Projection Theorem

Bryan S. Graham, UC - Berkeley & NBER

September 29, 2020

Optimization and approximation problems arise frequently in econometrics. Many of these problems are solvable using vector space methods. In particular by orthogonal projection in a vector space endowed with an inner product. While there is a fixed cost associated with developing a vector space approach, the long term pay off is considerable. Many properties of, for example, linear regression are easily derived using projection arguments. Projection arguments also play important roles in deriving the distributions of (complicated) sequences of statistics (cf., van der Vaart, 1998, Chapters 12 - 13) and in semiparametric efficiency bound analysis (cf., Newey, 1990). Our development of these tools will be informal, but of sufficient depth so as to allow for application to interesting problems.

Vector space

Let $\mathcal{H}$ denote a **vector space** over the field of real numbers. An element of this space is called a vector. Two examples of vector spaces that will feature prominently in what follows are (i) the $\mathbb{R}^N$ (Euclidean) and (ii) the L^2 spaces. An element of the Euclidean space is simply an $N \times 1$ vector (or list) of real numbers. An element of a L^2 space is a (function of a) random variable with finite variance. Vector spaces need to include a **null vector.** In Euclidean spaces the null vector is simply a list of zeros; in L^2 spaces the null vector is a degenerate random variable identically equal to zero. We will denote the null element of a vector space by a 0. We can add vectors in these spaces in the normal way (element-wise) and also multiply (i.e., rescale) them by scalars. Chapter 2 of Luenberger (1969) presents a formal development.

If we pair a vector space with an inner product defined on $\mathcal{H} \times \mathcal{H}$ we get what is called a (pre-) **Hilbert space**. A valid inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfies the conditions:

1. **Bi-linearity**: $\langle ah_1 + bh_2, ch_3 + dh_4 \rangle = ac \langle h_1, h_3 \rangle + ad \langle h_1, h_4 \rangle + bc \langle h_2, h_3 \rangle + bd \langle h_2, h_4 \rangle$ for a, b, c and d real scalars and h_1, h_2, h_3 and h_4 elements of $\mathcal{H}$;

2. **Symmetry**: $\langle h_1, h_2 \rangle = \langle h_2, h_1 \rangle$;

3. **Positivity** $\langle h_1, h_1 \rangle \geq 0$ with equality if, and only if, h_1 is a null vector.

Let $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$ and $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ be $N \times 1$ vectors of real numbers. For example $(X_1, Y_1), \ldots, (X_N, Y_N)$ may consist of pairs of years of completed schooling and adult earnings measures for a random *sample* of N adult male workers. Here $\mathbf{X}$ and $\mathbf{Y}$ are elements of the Euclidean space $\mathbb{R}^N$ and we will work with the inner product

$$\langle \mathbf{X}, \mathbf{Y} \rangle = \frac{\mathbf{X}'\mathbf{Y}}{N} = \frac{1}{N} \sum_{i=1}^{N} X_i Y_i. \tag{1}$$

It is an easy, but a useful exercise, to verify that (1) satisfies our three conditions for a valid inner product. Note that (1) is the familiar **dot product (**albeit divided by N).

Let X and Y denote years of completed schooling and earnings for a generic random draw from the *population* of adult male workers. Here X and Y may be regarded as elements of an L^2 space, where we will work with the inner product

$$\langle X, Y \rangle = \mathbb{E}[XY], \tag{2}$$

where $\mathbb{E}[X]$ denotes the expected value, or population mean, of the random variable X. Again, it is a useful exercise to verify that (2) satisfies our three conditions for a valid inner product. I will sometimes refer to (2) as the **covariance inner product**.

Associated with an inner product is a **norm** $\|h\| = \langle h, h \rangle^{1/2}$ which satisfies:

1. $\|h\| = 0$ if, and only if, $h = 0$;

2. $\|ah\| = |a| \, \|h\|$ for any scalar a;

3. **Triangle Inequality**: $\|h_1 + h_2\| \leq \|h_1\| + \|h_2\|$.

The first two properties of the norm are easy to verify. It is instructive to verify the third. To do this we will first prove.

