# Optimal prediction theory *

Jean-Marie Dufour †
McGill University

First version: May 1999
Revised: March 2022
This version: March 26, 2022
Compiled: March 26, 2022, 11:54

*This work was supported by the William Dow Chair in Political Economy (McGill University), the Bank of Canada (Research Fellowship), the Toulouse School of Economics (Pierre-de-Fermat Chair of excellence), the Universitad Carlos III de Madrid (Banco Santander de Madrid Chair of excellence), a Guggenheim Fellowship, a Konrad-Adenauer Fellowship (Alexander-von-Humboldt Foundation, Germany), 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, and the Fonds de recherche sur la société et la culture (Québec).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 414, 855 Sherbrooke Street West, Montréal, Québec H3A 2T7, Canada. TEL: (1) 514 398 6071; FAX: (1) 514 398 4800; e-mail: `jean-marie.dufour@mcgill.ca`. Web page: `http://www.jeanmariedufour.com`

# Contents



# List of Definitions, Assumptions, 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

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

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 :

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

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

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

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

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

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 $\mathsf{E}\left(W^2\right) < \infty$.

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

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

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

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

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

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

**Proposition 2.4** INVARIANCE.

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

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

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

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

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

*then*

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

**Proposition 2.7** MEAN SQUARE OPTIMALITY.

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

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

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

**Definition 2.1** CONDITIONAL COVARIANCE. *The conditional covariance matrix of $Y$ given $X$ is the matrix*

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

If we define

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

we see easily that

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

We can then write

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

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

**Proposition 2.9** VARIANCE DECOMPOSITION.

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

## 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

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

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

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

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

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

or

$$\mathsf{E}\left(xx'\right)b = \mathsf{E}(xY)\ .$$

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

$$\beta_0 + \mathsf{E}(X)'\gamma = \mathsf{E}(Y) \tag{3.3}$$

$$\mathsf{E}(Y)\beta_0 + \mathsf{E}\left(XX'\right)\gamma = \mathsf{E}(XY) \tag{3.4}$$

and

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

Further, by the basic properties of the expectation operator,

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

where

$$\mathsf{V}(X) = \mathsf{E}\left\{\mathsf{E}[X - \mathsf{E}(X)][X - \mathsf{E}(X)]'\right\}, \tag{3.5}$$

$$\mathsf{C}(X, Y) = \mathsf{E}\left\{[X - \mathsf{E}(X)][Y - \mathsf{E}(Y)]'\right\}. \tag{3.6}$$

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 $\beta_0$ and $\gamma$ 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

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

the optimal coefficients $\beta_0$ and $\gamma$ 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 $\mathscr{L}(X)$ the set of all linear transformations of $X$.

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

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

*and*

$$\begin{aligned} \varepsilon_L(X) \;\; : &= Y - \mathsf{P}_L(Y \mid X) \\ &= [Y - \mathsf{E}(Y)] - \mathsf{C}(Y,X)\mathsf{V}(X)^{-1}[X - \mathsf{E}(X)]\,. \end{aligned} \tag{4.2}$$

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

$$\mathsf{P}_L(AY \mid X) = A\,\mathsf{P}_L(Y \mid X)\,, \tag{4.3}$$

$$\mathsf{P}_L(YB \mid X) = \mathsf{P}_L(Y \mid X)\,B\,, \tag{4.4}$$

$$\mathsf{P}_L(Y + Z \mid X) = \mathsf{P}_L(Y \mid X) + \mathsf{P}_L(Z \mid X)\,. \tag{4.5}$$

**Proposition 4.3** INVARIANCE.

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

*where A and b are fixed matrices.*

Note that

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

$$\mathsf{C}\big(\varepsilon_L(X), X\big) = 0\,. \tag{4.6}$$

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

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

*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* $X = W$,

$$\mathsf{P}_L[\mathsf{P}_L(Y \mid X) \mid X] = \mathsf{P}_L(Y \mid X) \tag{4.7}$$

**Proposition 4.6** FRISCH-WAUGH THEOREM.

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

## 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 $\mathscr{L}^2(I)$ of linear transformations of a finite set of variables from $I$. Then we can define $\mathscr{H}(I)$ the smallest set of random variables in $L^2$ such that $\mathscr{H}(I)$ is closed, i.e. $\mathscr{H}(I)$ satisfies the following condition: if

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

then

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

entails

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

We call $\mathscr{H}(I)$ the "Hilbert space" generated by $I$.

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

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

The operator $\mathsf{P}_L(Y \,|\, I)$ enjoys properties sated in Propositions 4.2 to 4.6.

## 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., Translated by Quang Vuong.

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