Best Predictors and Linear Regression

Chi-Chao Hung [*]

last updated: March 17, 2025

Contents

[*]This note borrows heavily from Hansen (2022). I was also supported by ChatGPT. All errors are mine. If you find any errors or have any suggestions, please contact me via email at r13323021@ntu.edu.tw.

1 Notations

Let us begin by defining some notations that will be used throughout this note.

Definition 1.1 (Random Vector)**.** A **random vector** X is a collection of random variables arranged as a vector. Formally, an k-dimensional random vector is written as:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}$$

where each X_i is a random variable.

The **expectation** (mean) of X is a vector:

$$\mathrm{E}[X] = \begin{bmatrix} \mathrm{E}[X_1] \\ \mathrm{E}[X_2] \\ \vdots \\ \mathrm{E}[X_k] \end{bmatrix}.$$

Definition 1.2 (Inner Product)**.** Consider a **random vector** X and a corresponding **coefficient vector** β. The inner product $X'\beta$ is given by:

$$\begin{aligned} X'\beta &= \begin{bmatrix} X_1 & X_2 & \dots & X_k \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \\ &= X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k. \end{aligned}$$

We use the notation X' to represent the **transpose** of X, converting it from a column vector to a row vector. This transformation allows us to express the inner product as a *matrix multiplication*.
Squaring the inner product gives a **quadratic form**:

$$(X'\beta)^2 = (X_1\beta_1 + X_2\beta_2 + \cdots + X_k\beta_k)^2 .$$

Notice that using matrix notation, $X'\beta$ and $\beta'X$ are the same. Thus, squaring $X'\beta$ gives:

$$(X'\beta)^2 = \underset{(1\times 1)}{(\beta'X)}\underset{(1\times 1)}{(X'\beta)} = \underset{(1\times k)}{\beta'}\ \underset{(k\times k)}{(XX')}\ \underset{(k\times 1)}{\beta} .$$

Here, XX' is an outer product matrix:

$$XX' = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} \begin{bmatrix} X_1 & X_2 & \dots & X_k \end{bmatrix} = \begin{bmatrix} X_1X_1 & X_1X_2 & \dots & X_1X_k \\ X_2X_1 & X_2X_2 & \dots & X_2X_k \\ \vdots & \vdots & \ddots & \vdots \\ X_kX_1 & X_kX_2 & \dots & X_kX_k \end{bmatrix}.$$

Thus, XX' is an $k \times k$ *symmetric matrix* containing all possible pairwise products of the components of X. The **expectation** of XX' is a matrix:

$$\mathbf{Q}_{XX} := \mathrm{E}[XX'] = \begin{bmatrix} \mathrm{E}[X_1X_1] & \mathrm{E}[X_1X_2] & \dots & \mathrm{E}[X_1X_k] \\ \mathrm{E}[X_2X_1] & \mathrm{E}[X_2X_2] & \dots & \mathrm{E}[X_2X_k] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[X_kX_1] & \mathrm{E}[X_kX_2] & \dots & \mathrm{E}[X_kX_k] \end{bmatrix}.$$

Let $\tilde{X}$ be the de-meaned version of X:

$$\tilde{X} = X - \mathrm{E}[X].$$

Definition 1.3 (Covariance Matrix). The **covariance matrix** of X is defined as:

$$\begin{aligned} \mathrm{Var}[X] &= \mathrm{E}[\tilde{X}\tilde{X}'] = \mathrm{E}[(X - \mathrm{E}[X])(X - \mathrm{E}[X])'] \\ &= \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] & \dots & \mathrm{Cov}[X_1, X_k] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] & \dots & \mathrm{Cov}[X_2, X_k] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_k, X_1] & \mathrm{Cov}[X_k, X_2] & \dots & \mathrm{Var}[X_k] \end{bmatrix}. \end{aligned}$$

Finally, let Y be a random variable. XY is a random vector.

$$\underset{(k\times1)}{X}\underset{(1\times1)}{Y} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix} Y = \begin{bmatrix} X_1Y \\ X_2Y \\ \vdots \\ X_kY \end{bmatrix}.$$

The **expectation** of XY is a vector:

$$\mathbf{Q}_{XY} := \mathrm{E}[XY] = \begin{bmatrix} \mathrm{E}[X_1Y] \\ \mathrm{E}[X_2Y] \\ \vdots \\ \mathrm{E}[X_kY] \end{bmatrix}.$$

2 Best Predictor

Let X be a random vector of student characteristics, such as **age**, **past GPA**, and **the number of hours spent studying**. Professor Chen might be interested in *predicting* the midterm exam score Y of a student based on these characteristics.