Lemma 1. *(*Cauchy-Schwarz Inequality*) For all* $(h_1, h_2) \in \mathcal{H} \times \mathcal{H}$*,*

$$|\langle h_1, h_2 \rangle| \leq \|h_1\| \, \|h_2\|$$

with equality if, and only if, $h_1 = \alpha h_2$ *for some real scalar* α *or* $h_2 = 0$*.*

Proof. Begin by observing that for *all* scalars α

$$\begin{aligned} 0 &\leq \langle h_1 - \alpha h_2, h_1 - \alpha h_2 \rangle \\ &= \langle h_1, h_1 \rangle - \alpha \langle h_1, h_2 \rangle - \alpha \langle h_2, h_1 \rangle + \alpha^2 \langle h_2, h_2 \rangle \\ &= \|h_1\|^2 - 2\alpha \langle h_1, h_2 \rangle + \alpha^2 \|h_2\|^2 . \end{aligned} \tag{3}$$

Next set $\alpha = \frac{\langle h_1, h_2 \rangle}{\|h_2\|^2}$; substituting into (3) yields

$$0 \leq \|h_1\|^2 - \frac{\langle h_1, h_2 \rangle^2}{\|h_2\|^2},$$

which after re-arranging and taking square roots yields the result. □

With Lemma 1 in hand it is straightforward to prove the Triangle Inequality.

Lemma 2. *(*Triangle Inequality*) For all* $(h_1, h_2) \in \mathcal{H} \times \mathcal{H}$*,*

$$\|h_1 + h_2\| \leq \|h_1\| + \|h_2\| .$$

Proof. Applying the definition of the norm and using the bi-linearity property of the inner product yields

$$\begin{aligned} \|h_1 + h_2\|^2 &= \langle h_1 + h_2, h_1 + h_2 \rangle \\ &= \|h_1\|^2 + 2 \langle h_1, h_2 \rangle + \|h_2\|^2 \\ &\leq \|h_1\|^2 + 2 |\langle h_1, h_2 \rangle| + \|h_2\|^2 \\ &\leq \|h_1\|^2 + 2 \|h_1\| \|h_2\| + \|h_2\|^2 \\ &= (\|h_1\| + \|h_2\|)^2 , \end{aligned}$$

where the fourth line follows from the Cauchy-Schwarz inequality. Taking the square root of both sides gives the result. □

The Cauchy-Schwarz (CS) and Triangle (TI) inequalities are widely-used in probability, statistics and econometrics.

We say that the vectors X and Y are **orthogonal** if their inner product is zero. We denote this by $X \perp Y$. When two vectors are orthogonal the Triangle Inequality is tight, yielding Pythagoras' Theorem; the proof of which is left as an exercise.

Theorem 1. *(*Pythagorean Theorem*) If* $h_1 \perp h_2$*, then* $\|h_1 + h_2\|^2 = \|h_1\|^2 + \|h_2\|^2$*.*

The value of Hilbert spaces is that they provide a mechanism for generalizing geometric intuitions familiar from Euclidean spaces to more complicated situations. An example $\|h_1 - h_2\|$ measures the **distance** between h_1 and h_2. Another is that the notion of orthogonality corresponds to perpendicularity. A good way to develop intuition about some of the results in this note is to reflect on their geometric interpretations in 2 and 3 dimensions.

Projection Theorem

Let $\mathcal{L}$ be some linear subspace of $\mathcal{H}$ (i.e., if h_1 and h_2 are both in $\mathcal{L}$, then so is $ah_1 + bh_2$). In what follows, we will restrict ourselves to closed linear subspaces. This implies that if h_N is an element of $\mathcal{L}$ for all N and $h_N \to h$, then h is also in $\mathcal{L}$; Luenberger (1969) provides additional details. In what follows "subspace" refers to a closed linear subspace unless explicitly stated otherwise.

Let X and Y be two elements of $\mathcal{H}$. Then $\mathcal{L}$ might consist of all linear functions of X, or (almost) any function of X. It is of considerable interest to consider the projection of $Y \in \mathcal{H}$ onto the subspace $\mathcal{L}$. Specifically we define the **projection operator** $\Pi(\cdot \,|\, \mathcal{L}) : \mathcal{H} \to \mathcal{L}$ by: $\Pi(Y \,|\, \mathcal{L})$ is the element $\hat{Y} \in \mathcal{L}$ that achieves

$$\min_{\hat{Y} \in \mathcal{L}} \left\| Y - \hat{Y} \right\|. \tag{4}$$

In words we look for the element of $\mathcal{L}$, a subspace of $\mathcal{H}$, which is closest to Y (or "best approximates" Y). The notion of "best" is embodied in the chosen inner product and induced norm.

