can use variation to characterize the FOCs:

$$\begin{aligned}
& 0 = \arg\min_{\varepsilon \in \mathbb{R}} E\left[(y - (\Pi(y) + \varepsilon h(\mathbf{x})))^2\right] \\
& -2\, E\left[h(\mathbf{x})(y - (\Pi(y) + \varepsilon h(\mathbf{x})))\right]\big|_{\varepsilon=0} = 0 \\
& \Rightarrow E\left[h(\mathbf{x})u\right] = 0, \ \forall\, h(\mathbf{x}) \in L^2(P, \sigma(\mathbf{x}))
\end{aligned} \tag{6}$$

Relationship Between the Two Projections

- $\Pi_1(y)$ is the BLP of $\Pi_2(y)$ given $\mathbf{x}$, i.e., the BLPs of y and $\Pi_1(y)$ given $\mathbf{x}$ are the same.
- This is a straightforward application of the law of iterated projections.
- Explicitly, define

$$\beta_o = \arg\min_{\beta\in\mathbb{R}^k} E\left[\left(E\left[y|\mathbf{x}\right]-\mathbf{x}'\beta\right)^2\right] = \arg\min_{\beta\in\mathbb{R}^k} \int \left[\left(E\left[y|\mathbf{x}\right]-\mathbf{x}'\beta\right)^2\right] dF(\mathbf{x}).$$

- The FOCs for this minimization problem are

$$\begin{aligned} & E\left[-2\mathbf{x}\left(E\left[y|\mathbf{x}\right]-\mathbf{x}'\beta_o\right)\right] = \mathbf{0} \\ \Rightarrow\ & E\left[\mathbf{x}\mathbf{x}'\right]\beta_o = E\left[\mathbf{x}E\left[y|\mathbf{x}\right]\right] = E\left[\mathbf{x}y\right] \\ \Rightarrow\ & \beta_o = \left(E\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} E\left[\mathbf{x}y\right] = \beta_0 \end{aligned}$$

- In other words, β_0 is a (weighted) least squares approximation to the true model.
- If $E\left[y|\mathbf{x}\right]$ is not linear in $\mathbf{x}$, β_o depends crucially on the weighting function $F(\mathbf{x})$ or the distribution of $\mathbf{x}$.
- The weighting function ensures that frequently drawn $\mathbf{x}_i$ will yield small approximation errors at the cost of larger approximation errors for less frequently drawn $\mathbf{x}_i$.

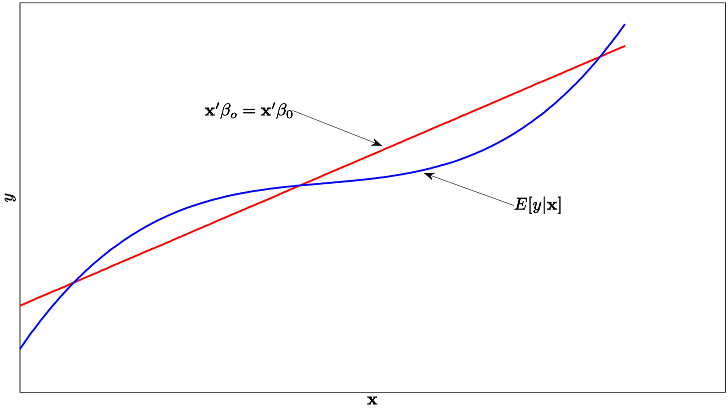

Figure: Linear Approximation of Conditional Expectation (I)

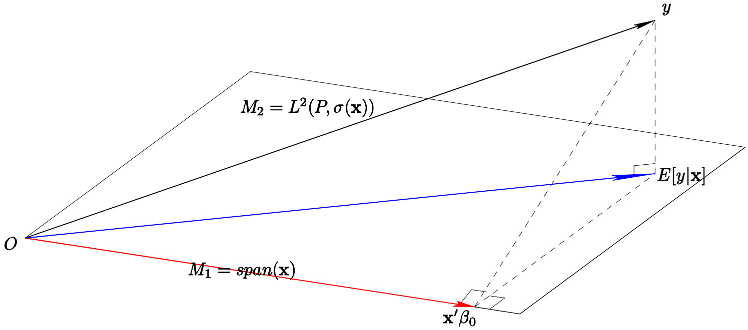

Figure: Linear Approximation of Conditional Expectation (II)

