Optimal prediction theory *

Jean-Marie Dufour †
McGill University

First version: May 1999
Revised: January 2008
This version: January 25, 2008
Compiled: January 25, 2008, 2:24pm

* This work was supported by the William Dow Chair in Political Economy (McGill University), the Canada Research Chair Program (Chair in Econometrics, Université de Montréal), the Bank of Canada (Research Fellowship), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), the Institut de finance mathématique de Montréal (IFM2), the Canadian Network of Centres of Excellence [program on *Mathematics of Information Technology and Complex Systems* (MITACS)], the Natural Sciences and Engineering Research Council of Canada, the Social Sciences and Humanities Research Council of Canada, the Fonds de recherche sur la société et la culture (Québec), and the Fonds de recherche sur la nature et les technologies (Québec), and a Killam Fellowship (Canada Council for the Arts).2

† William Dow Professor of Economics, McGill University, Centre interuniversitaire de recherche en analyse des organisations (CIRANO), and Centre interuniversitaire de recherche en économie quantitative (CIREQ). Mailing address: Department of Economics, McGill University, Leacock Building, Room 519, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 8879; FAX: (1) 514 398 4938; e-mail: jean-marie.dufour@mcgill.ca . Web page: http://www.jeanmariedufour.com

Contents

List of Definitions, Propositions and Theorems

1. Optimal mean square prediction

Let $Y,\ X_1,\ \ldots\ ,\ X_k$ be real random variables in L^2, and $X = (X_1,\ \ldots\ ,\ X_k)'$. We wish to find a function

$$g(X) = g(X_1,\ \ldots\ ,\ X_k)$$

such that

$$\mathsf{E}\big([Y - g(X)]^2 \big) \text{ is minimal.}$$

Given the mean square criterion, we also restrict $g(X)$ to be in L^2 :

$$\mathsf{E}\left[g(X)^2\right] < \infty\,.$$

Then it is easy to see that the optimal solution to this problem is

$$g(X) = M(X)$$

where

$$M(X) = \mathsf{E}(Y \mid X)\ .$$

In general, $M(X)$ is a nonlinear function of X. The optimality of $M(X)$ can easily be shown on observing that :

$$\begin{aligned}
\mathsf{E}\left\{[Y - g(X)]^2\right\} &= \mathsf{E}\left\{[Y - \mathsf{E}(Y \mid X) + \mathsf{E}(Y \mid X) - g(X)]^2\right\} \\
&= \mathsf{E}\big\{[Y - \mathsf{E}(Y \mid X)]^2 + [\,\mathsf{E}(Y \mid X) - g(X)]^2 \\
&\quad + 2\,[Y - \mathsf{E}(Y \mid X)]\,[\,\mathsf{E}(Y \mid X) - g(X)]\big\} \\
&= \mathsf{E}\left\{[Y - \mathsf{E}(Y \mid X)]^2\right\} + \mathsf{E}\left\{[\,\mathsf{E}(Y \mid X) - g(X)]^2\right\} \\
&\quad + 2E\left\{[\,\mathsf{E}(Y \mid X) - g(X)]\ \mathsf{E}[Y - \mathsf{E}(Y \mid X) \mid X]\right\} \\
&= \mathsf{E}\left\{[Y - \mathsf{E}(Y \mid X)]^2\right\} + \mathsf{E}\left\{[\,\mathsf{E}(Y \mid X) - g(X)]^2\right\}
\end{aligned}$$

from which it follows that the optimal solution is

$$g(X) = \mathsf{E}(Y \mid X)\ .$$

The set of random variables

$$M_0 = \left\{Z : Z = g(X) \text{ is a random variable and } \mathsf{E}\left(Z^2\right) < \infty\right\}$$

is a closed subspace of L^2. $M(X) = \mathsf{E}(Y \mid X)$ can be interpreted as the projection of Y on M_0 :

$$\mathsf{E}(Y \mid X) = P_{M_0} Y.$$

2. Properties of conditional expectations

Let

$$\begin{aligned} Y &= (Y_1, \dots, Y_q)' , \\ Z &= (Z_1, \dots, Z_q)' , \\ X &= (X_1, \dots, X_k) \end{aligned}$$

be random vectors whose components are all in L^2. By definition,

$$\mathsf{E}(Y \mid X) = \begin{bmatrix} \mathsf{E}(Y_1 \mid X) \\ \mathsf{E}(Y_2 \mid X) \\ \vdots \\ \mathsf{E}(Y_q \mid X) \end{bmatrix}$$

and similarly for $\mathsf{E}(Z \mid X)$.

