Lecture 1: Projection

Bryan S. Graham, UC - Berkeley & NBER

January 20, 2017

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 draw of) some (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 typically 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. See Chapter 2 of Luenberger (1969) for 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**. Let $X \in \mathcal{H}$ and $Y \in \mathcal{H}$, the inner product $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ satisfies the conditions of:

1. **Symmetry**: $\langle X, Y \rangle = \langle Y, X \rangle$;
2. **Bi-linearity**: $\langle aW + bX, cY + dZ \rangle = ac \langle W, Y \rangle + ad \langle W, Y \rangle + bc \langle X, Y \rangle + bd \langle X, Z \rangle$ for a, b, c and d real scalars and W, X, Y and Z elements of $\mathcal{H}$;
3. **Positivity** $\langle X, X \rangle \geq 0$ with equality if, and only if, X 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

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

It is 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 (**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.

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

1. $\|X\| = 0$ if, and only if, $X = 0$
2. $\|aX\| = |a| \|X\|$
3. **Triangle Inequality**: $\|X + Y\| \leq \|X\| + \|Y\|$.

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 $(X, Y) \in \mathcal{H} \times \mathcal{H}$,*

$$|\langle X, Y \rangle| \leq \|X\| \|Y\|$$

with equality if, and only if, $X = \alpha Y$ for some real scalar α or $Y = 0$.

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

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

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

$$0 \leq \|X\|^2 - \frac{\langle X, Y \rangle^2}{\|Y\|^2},$$

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* $(X, Y) \in \mathcal{H} \times \mathcal{H}$*,*

$$\|X + Y\| \leq \|X\| + \|Y\|.$$

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

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

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. These are good tools to have in your pocket.

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* $X \perp Y$*, then* $\|X + Y\|^2 = \|X\|^2 + \|Y\|^2$*.*

Theorem 1 is a generalization of the result familiar from high school geometry. As a special case consider the right triangle with vertices at $(0, 0)$, $(0, 3)$ and $(4, 0)$. The two vectors $\mathbf{X} = \begin{pmatrix} 4 \\ 0 \end{pmatrix}$ and $\mathbf{Y} = \begin{pmatrix} 0 \\ 3 \end{pmatrix}$ form a right angle, corresponding to orthogonality according to the dot product (1). Applying the associated Euclidean norm we have $\|\mathbf{X}\|^2 = \frac{1}{2}\sum_{i=1,2} X_i^2 = \frac{1}{2}(4^2 + 0^2) = \frac{16}{2}$, $\|\mathbf{Y}\|^2 = \frac{1}{2}\sum_{i=1,2} Y_i^2 = \frac{1}{2}(0^2 + 3^2) = \frac{9}{2}$ and $\|\mathbf{X} + \mathbf{Y}\|^2 = \frac{1}{2}\sum_{i=1,2}(X_i + Y_i)^2 = \frac{1}{2}(4^2 + 9^2) = \frac{25}{2}$ as needed.

Projection Theorem

Let $\mathcal{L}$ be some linear subspace of $\mathcal{H}$. Let X and Y be two elements of $\mathcal{H}$. Then $\mathcal{L}$ might consist of all linear functions of X, or (almost) any functions 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

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

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 Honduran 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** 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. $\hat{Y} = \Pi(Y|\mathcal{L})$ 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 $Y - \Pi\left(Y \middle| \mathcal{L}\right) \perp \tilde{Y}$ for all $\tilde{Y} \in \mathcal{L}$).

Proof. 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 technical argument and is often

[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.

more easily established directly in special cases. For a general result see Luenberger (1969, p. 51 - 52). Appendix B.10 of Bickel & Doksum (2015) provides a compact introduction to Hilbert space theory.

Two 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,

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

Rescaling and adding yields

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

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

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

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

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

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

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

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

References

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

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