Our goal is to find a **predictor** for Y. A predictor is a function $g(X)$ that takes the observed characteristics X and provides a prediction of Y. For example, we can predict that *"all students studying less then 10 hours would get a score of 50, and all students studying more than 10 hours would get a score of 80."* This is a very simple predictor.

Some questions naturally arise:

- How do we measure the quality of a predictor?
- What is the best predictor of Y?
- What are the properties of the best predictor?

It turns out that the conditional expectation function (CEF) is the best predictor of Y. So let us discuss some properties of the CEF functions.

> **Remark 1.** Here we are studying the **population**. This means we assume that the joint distribution of X and Y is known. We are not fitting a line to any data points, nor are we estimating the relationship between X and Y. These topics will be covered in Chapters 3 and 4 of Hansen (2022).

2.1 CEF Error

Let Y be a random variable, and let X be a set of observed variables. The **CEF error** is the deviation of Y from its conditional expectation:

$$e := Y - \mathrm{E}[Y \mid X].$$

This error term has several important properties:[1]

Property 2.1. The mean of CEF error is 0.

$$\mathrm{E}[e] = \mathrm{E}\left[Y - \mathrm{E}[Y \mid X]\right] = \mathrm{E}[Y] - \mathrm{E}[\mathrm{E}[Y \mid X]] = 0.$$

Property 2.2 (Mean Independence)**.** The expectation of the CEF error given X is zero:

$$\mathrm{E}[e \mid X] = \mathrm{E}[Y \mid X] - \mathrm{E}[Y \mid X] = 0.$$

[1] An implicit assumption is that $\mathrm{E}[|Y|] < \infty$. See Theorem 2.4 in Hansen (2022) for more details.

This implies that, on average, the error does not systematically deviate from zero for any given X.

Property 2.3. For any function $h(x)$ such that $\mathrm{E}[h(X)e]$ exists,

$$\mathrm{E}[h(X)e] = 0.$$

A special case is when $h(X) = X$, which gives: $\mathrm{E}[Xe] = 0$.

Notice that given property 2.1, $\mathrm{Cov}(h(X), e) = \mathrm{E}[h(X)e]$. So this property also implies that "the CEF error is **uncorrelated** with any function of X."

Proof.

$$\begin{aligned} \mathrm{E}[h(X)e] &= \mathrm{E}[\mathrm{E}[h(X)e \mid X]] = \mathrm{E}[h(X)\,\mathrm{E}[e \mid X]] \\ &= \mathrm{E}[h(X)0] = 0 \end{aligned}$$

□

2.2 CEF as the Best Predictor

Now let us return to the question of evaluating the quality of a predictor.

Definition 2.1 (Mean Squared Error)**.** The mean squared error (MSE) of a predictor $g(X)$ is given by:

$$\mathrm{MSE}(g) = \mathrm{E}\left[(Y - g(X))^2\right].$$

A **best predictor** is a function of X that *minimizes* the MSE. In fact, the CEF, $\mathrm{E}[Y \mid X]$, is the optimal predictor of Y in this sense.

Theorem 2.1 (Conditional Mean as Best Predictor)**.** If the second moment of Y exist[2], then for any predictor $g(X)$ the following holds:

$$\mathrm{E}\left[(Y - m(X))^2\right] \leq \mathrm{E}\left[(Y - g(X))^2\right],$$

where $m(X) := \mathrm{E}[Y \mid X]$ is the CEF.

That is, the CEF $\mathrm{E}[Y \mid X]$ minimizes the MSE among all predictors that are functions of X.

[2] $\mathrm{E}[Y^2] < \infty$. See Theorem 2.7 in Hansen (2022).