Let $L^2(X)$ be the set of random variables W such that $W = g(X)$ and $\mathsf{E}(W^2) < \infty$.

2.1 Proposition LINEARITY. *Let A an $m \times q$ fixed matrix and b an $m \times 1$ fixed vector. Then*

$$\begin{aligned} \mathsf{E}(AY + b \mid X) &= AE(Y \mid X) + b , \\ \mathsf{E}(Y + Z \mid X) &= \mathsf{E}(Y \mid X) + \mathsf{E}(Z \mid X) . \end{aligned}$$

2.2 Proposition POSITIVITY. *If $Y_i \geq 0$, for $i = 1, \dots, q$, then*

$$\mathsf{E}(Y_i \mid X) \geq 0 , \quad \textit{for} \quad i = 1, \dots, q .$$

2.3 Proposition MONOTONICITY. *If $Y_i \geq Z_i$, for $i = 1, \dots, q$, then*

$$\mathsf{E}(Y_i \mid X) \geq \mathsf{E}(Z_i \mid X) , \quad \textit{for} \quad i = 1, \dots, q .$$

2.4 Proposition INVARIANCE.

$$\begin{aligned} \mathsf{E}(Y \mid X) = Y \quad &\Leftrightarrow \quad Y \textit{ is a function of } X \\ &\Leftrightarrow \quad \textit{there is a function } g(x) \textit{ such that } Y = g(X) \\ &\qquad \textit{with probability } 1. \end{aligned}$$

2.5 Proposition ORTHOGONALITY. *If $g_1(X) \in L^2$ and $g_2(Y) \in L^2$, then*

$$\mathsf{E}\{g_1(X)[g_2(Y) - \mathsf{E}(g_2(Y) \mid X)]\} = 0 .$$

2.6 Proposition ITERATED CONDITIONINGS LAW. *If W is a random vector such that*

$$L^2(W) \subseteq L^2(X) \ ,$$

then

$$\begin{aligned} \mathsf{E}\left[\mathsf{E}(Y \mid X) \mid W\right] &= \mathsf{E}\left[\mathsf{E}(Y \mid W) \mid X\right] \\ &= \mathsf{E}(Y \mid W) \ . \end{aligned}$$

2.7 Proposition MEAN SQUARE OPTIMALITY.

$$\mathsf{E}\left[\left(Y_i - \mathsf{E}(Y_i \mid X)\right)^2\right] = \min_{g_i(X) \in L^2(X)} \mathsf{E}\left[\left(Y_i - g_i(X)\right)^2\right], \ i = 1, \ \ldots, q \, .$$

2.8 Proposition CHARACTERIZATION OF OPTIMALITY BY ORTHOGONALITY. *For any $i = 1, \ \ldots, \ q$,*

$$h_i(X) = \mathsf{E}(Y_i \mid X) \Leftrightarrow \mathsf{E}\left[g(X)\left(Y_i - h_i(X)\right)\right] = 0 \ , \ \forall g(X) \in L^2(X) \ .$$

2.9 Definition CONDITIONAL COVARIANCE. *The conditional covariance matrix of Y given X is the matrix*

$$\mathsf{V}(Y \mid X) = \mathsf{E}\left[\left(Y - \mathsf{E}(Y \mid X)\right)\left(Y - \mathsf{E}(Y \mid X)\right)' \mid X\right] \ .$$

If we define

$$\varepsilon(X) = Y - \mathsf{E}(Y \mid X) \ ,$$

we see easily that

$$\mathsf{V}\left[\varepsilon(X)\right] = \mathsf{E}\left[\mathsf{V}(Y \mid X)\right] \ .$$

We can then write

$$Y = \mathsf{E}(Y \mid X) + \varepsilon(X)$$

where $\mathsf{E}(Y \mid X)$ and $\varepsilon(X)$ are uncorrelated.