Linear Regression

- Linear regression is a special case of regression with $E[y|\mathbf{x}] = \mathbf{x}'\beta$.
- Regression and linear projection are implied by the definition of projection, but linear regression is a "model" where some structure (or restriction) is imposed.
- In the following figure, when we project y onto a larger space $M_2 = L^2(P, \sigma(\mathbf{x}))$, $\Pi(y)$ falls into a smaller space $M_1 = span(\mathbf{x})$ by coincidence, so there must be a restriction on the joint distribution of $(y, \mathbf{x})$ (what kind of restriction?).
- In summary, the **linear regression model** is

$$\begin{aligned} y &= \mathbf{x}'\beta + u, \\ E[u|\mathbf{x}] &= 0. \end{aligned}$$

- $E[u|\mathbf{x}] = 0$ is necessary for a causal interpretation of β.

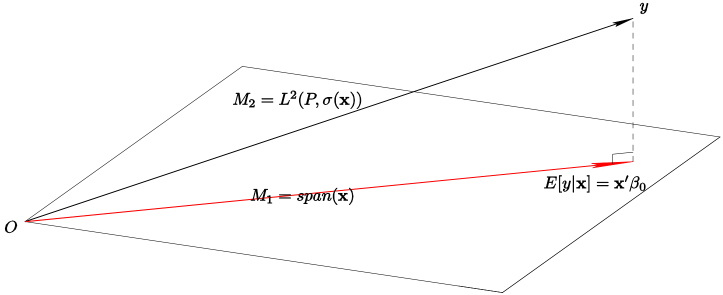

Figure: Linear Regression

Projection in $\mathbb{R}^n$

The LSE

- The projection in the L^2 space is treated as the population version.
- The projection in $\mathbb{R}^n$ is treated as the sample counterpart of the population version.
- The LSE is defined as

$$\widehat{\beta} = \arg\min_{\beta \in \mathbb{R}^k} SSR(\beta) = \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\beta\right)^2 = \arg\min_{\beta \in \mathbb{R}^k} E_n\left[\left(y - \mathbf{x}'\beta\right)^2\right],$$

where $E_n[\cdot]$ is the expectation under the empirical distribution of the data, and

$$SSR(\beta) \equiv \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i'\beta\right)^2 = \sum_{i=1}^{n} y_i^2 - 2\beta' \sum_{i=1}^{n} \mathbf{x}_i y_i + \beta' \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \beta$$

is the sum of squared residuals as a function of β.

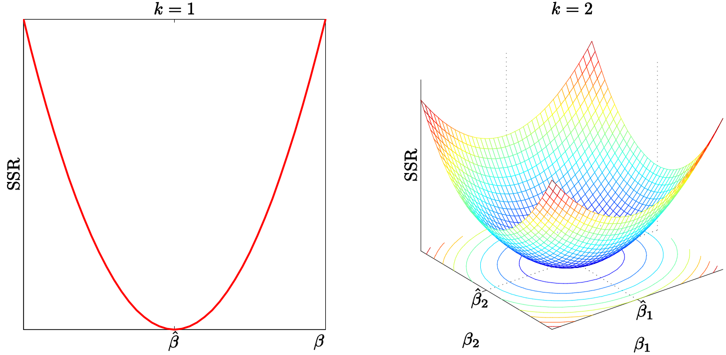

Figure: Objective Functions of OLS Estimation: $k = 1, 2$

Normal Equations

- $SSR(\beta)$ is a quadratic function of β, so the FOCs are also sufficient to determine the LSE.
- Matrix calculus[3] gives the FOCs for $\widehat{\beta}$:

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \beta} SSR(\widehat{\beta}) = -2\sum_{i=1}^{n} \mathbf{x}_i y_i + 2\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \widehat{\beta} \\ &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\widehat{\beta}, \end{aligned}$$

  which is equivalent to the **normal equations**

$$\mathbf{X}'\mathbf{X}\widehat{\beta} = \mathbf{X}'\mathbf{y}.$$

- So

$$\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

[3] $\frac{\partial}{\partial \mathbf{x}}(\mathbf{a}'\mathbf{x}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{a}) = \mathbf{a}$, and $\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}'\mathbf{A}\mathbf{x}) = (\mathbf{A}+\mathbf{A}')\mathbf{x}$.

