Chapter 2. Projection*

In this chapter, we explain projection in two Hilbert spaces (L^2 and $\mathbb{R}^n$) and integrate many estimators in one framework. Especially, the least squares estimator (LSE) can be treated as a projection in $\mathbb{R}^n$. Related materials can be found in Chapter 2 of Hansen (2007) and Chapter 2, 3, 4 and 6 of Ruud (2000). A more technical (but still readable) treatment of projection can be found in Yanai et al. (2011).

Although the LSE has many interpretations, e.g., as a MLE or a MoM estimator, the most intuitive interpretation is to employ the language of projection. Projection provides a geometric interpretation of the LSE.

1 Hilbert Space and Projection Theorem

Whenever we discuss projection, there must be an underlying Hilbert space since we must define "orthogonality".

Definition 1 (Hilbert Space) *A complete inner product space is called a* Hilbert space.[1] *An inner product is a bilinear operator $\langle \cdot, \cdot \rangle : H \times H \to \mathbb{R}$, where H is a real vector space,*[2] *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* $x = 0$.

*Email: pingyu@hku.hk

[1] 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 $m, n > N$, $d(x_m, x_n) < \varepsilon$.

[2] A real **vector space** is a set V together with two operations (addition and scalar multiplication) that satisfy the eight axioms listed below, where vectors are distinguished from scalars by boldface. (i) Associativity of addition: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$. (ii) Commutativity of addition: $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$. (iii) Identity element of addition: There exists an element $\mathbf{0} \in V$, called the *zero vector*, such that $\mathbf{v} + \mathbf{0} = \mathbf{v}$ for all $\mathbf{v} \in V$. (iv) Inverse elements of addition: For every $\mathbf{v} \in V$, there exists an element $-\mathbf{v} \in V$, called the *additive inverse* of $\mathbf{v}$, such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$. (v) Compatibility of scalar multiplication with multiplication in $\mathbb{R}$: $a(b\mathbf{v}) = (ab)\mathbf{v}$. (vi) Identity element of scalar multiplication: $1\mathbf{v} = \mathbf{v}$. (vii) Distributivity of scalar multiplication with respect to vector addition: $a(\mathbf{u} + \mathbf{v}) = a\mathbf{u} + a\mathbf{v}$. (viii) Distributivity of scalar multiplication with respect to addition in $\mathbb{R}$: $(a + b)\mathbf{v} = a\mathbf{v} + b\mathbf{v}$.

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

The inner product space is a special case of the normed space which is in turn a special case of the metric space. The complete normed space is called the **Banach space**. Essentially, the metric space defines "distance", the normed space defines "length", and the inner product space defines "angle".

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 $\langle\cdot,\cdot\rangle$.[3] 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 $[0, \pi]$. This is in analogy to the situation in two-dimensional Euclidean space as shown in Figure 1. If $\langle x, y\rangle = 0$, $\text{angle}(x, y) = \frac{\pi}{2}$; we call x is **orthogonal** to y and denote it as $x \perp y$.

The most popular Hilbert spaces are $L^2(P)$ and $\mathbb{R}^n$ associated with some inner products. Note that the same H endowed with different inner products are different Hilbert spaces.

Exercise 1 (Parallelogram Law) *Show that $\|x-y\|^2 + \|x+y\|^2 = 2\|x\|^2 + 2\|y\|^2$. What is the intuition for this law?*

The ingredients of a projection are $\{y, M, (H, \langle\cdot,\cdot\rangle)\}$, where M is a subspace of 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.

The following projection theorem is critical in the following discussion. To ease exposition, we first define the direct sum of two subspaces, the orthogonal space of a subspace and the projector.

Definition 2 *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 $V = M_1 \oplus M_2$.*

Exercise 2 *Show that for any $h \in V$, where $V = M_1 \oplus M_2$, it can be* uniquely *decomposed as $h = h_1 + h_2$, where $h_1 \in M_1$, and $h_2 \in M_2$.*

[3] This is why the inner product space is a special normed space. If we define $d(x, y) = \|x - y\|$, then we can see the normed space is a special metric space.

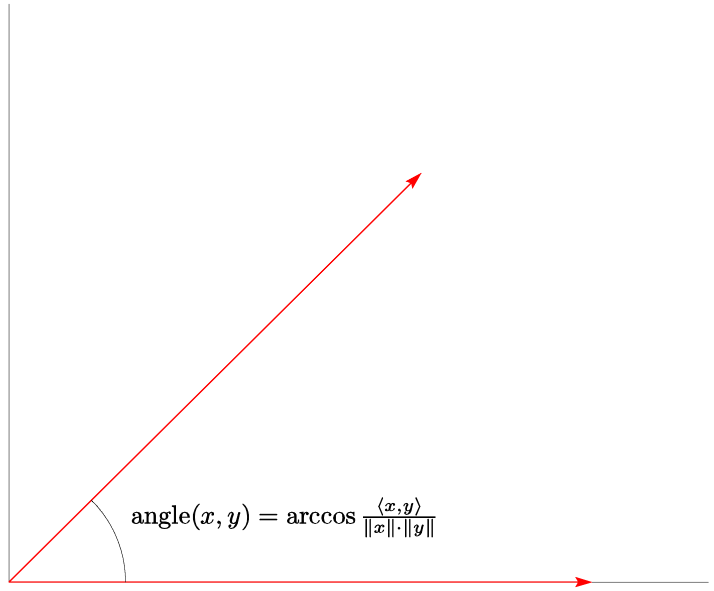
$\text{angle}(x,y) = \arccos \frac{\langle x,y\rangle}{\|x\|\cdot\|y\|}$

Figure 1: Angle in Two-dimensional Euclidean Space

Definition 3 *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 4 *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 $M_2 = M_1^{\perp}$,*[4] *we call P as an* orthogonal projector.

For intuitive illustration of the projector and the orthogonal projector, see Figure 2, where the right panel is an orthogonal projector and the left panel is an projector but is not an orthogonal projector. We label M_1 in the figure, but what is M_2 in the two cases?

Theorem 1 (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 $\langle y - x, M\rangle = 0$.*

Proof (*). 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.

[4] Generally, H cannot be expressed as $M_1 \oplus M_1^{\perp}$ unless M_1 is closed; see the following theorem.

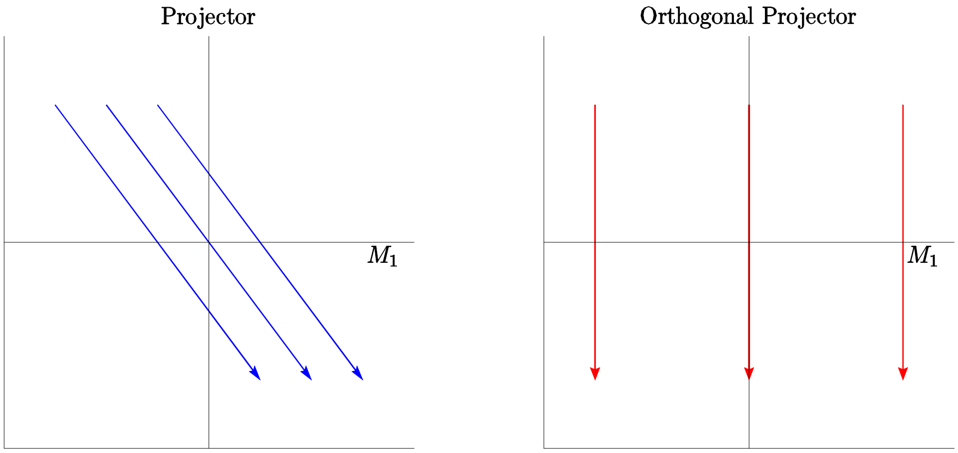

Figure 2: Projector and Orthogonal Projector

Existence of x: Let δ be the distance between y and M, $\{x_n\}$ a sequence in M such that the distance squared between y and x_i is below or equal to $\delta^2 + 1/n$. Let n and m be two integers, then the following equalities are true:

$$\|x_n - x_m\|^2 = \|x_n - y\|^2 + \|x_m - y\|^2 - 2\langle x_n - y, x_m - y\rangle$$

and

$$4\left\|\frac{x_n + x_m}{2} - y\right\|^2 = \|x_n - y\|^2 + \|x_m - y\|^2 + 2\langle x_n - y, x_m - y\rangle .$$

We have therefore

$$\begin{aligned} \|x_n - x_m\|^2 &= 2\|x_n - y\|^2 + 2\|x_m - y\|^2 - 4\left\|\frac{x_n + x_m}{2} - y\right\|^2 \\ &\leq 2\left(\delta^2 + \frac{1}{n}\right) + 2\left(\delta^2 + \frac{1}{m}\right) - 4\delta^2 = 2\left(\frac{1}{n} + \frac{1}{m}\right), \end{aligned}$$

where the inequality is from the assumption on $\{x_n\}$ and the fact that $\frac{x_n + x_m}{2} \in M$. The last inequality proves that $\{x_n\}$ is a Cauchy sequence. Since M is complete, the sequence is therefore convergent to a point x in M, whose distance from y is minimal.

Uniqueness of x: Let x_1 and x_2 be two minimizers. Then

$$\|x_2 - x_1\|^2 = 2\|x_1 - y\|^2 + 2\|x_2 - y\|^2 - 4\left\|\frac{x_1 + x_2}{2} - y\right\|^2 .$$

Since $\frac{x_1 + x_2}{2} \in M$, we have $\left\|\frac{x_1 + x_2}{2} - y\right\|^2 \geq \delta^2$ and therefore

$$\|x_2 - x_1\|^2 \leq 2\delta^2 + 2\delta^2 - 4\delta^2 = 0.$$

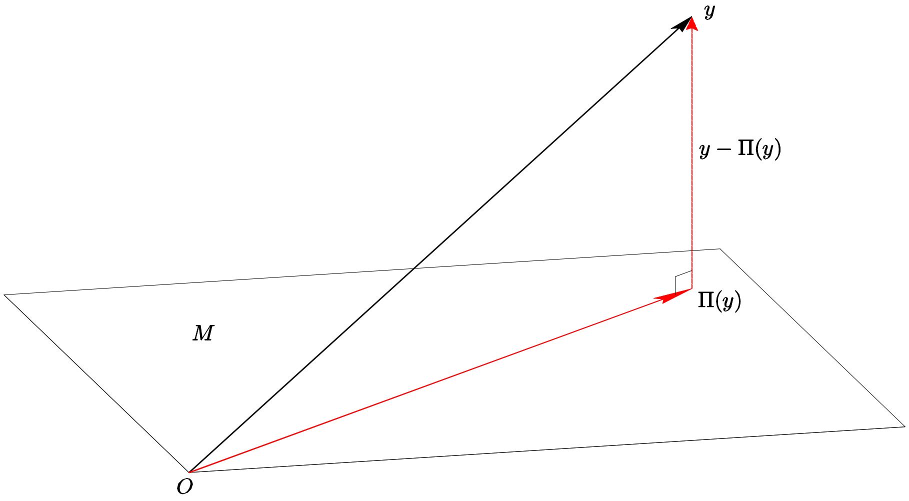

Figure 3: Projection

Hence $x_1 = x_2$, which proves unicity.

Orthogonal Conditions:

$\Longleftarrow$ Let $z \in M$ such that $\langle y - z, a \rangle = 0$ for all $a \in M$.

$$\|y - a\|^2 = \|y - z\|^2 + \|z - a\|^2 + 2\langle y - z, z - a \rangle = \|y - z\|^2 + \|z - a\|^2 \geq \|y - z\|^2,$$

which proves that z is a minimizer.

$\Longrightarrow$ Let $x \in M$ be the minimizer. Let $a \in M$ and $t \in \mathbb{R}$.

$$\|y - (x + ta)\|^2 - \|y - x\|^2 = -2t\langle y - x, a \rangle + t^2 \|a\|^2 = -2t\langle y - x, a \rangle + O\left(t^2\right) \geq 0.$$

Therefore, $\langle y - x, a \rangle = 0$. ■

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. The intuition for the Hilbert projection theorem is illustrated in Figure 3.

Exercise 3 (Pythagorean Theorem) *Prove that if* $x \perp y$ *in* $(H, \langle \cdot, \cdot \rangle)$, *then* $\|x + y\|^2 = \|x\|^2 + \|y\|^2$.

Sometimes, we need a sequential projection procedure. That is, 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 following theorem shows that such a sequential procedure is equivalent to projecting y onto M_1 directly.