2.10 Proposition VARIANCE DECOMPOSITION.

$$\begin{aligned} \mathsf{V}(Y) &= \mathsf{V}\left[\mathsf{E}(Y \mid X)\right] + \mathsf{V}\left[\varepsilon(X)\right] \\ &= \mathsf{V}\left[\mathsf{E}(Y \mid X)\right] + \mathsf{E}\left[\mathsf{V}(Y \mid X)\right] \ . \end{aligned}$$

3. Linear regression

Consider again the setup of Section 1. We now study the problem of finding a function of the form

$$\begin{aligned} L(X) &= b_0 + b_1 X_1 + \cdots + b_k X_k \\ &= \sum_{i=0}^{k} b_i X_i = b'x \end{aligned}$$

where

$$X_0 = 1, \;\; b = (b_0, b_1, \ldots, b_k)' \tag{3.1}$$

$$x = (X_0, X_1, \ldots, X_k)', \tag{3.2}$$

such that the mean square prediction error

$$\mathsf{E}\left\{[Y - L(X)]^2\right\} = \mathsf{E}\left[(Y - b'x)^2\right]$$

is minimal. In other words, we wish to minimize (with respect to b) the function

$$\begin{aligned} S(b) &= \mathsf{E}\left\{[Y - b'x]^2\right\} \\ &= \mathsf{E}\left(Y^2\right) - 2b'\mathsf{E}(xY) + b'\mathsf{E}(xx')\, b \, . \end{aligned}$$

It is easy to see that the optimal value of b must satisfy the equation

$$\mathsf{E}[x(Y - b'x)] = 0$$

or

$$\mathsf{E}(xx')\, b = \mathsf{E}(xY) \; .$$

If we write

$$b = \begin{pmatrix} \beta_0 \\ \gamma \end{pmatrix}, \;\; \gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_k \end{pmatrix}, \;\; X = \begin{pmatrix} X_1 \\ \vdots \\ X_k \end{pmatrix} \; ,$$

we see that

$$\begin{bmatrix} 1 & \mathsf{E}(X)' \\ \mathsf{E}(X) & \mathsf{E}(XX') \end{bmatrix} \begin{bmatrix} \beta_0 \\ \gamma \end{bmatrix} = \begin{bmatrix} \mathsf{E}(Y) \\ \mathsf{E}(XY) \end{bmatrix} \; ,$$

hence

$$\begin{aligned} \beta_0 + \mathsf{E}(X)'\gamma &= \mathsf{E}(Y) &(3.3)\\ \mathsf{E}(Y)\beta_0 + \mathsf{E}(XX')\gamma &= \mathsf{E}(XY) &(3.4)\end{aligned}$$

and

$$\beta_0 = \mathsf{E}(Y) - \mathsf{E}(X)'\gamma \,.$$

Further, by the basic properties of the expectation operator,

$$\begin{aligned} \mathsf{E}(XX') &= \mathsf{V}(X) + \mathsf{E}(X)\,\mathsf{E}(X)' \,,\\ \mathsf{E}(XY) &= \mathsf{C}(X,\,Y) + \mathsf{E}(X)\,\mathsf{E}(Y)\end{aligned}$$

where

$$\begin{aligned} \mathsf{V}(X) &= \mathsf{E}\left\{\mathsf{E}\left[X - \mathsf{E}(X)\right]\left[X - \mathsf{E}(X)\right]'\right\} \,, &(3.5)\\ \mathsf{C}(X,\,Y) &= \mathsf{E}\left\{\left[X - \mathsf{E}(X)\right]\left[Y - \mathsf{E}(Y)\right]'\right\} \,. &(3.6)\end{aligned}$$