Notations

- Matrices are represented using uppercase bold. In matrix notation the **sample** (**data**, or **dataset**) is $(\mathbf{y}, \mathbf{X})$, where $\mathbf{y}$ is an $n \times 1$ vector with ith entry y_i and $\mathbf{X}$ is a matrix with ith row $\mathbf{x}_i'$, i.e.,

$$\underset{(n\times 1)}{\mathbf{y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \text{ and } \underset{(n\times k)}{\mathbf{X}} = \begin{pmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{pmatrix},$$

- The first column of $\mathbf{X}$ is assumed to be ones if without further specification, i.e., the first column of $\mathbf{X}$ is

$$\mathbf{1} = (1, \cdots, 1)'.$$

- The bold zero, $\mathbf{0}$, denotes a vector or matrix of zeros.
- Reexpress $\mathbf{X}$ as

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{X}_k \end{pmatrix},$$

  where different from $\mathbf{x}_i$, $\mathbf{X}_j$, $j = 1, \cdots, k$, represents the jth column of $\mathbf{X}$ and is all the observations for jth variable.
- The linear regression model upon stacking all n observations is then

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u},$$

  where $\mathbf{u}$ is an $n \times 1$ column vector with ith entry u_i.

LSE as a Projection

- The above derivation of $\widehat{\beta}$ expresses the LSE using rows of the data matrices **y** and **X**. The following expresses the LSE using columns of **y** and **X**.
- $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X}_1, \cdots, \mathbf{X}_k \in \mathbb{R}^n$ are linearly independent, $M = span(\mathbf{X}_1, \cdots, \mathbf{X}_k) \equiv span(\mathbf{X})$,[4] $H = \mathbb{R}^n$ with the Euclidean inner product.[5]

$$\begin{aligned} \Pi(\mathbf{y}) &= \arg\min_{h \in M} \|\mathbf{y} - h\|^2 \\ &= \mathbf{X} \cdot \arg\min_{\beta \in \mathbb{R}^k} \|\mathbf{y} - \mathbf{X}\beta\|^2 \\ &= \mathbf{X} \cdot \arg\min_{\beta \in \mathbb{R}^k} \sum_{i=1}^n \left(y_i - \mathbf{x}_i'\beta\right)^2, \end{aligned} \tag{7}$$

where $\sum_{i=1}^n \left(y_i - \mathbf{x}_i'\beta\right)^2$ is exactly the objective function of OLS.

[4] $span(\mathbf{X}) = \left\{\mathbf{z} \in \mathbb{R}^n | \mathbf{z} = \mathbf{X}\alpha, \alpha \in \mathbb{R}^k\right\}$ is called the **column space** or **range space** of **X**.

[5] Recall that for $\mathbf{x} = (x_1, \cdots, x_n)$, and $\mathbf{z} = (z_1, \cdots, z_n)$, the Euclidean inner product of **x** and **z** is $\langle \mathbf{x}, \mathbf{z} \rangle = \sum_{i=1}^n x_i z_i$, so $\|\mathbf{x}\|^2 = \langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^n x_i^2$.

continue...

- As $\Pi(\mathbf{y}) = \mathbf{X}\widehat{\beta}$, we can solve out $\widehat{\beta}$ by premultiplying both sides by $\mathbf{X}$, that is,

$$\mathbf{X}'\Pi(\mathbf{y}) = \mathbf{X}'\mathbf{X}\widehat{\beta} \Rightarrow \widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Pi(\mathbf{y}),$$

  where $(\mathbf{X}'\mathbf{X})^{-1}$ exists because $\mathbf{X}$ is full rank.
- On the other hand, orthogonal conditions for this optimization problem are

$$\mathbf{X}'\widehat{\mathbf{u}} = \mathbf{0},$$

  where $\widehat{\mathbf{u}} = \mathbf{y} - \Pi(\mathbf{y})$.
- Since these orthogonal conditions are equivalent to normal equations (or the FOCs), $\widehat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- These two $\widehat{\beta}$'s are the same since $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\Pi(\mathbf{y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\widehat{\mathbf{u}} = 0$.
- Finally,

$$\Pi(\mathbf{y}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{P_X}\mathbf{y},$$

  where $\mathbf{P_X}$ is called the **projection matrix**.