Theorem 2 (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 $\left\langle \Pi_2^{\perp}(y), x \right\rangle = 0$ for any $x \in M_2$ and $M_1 \subset M_2$. ∎

Exercise 4 *Show that $\Pi(x+y) = \Pi(x) + \Pi(y)$ and $\Pi(ax) = a\Pi(x)$ for $a \in \mathbb{R}$ and $x, y \in H$.*

2 Projection in the L^2 Space

We will consider projection onto two subspaces of the L^2 space, where $x \in L^2(P)$ means $E[x^2] < \infty$. The first subspace is smaller and generates "linear projection"; the second subspace is larger and generates "regression". Usually, the projections in the L^2 space are treated as the population version of the counterparts in the sample space.[5]

Example 1 (Linear Projection) *$y \in L^2(P)$, $x_1, \cdots, x_k \in L^2(P)$, $M = span(x_1, \cdots, x_k) \equiv span(\mathbf{x})$,[6] $H = L^2(P)$ with $\langle \cdot, \cdot \rangle$ defined as $\langle x, y \rangle = E[xy]$. Now,*

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

is called the ***best linear predictor*** *(BLP)[7] of y given $\mathbf{x}$, or the linear projection of y onto $\mathbf{x}$, and is often denoted as $E^*[y|\mathbf{x}]$. Since this is a concave programming problem, FOCs are sufficient:*

$$-2E\left[\mathbf{x}\left(y - \mathbf{x}'\boldsymbol{\beta}_0\right)\right] = \mathbf{0} \Rightarrow E[\mathbf{x}u] = \mathbf{0} \tag{3}$$

where $u = y - \Pi(y)$ is the error, and $\boldsymbol{\beta}_0 = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} E\left[(y - \mathbf{x}'\boldsymbol{\beta})^2\right]$.[8] Now, the minimization problem (2) is reduced to orthogonal conditions (3) which are some moment conditions. Both of them are useful in solving this problem.

We could see that $\Pi(y)$ always exists and is unique, but $\boldsymbol{\beta}_0$ 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.). In this case, $\boldsymbol{\beta}_0$ is unique. If $\forall\ \mathbf{a} \neq 0$, $\mathbf{a}'\mathbf{x} \neq 0$, then $E\left[(\mathbf{a}'\mathbf{x})^2\right] > 0$ and $\mathbf{a}'E[\mathbf{x}\mathbf{x}']\mathbf{a} > 0$,

[5] Intuitively speaking, the population version is the version where the sample size is *infinite*, while the sample version is the version where the sample size is *finite*. This is why we need to study the bias and variance of the sample-version estimators; for the population version, both bias and variance are zero.

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

[7] BLP is also a short for Berry, Levinsohn and Pakes (1995) if your field is empirical IO.

[8] $\mathbf{x} = (1, x_2, \cdots, x_k)'$, then $E[\mathbf{x}u] = \mathbf{0}$ implies $E[u] = 0$ and $Cov(x_j, u) = 0$, $j = 2, \cdots, k$.

thus $E[\mathbf{x}\mathbf{x}'] > 0$. *So from (3),*

$$\boldsymbol{\beta}_0 = \left(E[\mathbf{x}\mathbf{x}']\right)^{-1} E[\mathbf{x}y] \text{ (why?)} \tag{4}$$

and $\Pi(y) = \mathbf{x}'\left(E[\mathbf{x}\mathbf{x}']\right)^{-1} E[\mathbf{x}y]$. *Note here that we use subscript 0 to denote the projection coefficient in the* L^2 *space; in the literature,* $\boldsymbol{\beta}_0$ *in (4) usually represents the true value of* $\boldsymbol{\beta}$. □

Exercise 5 *Show that* $\langle x, y\rangle = E[xy]$ *is an inner product in* $L^2(P)$.

Exercise 6 *Let* H *is* $L^2(P)$ *excluding constant random variables, i.e., the constant random variables are absorbed in the zero element of* H. *(i) Show that* $Cov(\cdot,\cdot)$ *is an inner product on* H. *(ii) Let* $M = span(x_1, \cdots, x_k)$ *be as in Example 1 with* $Var(\mathbf{x}) > 0$, *but now* $x_1, \cdots, x_k$ *couldn't be constant (otherwise,* $Var(\mathbf{x})$ *cannot be positive definite).* $\forall\ y \in H$, *find* $\Pi(y)$. *(Hint:* $\mathbf{x}'Var(\mathbf{x})^{-1}Cov(\mathbf{x}, y)$.*)*

Example 2 (Regression) *The setup is the same as Example 1 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* $L^2(P)$. *Now, the problem is*

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

Note that

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

so $\Pi(y) = E[y|\mathbf{x}]$,[9] *which is called the* ***population regression function*** *(PRF), where the error* $u = y - E[y|\mathbf{x}]$ *satisfies* $E[u|\mathbf{x}] = 0$ *(why?), and* $E[(y - E[y|\mathbf{x}])(E[y|\mathbf{x}] - h)] = 0$ *follows from the following exercise. Also, similar to (3), we can use variation to characterize the FOCs:*

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

So $E[h(\mathbf{x})u] = 0$ *is the FOC in this projection problem.* □

Exercise 7 *(i)* Law of Iterated Expectations *(LIE): Show that* $E\left[E[y|\mathbf{x}_1, \mathbf{x}_2]|\mathbf{x}_1\right] = E[y|\mathbf{x}_1]$ *if* y, $\mathbf{x}_1$ *and* $\mathbf{x}_2$ *are all in* $L^2(P)$. *When* $\mathbf{x}_1 = 1$, *what is the intuition for this result? (ii) Show that*

[9] (*) Strictly speaking, we do not need $y \in L^2(P)$ to define $E[y|\mathbf{x}]$. $y \in L^1(P)$, i.e., $E[|y|] < \infty$, is enough for the definition of $E[y|\mathbf{x}]$. Notwithstanding, given that $L^1(P)$ is not a Hilbert space (although it is a Banach space), we assume $y \in L^2(P)$ to use the projection structure of the $L^2(P)$ space. See the technical appendix of this chapter for the formal definition of conditional expectation.

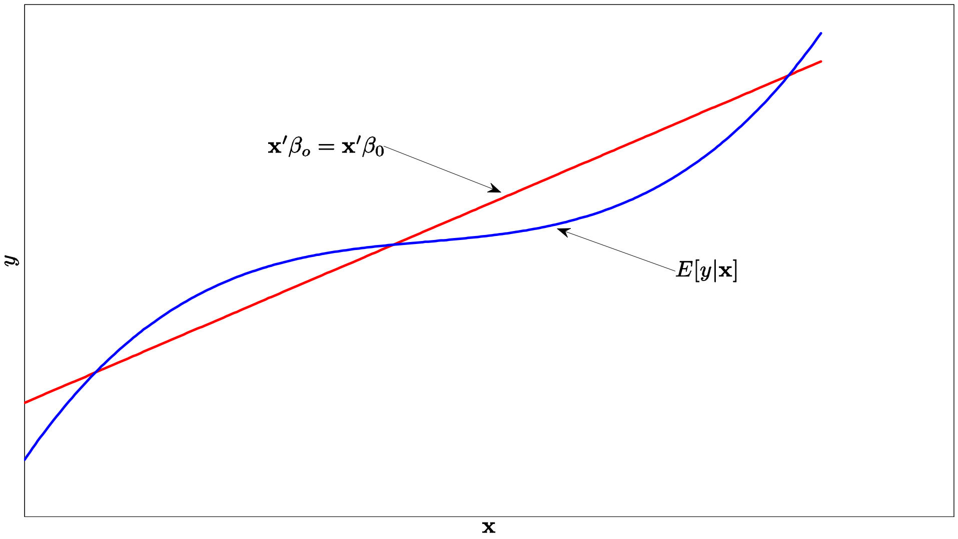

Figure 4: Linear Approximation of Conditional Expectation (I)

$E[h(\mathbf{x})u|\mathbf{x}] = h(\mathbf{x})E[u|\mathbf{x}]$. *What is the intuition for this result? (iii) Based on the results in (i) and (ii), show that if* $E[u|\mathbf{x}] = 0$*, then* $E[u] = 0$*,* $E[\mathbf{x}u] = \mathbf{0}$ *and* $E[h(\mathbf{x})u] = 0$ *for any function* $h(\cdot)$*. (iv) Show that the conditional variance* $\sigma^2(\mathbf{x}) = E\left[y^2|\mathbf{x}\right] - m(\mathbf{x})^2$*, and the unconditional variance* $Var(y) = Var(m(\mathbf{x})) + E\left[\sigma^2(\mathbf{x})\right]$*, where* $m(\mathbf{x}) = E\left[y|\mathbf{x}\right]$*.*

In Example 1, $\Pi_1(y) = \mathbf{x}'\left(E\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} E\left[\mathbf{x}y\right]$, and in Example 2, $\Pi_2(y) = E\left[y|\mathbf{x}\right]$. What is the relationship between $\Pi_1(y)$ and $\Pi_2(y)$? We use Figure 4 and 5 to illustrate this relationship. Basically, $\Pi_1(y)$ 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. Nevertheless, we conduct the following explicit calculation to get more insights. Define

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

The FOCs for this minimization problem are

$$\begin{aligned} & E\left[-2\mathbf{x}\left(E\left[y|\mathbf{x}\right] - \mathbf{x}'\boldsymbol{\beta}_o\right)\right] = \mathbf{0} \\ \Rightarrow\ & E\left[\mathbf{x}\mathbf{x}'\right]\boldsymbol{\beta}_o = E\left[\mathbf{x}E\left[y|\mathbf{x}\right]\right] = E\left[\mathbf{x}y\right] \\ \Rightarrow\ & \boldsymbol{\beta}_o = \left(E\left[\mathbf{x}\mathbf{x}'\right]\right)^{-1} E\left[\mathbf{x}y\right] = \boldsymbol{\beta}_0 \end{aligned}$$

In other words, $\boldsymbol{\beta}_0$ is a (weighted) least squares approximation to the true model. If $E\left[y|\mathbf{x}\right]$ is not linear in $\mathbf{x}$, $\boldsymbol{\beta}_o$ 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 $\mathbf{x}_i$. This approximation was investigated

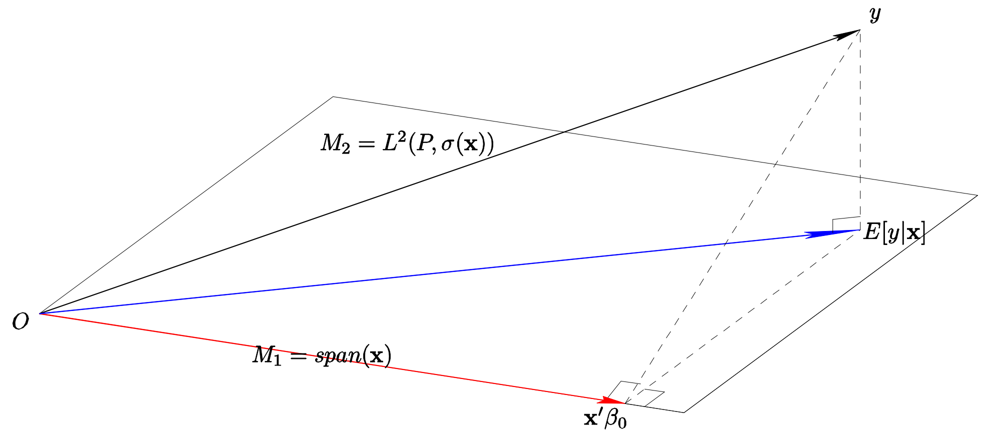

Figure 5: Linear Approximation of Conditional Expectation (II)

in White (1980a).

Exercise 8 *Show that* $E\left[(y - \mathbf{x}'\boldsymbol{\beta}_o)^2\right] = E\left[(E[y|\mathbf{x}] - \mathbf{x}'\boldsymbol{\beta}_o)^2\right] + E[u^2]$*, where* $u = y - E[y|\mathbf{x}]$*. Interpret this result using Figure 5.*

Exercise 9 *Let* x *and* y *have the joint density* $f(x, y) = \frac{3}{2}(x^2 + y^2)$ *on* $0 \leq x \leq 1$*,* $0 \leq y \leq 1$*. Compute the coefficients of the best linear predictor* $\Pi_1(y) = E^*[y|x] = \beta_1 + \beta_2 x$*. Compute the conditional mean* $\Pi_2(y) = E[y|x]$*. Are they different?*

Exercise 10 (Hájek (1968) Projection *) *Let* $x_1, \cdots, x_k$ *be independent random variables*[10]*, and* $M = \left\{\sum_{i=1}^{k} g_i(x_i) : E\left[g_i^2(x_i)\right] < \infty,\ i = 1, \cdots, k\right\}$*.* $\forall\ y \in L^2(P)$*, find* $\Pi(y)$*. (Hint:* $\Pi(y) = \sum_{i=1}^{k} E\left[y|x_i\right] - (k-1)E[y]$*.)*

Linear regression is a special case of regression with $E[y|\mathbf{x}] = \mathbf{x}'\boldsymbol{\beta}$. From the above discussion, we know regression and linear projection are implied by the definition of projection, but linear regression is a "model" where some structure (or restriction) is imposed. Figure 6 clarifies this point. In Figure 6, 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

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

[10] Note that $x_1, \cdots, x_k$ in Example 1, 2 and Exercise 6 could be dependent.