Proof.

$$\begin{aligned}
&\mathrm{E}\left[(Y-g(X))^2\right] && \text{(MSE of } g(X)\text{)}\\
&=\mathrm{E}\left[(Y-m(X)+m(X)-g(X))^2\right] && \text{(Add and Subtract)}\\
&=\mathrm{E}\left[(e+m(X)-g(X))^2\right] && \text{(Definition of CEF error)}\\
&=\mathrm{E}\left[e^2+2e(m(X)-g(X))+(m(X)-g(X))^2\right] &&\\
&=\mathrm{E}\left[e^2\right]+2\,\mathrm{E}\left[e(m(X)-g(X))\right]+\mathrm{E}\left[(m(X)-g(X))^2\right] &&\\
&=\mathrm{E}\left[e^2\right]+\mathrm{E}\left[(m(X)-g(X))^2\right] && \text{(Property 2.3)}\\
&\geq \mathrm{E}\left[e^2\right]=\mathrm{E}\left[(Y-m(X))^2\right] && \text{(MSE of } m(X)\text{)}
\end{aligned}$$

□

Remark 2. Notice that if X is a constant, then the CEF is simply $\mathrm{E}[Y]$. The best predictor of Y is then the **unconditonal mean**. $Y = \mathrm{E}[Y] + e$ is called the intercept-only model.

3 Best Linear Predictor

3.1 Deriving the Best Linear Predictor

Although we have proven that the CEF is the optimal predictor of Y Y, its functional form is generally unknown. Therefore, it is often useful to restrict our focus to a class of predictors with a simpler structure. Here, we consider one of the simplest cases: linear predictors.

The **linear predictor** of Y given X takes the form

$$\mathcal{P}(Y \mid X) = X'\beta,$$

where

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_{k-1} \\ 1 \end{bmatrix}$$

and β is an $k \times 1$ coefficient vector.

Definition 3.1 (Best Linear Predictor)**.** A function $\mathcal{P}(Y \mid X) = X'\beta$ is the best linear predictor (BLP) of Y given X if[3]

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

[3]See Definition 2.5 in Hansen (2022).

Now let us derive the BLP. First define

$$S(\beta) := \mathrm{E}[e^2] = \mathrm{E}\left[(Y - X'\beta)^2\right].$$

Expanding the square:

$$\begin{aligned} S(\beta) &= \mathrm{E}\left[Y^2 - 2YX'\beta + \beta' X X' \beta\right] \\ &= \mathrm{E}[Y^2] - 2\beta' \mathrm{E}[XY] + \beta' \mathrm{E}[XX']\beta. \end{aligned}$$

Taking the derivative with respect to β,[4] the first-order condition is

$$\underset{(k\times 1)}{\frac{\partial}{\partial \beta} S(\beta)} = -2\underset{(k\times 1)}{\mathrm{E}[XY]} + 2\underset{(k\times k)}{\mathrm{E}[XX']}\underset{(k\times 1)}{\beta} = \underset{(k\times 1)}{0}.$$

Solving for β:

$$\underset{(k\times k)}{\mathrm{E}[XX']}\underset{(k\times 1)}{\beta} = \underset{(k\times 1)}{\mathrm{E}[XY]}$$

If $\mathrm{E}[XX']$ is invertible, the solution for the best linear predictor is:

$$\beta = \underset{(k\times k)}{\mathrm{E}[XX]^{-1}}\underset{(k\times 1)}{\mathrm{E}[XY]}.$$

Using the notation $\mathbf{Q}_{XX} = \mathrm{E}[XX']$ and $\mathbf{Q}_{XY} = \mathrm{E}[XY]$, we rewrite:

$$\beta = \mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY}.$$

The solution $X'\beta = X'(\mathrm{E}[XX'])^{-1}\mathrm{E}[XY]$ is often called the **linear projection** of Y on X, and β is called the **projection coefficient**.[5] Some authors refer to the BLP as the *population* regression.(Angrist and Pischke, 2009)

3.2 Univariate Case

To illustrate the BLP, let us consider the univariate case where X_1 is the only random variable included.

$$X = \begin{bmatrix} 1 \\ X_1 \end{bmatrix}$$

The BLP is

$$\mathcal{P}(Y \mid X) = X'\beta$$

[4]See Appendix for technical details.