It is instructive to consider two examples. Let $\mathbf{Y}$ and $\mathbf{X}$ be the vectors consisting of earnings-schooling pairs for a sample of adult male workers. Let $\mathcal{L}$ be the linear span of $\mathbf{1}, \mathbf{X}$ (i.e., vectors of the form $\alpha\mathbf{1} + \beta\mathbf{X}$). In that case finding $\Pi(Y \,|\, \mathcal{L})$ corresponds to computing $\hat{\alpha}$ and $\hat{\beta}$, the solutions to

$$\min_{(\alpha,\beta)\in\mathbb{R}^2} \|\mathbf{Y} - \alpha\mathbf{1} - \beta\mathbf{X}\|^2 = \min_{(\alpha,\beta)\in\mathbb{R}^2} \frac{1}{N} \sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2. \tag{5}$$

This corresponds to finding the **ordinary least squares** (OLS) fit of $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_N)'$ onto a constant and $\mathbf{X} = (X_1, X_2, \ldots, X_N)'$.

Alternatively let (X, Y) denote a schooling-earnings pair for a generic random draw from the population of adult male workers. Let $\mathcal{L}$ consist of all linear functions of X; using the norm associated with our L^2 Hilbert space we have that $\Pi(Y \,|\, \mathcal{L})$ corresponds to computing

α_0 and β_0, the solutions to

$$\min_{(\alpha,\beta)\in\mathbb{R}^2} \|Y-\alpha-\beta X\|^2 = \min_{(\alpha,\beta)\in\mathbb{R}^2} \mathbb{E}\left[(Y-\alpha-\beta X)^2\right]. \tag{6}$$

This corresponds to finding the best (i.e., mean squared error minimizing) **linear predictor** (LP) of Y given X.

Both (5) and (6) correspond to prediction problems. It turns out that, using the elementary Hilbert space theory outlined above, we can provide a generic solution to both of them (and indeed many other problems). The solution is a generalization of the idea familiar from elementary school geometry that one can finding the shortest distance between a point and a line by "dropping the perpendicular".

Theorem 2. *(*Projection Theorem*) Let $\mathcal{H}$ be a vector space with an inner product and associated norm and $\mathcal{L}$ a subspace of $\mathcal{H}$, then for Y an arbitrary element of $\mathcal{H}$, if there exists a vector $\hat{Y}\in\mathcal{L}$ such that*

$$\left\|Y-\hat{Y}\right\| \leq \left\|Y-\tilde{Y}\right\| \tag{7}$$

for all $\tilde{Y}\in\mathcal{L}$, then

1. $\hat{Y}=\Pi\left(Y\,|\,\mathcal{L}\right)$ is unique
2. A necessary and sufficient condition for $\hat{Y}$ to be the uniquely minimizing vector in $\mathcal{L}$ is the orthogonality condition

$$\left\langle Y-\hat{Y},\tilde{Y}\right\rangle = 0 \text{ for all } \tilde{Y}\in\mathcal{L}$$

(or $Y-\Pi\left(Y\,|\,\mathcal{L}\right)\perp\tilde{Y}$ for all $\tilde{Y}\in\mathcal{L}$).

Proof. See Luenberger (1969, Theorem 1). We begin by verifying that orthogonality is a necessary condition for $\hat{Y}$ to be norm minimizing. Suppose there exists a vector $\tilde{Y}$ which is not orthogonal to the prediction error $Y-\hat{Y}$. This implies that $\left\langle Y-\hat{Y},\tilde{Y}\right\rangle=\alpha\neq 0$. We

can, without loss of generality assume that $\left\|\tilde{Y}\right\| = 1$,[1] and evaluate

$$\begin{aligned}\left\|Y-\hat{Y}-\alpha\tilde{Y}\right\|^2 &= \left\langle Y-\hat{Y}-\alpha\tilde{Y}, Y-\hat{Y}-\alpha\tilde{Y}\right\rangle \\ &= \left\|Y-\hat{Y}\right\|^2 - \left\langle Y-\hat{Y}, \alpha\tilde{Y}\right\rangle - \left\langle \alpha\tilde{Y}, Y-\hat{Y}\right\rangle + \alpha^2\left\|\tilde{Y}\right\|^2 \\ &= \left\|Y-\hat{Y}\right\|^2 - \alpha^2,\end{aligned}$$

which implies the contradiction $\left\|Y-\hat{Y}-\alpha\tilde{Y}\right\| \leq \left\|Y-\hat{Y}\right\|$. Next we show that if $Y-\hat{Y} \perp \mathcal{L}$, then $\hat{Y}$ is the unique minimizing vector. Let $\tilde{Y}$ be some arbitrary element of $\mathcal{L}$; we have that