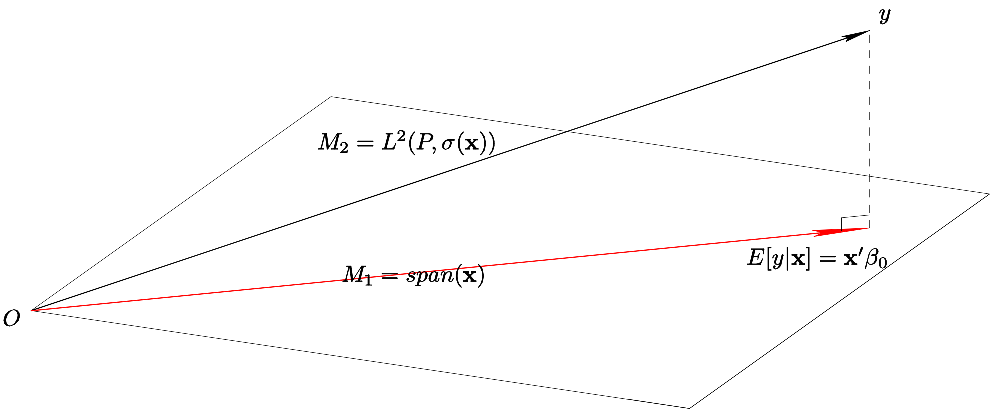
y
$M_2 = L^2(P, \sigma(\mathbf{x}))$
$M_1 = span(\mathbf{x})$
$E[y|\mathbf{x}] = \mathbf{x}'\beta_0$
O

Figure 6: Linear Regression

Recall that $E[u|\mathbf{x}] = 0$ is necessary for a causal interpretation of $\boldsymbol{\beta}$.

Exercise 11 *Suppose that y is discrete-valued, taking values only on the non-negative integers, and the conditional distribution of y given $\mathbf{x}$ is Poisson:*

$$P(y = k|\mathbf{x}) = \frac{\exp(-\mathbf{x}'\boldsymbol{\beta})(\mathbf{x}'\boldsymbol{\beta})^j}{j!}, j = 0, 1, 2, \cdots$$

Compute $E[y|\mathbf{x}]$. Does this justify a linear regression model of the form $y = \mathbf{x}'\boldsymbol{\beta} + u$? (Hint: If $P(y = j) = \frac{\exp(-\lambda)\lambda^j}{j!}$, then $E[y] = \lambda$.)

3 Projection in $\mathbb{R}^n$

We studies two projections in $\mathbb{R}^n$. The first projection generates the **ordinary least squares** (OLS) **estimator** or the LSE,[11] and the second one generates the **generalized least squares** (GLS) **estimator**. The GLS estimator is sometimes called the **Aitken estimator**. Usually, the projection in $\mathbb{R}^n$ is treated as the sample counterpart of the population version.

From elementary econometrics, the least squares estimator (LSE) is defined as

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

[11] The least-squares method is usually credited to Gauss (1809), but it was first published as an appendix to Legendre (1805) which is on the paths of comets. Nevertheless, Gauss claimed that he had been using the method since 1795.

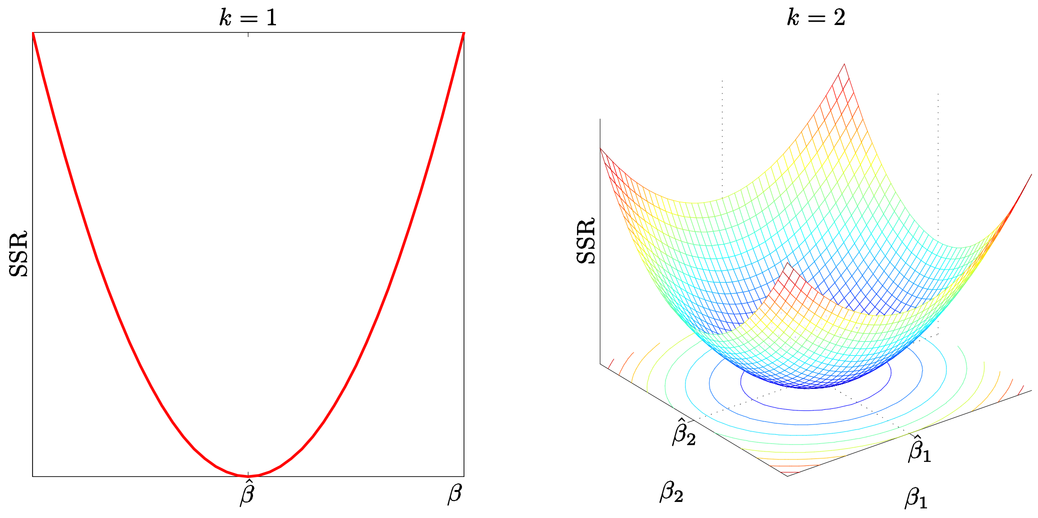

Figure 7: Objective Functions of OLS Estimation: $k = 1, 2$

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

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

is the sum of squared residuals as a function of $\boldsymbol{\beta}$. Figure 7 shows a typical $SSR(\boldsymbol{\beta})$ when $k = 1$ and 2.

$SSR(\boldsymbol{\beta})$ is a quadratic function of $\boldsymbol{\beta}$, so the FOCs are also sufficient to determine the LSE. Matrix calculus[12] gives the FOCs for $\widehat{\boldsymbol{\beta}}$:

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

which is equivalent to the **normal equations**[13]

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

So

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

The above derivation of $\widehat{\boldsymbol{\beta}}$ expresses the LSE using rows of the data matrices $\mathbf{y}$ and $\mathbf{X}$. The following example expresses the LSE using columns of $\mathbf{y}$ and $\mathbf{X}$.

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

[13] The term "normal equations" comes from the fact that they characterize the *normal vector* $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ to $span(\mathbf{X})$ which is defined below.

Example 3 (OLS) *$\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X}_1, \cdots, \mathbf{X}_k \in \mathbb{R}^n$ are linearly independent, $M = span(\mathbf{X}_1, \cdots, \mathbf{X}_k) \equiv span(\mathbf{X})$,[14] $H = \mathbb{R}^n$ with the Euclidean inner product.[15] Now,*

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

where $\sum_{i=1}^n (y_i - \mathbf{x}_i'\boldsymbol{\beta})^2$ is exactly the objective function of OLS. As $\Pi(\mathbf{y}) = \mathbf{X}\widehat{\boldsymbol{\beta}}$, we can solve out $\widehat{\boldsymbol{\beta}}$ by premultiplying both sides by $\mathbf{X}$, that is,

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

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

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

where $\mathbf{P}_{\mathbf{X}}$ is called the ***projection matrix****.* □

Exercise 12 *(i) Someone claims that $\sum_{i=1}^n \widehat{u}_i = 0$. Is this true or not? When is this true? (Hint: $\mathbf{1} \in span(\mathbf{X})$.) (ii) Check that $\Pi(\mathbf{y})'\widehat{\mathbf{u}} = 0$ by explicit calculation.*

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

Exercise 13 *A dummy variable takes on only the values 0 and 1. It is used for categorical data, such as an individual's gender. Let $\mathbf{d}_1$ and $\mathbf{d}_2$ be vectors of 1's and 0's, with the i'th element of $\mathbf{d}_1$ equaling 1 and that of $\mathbf{d}_2$ equaling 0 if the person is a man, and the reverse if the person is a woman. Suppose that there are n_1 men and n_2 women in the sample. Consider the three regressions*

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{d}_1\alpha_1 + \mathbf{d}_2\alpha_2 + \mathbf{u}, \tag{8}$$

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

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

$$\mathbf{y} = \mathbf{d}_1\alpha_1 + \mathbf{d}_2\alpha_2 + \mathbf{u}, \tag{9}$$

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{d}_1\phi + \mathbf{u}, \tag{10}$$

(a) *Can all three regressions (8), (9), and (10) be estimated by OLS? Explain if not.*

(b) *Compare regressions (9) and (10). Is one more general than the other? Explain the relationship between the parameters in (9) and (10).*

(c) *Compute* $\mathbf{1}'\mathbf{d}_1$ *and* $\mathbf{1}'\mathbf{d}_2$*.*

(d) *Letting* $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$*, write equation (9) as* $\mathbf{y} = \mathbf{X}\boldsymbol{\alpha} + \mathbf{u}$*. Consider the assumption* $E[\mathbf{x}_i u_i] = \mathbf{0}$*. Is there any content to this assumption in this setting?*

Example 4 (GLS) *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. The projection*

$$\Pi(\mathbf{y}) = \mathbf{X} \cdot \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^k} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_{\mathbf{W}}^2 . \tag{11}$$

FOCs are

$$\langle \mathbf{X}, \widetilde{\mathbf{u}} \rangle_{\mathbf{W}} = 0 \textit{ (orthogonal conditions)}$$

where $\widetilde{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widetilde{\boldsymbol{\beta}}$*, that is,*

$$\langle \mathbf{X}, \mathbf{X} \rangle_{\mathbf{W}} \widetilde{\boldsymbol{\beta}} = \langle \mathbf{X}, \mathbf{y} \rangle_{\mathbf{W}} \Rightarrow \widetilde{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}\mathbf{y}.$$

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.* □

Exercise 14 *(i) Show that* $\langle \cdot, \cdot \rangle_{\mathbf{W}}$ *is an inner product. (ii) Is* $\mathbf{P}_{\mathbf{X}\perp\mathbf{W}\mathbf{X}}$ *idempotent*[16]*? Why? (iii) Is* $\mathbf{P}_{\mathbf{X}\perp\mathbf{W}\mathbf{X}}$ *symmetric? (iv) Check that* $\Pi(\mathbf{y})'\mathbf{W}\widetilde{\mathbf{u}} = 0$ *by explicit calculation.*

(*) The sample counterpart of regression is out of the scope of this course. Roughly speaking, there are two methods to estimate $E[y|\mathbf{x}]$: the kernel method and the sieve method. See Pagan and Ullah (1999) and Li and Racine (2007) for an introduction on these methods. It is quite natural to interpret the sieve estimator as a projection. As to the kernel estimator, see Mammen et al. (2001) for such an interpretation.

3.1 Projection Matrices

Since $\Pi(\mathbf{y}) = \mathbf{P}_{\mathbf{X}}\mathbf{y}$ is the orthogonal projection onto *span*$(\mathbf{X})$, $\mathbf{P}_{\mathbf{X}}$ is the orthogonal projector onto *span*$(\mathbf{X})$. Similarly, $\widehat{\mathbf{u}} = \mathbf{y} - \Pi(\mathbf{y}) = (\mathbf{I}_n - \mathbf{P}_{\mathbf{X}})\,\mathbf{y} \equiv \mathbf{M}_{\mathbf{X}}\mathbf{y}$ is the orthogonal projection onto

[16] A matrix $\mathbf{A}$ is **idempotent** if $\mathbf{A}$ is square and $\mathbf{A}\mathbf{A} = \mathbf{A}$.

span$^\perp(\mathbf{X})$, so $\mathbf{M}_\mathbf{X}$ is the orthogonal projector onto *span*$^\perp(\mathbf{X})$. Check that

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

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

$$\widehat{\mathbf{u}} = \mathbf{M}_\mathbf{X}\mathbf{y} = \mathbf{M}_\mathbf{X}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{M}_\mathbf{X}\mathbf{u}.$$

Also, it is easy to check $\mathbf{M}_\mathbf{X}\mathbf{P}_\mathbf{X} = \mathbf{0}$, so $\mathbf{M}_\mathbf{X}$ and $\mathbf{P}_\mathbf{X}$ are orthogonal.