Multicollinearity

- In the above calculation, we first project **y** on *span*$(\mathbf{X})$ and then find $\widehat{\beta}$ by solving $\Pi(\mathbf{y}) = \mathbf{X}\widehat{\beta}$.
- The two steps involve very different operations: optimization versus solving linear equations.
- Furthermore, although $\Pi(\mathbf{y})$ is unique, $\widehat{\beta}$ may not be. When $\text{rank}(\mathbf{X}) < k$ or **X** is *rank deficient*, there are more than one (actually, infinite) $\widehat{\beta}$ such that $\mathbf{X}\widehat{\beta} = \Pi(\mathbf{y})$.
- This is called **multicollinearity** and will be discussed in more details in the next chapter.
- In the following discussion, we always assume $\text{rank}(\mathbf{X}) = k$ or **X** is *full-column rank*; otherwise, some columns of **X** can be deleted to make it so.

Generalized Least Squares

- All are the same as in the last example except $\langle \mathbf{x}, \mathbf{z} \rangle_{\mathbf{W}} = \mathbf{x}'\mathbf{W}\mathbf{z}$, where the weight matrix $\mathbf{W}$ is positive definite and denoted as $\mathbf{W} > 0$.
- The projection
$$\Pi(\mathbf{y}) = \mathbf{X} \cdot \arg\min_{\beta \in \mathbb{R}^k} \|\mathbf{y} - \mathbf{X}\beta\|^2_{\mathbf{W}}. \tag{8}$$
- FOCs are
$$\langle \mathbf{X}, \widetilde{\mathbf{u}} \rangle_{\mathbf{W}} = 0 \text{ (orthogonal conditions)}$$
where $\widetilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widetilde{\beta}$, that is,
$$\langle \mathbf{X}, \mathbf{X} \rangle_{\mathbf{W}} \widetilde{\beta} = \langle \mathbf{X}, \mathbf{y} \rangle_{\mathbf{W}} \Rightarrow \widetilde{\beta} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$
- Thus
$$\Pi(\mathbf{y}) = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y} = \mathbf{P}_{\mathbf{X}\perp\mathbf{W}\mathbf{X}}\mathbf{y}$$
where the notation $P_{\mathbf{X}\perp\mathbf{W}\mathbf{X}}$ will be explained later.

Projection Matrices

- Since $\Pi(\mathbf{y}) = \mathbf{P_X y}$ is the orthogonal projection onto *span*$(\mathbf{X})$, $\mathbf{P_X}$ is the orthogonal projector onto *span*$(\mathbf{X})$.
- Similarly, $\widehat{\mathbf{u}} = \mathbf{y} - \Pi(\mathbf{y}) = (\mathbf{I}_n - \mathbf{P_X})\mathbf{y} \equiv \mathbf{M_X y}$ is the orthogonal projection onto *span*$^\perp(\mathbf{X})$, so $\mathbf{M_X}$ is the orthogonal projector onto *span*$^\perp(\mathbf{X})$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix.
- Since

$$\begin{aligned}\mathbf{P_X X} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{X}, \\ \mathbf{M_X X} &= (\mathbf{I}_n - \mathbf{P_X})\mathbf{X} = \mathbf{0};\end{aligned}$$

  we say $\mathbf{P_X}$ *preserves span*$(\mathbf{X})$, $\mathbf{M_X}$ *annihilates span*$(\mathbf{X})$, and $\mathbf{M_X}$ is called the **annihilator**.
- This implies another way to express $\widehat{\mathbf{u}}$:

$$\widehat{\mathbf{u}} = \mathbf{M_X y} = \mathbf{M_X}(\mathbf{X}\beta + \mathbf{u}) = \mathbf{M_X u}.$$

- Also, it is easy to check $\mathbf{M_X P_X} = \mathbf{0}$, so $\mathbf{M_X}$ and $\mathbf{P_X}$ are orthogonal.

continue...

- $\mathbf{P_X}$ is symmetric: $\mathbf{P}_\mathbf{X}' = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P_X}$.
- $\mathbf{P_X}$ is idempotent[6](intuition?): $\mathbf{P}_\mathbf{X}^2 = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = \mathbf{P_X}$.
- $\mathbf{P_X}$ is positive semidefinite: for any $\alpha \in \mathbb{R}^n$, $\alpha'\mathbf{P_X}\alpha = (\mathbf{X}'\alpha)'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\alpha \geq 0$,
- "Positive semidefinite" cannot be strengthen to "positive definite".
- Why? For an idempotent matrix, the rank equals the trace[7].