$$\begin{aligned}\left\|Y-\tilde{Y}\right\|^2 &= \left\|Y-\hat{Y}+\hat{Y}-\tilde{Y}\right\|^2 \\ &= \left\|Y-\hat{Y}\right\|^2 + 2\left\langle Y-\hat{Y}, \hat{Y}-\tilde{Y}\right\rangle + \left\|\hat{Y}-\tilde{Y}\right\|^2 \\ &= \left\|Y-\hat{Y}\right\|^2 + \left\|\hat{Y}-\tilde{Y}\right\|^2,\end{aligned}$$

where the last equality follows from the fact that $\hat{Y}-\tilde{Y} \in \mathcal{L}$ and $Y-\hat{Y}$ is orthogonal to any element of $\mathcal{L}$. Next, by the properties of the norm, $\left\|\hat{Y}-\tilde{Y}\right\| \geq 0$, with equality if, and only if, $\hat{Y}=\tilde{Y}$. This implies (7) for all $\tilde{Y} \in \mathcal{L}$ with the equality only holding if $\hat{Y}=\tilde{Y}$. This gives sufficiency and uniqueness. □

Observe that we have not shown existence of a solution to (4). We have shown that conditional on the existence of a solution, that the solution is unique and that the prediction error $Y-\hat{Y}$ is orthogonal to the subspace $\mathcal{L}$. Proving existence is a more technical argument. For a general result, which applies to closed linear subspaces (and all the cases considered here), see Luenberger (1969, p. 51 - 52).

Three additional properties of projections will prove useful to us. First, they are **linear operators**. To see this note that we can write, using the Projection Theorem,

$$\begin{aligned}X &= \Pi\left(X|\,\mathcal{L}\right) + U_X,\ \ U_X \perp \mathcal{L} \\ Y &= \Pi\left(Y|\,\mathcal{L}\right) + U_Y,\ \ U_Y \perp \mathcal{L}.\end{aligned}$$

[1] To see this note we could always work with the normalized vector $\tilde{Y}^* = \tilde{Y}/\left\|\tilde{Y}\right\|$ and constant $\alpha^* = \alpha/\left\|\tilde{Y}\right\|$ in what follows.

Rescaling and adding yields

$$aX + bY = a\Pi\left(X\middle|\mathcal{L}\right) + b\Pi\left(Y\middle|\mathcal{L}\right) + aU_X + bU_Y.$$

Now observe that, using bi-linearity of the inner product, for all $W \in \mathcal{L}$

$$\left\langle aU_X + bU_Y, W\right\rangle = a\left\langle U_X, W\right\rangle + b\left\langle U_Y, W\right\rangle = 0$$

and hence

$$\Pi\left(aX + bY\middle|\mathcal{L}\right) = a\Pi\left(X\middle|\mathcal{L}\right) + b\Pi\left(Y\middle|\mathcal{L}\right). \tag{8}$$

Linearity of the projection operator will be useful for establishing several properties of linear regression.

A second property of the projection operator is **idempotency**. Idempotency of an operator means that it can be applied multiple times without changing the result beyond the one found after the initial application. In the context of projections this property implies that

$$\Pi\left(\Pi\left(Y\middle|\mathcal{L}\right)\middle|\mathcal{L}\right) = \Pi\left(Y\middle|\mathcal{L}\right). \tag{9}$$

The projection of a projection is itself (assuming the same subspaces are projected onto in both cases). To see this observe that

$$\begin{aligned}
0 &= \left\langle \Pi\left(\Pi\left(Y\middle|\mathcal{L}\right)\middle|\mathcal{L}\right) - \Pi\left(Y\middle|\mathcal{L}\right), \tilde{Y}\right\rangle \\
&= \left\langle Y - \Pi\left(Y\middle|\mathcal{L}\right), \tilde{Y}\right\rangle - \left\langle Y - \Pi\left(\Pi\left(Y\middle|\mathcal{L}\right)\middle|\mathcal{L}\right), \tilde{Y}\right\rangle \\
&= 0 - \left\langle Y - \Pi\left(\Pi\left(Y\middle|\mathcal{L}\right)\middle|\mathcal{L}\right), \tilde{Y}\right\rangle,
\end{aligned}$$

but the last line is the necessary and sufficient condition for $\Pi\left(\Pi\left(Y\middle|\mathcal{L}\right)\middle|\mathcal{L}\right)$ to be the unique projection of Yonto $\mathcal{L}$. This gives (9) above.