Check that

$$\mathbf{P}'_\mathbf{X} = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}_\mathbf{X},$$

so $\mathbf{P}_\mathbf{X}$ is symmetric. Check also that

$$\mathbf{P}^2_\mathbf{X} = \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) = \mathbf{P}_\mathbf{X},$$

so $\mathbf{P}_\mathbf{X}$ is idempotent. Furthermore, for any $\boldsymbol{\alpha} \in \mathbb{R}^n$,

$$\boldsymbol{\alpha}'\mathbf{P}_\mathbf{X}\boldsymbol{\alpha} = \left(\mathbf{X}'\boldsymbol{\alpha}\right)'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\alpha} \geq 0,$$

so $\mathbf{P}_\mathbf{X}$ is positive semidefinite. Symmetry, idempotentness and positive semidefiniteness are actually necessary properties for any orthogonal projector in $\mathbb{R}^n$.

Exercise 15 *(i) Show that* $\mathbf{M}_\mathbf{X}$ *is symmetric, idempotent and positive semidefinite. (ii) In the GLS estimation, define* $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$ *and* $\mathbf{M} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{W}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}$*. Show that* $\mathbf{P}'\mathbf{W}\mathbf{P} = \mathbf{W}\mathbf{P}$ *and* $\mathbf{M}'\mathbf{W}\mathbf{M} = \mathbf{W}\mathbf{M}$*.*

Theorem 3 (*) *If* $\mathbf{P}$ *is an orthogonal projector onto the subspace* M *of* $\mathbb{R}^n$*, then* $\mathbf{P}$ *is unique, symmetric, idempotent, and positive semidefinite and* $\mathbf{I} - \mathbf{P}$ *is an orthogonal projector onto* $M^\perp$*.*

Proof. Unique: Although uniqueness of $\mathbf{P}$ is implied by the Hilbert projection theorem, we use some special structure of $\mathbb{R}^n$ in the following proof. Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be two orthogonal projectors onto M. By the definition of orthogonal projectors, the orthogonal projection is unique: $\mathbf{P}_1\mathbf{z} = \mathbf{P}_2\mathbf{z}$ for all $\mathbf{z} \in \mathbb{R}^n$. Setting $\mathbf{z}$ equal to each of the natural basis vector in $\mathbf{I}_n$, we have the matrix equality $\mathbf{P}_1\mathbf{I}_n = \mathbf{P}_2\mathbf{I}_n$ or $\mathbf{P}_1 = \mathbf{P}_2$. **Symmetric**: By definition, $\mathbf{P}\mathbf{z} \perp (\mathbf{I}_n - \mathbf{P})\mathbf{z}$ for all $\mathbf{z} \in \mathbb{R}^n$. That is,

$$\mathbf{0} = [(\mathbf{I}_n - \mathbf{P})\mathbf{z}]'(\mathbf{P}\mathbf{z}) = \mathbf{z}'(\mathbf{P} - \mathbf{P}'\mathbf{P})\mathbf{z}.$$

Because this is true for *all* $\mathbf{z}$, $\mathbf{P} - \mathbf{P}'\mathbf{P} = \mathbf{0}$ or $\mathbf{P} = \mathbf{P}'\mathbf{P}$. Because $\mathbf{P}'\mathbf{P}$ is symmetric, so is $\mathbf{P}$. **Idempotent**: By definition, $\mathbf{P}\mathbf{z} \in M$ for all $\mathbf{z} \in \mathbb{R}^n$. Also by definition, $\mathbf{P}(\mathbf{P}\mathbf{z}) = \mathbf{P}\mathbf{z}$ for all $\mathbf{z}$. Therefore, $\mathbf{P}\mathbf{P} = \mathbf{P}$. This is actually a simple implication of the law of iterated projections.

Positive semidefinite: For any $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{z}'\mathbf{P}\mathbf{z} = \mathbf{z}'\mathbf{P}\mathbf{P}\mathbf{z} = \mathbf{z}'\mathbf{P}'\mathbf{P}\mathbf{z} = (\mathbf{P}\mathbf{z})'(\mathbf{P}\mathbf{z}) = \|\mathbf{P}\mathbf{z}\|^2 \geq 0$. **Duality**: Let $\mathbf{z} = \mathbf{z}_1 + \mathbf{z}_2$, where $\mathbf{z}_1 \in M$ and $\mathbf{z}_2 \in M^\perp$. Then $(\mathbf{I}_n - \mathbf{P})\mathbf{z} = \mathbf{z} - \mathbf{z}_1 = \mathbf{z}_2 \in M^\perp$ so that $\mathbf{I}_n - \mathbf{P}$ is the orthogonal projector onto $M^\perp$. ■

For a general "nonorthogonal" projector $\mathbf{P}$, it is still unique and idempotent, but need not be symmetric (let alone positive semidefiniteness). For example, $\mathbf{P}_{\mathbf{X}\perp\mathbf{W}\mathbf{X}}$ in Example 4 is not symmetric (see Exercise 14). We will provide more discussion on the symmetry of a projector in the next section.

Another caution is that we cannot strengthen "positive semidefinite" to "positive definite". Recall that for an idempotent matrix, the rank equals the trace.

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

and

$$\operatorname{tr}(\mathbf{M}_\mathbf{X}) = \operatorname{tr}(\mathbf{I}_n - \mathbf{P}_\mathbf{X}) = \operatorname{tr}(\mathbf{I}_n) - \operatorname{tr}(\mathbf{P}_\mathbf{X}) = n - k < n.$$

When $k = n$, $\mathbf{P}_\mathbf{X} = \mathbf{I}_n$ and the projector is trivial.

4 Partitioned Fit and Residual Regression

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

$$\mathbf{X}\boldsymbol{\beta} = \left[\mathbf{X}_1 \vdots \mathbf{X}_2\right] \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \begin{matrix} k_1 \\ k_2 \end{matrix}$$

with $\operatorname{rank}(\mathbf{X}) = k$.[17] We will show that $\widehat{\boldsymbol{\beta}}_1$ 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. We could state this theorem in both Hilbert spaces, but for practical purpose, we only state the theorem in $\mathbb{R}^n$, and the L^2 case is left as exercise. To simplify notation, $\mathbf{P}_j \equiv \mathbf{P}_{\mathbf{X}_j}$, $\mathbf{M}_j \equiv \mathbf{M}_{\mathbf{X}_j}$, $\Pi_j(\mathbf{y}) = \mathbf{X}_j\widehat{\boldsymbol{\beta}}_j$, $j = 1, 2$.

Theorem 4 *$\widehat{\boldsymbol{\beta}}_1$ 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,*

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

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 $\widehat{\boldsymbol{\beta}}_1$ 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**.

[17] (*) Partitioned fit can also reduce the dimensionality of a fitting problem when the dimension of *span*$(\mathbf{X}_2)$ is huge, e.g., in the fixed effect panel data model, $\mathbf{X}_2$ collects the indicators for each individual, so $\dim(\mathit{span}(\mathbf{X}_2)) = \mathit{rank}(\mathbf{X}_2) = n$.

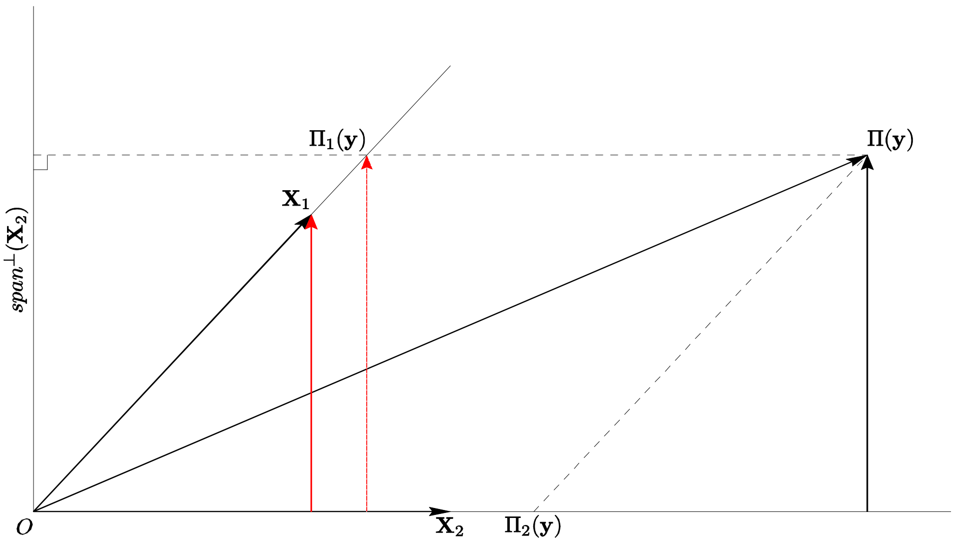

Figure 8: The FWL Theorem

Exercise 16 *When can we get $\widehat{\boldsymbol{\beta}}_1$ by regressing $\mathbf{y}$ only on $\mathbf{X}_1$? What's the intuition here? (Hint: $\mathbf{X}_1$ is orthogonal to $\mathbf{X}_2$.)*

Exercise 17 *Show that $\mathbf{X}'_{1\perp 2}\mathbf{X}_2 = \mathbf{0}$, $\mathbf{y}'_{\perp 2}\mathbf{X}_2 = \mathbf{0}$, and we regress $\mathbf{y}_{\perp 2}$ on $(\mathbf{X}_{1\perp 2}, \mathbf{X}_2)$, the coefficients for $\mathbf{X}_2$ are zero. What is the intuition in these results?*

Corollary 1

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

Exercise 18 *Show that $\mathbf{P}_{12}$ preserves span$(\mathbf{X}_1)$ and annihilates span$(\mathbf{X}_2)$, and which implies $\mathbf{P}_{12}(\Pi_1(\mathbf{y})) = \Pi_1(\mathbf{y})$, $\mathbf{P}_{12}(\Pi_2(\mathbf{y})) = \mathbf{0}$ and $\mathbf{P}_{12}(\Pi(\mathbf{y})) = \Pi_1(\mathbf{y})$.*

Figure 8 shows the intuition underlying the FWL theorem. In the figure, we use $\Pi(\mathbf{y})$ instead of $\mathbf{y}$ or we assume there is a perfect fit. But from the discussion in the last section or the corollary, this will not affect the result. The residual of $\Pi(\mathbf{y})$ regressing on $\mathbf{X}_2$ is equal to the dotted red arrow in the figure, and the residual of $\mathbf{X}_1$ regressing on $\mathbf{X}_2$ is equal to the solid red arrow in the figure. $\widehat{\boldsymbol{\beta}}_1$ is equal to the length of $\Pi_1(\mathbf{y})$ divided by the length of $\mathbf{X}_1$, which is exactly equal to the length of the dotted red arrow divided by the length of the solid red arrow by similarity of the two triangles.

To understand the projector $\mathbf{P}_{12}$, we write it out explicitly as

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

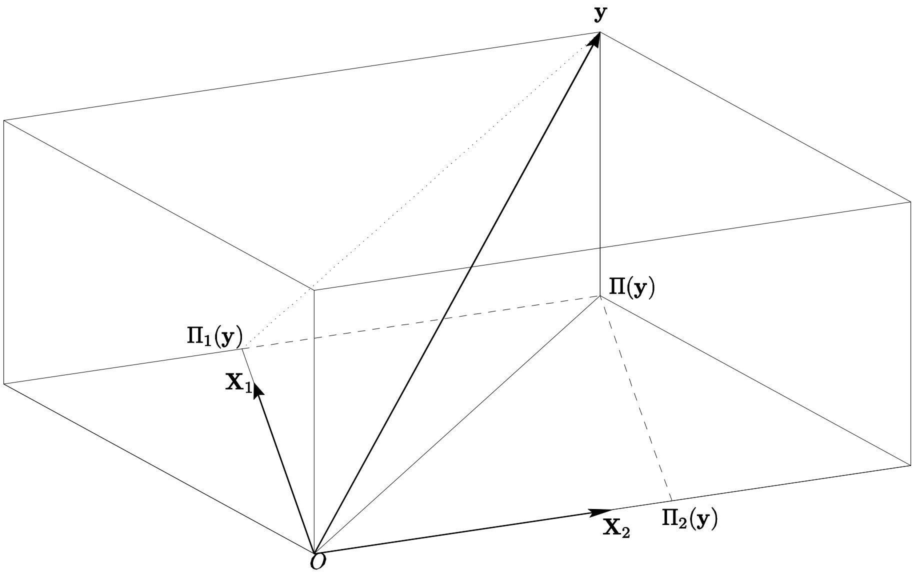

Figure 9: Projection by $\mathbf{P}_{12}$

$\mathbf{I} - \mathbf{P}_2$ 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: $\mathbf{P}_{12}\mathbf{X}_1 = \mathbf{X}_1$.