$$\mathrm{tr}(\mathbf{P_X}) = \mathrm{tr}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') = \mathrm{tr}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}) = \mathrm{tr}(\mathbf{I}_k) = k < n,$$

and

$$\mathrm{tr}(\mathbf{M_X}) = \mathrm{tr}(\mathbf{I}_n - \mathbf{P_X}) = \mathrm{tr}(\mathbf{I}_n) - \mathrm{tr}(\mathbf{P_X}) = n - k < n.$$

- For a general "nonorthogonal" projector $\mathbf{P}$, it is still unique and idempotent, but need not be symmetric (let alone positive semidefiniteness).
- For example, $\mathbf{P}_{\mathbf{X}\perp\mathbf{WX}}$ in the GLS estimation is not symmetric.

---

[6]A square matrix $\mathbf{A}$ is **idempotent** if $\mathbf{A}^2 = \mathbf{AA} = \mathbf{A}$.

[7]Trace of a square matrix is the sum of its diagonal elements. $\mathrm{tr}(\mathbf{A}+\mathbf{B}) = \mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B})$ and $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$.

Partitioned Fit and Residual Regression

Partitioned Fit

- It is of interest to understand the meaning of part of $\widehat{\beta}$, say, $\widehat{\beta}_1$ in the partition of $\widehat{\beta} = (\widehat{\beta}_1', \widehat{\beta}_2')'$, where we partition

$$\mathbf{X}\beta = \left[\mathbf{X}_1 \vdots \mathbf{X}_2 \right] \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

  with $\text{rank}(\mathbf{X}) = k$.
- We will show that $\widehat{\beta}_1$ is the "net" effect of $\mathbf{X}_1$ on $\mathbf{y}$ when the effect of $\mathbf{X}_2$ is removed from the system. This result is called the Frisch-Waugh-Lovell (FWL) theorem due to Frisch and Waugh (1933) and Lovell (1963).
- The FWL theorem is an excellent implication of the projection property of least squares.
- To simplify notation, $\mathbf{P}_j \equiv \mathbf{P}_{\mathbf{X}_j}$, $\mathbf{M}_j \equiv \mathbf{M}_{\mathbf{X}_j}$, $\Pi_j(\mathbf{y}) = \mathbf{X}_j \widehat{\beta}_j$, $j = 1, 2$.

The FWL Theorem

Theorem

$\widehat{\beta}_1$ could be obtained when the residuals from a regression of $\mathbf{y}$ on $\mathbf{X}_2$ alone are regressed on the set of residuals obtained when each column of $\mathbf{X}_1$ is regressed on $\mathbf{X}_2$. In mathematical notations,

$$\widehat{\beta}_1 = \left(\mathbf{X}'_{1\perp 2}\mathbf{X}_{1\perp 2}\right)^{-1}\mathbf{X}'_{1\perp 2}\mathbf{y}_{\perp 2} = \left(\mathbf{X}'_1\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}'_1\mathbf{M}_2\mathbf{y}.$$

where $\mathbf{X}_{1\perp 2} = (\mathbf{I}-\mathbf{P}_2)\mathbf{X}_1 = \mathbf{M}_2\mathbf{X}_1$, $\mathbf{y}_{\perp 2} = (\mathbf{I}-\mathbf{P}_2)\mathbf{y} = \mathbf{M}_2\mathbf{y}$.

- This theorem states that $\widehat{\beta}_1$ can be calculated by the OLS regression of $\widetilde{\mathbf{y}} = \mathbf{M}_2\mathbf{y}$ on $\widetilde{\mathbf{X}}_1 = \mathbf{M}_2\mathbf{X}_1$. This technique is called **residual regression**.

Corollary

$$\Pi_1(\mathbf{y}) \equiv \mathbf{X}_1\widehat{\beta}_1 = \mathbf{X}_1\left(\mathbf{X}'_{1\perp 2}\mathbf{X}_1\right)^{-1}\mathbf{X}'_{1\perp 2}\mathbf{y} \equiv \mathbf{P}_{12}\mathbf{y} = \mathbf{P}_{12}(\Pi(\mathbf{y})).$$