A third property of projections is that they are **norm reducing**. Let 1 denote a constant vector. In a Euclidean space the constant vector equals an $N \times 1$ vector of ones, denoted by $\mathbf{1}$. In the L^2 space it is a degenerate random variable always equal to 1. In the Euclidean case we have that $\Pi\left(\mathbf{Y}\middle|\mathbf{1}\right) = \frac{1}{N}\sum_{i=1}^N Y_i = \bar{Y}$, the sample mean. In the L^2 case $\Pi\left(Y\middle|1\right) = \mathbb{E}\left[Y\right]$. It is a useful exercise to verify these claims. Observe further that, for the dot product and covariance inner products respectively,

$$\left\|\mathbf{Y} - \Pi\left(\mathbf{Y}\middle|\mathbf{1}\right)\right\|^2 = \frac{1}{N}\sum_{i=1}^N \left(Y_i - \bar{Y}\right)^2, \ \left\|Y - \Pi\left(Y\middle|1\right)\right\|^2 = \mathbb{V}\left(Y\right),$$

which correspond to the sample and population variance respectively.

We can use the results above, in conjunction with the Pythagorean Theorem, to develop a more general version of the familiar analysis of variance decomposition (e.g., Goldberger, 1991, p. 48).

Lemma 3. *(*Analysis of Variance*) If the linear subspace $\mathcal{L}$ contains the constant vector, then*

$$\|Y - \Pi(Y|1)\|^2 = \|Y - \Pi(Y|\mathcal{L})\|^2 + \|\Pi(Y|\mathcal{L}) - \Pi(Y|1)\|^2.$$

Proof. By the bi-linearity property of the inner product:

$$\begin{aligned}\|Y - \Pi(Y|\mathcal{L})\|^2 = &\|Y - \Pi(Y|1)\|^2 - 2\langle Y - \Pi(Y|1), \Pi(Y|\mathcal{L}) - \Pi(Y|1)\rangle \\ &+ \|\Pi(Y|\mathcal{L}) - \Pi(Y|1)\|^2.\end{aligned}$$

The middle term in the expression above can be further manipulated:

$$\begin{aligned}\langle Y - \Pi(Y|1), \Pi(Y|\mathcal{L}) - \Pi(Y|1)\rangle &= \langle Y - \Pi(Y|\mathcal{L}) + \Pi(Y|\mathcal{L}) - \Pi(Y|1), \Pi(Y|\mathcal{L}) - \Pi(Y|1)\rangle \\ &= \langle Y - \Pi(Y|\mathcal{L}), \Pi(Y|\mathcal{L}) - \Pi(Y|1)\rangle \\ &\quad + \|\Pi(Y|\mathcal{L}) - \Pi(Y|1)\|^2 \\ &= \|\Pi(Y|\mathcal{L}) - \Pi(Y|1)\|^2\end{aligned}$$

where $\langle Y - \Pi(Y|\mathcal{L}), \Pi(Y|\mathcal{L}) - \Pi(Y|1)\rangle = 0$ by the Projection Theorem since $\Pi(Y|\mathcal{L}) - \Pi(Y|1) \in \mathcal{L}$ whenever $\mathcal{L}$ contains the constant vector. Re-arranging gives the result. □

An immediate implication of Lemma 3 is that

$$\|Y - \Pi(Y|\mathcal{L})\| \leq \|Y - \Pi(Y|1)\|$$

(i.e., $\Pi(Y|\mathcal{L})$ is norm reducing).

The least squares fit as a projection

You are likely already familiar with one important projection: the **least squares fit**. Let $\mathbf{Y}$ be an $N \times 1$ vector of log earnings measures for a simple random sample of N adult males. Let $\mathbf{X}$ be a corresponding $N \times K$ matrix of covariates. The first column of this matrix consists of a vector of ones. The remaining columns includes measures of various respondent attributes (years of completed schooling, Armed Forces Qualification Test (AFQT) score, ethnicity

dummies, parents' school etc.). Assume that the columns of $\mathbf{X}$ are linearly independent such that the rank of $\mathbf{X}$ is K.