By the equations (3.3)-(3.6), we then see easily that

$$\begin{aligned} \mathsf{E}(X)\beta_0 + \mathsf{E}(X)\,\mathsf{E}(X)'\gamma &= \mathsf{E}(X)\,\mathsf{E}(Y) \,,\\ \mathsf{E}(X)\beta_0 + \mathsf{V}(X)\gamma + \mathsf{E}(X)\,\mathsf{E}(X)'\gamma &= \mathsf{C}(X,\,Y) + \mathsf{E}(X)\,\mathsf{E}(Y)\end{aligned}$$

hence

$$\mathsf{V}(X)\gamma = \mathsf{C}(X,\,Y) \,.$$

Thus,

$$\begin{aligned} \beta_0 &= \mathsf{E}(Y) - \mathsf{E}(X)'\gamma \,, &(3.7)\\ \mathsf{V}(X)\gamma &= \mathsf{C}(X,\,Y) \,. &(3.8)\end{aligned}$$

The function

$$L(X) = \beta_0 + X'\gamma$$

is called the

linear regression of X on Y

or the

affine projection of Y on X. (3.9)

We write

$$L(X) = P_L(Y \mid X) = \beta_0 + X'\gamma$$

where β_0 and γ are any solution of the normal equations:

$$\begin{aligned} \mathsf{V}(X)\,\gamma &= \mathsf{C}(X,\,Y)\ , \\ \beta_0 &= \mathsf{E}(Y) - \mathsf{E}(X)'\,\gamma\ . \end{aligned}$$

If we denote by

$$\varepsilon = Y - P_L(Y \mid X)$$

the prediction error, we see easily that:

$$\begin{aligned} \mathsf{E}(\varepsilon) &= 0\ , \\ \mathsf{C}(X,\,\varepsilon) &= 0\ . \end{aligned}$$

In the language of Hilbert space theory, we can also write

$$L(X) = P_M Y = P_L(Y \mid X)$$

where

$$M = \overline{sp}\,\{1,\,X\} = \overline{sp}\,\{1,\,X_1,\ \ldots\ ,\ X_k\}\ .$$

If

$$\det\left[\mathsf{V}(X)\right] \neq 0\ ,$$

the optimal coefficients β_0 and γ are uniquely defined :

$$\gamma = \mathsf{V}(X)^{-1}\,\mathsf{C}(X,\,Y)\,,\ \ \beta_0 = \mathsf{E}(Y) - \mathsf{E}(X)'\,\gamma\ .$$

4. Properties of the projection operator

Let

$$\begin{aligned} Y &= (Y_1,\,\ldots\,,\,Y_q)'\ , \\ Z &= (Z_1,\,\ldots\,,\,Z_q)'\ , \\ X &= (X_1,\,\ldots\,,\,X_k) \end{aligned}$$

be random vectors whose components are all in L^2. By definition,

$$\mathsf{P}_L(Y \mid X) = \begin{bmatrix} \mathsf{P}_L(Y_1 \mid X) \\ \mathsf{P}_L(Y_2 \mid X) \\ \vdots \\ \mathsf{P}_L(Y_q \mid X) \end{bmatrix}$$

We call $\mathcal{L}(X)$ the set of all linear transformations of X.

4.1 Proposition *If* $\det[V(X)] \neq 0$,

$$
\begin{aligned}
\mathsf{P}_L(Y \mid X) &= \mathsf{E}(Y) + \mathsf{C}(Y, X)\mathsf{V}(X)^{-1}(X - \mathsf{E}(X)) \\
&= [\mathsf{E}(Y) - \mathsf{C}(Y, X)\mathsf{V}(X)^{-1}\mathsf{E}(X)] + \mathsf{C}(Y, X)\mathsf{V}(X)^{-1}X\,. \qquad (4.1)
\end{aligned}
$$

4.2 Proposition LINEARITY. *Let A and B be two fixed matrices of dimensions $n \times q$ and $1 \times n$ respectively. Then*

$$
\begin{aligned}
\mathsf{P}_L(AY \mid X) &= A\,\mathsf{P}_L(Y \mid X)\,, && (4.2)\\
\mathsf{P}_L(YB \mid X) &= \mathsf{P}_L(Y \mid X)\,B\,, && (4.3)\\
\mathsf{P}_L(Y + Z \mid X) &= \mathsf{P}_L(Y \mid X) + \mathsf{P}_L(Z \mid X)\,. && (4.4)
\end{aligned}
$$

4.3 Proposition INVARIANCE.

$$
\begin{aligned}
\mathsf{P}_L(Y \mid X) = Y \quad &\Leftrightarrow \quad Y \text{ is a linear function of } X \\
&\Leftrightarrow \quad Y = AX + b \text{ with probability 1}
\end{aligned}
$$

where A and b are fixed matrices.