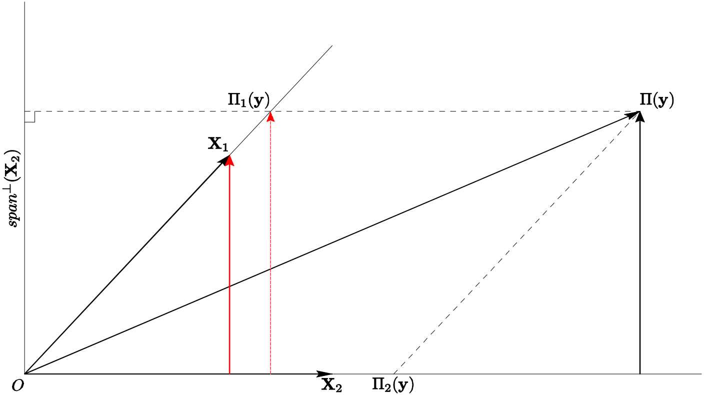

Figure: The FWL Theorem

$\mathbf{P}_{12}$

$$\mathbf{P}_{12} = \underbrace{\mathbf{X}_1}_{\text{trailing term}} \left(\mathbf{X}_1'(\mathbf{I}-\mathbf{P}_2)\mathbf{X}_1\right)^{-1} \underbrace{\mathbf{X}_1'(\mathbf{I}-\mathbf{P}_2)}_{\text{leading term}}.$$

- $\mathbf{I}-\mathbf{P}_2$ in the leading term annihilates $span(\mathbf{X}_2)$ so that $\mathbf{P}_{12}(\Pi_2(\mathbf{y})) = 0$. The leading term sends $\Pi(\mathbf{y})$ toward $span^{\perp}(\mathbf{X}_2)$.
- But the trailing $\mathbf{X}_1$ ensures that the final result will lie in $span(\mathbf{X}_1)$.
- The rest of the expression for $\mathbf{P}_{12}$ ensures that $\mathbf{X}_1$ is preserved under the transformation: $\mathbf{P}_{12}\mathbf{X}_1 = \mathbf{X}_1$.
- Why $\mathbf{P}_{12}\mathbf{y} = \mathbf{P}_{12}(\Pi(\mathbf{y}))$? We can treat the projector $\mathbf{P}_{12}$ as a sequential projector: first project $\mathbf{y}$ onto $span(\mathbf{X})$ to get $\Pi(\mathbf{y})$, and then project $\Pi(\mathbf{y})$ to $span(\mathbf{X}_1)$ along $span(\mathbf{X}_2)$ to get $\Pi_1(\mathbf{y})$.
- $\widehat{\beta}_1$ is calculated from $\Pi_1(\mathbf{y})$ by

$$\widehat{\beta}_1 = (\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'\Pi_1(\mathbf{y}).$$

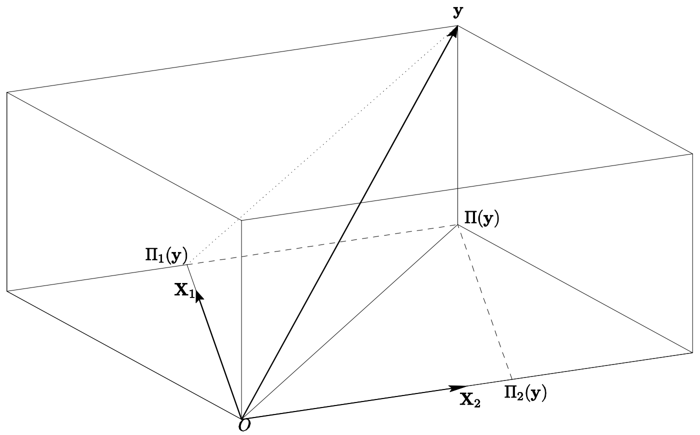

Figure: Projection by $\mathbf{P}_{12}$

Proof I of the FWL Theorem (brute-force)

- Calculate $\widehat{\beta}_1$ explicitly in the residual regression and check whether it is equal to the LSE of β_1.
- Residual regression includes the following three steps.