The **column space** of $\mathbf{X}$ is the span – or set of all possible linear combinations – of its column vectors. The set of all $N \times 1$ vectors expressable as linear combinations of the K columns of $\mathbf{X}$ is

$$\mathcal{L} = \mathcal{C}(\mathbf{X}) \stackrel{def}{=} \{\mathbf{X}\beta \,:\, \beta \text{ is a } K \times 1 \text{ vector of real numbers}\}. \tag{10}$$

The projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$ satisfies, by the Projection Theorem introduced earlier, the following necessary and sufficient condition

$$\langle \mathbf{Y} - \Pi(\mathbf{Y}|\,\mathcal{L}), \mathbf{X}\beta \rangle = \left\langle \mathbf{Y} - \mathbf{X}\hat{\beta}, \mathbf{X}\beta \right\rangle = 0. \tag{11}$$

Note that both $\mathbf{Y}$ and (any element of) $\mathcal{C}(\mathbf{X})$ are Euclidean vectors; hence we work with inner product (1) such that (11) coincides with

$$\frac{1}{N}\sum_{i=1}^{N}\left(Y_i - X_i'\hat{\beta}\right)X_i'\beta = \frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\beta_k X_{ik}\left(Y_i - X_i\hat{\beta}\right) = 0.$$

By setting $\beta_k = 1$ and $\beta_j = 0$ for $j \neq k$ we get the implication

$$\frac{1}{N}\sum_{i=1}^{N} X_{ik}\left(Y_i - X_i\hat{\beta}\right) = 0$$

for $k = 1, \ldots, K$ or, stacking conditions into a vector,

$$\frac{1}{N}\sum_{i=1}^{N} X_i\left(Y_i - X_i\hat{\beta}\right) = 0. \tag{12}$$

Hence (11) implies (12). The converse, that (12) implies (11) for any $\beta \in \mathbb{R}^K$, follows directly. This implies equivalence of the two conditions. We can therefore use (12) to find the projection $\Pi(\mathbf{Y}|\,\mathcal{L}) = \mathbf{X}\hat{\beta}$.

Condition (12) is a system of K linear equations

$$\underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right]}_{K\times 1} - \underbrace{\left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i\right]}_{K\times K}\underbrace{\hat{\beta}}_{K\times 1} = 0.$$

Solving this system for $\hat{\beta}$ yields the familiar ordinary least squares (OLS) estimator:

$$\begin{aligned}\hat{\beta} &= \left[\frac{1}{N}\sum_{i=1}^{N} X_i X_i'\right]^{-1} \times \left[\frac{1}{N}\sum_{i=1}^{N} X_i Y_i\right] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}.\end{aligned} \tag{13}$$

Hence the projection of $\mathbf{Y}$ onto the column space of $\mathbf{X}$ equals

$$\Pi(\mathbf{Y}|\,\mathcal{L}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\beta}.$$

This is also called the least squares regression fit or simply the **least squares fit**.

Space of real matrices

The discussion above provides a projection interpretation of least squares. It is useful to extend our Euclidean vector space to accommodate matrix prediction. This extension, as will we see later, has numerous applications in the area of panel data (where multiple outcomes for each sampling unit are observed). It also provides an interesting perspective on some basic results in matrix analysis.

Let X_i *and* Y_i be $K \times 1$ vectors of real numbers. For example, $Y_i = (Y_{i1}, \ldots, Y_{iT})'$ might be realized log earnings for individual i in years $t = 1, \ldots, T$ (here $T = K$). Define the real $N \times K$ matrices $\mathbf{X} = (X_1, \ldots, X_N)'$ and $\mathbf{Y} = (Y_1, \ldots, Y_N)'$. Here $\mathbf{X}$ and $\mathbf{Y}$ are elements of the space of real $N \times K$ matrices, say $M_{N\times K}(\mathbb{R})$. For this space we will work with the inner product

$$\langle \mathbf{X}, \mathbf{Y}\rangle = \frac{1}{N}\mathrm{Tr}(\mathbf{X}\mathbf{Y}') = \frac{1}{N}\sum_{i=1}^{N} X_i' Y_i. \tag{14}$$

Note that, for $K = 1$, (14) coincides with our inner product definition for Euclidean vector spaces, equation (1) above. Sometimes (14) is called the **Frobenius Inner Product** and denoted by $\langle \mathbf{X}, \mathbf{Y}\rangle_F$. The division by N in (14) is non-standard.

The associated norm is

$$\begin{aligned}\|\mathbf{X}\| &= \left[\frac{Tr\left(\mathbf{X}\mathbf{X}'\right)}{N}\right]^{1/2} \\ &= \left[\frac{1}{N}\sum_{i=1}^{N} X_i' X_i\right]^{1/2} \\ &= \left[\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N} |X_{ij}|^2\right]^{1/2} \\ &= \|\mathbf{X}\|_F .\end{aligned} \tag{15}$$

Equation (15) is the **Frobenius norm** (the division by N inside the $[\cdot]$ is typically omitted).