Note that

4.4 Proposition ORTHOGONALITY. *If* $\varepsilon_L(X) = Y - \mathsf{P}_L(Y \mid X)$,

$$
\mathsf{C}\big(\varepsilon_L(X), X\big) = 0\,. \qquad (4.5)
$$

4.5 Proposition LAW OF ITERATED PROJECTIONS. *If W is a random vector such that*

$$
\mathcal{L}(W) \subseteq \mathcal{L}(X)\,,
$$

then

$$
\begin{aligned}
\mathsf{P}_L[\mathsf{P}_L(Y \mid X) \mid W] &= \mathsf{P}_L[\mathsf{P}_L(Y \mid W) \mid X] \\
&= \mathsf{P}_L(Y \mid W)\,.
\end{aligned}
$$

In particular, if $X = W$,

$$
\mathsf{P}_L[\mathsf{P}_L(Y \mid X) \mid X] = \mathsf{P}_L(Y \mid X) \qquad (4.6)
$$

4.6 Proposition PROJECTION ON UNCORRELATED VECTORS. *If X and W are uncorrelated, then*

$$\mathsf{P}_L(Y \mid X, W) = \mathsf{P}_L(Y \mid X) + \mathsf{P}_L(Y \mid X) - \mathsf{E}(Y). \tag{4.7}$$

4.7 Proposition FRISCH-WAUGH THEOREM.

$$\mathsf{P}_L(Y \mid X, W) = \mathsf{P}_L(Y \mid X) + \mathsf{P}_L\big(Y \mid W - \mathsf{P}_L(W \mid X)\big) - \mathsf{E}(Y). \tag{4.8}$$

5. Prediction based on an infinite number of variables

It is possible to generalized the concept of projection to the case where X contains an infinite number of variables

$$X \equiv \overline{X}_{t-1} = (X_{t-1}, X_{t-2}, \ldots) = (X_{t-k} : k \geq 1). \tag{5.1}$$

Let Y a scalar random variable. If we consider a potentially infinite set I of random variables such that the variables in I have finite second order moments, we can define the set $\mathcal{L}^2(I)$ of linear transformations of a finite set of variables from I. Then we can define $\mathcal{H}(I)$ the smallest set of random variables in L^2 such that $\mathcal{H}(I)$ is closed, i.e. $\mathcal{H}(I)$ satisfies the following condition: if

$$\{Y_n : n \in \mathbb{Z}\} \subseteq \mathcal{H}(I) \tag{5.2}$$

then

$$\mathsf{E}\left[(Y_m - Y_n)^2\right] \longrightarrow 0 \text{ when } m, n \longrightarrow \infty \tag{5.3}$$

entails

$$\text{there exists } Y \in \mathcal{H}(I) \text{ such that } \mathsf{E}\big[\,(Y_n - Y)^2\,\big] \underset{n\to\infty}{\longrightarrow} 0\,. \tag{5.4}$$

We call $\mathcal{H}(I)$ the “Hilbert space” generated by I.

5.1 Theorem *There exists a unique random variable $\widehat{Y}_{|t-1} \equiv \mathsf{P}_L(Y \mid I)$ such that*

$$\mathsf{E}\big[\big(Y - \widehat{Y}_{|t-1}\big)^2\big] = \inf_{Z \in \mathcal{H}(I)} \mathsf{E}\big[\big(Y - Z\big)^2\big]. \tag{5.5}$$

The operator $\mathsf{P}_L(Y \mid I)$ enjoys properties sated in Propositions **4.2** to **4.7**.

6. Bibliographic notes

On the properties of conditional expectations, see Gouriéroux and Monfort (1995, Appendix B) and Williams (1991).

References

GOURIÉROUX, C., AND A. MONFORT (1995): *Statistics and Econometric Models, Volumes One and Two*. Cambridge University Press, Cambridge, U.K.

WILLIAMS, D. (1991): *Probability with Martingales*. Cambridge University Press, Cambridge, U.K.