[5]Loosely speaking, "linear projection" means to find the random variable linear in X that has the smallest distance to Y. Here MSE can be seen as a measure of distances between two random variables.

where

$$\beta = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix} = \begin{bmatrix}1 & \mathrm{E}[X_1] \\ \mathrm{E}[X_1] & \mathrm{E}[X_1^2]\end{bmatrix}^{-1} \begin{bmatrix}\mathrm{E}[Y] \\ \mathrm{E}[X_1 Y].\end{bmatrix}$$

Since

$$\begin{bmatrix}1 & \mathrm{E}[X_1] \\ \mathrm{E}[X_1] & \mathrm{E}[X_1^2]\end{bmatrix}^{-1} = \frac{1}{\mathrm{E}[X_1^2] - \mathrm{E}[X_1]^2} \begin{bmatrix}\mathrm{E}[X_1^2] & -\mathrm{E}[X_1] \\ -\mathrm{E}[X_1] & 1\end{bmatrix}$$

we can derive the projection coefficients β_1 and β_0.

$$\beta_1 = \frac{(\mathrm{E}[X_1 Y] - \mathrm{E}[X_1]\,\mathrm{E}[Y])}{\mathrm{E}[X_1^2] - \mathrm{E}[X_1]^2} = \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}$$

and

$$\begin{aligned}\beta_0 &= \frac{\mathrm{E}[Y]\,\mathrm{E}[X_1^2] - \mathrm{E}[X_1 Y]\,\mathrm{E}[X_1]}{\mathrm{Var}(X_1)} \\ &= \mathrm{E}[Y] - \frac{\mathrm{Cov}(X_1, Y)}{\mathrm{Var}(X_1)}\,\mathrm{E}[X_1] \\ &= \mathrm{E}[Y] - \beta_1\,\mathrm{E}[X_1]\end{aligned}$$

The BLP of Y is then

$$\mathcal{P}(Y \mid X_1) = \beta_0 + \beta_1 X_1$$

3.3 Projection Error

The **projection error** is the deviation of Y from its linear projection:

$$e := Y - \mathcal{P}(Y \mid X).$$

This error term has several important properties.

Property 3.1. $\mathrm{E}[Xe] = 0$.

Proof.

$$\begin{aligned}\mathrm{E}[Xe] &= \mathrm{E}\left[X\left(Y - X'\beta\right)\right] \\ &= \mathrm{E}[XY] - \mathrm{E}[XX']\left(\mathrm{E}[XX']\right)^{-1}\mathrm{E}[XY] = 0\end{aligned}$$

□

This is sometimes called the **orthogonality condition**.

Property 3.2. $\mathrm{E}[e] = 0$ if X includes a constant. Notice that property 3.1 is a set of k equations:

$$\mathrm{E}[X_j e] = 0 \text{ for all j = 1,...,k.}$$

Whenever the constant 1 is included in X, it follows immediately from property 3.1 that $\mathrm{E}[e] = 0$. This also implies that $\mathrm{Cov}(X_j, e) = 0$ since $\mathrm{Cov}(X_j, e) = \mathrm{E}[X_j e] = \mathrm{E}[X_j]\,\mathrm{E}[e]$.

3.4 Linear CEF

In general, BLP is not CEF. However, if we consider the special case that we know (or assume) that CEF is linear in X, then the BLP and the CEF coincide.

Theorem 3.1 (Linear CEF)**.** If the CEF is linear in X, then the CEF is the BLP.

The proof is trivial since we already know that the CEF is the best predictor among *all* predictors. Here we present another way to understand this result.

Proof. Let $e := Y - \mathrm{E}[Y \mid X]$. Under the linear CEF assumption, we have

$$Y = \mathrm{E}[Y \mid X] + e = X'\beta + e$$

Using property 2.1 of the CEF error, we have

$$\mathrm{E}[XY] = \mathrm{E}[X(X'\beta + e)] = \mathrm{E}[XX']\beta + \mathrm{E}[Xe] = \mathrm{E}[XX']\beta.$$

As a result, $\beta = \mathrm{E}[XX']^{-1}\,\mathrm{E}[XY]$ as in the BLP. $\square$

Definition 3.2 (Linear CEF Model)**.**

$$\mathrm{E}[Y \mid X] = X'\beta \quad (1)$$
$$Y = E[Y \mid X] + e \quad (2)$$
$$\mathrm{E}[e \mid X] = 0 \quad (3)$$

Remark 3. Notice that a linear CEF is a *model assumption.* If the CEF is not linear, we say the linear CEF model is **misspecified**. However, the BLP is still the best linear approximation of the CEF.[a] This is a motivation for using the BLP.