Linear regression/prediction

Let Y equal log earnings for a generic random draw from the population of adult males. Let X be a corresponding $K \times 1$ vector of respondent attributes (the first element of X is a constant). The linear subspace spanned by X equals

$$\mathcal{L} = \{X'\beta \,:\, \beta \text{ is a } K \times 1 \text{ vector of real numbers}\}. \tag{16}$$

Let $Z = (X', Y)'$ denote a generic random draw from the population of interest (with distribution function F_0 on support $Z \in \mathbb{Z} \subseteq \mathbb{R}^{K+1}$). Let $\|X\|_2 = \left[\sum_{k=1}^{K} X_k^2\right]^{1/2}$ denote the Euclidean Norm of X and impose the following regularity condition on the population distribution function, F_0.

Assumption 1. *(i)* $\mathbb{E}[Y^2] < \infty$, *(ii)* $\mathbb{E}\left[\|X\|_2^2\right] < \infty$, *and (iii)* $\mathbb{E}\left[(\alpha' X)^2\right] > 0$ *for any non-zero* $\alpha \in \mathbb{R}^K$.

We wish to compute the projection of Y onto $\mathcal{L}$:

$$\min_{\beta \in \mathbb{R}^K} \|Y - X'\beta\|^2 = \min_{\beta \in \mathbb{R}^K} \mathbb{E}\left[(Y - X'\beta)^2\right].$$

Next consider the space of all 1-dimensional random functions of Z with finite variance (L^2 space). Condition (i) of Assumption 1 implies that Y is an element of this space. By the Hölder Inequality (HI) inequality we have

$$\mathbb{E}\left[\|X'\beta\|_2\right] \leq \mathbb{E}\left[\|X\|_2^2\right]^{1/2} \|\beta\|_2 < \infty$$

by part (ii) of Assumption 1. Hence any element of $\mathcal{L}$ is also in L^2.

From the Projection Theorem we get the necessary and sufficient condition

$$\langle Y - X'\beta_0, X\beta \rangle = \mathbb{E}\left[(Y - X'\beta_0)\, X\beta\right] = 0$$

for all $\beta \in \mathbb{R}^K$. Using an argument similar to one used in the analysis of least squares earlier, we work with the equivalent $K \times 1$ vector of conditions:

$$\mathbb{E}\left[X\left(Y - X'\beta_0\right)\right] = 0. \tag{17}$$

Re-arranging (17) yields the system of K linear equations:

$$\mathbb{E}\left[XX'\right]\beta_0 - \mathbb{E}\left[XY\right] = 0.$$

Part (iii) of Assumption 1 requires that no single predictor corresponds to a linear combination of the others (i.e., that the elements of X be linearly independent). This condition ensures invertibility of $\mathbb{E}\left[XX'\right]$ since

$$\mathbb{E}\left[\left(\alpha' X\right)^2\right] = \alpha' \mathbb{E}\left[XX'\right]\alpha,$$

condition (iii) implies positive-definiteness of $\mathbb{E}\left[XX'\right]$. This, in turn, implies that the determinant of $\mathbb{E}\left[XX'\right]$ is non-zero (non-singularity) and hence that $\mathbb{E}\left[XX'\right]^{-1}$ is well-defined. Since $\mathbb{E}\left[XX'\right]$ is invertible we can directly solve for β_0:

$$\beta_0 = \mathbb{E}\left[XX'\right]^{-1} \times \mathbb{E}\left[XY\right]. \tag{18}$$

The corresponding best (i.e., MSE-minimizing) linear predictor (LP) of Y given $X = x$ is

$$\Pi\left(Y \middle| \mathcal{L}\right) = \mathbb{E}^*\left[Y \middle| X\right] = X'\beta_0. \tag{19}$$

Here $\mathbb{E}^*\left[Y \middle| X\right]$ is special notation for the best linear predictor of Y given X. Unless stated otherwise, I assume that a constant is a component of X.

Define $U = Y - X'\beta_0$ to be the prediction error associated with (19). From the first order conditions to (17) we get

$$\mathbb{E}\left[XU\right] = 0. \tag{20}$$

Equation (20) indicates that β_0 is chosen to ensure that the covariance between X and U is

zero. Recall that the first element of X is a constant so that (20) implies

$$\mathbb{E}[U] = 0$$

or zero average prediction error.

Multivariate regression

Let $\mathcal{H}$ be a Hilbert space consisting of J-dimensional random functions with finite variance. That is,

$$h : \mathbb{Z} \to \mathbb{R}^J$$

where $h(Z) = (h_1(Z), \ldots, h_J(Z))'$ is such that

$$\mathbb{E}\left[h(Z)' h(Z)\right] < \infty.$$

The null vector in this space is the degenerate random variable identically equal to a $J \times 1$ vector of zeros. In what follows I use h to denote $h(Z)$, with the dependence on the random variable Z left implicit. Call this space the space of K-dimensional random functions with finite variance.