Since $\mathbf{P}_{12}\mathbf{y} = \mathbf{P}_{12}(\Pi(\mathbf{y}))$, 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 $\Pi_1(\mathbf{y})$. $\widehat{\boldsymbol{\beta}}_1$ is calculated from $\Pi_1(\mathbf{y})$ by

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

Figure 9 shows the intuition in this sequential projection.

Proof of Theorem. We use three methods to prove this theorem.

Method I: The crude (or brute-force) method is to calculate $\widehat{\boldsymbol{\beta}}_1$ explicitly in the residual regression and check whether it is equal to the LSE of $\boldsymbol{\beta}_1$. This calculation need the partitioned inverse formula,

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

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

Recall that 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 $\boldsymbol{\beta}_1$

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

From (12), we have

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

so

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

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

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

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

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

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 $\mathbf{M}_2\mathbf{M} = \mathbf{M}$). We still need to check $\left(\mathbf{X}_{1\perp 2}'\mathbf{X}_{1\perp 2}\right)^{-1}$ exists, or $\mathbf{X}_{1\perp 2}$ is full-column rank. To see that consider

$$\mathbf{X} = (\mathbf{M}_2 + \mathbf{P}_2)\mathbf{X} = \left[\mathbf{X}_{1\perp 2} + \mathbf{P}_2\mathbf{X}_1, \mathbf{X}_2\right].$$

The columns of $\mathbf{P}_2\mathbf{X}_1$ are linearly dependent on $\mathbf{X}_2$. But $\mathbf{X}$ is full column rank, so that $\mathbf{X}_{1\perp 2}$ must also be full-column rank.

Method III: In this method, we try to show that for any $\boldsymbol{\beta}_1$,

$$\min_{\boldsymbol{\beta}_2} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y}_{\perp 2}-\mathbf{X}_{1\perp 2}\boldsymbol{\beta}_1\|^2 ,$$

which implies our target result. To generate $\mathbf{y}_{\perp 2}$ and $\mathbf{X}_{1\perp 2}$, we decompose $\mathbf{y}-\mathbf{X}\boldsymbol{\beta}$ as

$$\begin{aligned}
\mathbf{y}-\mathbf{X}\boldsymbol{\beta} &= \mathbf{M}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) + \mathbf{P}_2(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\
&= [\mathbf{M}_2\mathbf{y} - \mathbf{M}_2\mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{M}_2\mathbf{X}_2\boldsymbol{\beta}_2] + [\mathbf{P}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) - \mathbf{P}_2\mathbf{X}_2\boldsymbol{\beta}_2] \\
&= [\mathbf{y}_{\perp 2} - \mathbf{X}_{1\perp 2}\boldsymbol{\beta}_1] + [\mathbf{P}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) - \mathbf{X}_2\boldsymbol{\beta}_2] .
\end{aligned}$$

Since the two terms are orthogonal (why?),

$$\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \|\mathbf{y}_{\perp 2}-\mathbf{X}_{1\perp 2}\boldsymbol{\beta}_1\|^2 + \|\mathbf{P}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) - \mathbf{X}_2\boldsymbol{\beta}_2\|^2 .$$

We need to show the second term is zero. Note that the first term does not depend on $\boldsymbol{\beta}_2$, so the minimizer of $\boldsymbol{\beta}_2$ is determined only by the second term. However, since $\mathbf{P}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) \in \mathit{span}(\mathbf{X}_2)$, we can always find $\boldsymbol{\beta}_2$ such that $\mathbf{P}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) = \mathbf{X}_2\boldsymbol{\beta}_2$, so the second term does not contribute to the objective function. ■

Exercise 19 *Show (12). When A_{11}, A_{12}, A_{21} and A_{22} are scalars, check whether (12) matches the simpler formula* $\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} = \frac{1}{A_{11}A_{22}-A_{12}A_{21}} \begin{pmatrix} A_{22} & -A_{12} \\ -A_{21} & A_{11} \end{pmatrix}$. *(Hint: Check the right hand side of (12) times* $\begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix}$ *equals the identity matrix)*

Exercise 20 *Show that if rank$(\mathbf{X}) = k$, rank$(\mathbf{X}_{1\perp 2}) = k_1$. (Hint: Use the definition of full column rank)*

Method III uses a tool called *concentrating* the SSR function. Suppose $\arg\min_{\boldsymbol{\beta}_2} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|^2 = \widehat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_1)$; then the *profile* or *concentrated* SSR function is

$$\left\|\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1 - \mathbf{X}_2\widehat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_1)\right\|^2 ,$$

which is shown to be $\|\mathbf{y}_{\perp 2}-\mathbf{X}_{1\perp 2}\boldsymbol{\beta}_1\|^2$. Also, from Method III,

$$\begin{aligned}
\|\mathbf{y}_{\perp 2}-\mathbf{X}_{1\perp 2}\boldsymbol{\beta}_1\|^2 &= \|\mathbf{M}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1)\|^2 \\
&= (\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1)'\mathbf{M}_2'\mathbf{M}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) \\
&= (\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1)'\mathbf{M}_2(\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1) \\
&= \|\mathbf{y}-\mathbf{X}_1\boldsymbol{\beta}_1\|^2_{\mathbf{M}_2} .
\end{aligned}$$

In other words, $\widehat{\boldsymbol{\beta}}_1$ is the coefficient of projecting $\mathbf{y}$ on $\mathit{span}(\mathbf{X}_1)$ under the inner product $\langle\cdot,\cdot\rangle_{\mathbf{M}_2}$.

Exercise 21 *(i) Suppose the model is* $y = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 + u$, $E[\mathbf{x}u] = 0$. *State and prove the FWL theorem in the* L^2 *space. (ii) If* $x_2 = 1$, *what is* $\boldsymbol{\beta}_1$*? Compare this result with that in Exercise 6. (iii) In the FWL theorem in* $\mathbb{R}^n$, *if* $X_2 = \mathbf{1}$, *what is* $\widehat{\boldsymbol{\beta}}_1$*?*

Exercise 22 *Let* $\mathbf{d}_1$ *and* $\mathbf{d}_2$ *be defined as in Exercise 13.*

(a) *In the OLS regression*

$$\mathbf{y} = \mathbf{d}_1\widehat{\alpha}_1 + \mathbf{d}_2\widehat{\alpha}_2 + \widehat{\mathbf{u}}_y,$$

show that $\widehat{\alpha}_1$ *is sample mean of the dependent variable among the men of the sample* $(\overline{y}_1)$, *and that* $\widehat{\alpha}_2$ *is the sample mean among the women* $(\overline{y}_2)$.

(b) *Describe in words the transformations*

$$\begin{aligned} \mathbf{y}^* &= \mathbf{y} - \mathbf{d}_1\overline{y}_1 - \mathbf{d}_2\overline{y}_2 \\ \mathbf{X}^* &= \mathbf{X} - \mathbf{d}_1\overline{\mathbf{X}}_1 - \mathbf{d}_2\overline{\mathbf{X}}_2. \end{aligned}$$

(c) *Compare* $\widetilde{\boldsymbol{\beta}}$ *from the OLS regresion*

$$\mathbf{y}^* = \mathbf{X}^*\widetilde{\boldsymbol{\beta}} + \widetilde{\mathbf{u}}$$

with $\widehat{\boldsymbol{\beta}}$ *from the OLS regression*

$$\mathbf{y} = \mathbf{d}_1\widehat{\alpha}_1 + \mathbf{d}_2\widehat{\alpha}_2 + \mathbf{X}\widehat{\boldsymbol{\beta}} + \widehat{\mathbf{u}}.$$

Exercise 23 (*) *If* $y = \mathbf{x}_1'\boldsymbol{\beta}_1 + g(\mathbf{x}_2) + u$, *where* g *is any function such that* $E\left[g(\mathbf{x}_2)^2\right] < \infty$ *and* $E[u|\mathbf{x}_1, \mathbf{x}_2] = 0$, *then what is true value of* $\boldsymbol{\beta}_1$*? (Hint: Use the FWL theorem; see Robinson (1988))*

4.1 Projection along a Subspace

$\mathbf{P}_{12}$ takes an interesting form which deserves further exploration.

Lemma 1 *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*

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

Proof (*). That $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}$ is idempotent can be proved similarly as in proving that the orthogonal projector is idempotent in Section 3.

From the following exercise, $\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'$ preserves $span(\mathbf{X})$ and annihilates $span^{\perp}(\mathbf{Z})$. So from the definition of $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}$, we need only to show that $\mathbf{Z}'\mathbf{X}$ is nonsingular is equivalent to that $\mathbb{R}^n = span(\mathbf{X}) \oplus span^{\perp}(\mathbf{Z})$.

$\Longleftarrow$ Given that $\mathbf{Z}'\mathbf{X}$ is square, we need only to show that if $\mathbf{Z}'\mathbf{X}\mathbf{a} = 0$ then $\mathbf{a} = 0$. If $\mathbb{R}^n = span(\mathbf{X}) \oplus span^{\perp}(\mathbf{Z})$, then from the definition of $\oplus$, $span(\mathbf{X}) \cap span^{\perp}(\mathbf{Z}) = \{\mathbf{0}\}$. That is, $\mathbf{Z}'\mathbf{X}\mathbf{a} = 0$ if and only if $\mathbf{X}\mathbf{a} = 0$. Because $\mathbf{X}$ is full-column rank, $\mathbf{X}\mathbf{a} = 0$ if and only if $\mathbf{a} = 0$.

$\Longrightarrow$ If $\mathbf{Z}'\mathbf{X}$ is nonsingular, then for every $\mathbf{w} \in \mathbb{R}^n$,

$$\mathbf{w} = \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{w} + (\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}')\mathbf{w}, \tag{13}$$

where $\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{w} \in \mathit{span}(\mathbf{X})$ and $(\mathbf{I} - \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}')\mathbf{w} \in \mathit{span}^{\perp}(\mathbf{Z})$. Now we show that this decomposition is unique. For any other $\mathbf{w}_1 \in \mathit{span}(\mathbf{X})$ and $\mathbf{w}_2 \in \mathit{span}^{\perp}(\mathbf{Z})$ such that $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$, it follows that $\mathbf{w}_1 = \mathbf{X}\boldsymbol{\alpha}$ for some $\boldsymbol{\alpha}$ and $\mathbf{Z}'\mathbf{w}_2 = 0$ so that

$$\begin{aligned} \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{w} &= \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'(\mathbf{w}_1 + \mathbf{w}_2) \\ &= \mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{X}\alpha = \mathbf{X}\alpha = \mathbf{w}_1 \end{aligned}$$

returning the original decomposition. Therefore, (13) is a unique decomposition so that $\mathbb{R}^n = \mathit{span}(\mathbf{X}) \oplus \mathit{span}^{\perp}(\mathbf{Z})$. ■

Exercise 24 *In the setup of the above lemma, show that (i)* $\mathbf{X}(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'$ *preserves span*$(\mathbf{X})$ *and annihilates span*$^{\perp}(\mathbf{Z})$*; (ii) if* $\mathbf{Z}'\mathbf{X}$ *is nonsingular, then rank*$(\mathbf{X})$ *=rank*$(\mathbf{Z}) = k$*. (iii) If* $\mathbf{Z}'\mathbf{X}$ *is nonsingular, then no column of* $\mathbf{Z}$ *falls in span*$^{\perp}(\mathbf{X})$*.*

For orthogonal projectors, $\mathbf{P}_{\mathbf{X}} = \mathbf{P}_{\mathbf{X}\perp\mathbf{X}}$. From this lemma, $\mathbf{P}_{12} = \mathbf{P}_{\mathbf{X}_1\perp\mathbf{X}_{1\perp2}}$. 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. Also, the rank of $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{X}\perp\mathbf{Z}}$ is the dimension of *span*$(\mathbf{X})$ on which we are projecting onto (why? rank = trace).