Step 1: Projecting $\mathbf{y}$ on $\mathbf{X}_2$, we have the residuals

$$\widehat{\mathbf{u}}_y = \mathbf{y} - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y} = \mathbf{M}_2\mathbf{y}.$$

Step 2: Projecting $\mathbf{X}_1$ on $\mathbf{X}_2$, we have the residuals

$$\widehat{\mathbf{U}}_{\mathbf{x}_1} = \mathbf{X}_1 - \mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1 = \mathbf{M}_2\mathbf{X}_1.$$

Step 3: Projecting $\widehat{\mathbf{u}}_y$ on $\widehat{\mathbf{U}}_{\mathbf{x}_1}$, we get the residual regression estimator of β_1

$$\begin{aligned}
\widetilde{\beta}_1 &= \left(\widehat{\mathbf{U}}_{\mathbf{x}_1}'\widehat{\mathbf{U}}_{\mathbf{x}_1}\right)^{-1}\widehat{\mathbf{U}}_{\mathbf{x}_1}'\widehat{\mathbf{u}}_y = (\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}(\mathbf{X}_1'\mathbf{M}_2\mathbf{y}) \\
&= \left[\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{X}_1\right]^{-1}\left[\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{y}\right] \\
&\equiv \mathbf{W}^{-1}\left[\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{y}\right]
\end{aligned}$$

continue...

- On the other hand,

$$\widehat{\beta} = \begin{pmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix}$$
$$= \begin{pmatrix} \mathbf{W}^{-1} & -\mathbf{W}^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1} \\ * & * \end{pmatrix} \begin{pmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{pmatrix},$$

and

$$\widehat{\beta}_1 = \mathbf{W}^{-1}\mathbf{X}_1'\mathbf{y} - \mathbf{W}^{-1}\mathbf{X}_1'\mathbf{X}_2(\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$$
$$= \mathbf{W}^{-1}\left[\mathbf{X}_1'\mathbf{y} - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{y}\right] = \widetilde{\beta}_1.$$

- The partitioned inverse formula:

$$\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \widetilde{\mathbf{A}}_{11}^{-1} & -\widetilde{\mathbf{A}}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\widetilde{\mathbf{A}}_{11}^{-1} & \mathbf{A}_{22}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\widetilde{\mathbf{A}}_{11}^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{pmatrix} \quad (9)$$

where $\widetilde{\mathbf{A}}_{11} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}$.

Proof II of the FWL Theorem

- To show $\widehat{\beta}_1 = \left(\mathbf{X}_{1\perp 2}'\mathbf{X}_{1\perp 2}\right)^{-1}\mathbf{X}_{1\perp 2}'\mathbf{y}_{\perp 2}$, we need only show that

$$\mathbf{X}_1'\mathbf{M}_2\mathbf{y} = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)\widehat{\beta}_1.$$

- Multiplying $\mathbf{y} = \mathbf{X}_1\widehat{\beta}_1 + \mathbf{X}_2\widehat{\beta}_2 + \widehat{\mathbf{u}}$ by $\mathbf{X}_1'\mathbf{M}_2$ on both sides, we have

$$\mathbf{X}_1'\mathbf{M}_2\mathbf{y} = \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\widehat{\beta}_1 + \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_2\widehat{\beta}_2 + \mathbf{X}_1'\mathbf{M}_2\widehat{\mathbf{u}} = \mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\widehat{\beta}_1,$$

where the last equality is from $\mathbf{M}_2\mathbf{X}_2 = 0$, and $\mathbf{X}_1'\mathbf{M}_2\widehat{\mathbf{u}} = \mathbf{X}_1'\widehat{\mathbf{u}} = 0$ (why the first equality hold? $\widehat{\mathbf{u}} = \mathbf{M}\mathbf{u}$ and $\mathbf{M}_2\mathbf{M} = \mathbf{M}$).

$\mathbf{P}_{12}$ as a Projector along a Subspace

Lemma

Define $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}$ *as the projector onto span*$(\mathbf{X})$ *along span*$^{\perp}(\mathbf{Z})$*, where* $\mathbf{X}$ *and* $\mathbf{Z}$ *are* $n \times k$ *matrices and* $\mathbf{Z}'\mathbf{X}$ *is nonsingular. Then* $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}$ *is idempotent, and*

$$\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}} = \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'.$$

- For orthogonal projectors, $\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}\perp\mathbf{X}}$.
- To see the difference between $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}$, we check Figure 2 again.
- In the left panel, $\mathbf{X} = (1,0)'$ and $\mathbf{Z} = (1,1)'$; in the right panel, $\mathbf{X} = (1,0)'$. (why?)
- It is easy to check that

$$\mathbf{P}_{\mathbf{X}} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \text{ and } \mathbf{P}_{\mathbf{X}\perp\mathbf{Z}} = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}.$$

  So an orthogonal projector must be symmetric, while an projector need not be.
- $\mathbf{P}_{12} = \mathbf{P}_{\mathbf{X}_1\perp\mathbf{X}_{1\perp 2}}$.