We extend the covariance inner product, equation (2) above, to accommodate multidimensional dimensional h:

$$\langle h_1, h_2 \rangle = \mathbb{E}\left[h_1(Z)' h_2(Z)\right].$$

Next let $g(Z)$ be a $K \times 1$ vector of random functions with $\mathbb{E}[g'g] < \infty$. The linear subspace spanned by $g(Z)$ equals

$$\mathcal{L} = \{\Pi g : \Pi \text{ is a } J \times K \text{ matrix of real numbers}\}. \tag{21}$$

Assume that the random functions $g_1(Z), g_2(Z), \ldots, g_K(Z)$ are linearly independent such that the inverse $\mathbb{E}\left[g(Z) g(Z)'\right]^{-1}$ exists.

To find the projection of $h \in \mathcal{H}$ onto $\mathcal{L}$ we use the necessary and sufficient condition from the Projection Theorem introduced earlier:

$$\langle h - \Pi_0 g, \Pi g \rangle = 0, \text{ for all } \Pi \in \mathbb{R}^{J \times K}.$$

Under the covariance inner product this condition implies that

$$\mathbb{E}\left[(h-\Pi_0 g)'\,\Pi g\right]=0. \tag{22}$$

To find the form of Π_0 observe that condition (22) is equivalent to

$$\mathbb{E}\left[g\,(h-\Pi_0 g)'\right]=\mathbf{0}_{K\times J}. \tag{23}$$

To show the equivalence of (22) and (23) note that, using linearity of the expectation and trace operators,

$$\begin{aligned}\mathbb{E}\left[(h-\Pi_0 g)'\,\Pi g\right] &= \mathbb{E}\left[\mathrm{Tr}\left((h-\Pi_0 g)'\,\Pi g\right)\right]\\ &= \mathbb{E}\left[\mathrm{Tr}\left(\Pi v\,(h-\Pi_0 g)\right)'\right]\\ &= \sum_j\sum_k \pi_{jk}\mathbb{E}\left[g_j\,(h-\Pi_0 g)_k\right],\end{aligned}$$

where π_{jk} is the jk^{th} element of the arbitrary real $J\times K$ matrix Π. For any pair j,k we can set $\pi_{jk}=1$ and the rest of the elements of Π to zero. This implies

$$\mathbb{E}\left[g_j\,(h-\Pi_0 g)_k\right]=0$$

for all j,k or (23) . The converse, that (23) implies (22) for any $\Pi\in\mathbb{R}^{J\times K}$, follows directly. Solving (23) for Π_0 indicates that the unique projection of h onto $\mathcal{L}$ is:

$$\Pi\left(h|\,\mathcal{L}\right)=\Pi_0 g,\ \ \Pi_0=\mathbb{E}\left[hg'\right]\times\mathbb{E}\left[gg'\right]^{-1}. \tag{24}$$

This is the multivariate linear predictor (or multivariate linear regression) of $h(Z)$ onto $g(Z)$. Letting $Z=(Y_1,\ldots,Y_J,X_1,\ldots,X_K)'$, $h(Z)=Y=(Y_1,\ldots Y_J)'$ and $g(Z)=X=(X_1,\ldots,X_K)'$ yields

$$\Pi\left(Y|\,\mathcal{L}\right)=\Pi_0 X,\ \ \Pi_0=\mathbb{E}\left[YX'\right]\times\mathbb{E}\left[XX'\right]^{-1}. \tag{25}$$

Bibliographic notes

The standard reference on vector space optimization is Luenberger (1969). These notes draw extensively from this reference (glossing over details!). Appendix B.10 of Bickel & Doksum (2015) and van der Vaart (1998, Chapter 10) provide compact introductions targeted toward

statistical applications. I have also found the presentation of Hilbert Space theory in Tsiatis (2006) very helpful.

References

Bickel, P. J. & Doksum, K. A. (2015). *Mathematical Statistics*, volume 1. Boca Raton: Chapman & Hall, 2nd edition.

Goldberger, A. S. (1991). *A Course in Econometrics*. Cambridge, MA: Harvard University Press.

Luenberger, D. G. (1969). *Optimization by Vector Space Methods*. New York: John Wiley & Sons, Inc.

Newey, W. K. (1990). Semiparametric efficiency bounds. *Journal of Applied Econometrics*, 5(2), 99 – 135.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