Exercise 25 (*) *Show that (i) if* $\mathbf{A}$ *is idempotent, then* $\mathbf{A}$ *is a projector; (ii) if* $\mathbf{A}$ *is idempotent and symmetric, then* $\mathbf{A}$ *is an orthogonal projector.*

Lemma 2 $\mathbf{P}_{12}$ *is the unique projector onto span*$(\mathbf{X}_1)$ *along span*$(\mathbf{X}_2) \oplus$ *span*$^{\perp}(\mathbf{X})$*.*

Proof. Because $\mathbf{X}$ is full rank,

$$\begin{aligned} \mathbb{R}^n &= \mathit{span}(\mathbf{X}) \oplus \mathit{span}^{\perp}(\mathbf{X}) \\ &= \mathit{span}(\mathbf{X}_1) \oplus \left[\mathit{span}(\mathbf{X}_2) \oplus \mathit{span}^{\perp}(\mathbf{X})\right]. \end{aligned}$$

Exercise 18 shows that $\mathbf{P}_{12}$ preserves *span*$(\mathbf{X}_1)$ and annihilates *span*$(\mathbf{X}_2)$, so we need only to show that $\mathbf{P}_{12}$ also annihilates *span*$^{\perp}(\mathbf{X})$. To see this, note that for all $\mathbf{z} \in \mathit{span}^{\perp}(\mathbf{X})$,

$$\mathbf{X}_{1\perp2}'\mathbf{z} = \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)'\mathbf{z} = \mathbf{X}_1'(\mathbf{I} - \mathbf{P}_2)\mathbf{z} = \mathbf{X}_1'\mathbf{z} - \mathbf{X}_1'\mathbf{P}_2\mathbf{z} = 0$$

where the last equality is from the fact that $\mathbf{P}_2\mathbf{z} = 0$ and $\mathbf{X}_1'\mathbf{z} = 0$ for $\mathbf{z} \in span^{\perp}(\mathbf{X})$. The uniqueness of $\mathbf{P}_{12}$ can be similarly proved as in the proof of the uniqueness of the orthogonal projector. ■

Proof of Corollary 1. First,

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

It remains to show that $\mathbf{P}_{12}\mathbf{y} = \mathbf{P}_{12}(\Pi(\mathbf{y}))$, which is equivalent to $\mathbf{P}_{12}(\mathbf{y} - \Pi(\mathbf{y})) = \mathbf{0}$. From the last lemma, this indeed holds since $\mathbf{y} - \Pi(\mathbf{y}) \in span^{\perp}(\mathbf{X})$. ■

5 Restricted Least Squares (*)

In least squares, we often want to impose some restrictions on $\boldsymbol{\beta}$ from economic theory or experience, which results in **restricted** (or **constrained**) **least squares** (RLS). For example, the Cobb-Douglas production function $Y = AL^{\beta_1}K^{\beta_2}$ implies $y = \alpha + \beta_1 l + \beta_2 k$, where lower case letters are log of upper case letters and $\alpha = \log A$. The restriction from the economic theory that the production function is constant return to scale is $\beta_1 + \beta_2 = 1$. In other examples, we want to assume some parameters are zero (*exclusion restrictions*), or several parameters are equal (*equality restrictions*). All these restrictions are linear and can be written in the form

$$\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\gamma} + \mathbf{s}, \tag{14}$$

where $\mathbf{S}$ is a $k \times m$ matrix of known constants, $\mathbf{s}$ is a $k \times 1$ vector of known constants, and $\boldsymbol{\gamma}$ is a $m \times 1$ vector of unknown parameters. $m < k$ implies $k - m$ restrictions are imposed on $\boldsymbol{\beta}$. By imposing the constraint we hope to improve estimation efficiency. The RLS estimator of $\boldsymbol{\beta}$ is defined as

$$\widehat{\boldsymbol{\beta}}_R \equiv \arg \min_{\{\boldsymbol{\beta}|\boldsymbol{\beta}=\mathbf{S}\boldsymbol{\gamma}+\mathbf{s}\}} \left\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\right\|^2.$$

Exercise 26 *Consider the model* $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$. *Express the following restrictions in the form of (14) and find* $\widehat{\boldsymbol{\beta}}_R$*: (i)* $\boldsymbol{\beta}_2 = \mathbf{0}$*; (ii)* $\boldsymbol{\beta}_1 = \mathbf{c}$*; (iii)* $\boldsymbol{\beta}_1 = -\boldsymbol{\beta}_2$.

The following theorem states the relationship between $\widehat{\boldsymbol{\beta}}_R$ and $\widehat{\boldsymbol{\beta}}$.

Theorem 5 *The RLS fitted vector is the orthogonal projection of* $\mathbf{y}$ *plus the translation given in*

$$\begin{aligned}\Pi_R(\mathbf{y}) &= \mathbf{P}_{\mathbf{XS}}\mathbf{y} + (\mathbf{I} - \mathbf{P}_{\mathbf{XS}})\mathbf{X}\mathbf{s} \\ &= \mathbf{P}_{\mathbf{XS}}\Pi(\mathbf{y}) + (\mathbf{I} - \mathbf{P}_{\mathbf{XS}})\mathbf{X}\mathbf{s},\end{aligned}$$

where $\Pi_R(\cdot)$ *is the projector onto span*$\{\mathbf{X}\boldsymbol{\beta}|\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\gamma} + \mathbf{s}\}$. *The RLS coefficient vector is*

$$\begin{aligned}\widehat{\boldsymbol{\beta}}_R &= \mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{X}\mathbf{S}(\mathbf{y} - \mathbf{X}\mathbf{s}) + \mathbf{s} \\ &= \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}\mathbf{S}}\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}\mathbf{S}})\mathbf{s}.\end{aligned}$$

Proof. Since $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\gamma} + \mathbf{s}$,

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{X}(\mathbf{S}\boldsymbol{\gamma} + \mathbf{s}) = \mathbf{X}\mathbf{S}\boldsymbol{\gamma} + \mathbf{X}\mathbf{s}.$$

As a result,

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = \|(\mathbf{y} - \mathbf{X}\mathbf{s}) - \mathbf{X}\mathbf{S}\boldsymbol{\gamma}\|^2,$$

that is, a constrained problem is reduced to an unconstrained problem - projecting $(\mathbf{y} - \mathbf{X}\mathbf{s})$ onto $\mathbf{X}\mathbf{S}$. So

$$\widehat{\boldsymbol{\gamma}} = (\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{X}\mathbf{S}\,(\mathbf{y} - \mathbf{X}\mathbf{s}),$$

and

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_R &= \mathbf{S}\widehat{\boldsymbol{\gamma}} + \mathbf{s} = \mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\,(\mathbf{y} - \mathbf{X}\mathbf{s}) + \mathbf{s}, \\
\Pi_R(\mathbf{y}) &= \mathbf{X}\widehat{\boldsymbol{\beta}}_R = \mathbf{X}\left[\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\,(\mathbf{y} - \mathbf{X}\mathbf{s}) + \mathbf{s}\right] \\
&= \mathbf{P}_{\mathbf{X}\mathbf{S}}\,(\mathbf{y} - \mathbf{X}\mathbf{s}) + \mathbf{X}\mathbf{s} = \mathbf{P}_{\mathbf{X}\mathbf{S}}\mathbf{y} + (\mathbf{I} - \mathbf{P}_{\mathbf{X}\mathbf{S}})\mathbf{X}\mathbf{s}.
\end{aligned}$$

From Theorem 2, we know

$$\Pi_R(\mathbf{y}) = \Pi_R(\Pi(\mathbf{y})),$$

where $M_2 = \{\mathbf{X}\boldsymbol{\beta}\}$, $M_1 = \{\mathbf{X}\boldsymbol{\beta} | \boldsymbol{\beta} = \mathbf{S}\boldsymbol{\gamma} + \mathbf{s}\}$ and $\Pi(\mathbf{y}) = \mathbf{P}_{\mathbf{X}}\mathbf{y} = \mathbf{X}\widehat{\boldsymbol{\beta}}$. So

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_R &\equiv \arg \min_{\{\boldsymbol{\beta} | \boldsymbol{\beta} = \mathbf{S}\boldsymbol{\gamma} + \mathbf{s}\}} \|\Pi(\mathbf{y}) - \mathbf{X}\boldsymbol{\beta}\|^2 \\
&= \mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\,(\Pi(\mathbf{y}) - \mathbf{X}\mathbf{s}) + \mathbf{s} \\
&= \mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \mathrm{s}) + \mathbf{s} \\
&= \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}\mathbf{S}}\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}\mathbf{S}})\mathbf{s},
\end{aligned}$$

and

$$\begin{aligned}
\Pi_R(\mathbf{y}) &= \mathbf{X}\left[\mathbf{S}(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S})^{-1}\mathbf{S}'\mathbf{X}'\,(\Pi(\mathbf{y}) - \mathbf{X}\mathbf{s}) + \mathbf{s}\right] \\
&= \mathbf{P}_{\mathbf{X}\mathbf{S}}\Pi(\mathbf{y}) + (\mathbf{I} - \mathbf{P}_{\mathbf{X}\mathbf{S}})\mathbf{X}\mathbf{s}.
\end{aligned}$$

■

Exercise 27 *In this exercise, we will show that the RLS solution also has the general form*

$$\widehat{\boldsymbol{\beta}}_R = \arg \min_{\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2, \tag{15}$$

where $\mathbf{R}$ *is a* $k \times q$ *matrix with* $q = k - m$, $\mathbf{c}$ *is a* $q \times 1$ *vector of known constants, and rank*$(\mathbf{R}) = q$ *so that there are no redundant or mutually exclusive restrictions.*

(a) *Show that* $\mathbf{R}'\boldsymbol{\beta} = \mathbf{c}$ *can always be written in the form* $\boldsymbol{\beta} = \mathbf{S}\boldsymbol{\gamma} + \mathbf{s}$. *(Hint: Show that we can always order and partition the elements of* $\boldsymbol{\beta}$ *and* $\mathbf{R}$ *so that* $\mathbf{R}'\boldsymbol{\beta} = \mathbf{R}_1'\boldsymbol{\beta}_1 + \mathbf{R}_2'\boldsymbol{\beta}_2 = \mathbf{c}$, $\mathbf{R}_1$ *is nonsingular, and* $\mathbf{R}_2$ *has* m *elements.)*

(b) *Let* $\mathbf{c} = \mathbf{0}$ *and show that (15) can also be written as*

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_R &= \arg \min_{\boldsymbol{\beta} \in span^{\perp}(\mathbf{R})} \left\| \Pi\left(\mathbf{y}\right) - \mathbf{X}\boldsymbol{\beta} \right\|^2 \\ &= \arg \min_{\boldsymbol{\beta} \in span^{\perp}(\mathbf{R})} \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|^2_{\mathbf{X}'\mathbf{X}} . \end{aligned}$$

(c) *Show that*

$$\left(\mathbf{I} - \mathbf{P}_{\mathbf{A}^{-1}\mathbf{X}\perp\mathbf{X}}\right)\mathbf{y} = \arg \min_{\mathbf{z} \in span^{\perp}(\mathbf{X})} \left\| \mathbf{y} - \mathbf{z} \right\|^2_{\mathbf{A}} .$$

Use this result to show that

$$\widehat{\boldsymbol{\beta}}_R = \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\mathbf{R}'\widehat{\boldsymbol{\beta}}.$$

(d) *When* $\mathbf{R}' = [\mathbf{0}, \mathbf{I}_q]$*, show that* $\widehat{\boldsymbol{\beta}}_R$ *in (c) reduces to* $\left(\widehat{\mathbf{Q}}'_{1\mathbf{y}}\widehat{\mathbf{Q}}^{-1}_{11}, \mathbf{0}'\right)'$*, where*

$$\widehat{\mathbf{Q}}_{11} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{1i}\mathbf{x}'_{1i}, \widehat{\mathbf{Q}}_{1\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_{1i}\mathbf{y}_i.$$

(e) *When* $\mathbf{c} \neq \mathbf{0}$*, show that*

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_R &= \arg \min_{\boldsymbol{\beta} \in span^{\perp}(\mathbf{R}) + \mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{c}} \left\| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \right\|^2_{\mathbf{X}'\mathbf{X}} \\ &= (\mathbf{I} - \mathbf{P}_{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\perp\mathbf{R}})\widehat{\boldsymbol{\beta}} + \mathbf{P}_{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\perp\mathbf{R}}\mathbf{R}(\mathbf{R}'\mathbf{R})^{-1}\mathbf{c} \\ &= \widehat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right). \end{aligned}$$

Verify that $\mathbf{R}'\widehat{\boldsymbol{\beta}}_R = \mathbf{c}$.