[a] See Section 2.25 in Hansen (2022).

4 Linear Regression

4.1 What is Linear Regression?

Regression is a method that allows researchers to summarize how predictions or average values of an outcome vary across individuals defined by a set of predictors.(Gelman et al., 2021) Generally, a linear regression model takes the form

$$Y = \underbrace{X'\beta}_{\text{predictor}} + \underbrace{e}_{\text{error}}$$

In Hansen (2022), the term "linear regression" refers to the linear CEF model. In some other cases, $X'\beta$ is viewed as the best linear predictor (may not be CEF) and e is the projection error. An important thing to remember is that in both cases the error term e has no life of its own. It is always defined as the difference between the outcome and the predictor.

Table 1: Comparison of the CEF error and Projection Error

	CEF Error	Projection Error
Definition	$e := Y - \mathrm{E}[Y \mid X]$	$e := Y - \mathcal{P}(Y \mid X)$
Properties	$\mathrm{E}[e \mid X] = 0$	
	$\mathrm{E}[h(X)e] = 0$	$\mathrm{E}[Xe] = 0$
	$\mathrm{E}[e] = 0$	$\mathrm{E}[e] = 0$ when 1 is included in X

Remark 4. In older textbooks, the assumption $\mathrm{E}[e \mid X]$ is called the "exogeneity" assumption. This is a very *misleading* terminology since "exogeneity" is a causal concept while the CEF itself has no causal interpretation.

4.2 What is Linear Regression used for?

1. **Prediction**: Either linear CEF or linear projection serves as the best linear predictor.

2. **Comparison**: Linear regression can be used to compare the average values of Y across different groups defined by X. This is escpecially useful when X is binary or categorical.

3. **Exploring Assotiations**: Linear regression can be used to explore the association between Y and X.

4. **Causal Inference**: So far we said nothing about causality. In fact, linear regression is not a causal model. The linear function $Y = X'\beta$ does not tell us wether it is X causing Y or Y causing X. We will discuss the **Potential Outcome Framework** at the end this semester, which provides a framework for causal inference. Once a causal model is established, linear regression may be used as a tool.

5 More Details

5.1 Independence, Mean Independence, and Uncorrelatedness

1. X and Y are **statistically independent** if and only if their joint distribution equals the product of their marginals.

2. Y is **mean independent** of X if and only if $\mathrm{E}[Y \mid X] = \mathrm{E}[Y]$.

3. X and Y are **uncorrelated** if and only if $\mathrm{Cov}(X, Y) = 0$. If $\mathrm{E}[X]$ or $\mathrm{E}[Y]$ is zero, then $\mathrm{E}[XY] = 0$.

Theorem 5.1. If X and Y are statistically independent, then X is mean independent of Y and Y is mean independent of X. Moreover, X and Y are uncorrelated. The converse is not true.

Theorem 5.2. If X is mean independent of Y, then X and Y are uncorrelated. The converse is not true.

The proofs is left for the readers. See DiTraglia (2023) for more details.

5.2 Existence and Uniqueness of the BLP

Though we made no assumptions on the CEF when deriving the BLP, the existence of the linear projection requires some conditions.

1. $\mathrm{E}[Y^2] < \infty$

2. $\mathrm{E}\,||X||^2 < \infty$

3. $\mathbf{Q}_{XX} = \mathrm{E}[XX']$ is invertible.

The first two conditions says that both X (vector) and Y (scalar) have finite second moments. The third condition is the most important. It says that the predictors in X are not perfectly collinear. If $\mathbf{Q}_{XX}$ is not invertible, then the unique linear projection does not exist.

References

Angrist, Joshua David and Jörn-Steffen Pischke, *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton: Princeton University Press, 2009.

DiTraglia, Frank, "Why Econometrics Is Confusing Part II: The Independence Zoo," https://www.econometrics.blog/post/why-econometrics-is-confusing-part-ii-the-independence-zoo/ January 2023.

Gelman, Andrew, Jennifer Hill, and Aki Vehtari, *Regression and Other Stories* Analytical Methods for Social Research, Cambridge: Cambridge University Press, 2021.

Hansen, Bruce E., *Econometrics*, Princeton: Princeton University Press, 2022.

Acronyms

BLP best linear predictor. 6, 7
CEF conditional expectation function. 4–6
MSE mean squared error. 5