(f) *We can solve (15) by the method of Lagrange. Let* $\boldsymbol{\lambda}$ *be a vector of* q *Lagrange multipliers for the* q *restrictions. The Lagrangian is*

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\lambda}) = \frac{1}{2}(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}'\boldsymbol{\beta}) + \boldsymbol{\lambda}'(\mathbf{R}'\boldsymbol{\beta} - \mathbf{c}).$$

Show that the solution for $\boldsymbol{\beta}$ *is the same as in (e) and*

$$\widehat{\boldsymbol{\lambda}} = \left[\mathbf{R}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}\right]^{-1}\left(\mathbf{R}'\widehat{\boldsymbol{\beta}} - \mathbf{c}\right).$$

(g) $\widehat{\boldsymbol{\lambda}}$ *is usually interpreted as "shadow prices" of constraints. Show that the shadow prices of the constraints that are satisfied by* $\widehat{\boldsymbol{\beta}}$ *are zero.*

5.1 RLS as a Projection

Suppose first that $\mathbf{s}=\mathbf{0}$. Note from the proof of Theorem 5 that

$$\widehat{\boldsymbol{\beta}}_R \equiv \arg\min_{\{\boldsymbol{\beta}|\boldsymbol{\beta}=\mathbf{S}\boldsymbol{\gamma}\}} \left\|\mathbf{X}\widehat{\boldsymbol{\beta}}-\mathbf{X}\boldsymbol{\beta}\right\|^2 = \arg\min_{\boldsymbol{\beta}\in span(\mathbf{S})} \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2_{\mathbf{X}'\mathbf{X}},$$

where the inner product is defined as $\langle \mathbf{x},\mathbf{z}\rangle_{\mathbf{Q}} = \mathbf{x}'\mathbf{Q}\mathbf{z}$ for a positive definite matrix $\mathbf{Q}$, and the associated norm is defined as $\|\mathbf{x}\|^2_{\mathbf{Q}} = \mathbf{x}'\mathbf{Q}\mathbf{x}$.[18] The RLS estimator is a special **minimum distance** (MD) **estimator** or **minimum chi-square estimator** of Wolfowitz (1953, 1957), which tries to find a constrained parameter value that is as close as possible to the unconstrained estimate (see Ferguson (1958), Malinvaud (1970) and Rothenberg (1973)). Specifically, the MD estimator under the constraints (14) is

$$\widehat{\boldsymbol{\beta}}_{MD} = \arg\min_{\boldsymbol{\beta}=\mathbf{S}\boldsymbol{\gamma}+\mathbf{s}} J_n(\boldsymbol{\beta}),$$

where

$$J_n(\boldsymbol{\beta}) = n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)'\mathbf{W}_n\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right),$$

where $\mathbf{W}_n > 0$ is a $k\times k$ weight matrix.

Exercise 28 *Show that when* $\mathbf{s}=\mathbf{0}$*, the expression of* $\widehat{\boldsymbol{\beta}}_R$ *in Theorem 5 is equal to* $\arg\min_{\boldsymbol{\beta}\in span(\mathbf{S})} \left\|\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right\|^2_{\mathbf{X}'\mathbf{X}}$ *by using the orthogonal conditions directly.*

The result for $\widehat{\boldsymbol{\beta}}_R$ is generally correct, as shown in the following theorem.

Theorem 6 *Let* $\mathbf{X}$ *be full-column rank. Then*

$$\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}\mathbf{y} = \arg\min_{\boldsymbol{\mu}\in span(\mathbf{X})} \|\mathbf{y}-\boldsymbol{\mu}\|^2_{\mathbf{Q}}.$$

Proof. We need only to show that $\langle \mathbf{y}-\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}\mathbf{y},\boldsymbol{\mu}\rangle_{\mathbf{Q}} = 0$ for any $\boldsymbol{\mu}\in span(\mathbf{X})$. $\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}$ is the unique projector onto $span(\mathbf{X})$ along $span^{\perp}(\mathbf{Q}\mathbf{X})$. Therefore, $\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}\mathbf{y}\in span(\mathbf{X})$ and

$$\mathbf{y}-\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}\mathbf{y} = (\mathbf{I}-\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}})\mathbf{y}\in span^{\perp}(\mathbf{Q}\mathbf{X}).$$

For any $\boldsymbol{\mu}\in span(\mathbf{X})$, $\langle \mathbf{y}-\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}\mathbf{y},\mathbf{Q}\boldsymbol{\mu}\rangle = \langle \mathbf{y}-\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}\mathbf{y},\boldsymbol{\mu}\rangle_{\mathbf{Q}} = 0$. ■
This theorem shows that the projector $\mathbf{P}_{\mathbf{X}\perp\mathbf{Q}\mathbf{X}}$ is the orthogonal projector onto $span(\mathbf{X})$ under the inner product $\langle\cdot,\cdot\rangle_{\mathbf{Q}}$.

Geometrically, the difference between $\langle\cdot,\cdot\rangle$ and $\langle\cdot,\cdot\rangle_{\mathbf{Q}}$ is the difference between spheres and ellipses. We use Example 4.8 of Ruud (2000) to illustrate this point.

Example 5 *Consider the following simple setup,*

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \mathbf{S} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \mathbf{s} = 0,$$

[18] This norm is introduced in statistics by Mahalanobis (1936).

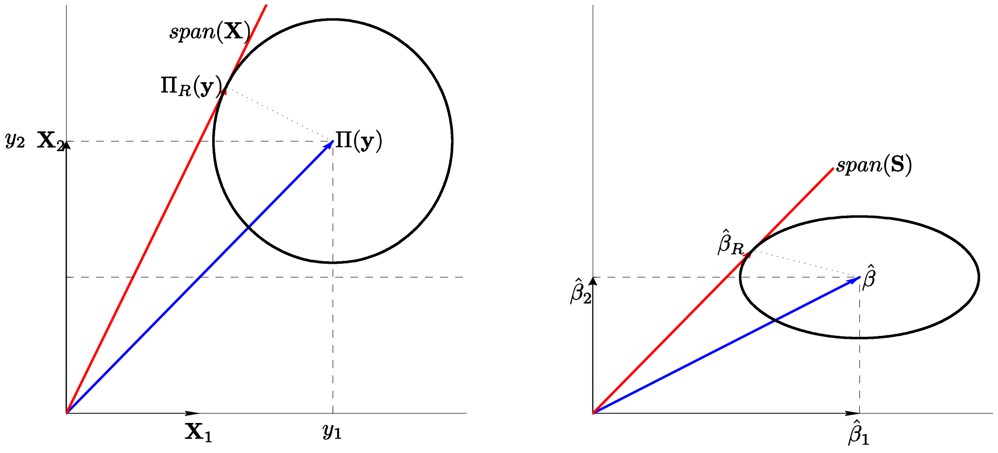

Figure 10: RLS as a Projection of OLS: $(y_1, y_2) = (1, 1)$

where $n = 2$, $k = 2$ and $m = 1$. Since $n = k$, there is a perfect fit with $\widehat{\beta}_1 = y_1$ and $\widehat{\beta}_2 = y_2/2$. The restricted estimator is

$$\widehat{\gamma} = \frac{y_1 + 2y_2}{5} = \widehat{\beta}_{R1} = \widehat{\beta}_{R2}. \text{ (Check!)}$$

The OLS minimization is depicted in the left panel of Figure 10, where we assume $(y_1, y_2) = (1, 1)$ so $\Pi_R(\mathbf{y}) = (0.6, 1.2)'$. The minimization in the parameter space (β_1, β_2) is depicted in the right panel, where the ellipse, which achieves the minimum, is

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2_{\mathbf{X}'\mathbf{X}} = \left(\widehat{\beta}_1 - \beta_1\right)^2 + 4\left(\widehat{\beta}_2 - \beta_2\right)^2 = 0.2.$$

From Figure 10, we can see $\mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{XS}}$ maps $\widehat{\boldsymbol{\beta}}$ to span$(\mathbf{S})$, but not orthogonally because distance is measured elliptically rather than spherically.

When $\mathbf{s} \neq \mathbf{0}$, the second term in $\widehat{\boldsymbol{\beta}}_R$, $(\mathbf{I} - \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{XS}})\mathbf{s}$, is a translation of $\mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{XS}}\widehat{\boldsymbol{\beta}}$. This translation arises from a translation that appears in the RLS program itself, which we now write as

$$\begin{aligned}
\widehat{\boldsymbol{\beta}}_R &= \arg\min_{\boldsymbol{\beta}\in span(\mathbf{S})+\mathbf{s}} \left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2_{\mathbf{X}'\mathbf{X}} \\
&= \arg\min_{\mathbf{b}\in span(\mathbf{S})} \left\|\widehat{\boldsymbol{\beta}} - \mathbf{b} - \mathbf{s}\right\|^2_{\mathbf{X}'\mathbf{X}} + \mathbf{s} \\
&= \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{XS}}\left(\widehat{\boldsymbol{\beta}} - \mathbf{s}\right) + \mathbf{s} \\
&= \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{XS}}\widehat{\boldsymbol{\beta}} + (\mathbf{I} - \mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{XS}})\mathbf{s},
\end{aligned}$$

where the second equality is from the transformation $\boldsymbol{\beta} = \mathbf{b} + \mathbf{s}$.

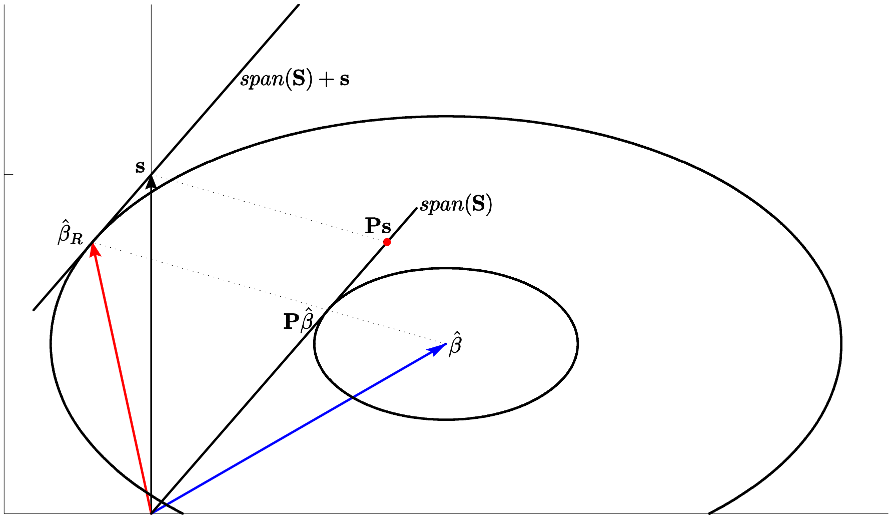

Figure 11: RLS as a Projection and Translation of OLS

Figure 11 illustrates the translation $\mathbf{s} = (0,1)'$ in the setup of the previous example, where $\widehat{\boldsymbol{\beta}}_R = (-0.2, 0.8)'$, $\mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}}\mathbf{s} = (0.8, 0.8)'$ is denoted as $\mathbf{Ps}$, $\mathbf{P}_{\mathbf{S}\perp\mathbf{X}'\mathbf{X}}\widehat{\boldsymbol{\beta}} = (0.6, 0.6)'$ is denoted as $\mathbf{P}\widehat{\boldsymbol{\beta}}$, and the larger ellipse, which achieves the minimum in the translated problem, is

$$\left\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right\|^2_{\mathbf{X}'\mathbf{X}} = \left(\widehat{\beta}_1 - \beta_1\right)^2 + 4\left(\widehat{\beta}_2 - \beta_2\right)^2 = 1.8.$$

From the figure, $\widehat{\boldsymbol{\beta}}_R = \mathbf{P}\widehat{\boldsymbol{\beta}} + (\mathbf{s} - \mathbf{Ps})$.

Exercise 29 (Empirical) *The data file cps85.dat contains a random sample of 528 individuals from the 1985 Current Population Survey by the U.S. Census Bureau. The file contains observations on nine variables, listed as below.*

V1 = education (in years)
V2 = region of residence (coded 1 if South, 0 otherwise)
V3 = (coded 1 if nonwhite and non-Hispanic, 0 otherwise)
V4 = (coded 1 if Hispanic, 0 otherwise)
V5 = gender (coded 1 if female, 0 otherwise)
V6 = marital status (coded 1 if married, 0 otherwise)
V7 = potential labor market experience (in years)
V8 = union status (coded 1 if in union job, 0 otherwise)
V9 = hourly wage (in dollars)

Estimate a regression of wage y_i *on education* x_{1i}, *experience* x_{2i}, *and experienced-squared* $x_{3i} = x_{2i}^2$ *(and a constant). Report the OLS estimates.*

Let $\widehat{u}_i$ be the OLS residual and $\widehat{y}_i$ the predicted value from the regression. Numerically calculate the following:

(a) $\sum_{i=1}^{n} \widehat{u}_i$

(b) $\sum_{i=1}^{n} x_{1i}\widehat{u}_i$

(c) $\sum_{i=1}^{n} x_{2i}\widehat{u}_i$

(d) $\sum_{i=1}^{n} x_{1i}^2\widehat{u}_i$

(e) $\sum_{i=1}^{n} x_{2i}^2\widehat{u}_i$

(f) $\sum_{i=1}^{n} \widehat{y}_i\widehat{u}_i$

Are these calculations consistent with the theoretical properties of OLS? Explain.

(g) *Re-estimate the slope on education using the residual regression approach. Regress y_i on $(1, x_{2i}, x_{2i}^2)$, regress x_{1i} on $(1, x_{2i}, x_{2i}^2)$, and regress the residuals on the residuals. Report the estimate from this regression. Does it equal the value from the first OLS regression? Explain.*

Technical Appendix: Conditional Expectation

In Section 2, $E[y|\mathbf{x}]$ is interpreted as the projection of y on $L^2(P, \sigma(\mathbf{x}))$. We restrict $y \in L^2(P)$ because $L^2(P)$ is a Hilbert space and has nice geometric structures. Actually, we only need $y \in L^1(P)$ to define $E[y|\mathbf{x}]$.

The first thing about conditional expectation (CE) is that any CE is defined with respect to (w.r.t.) some sigma algebra (σ-algebra, or Borel field). A collection of subsets of the sample space Ω (which is the collection of all possible outcomes), denoted by $\mathcal{F}$, is called a **sigma algebra** if the empty set $\emptyset \in \mathcal{F}$, $A \in \mathcal{F}$ implies its complement $A^c \in \mathcal{F}$, and $A_1, A_2, \cdots \in \mathcal{F}$ implies $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$. It is better to first generate a σ-algebra from a partition although every σ-algebra cannot be generated from a partition.

Roughly speaking, σ-algebras represent information and a finer σ-algebra represents more information. One example may clarify this point. If we toss a die and can observe every possible outcome, then our information is $\mathcal{F} = \sigma(\{\omega_1\}, \{\omega_2\}, \cdots, \{\omega_6\})$, where $\sigma(\cdot)$ means the σ-algebra generated from a collection of sets, ω_i, $i = 1, \cdots, 6$, means the outcome i, and $\Omega = \{\omega_1, \omega_2, \cdots, \omega_6\}$. Suppose now we cannot observe the exact outcome, but we are told that the outcome is either even or odd. Our current information is $\mathcal{A} = \sigma(\Omega_1, \Omega_2) \equiv \sigma(\{\omega_1, \omega_3, \omega_5\}, \{\omega_2, \omega_4, \omega_6\})$. Obviously, $\mathcal{A} \subset \mathcal{F}$, i.e., $\mathcal{A}$ contains less information than $\mathcal{F}$. The roughest σ-algebra is $\{\emptyset, \Omega\}$ which means that we do not have any information.

If we are asked to give a prediction for the outcome when the outcome is told to be even, our answer should be

$$2 \times \frac{1}{3} + 4 \times \frac{1}{3} + 6 \times \frac{1}{3} = 4.$$

How did we get this result?

$$2 \times \frac{1/6}{1/2} + 4 \times \frac{1/6}{1/2} + 6 \times \frac{1/6}{1/2} = 4.$$

In other words, we rescale the probability on Ω_2 and then take average. This is the essence of CE: on each piece of the partition, rescale the probability and calculate the average, and use this average (a constant) as the prediction on that piece. This will generate a new random variable (r.v.), which is constant on each piece of the partition; if use the jargon of probability, this new random variable is measurable w.r.t. the σ-algebra generated by the partition. In the above example, we predict the random variable X, which is defined as $X(\omega_i) = i$, $i = 1, \cdots, 6$, by

$$E[X|\mathcal{A}] = \begin{cases} X(\omega_i) = 3, & i = 1, 3, 5 \\ X(\omega_i) = 4, & i = 2, 4, 6 \end{cases}$$

Extend a little further: suppose $\mathcal{A} = \sigma(A_i, i = 1, 2, \cdots)$, what is $E[X|\mathcal{A}]$? It should be

$$E[X|\mathcal{A}] = \sum_{i=1}^{\infty} E\left[X \frac{\mathbf{1}_{A_i}}{P(A_i)}\right] \cdot \mathbf{1}_{A_i} = \sum_{i=1}^{\infty} \frac{E[X\mathbf{1}_{A_i}]}{P(A_i)} \cdot \mathbf{1}_{A_i},$$

where $\mathbf{1}_{A_i}$ means the indicator function of A_i. From the formula of $E[X|\mathcal{A}]$, we can see (i) $\int_{A_i} E[X|\mathcal{A}]\, dP = \int_{A_i} X dP$, $\forall\ A_i \in \mathcal{A}$; (ii) $E[E[X|\mathcal{A}]] = E[X]$ by adding all equations in (i) together. Result (i) is the definition of CE, and result (ii) is the famous LIE. In the above example, the LIE can be check as follows:

$$3 \times \frac{1}{2} + 4 \times \frac{1}{2} = \frac{1}{6}(1 + 2 + \cdots + 6) = \frac{7}{2}.$$

We are now ready to give out the abstract definition of CE.

Definition 5 (Conditional Expectation) *Let $(\Omega, \mathcal{F}, P)$ be a probability space, $\mathcal{A} \subset \mathcal{F}$ be a sub-σ algebra, and X be a $L^1(P)$ random variable on $(\Omega, \mathcal{F}, P)$, the* conditional expectation *of X given $\mathcal{A}$ is a random variable Y that satisfies the following two conditions:*

(i) *Y is $\mathcal{A}$-measurable;*

(ii) *$\int_A Y dP = \int_A X dP$, $\forall\ A \in \mathcal{A}$.*

Obviously, if we change Y on a null set (i.e., a set with probability zero), it still satisfies conditions (i) and (ii), so CE is unique only almost surely (a.s.).

Given the definition of conditional expectation, we can understand the meaning of some popular terms.

(1) For a random variable X, $E[Y|X] = E[Y|\sigma(X)]$, where $\sigma(X) = \sigma(\{\omega | X(\omega) \leq a\}, a \in \mathbb{R})$ means the σ-algebra generated by X. If $X = c$, a constant, then $E[Y|c] = E[Y|\{\emptyset, \Omega\}] = E[Y]$.

(2) Conditional probability is a special case of conditional expectation: $P(B|\mathcal{A}) = E[\mathbf{1}_B|\mathcal{A}]$. We can check this definition coincides with the definition of conditional probability in elementary econometric books: $\forall\ A \in \mathcal{A}$ with $P(A) > 0$, $P(B|A) = \frac{E[\mathbf{1}_B \mathbf{1}_A]}{P(A)} = \frac{P(A\cap B)}{P(A)}$.

(3) $E[Y|X = x]$ can be understood as follows. First, calculate $E[Y|X]$. Since it is measurable w.r.t. $\sigma(X)$, there is a Borel function g such that $E[Y|X](\omega) = g(X(\omega))$. We define $E[Y|X = x] \equiv g(x)$.

Although most intuitive understandings of conditional expectation are correct, there are indeed some cases deserving careful inspection; the famous one is the **Borel paradox**, which can be found in some probability book and is not discussed here.

Example 6 *Let $\Omega = \mathbb{R}$, $\mathcal{F} = B(\mathbb{R}) = \sigma((-\infty, a], a \in \mathbb{R})$, the Borel σ-algebra, and P is defined by the cdf*

$$F(a) = \int_{-\infty}^{a} e^{-x}\mathbf{1}_{\{x\geq 0\}}dx$$

Define $X : \Omega \to \mathbb{R}$ by $X(\omega) = \omega$ and $Y = 3 \cdot \mathbf{1}_{\{X\leq 10\}} + 27 \cdot \mathbf{1}_{\{X>10\}}$. Find $E[X|Y]$.

Solution: First, $\sigma(Y) = \sigma(\{\omega > 10\})$, that is, $\mathcal{A}$ is a partition, so we need only calculate two values:

- $\omega > 10$: $E[X|Y](\omega) = \frac{\int_{10}^{\infty} xe^{-x}dx}{\int_{10}^{\infty} e^{-x}dx} = 11$;
- $\omega \leq 10$: $E[X|Y](\omega) = \frac{\int_{0}^{10} xe^{-x}dx}{\int_{0}^{10} e^{-x}dx} = \frac{1-11\cdot e^{-10}}{1-e^{-10}}$,

In summary, $E[X|Y](\omega) = \begin{cases} \frac{1-11\cdot e^{-10}}{1-e^{-10}}, & \omega \leq 10 \\ 11, & \omega > 10 \end{cases}$, a.s.. □

Exercise 30 *In the example above, define $Z = X \wedge c \equiv \min(X, c)$, where c is a constant. Find $E[X|Z]$.*

Exercise 31 (Bayes Formula) *If $dP = LdQ$, prove*

$$E^P[Z|\mathcal{A}] = \frac{E^Q[ZL|\mathcal{A}]}{E^Q[L|\mathcal{A}]},$$

where for a probability measure R, $E^R[\cdot]$ means the expection under R.

Example 7 *Suppose X and Y are independent, $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is a measurable function such that $\phi(X, Y)$ is integrable. Define $\psi(x) = E[\phi(x, Y)]$. Prove $E[\phi(X, Y)|X] = \psi(X)$.*

Proof. $\psi(X)$ is $\sigma(X)$ measurable, so we need only check condition (ii) in the definition of conditional expectation. First prove the case with $\phi(X, Y) = \mathbf{1}_A(X)\mathbf{1}_B(Y)$, where $A \in B(\mathcal{X})$, and

$B \in B(\mathcal{Y})$. Define $\psi_{A\times B}(x) = E\left[\mathbf{1}_{A\times B}(x, Y)\right] = \mathbf{1}_A(x)E\left[\mathbf{1}_B(Y)\right]$. For any $C \in B(\mathcal{X})$,

$$\begin{aligned} E\left[\mathbf{1}_C\phi(X,Y)\right] &= E\left[\mathbf{1}_C\mathbf{1}_A(X)\mathbf{1}_B(Y)\right] \\ &= E\left[\mathbf{1}_C\mathbf{1}_A(X)\right]E\left[\mathbf{1}_B(Y)\right] \\ &= E\left[\mathbf{1}_C\mathbf{1}_A(X)E\left[\mathbf{1}_B(Y)\right]\right] = E\left[\mathbf{1}_C\psi_{A\times B}(X)\right]. \end{aligned}$$

Then by some technical steps, we could prove this is true for any $\phi(X,Y)$. ■

Exercise 32 *Let $Y : \Omega \to \mathbb{N}$ be independent of the i.i.d. random variables $\{X_i\}_{i=1}^{\infty}$, where $\mathbb{N}$ is the set of natural number. Prove*

$$E\left[\sum_{i=1}^{Y} X_i \middle| Y\right] = Y \cdot E[X_1].$$

We conclude by collecting some popular properties of conditional expectation.

(1) If X is $\mathcal{A}$ measurable, then $E[X|\mathcal{A}] = X$. For example, $E[h(X)|X] = h(X)$ for any Borel measurable function $h(\cdot)$, and $E[c|\mathcal{A}] = c$ for any σ-algebra $\mathcal{A}$.

(2) If X is independent of Y, then $E[X|Y] = E[X]$.

(3) If Z is $\mathcal{A}$ measurable, $X \in L^1$, $XZ \in L^1$, then

$$E[XZ|\mathcal{A}] = ZE[X|\mathcal{A}]$$

(4)(Linearity) $E[\alpha X + \beta Y|\mathcal{A}] = \alpha E[X|\mathcal{A}] + \beta E[Y|\mathcal{A}]$.

(5)(Monotonicity) If $X \leq Y$, then $E[X|\mathcal{A}] \leq E[Y|\mathcal{A}]$.

(6)(LIE) If $\mathcal{A} \subset \mathcal{G} \subset \mathcal{F}$, $X \in L^1$, then

$$E[E[X|\mathcal{A}]|\mathcal{G}] = E[E[X|\mathcal{G}]|\mathcal{A}] = E[X|\mathcal{A}]$$

(7)(Contraction) If $\mathcal{A} \subset \mathcal{G}$, then $E[X] \leq \|E[X|\mathcal{A}]\|_p \leq \|E[X|\mathcal{G}]\|_p \leq \|X\|_p$, $\forall\, p \geq 1$, where for a random variable Z, $\|Z\|_p = \left(E\left[\|Z\|^p\right]\right)^{1/p}$. Especially, $(E[X])^2 \leq E[(E[X|\mathcal{A}])^2] \leq E[X^2]$.

(8) If $\mathcal{A} \subset \mathcal{G}$, then $\|X - E[X|\mathcal{A}]\|_p \geq \|X - E[X|\mathcal{G}]\|_p$, $\forall\, p \geq 1$.