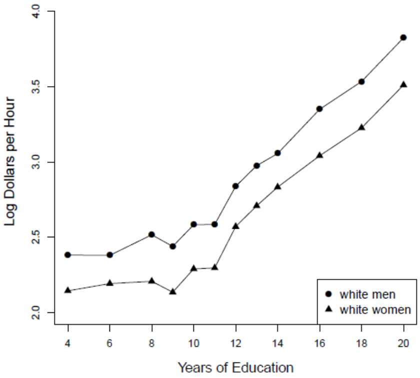
Refresh Linear Regression in Moderately High Dimensions

Alexander Quispe
The World Bank, PUCP
alexander.quispe@pucp.edu.pe

April 3, 2024

Understanding CEF in Econometrics

Figure: Expected LogWage as a Function of Years of Education



Understanding CEF in Econometrics

Key Concept

The **Conditional Expectation Function** relates the expected value of a dependent variable, conditioned on certain values of independent variables.

Example in Wage Study

The expectation of $\log(\text{wage})$ given gender, race, and education can be expressed as:

$$E[\log(\text{wage})|\text{gender, race, education}]$$

For instance,

$$E[\log(\text{wage})|\text{gender} = \text{man}, \text{race} = \text{white}, \text{education} = 12] = 2.84$$

Conditional Expectation Function (CEF)

Notation

The general notation for CEF is:

$$E[Y|X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k] = m(x_1, x_2, \ldots, x_k)$$

where Y is the dependent variable and $X_1, X_2, \ldots, X_k$ are conditioning variables.

Plot Interpretation

A plot of $\log(\text{wage})$ as a function of education shows the difference in conditional expectations between genders across education levels, highlighting the wage gap consistency.

Law of Iterated Expectations

Theorem (Simple Law of Iterated Expectations)

If $E|Y| < \infty$ *then for any random vector* X*,*

$$E[E[Y|X]] = E[Y].$$

Theorem (Law of Iterated Expectations)

If $E|Y| < \infty$ *then for any random vectors* X_1 *and* X_2*,*

$$E[E[Y|X_1, X_2]|X_1] = E[Y|X_1].$$

Law of Iterated Expectations

Theorem (Conditioning Theorem)

If $E|Y| < \infty$ *then*

$$E[g(X)Y|X] = g(X)E[Y|X].$$

If in addition $E|g(X)| < \infty$ *then*

$$E[g(X)Y] = E[g(X)E[Y|X]].$$

- These theorems establish that the expectation of the conditional expectation is the unconditional expectation.
- They are crucial for understanding the structure of regressions and the behavior of expectations in the presence of conditioning information.

Conditional Expectation Function (CEF)

Definition

We define the function that provides the expected value of Y given X as the **Conditional Expectation Function (CEF)**:

$$m(X) = E[Y|X]$$

Decomposition

Any random variable (RV) can be decomposed into CEF and a mean independent residual:

$$Y = E[Y|X] + \varepsilon = m(X) + \varepsilon$$

Mean Independence

The mean independent residual is characterized by the property:

$$E[\varepsilon|X] = E[Y - m(X)|X] = m(X) - m(X) = 0$$

Why Conditional Expectation Functions?

Function of Expected Values

The decomposition allows us to understand the preference for the CEF:

$$Y = m(X) + \varepsilon$$

where $m(X)$ is the CEF.

Error Minimization

Any function $g(X)$ results in an error $Y - g(X)$. The CEF is the function that minimizes the expected squared error:

$$m(X) = \underset{g(X)}{\arg\min}\, E[(Y - g(X))^2]$$

Optimal Estimation

$\Rightarrow$ The **CEF** provides the best possible estimate for the outcome value in the population.

How to model and estimate CEF?

1. Parametric model - Linear Model, Probit
2. Nonparametric model - Kernel Regression
3. Semiparametric model - Partially linear model

Applications of Conditional Expectation Function

Example 1: Linear Model

The linear model can be expressed as:

$$m(X;\beta) = X'\beta$$

where β is often estimated using Ordinary Least Squares (OLS).

Example 2: Probit Model for Binary Outcome

For a binary outcome Y taking values in $\{0, 1\}$, the Probit model is:

$$P[Y = 1] = E[Y|X] = m(X;\beta) = \Phi(X'\beta)$$

where Φ denotes the cumulative distribution function of the standard normal distribution.

Best Linear Predictor: In Population

1. We can compute an optimal β by solving the First Order Conditions (FOC) for the BLP problem, called Normal Equations:

$$E(Y - \beta' X)X = 0 \tag{1}$$

Any optimal $b = \beta$ satisfies the Normal Equations. Defining the regression error as

$$\varepsilon := Y - b' X \tag{2}$$

we have the simple decomposition of Y:

$$Y = \beta' X + \varepsilon, \quad EX\varepsilon = 0 \tag{3}$$

2. $\beta' X$ is the part of Y that can be predicted and
3. ε is the unexplained or residual part.

Best Linear Predictor: In Sample

1. In applications the researcher does not have access to the population in total, but observes only a sample

$$(Y_i, X_i)_1^n = ((Y_1, X_1), \ldots, (Y_n, X_n)) \tag{4}$$

2. Best Linear Prediction Problem in the Sample:

$$\min_{b \epsilon \mathbb{R}^p} \mathbb{E}_n (Y_i - b' X_i)^2 \tag{5}$$

where β is any solution to the BLP problem in the sample. The βs are called the sample regression coefficients.

3. Again from FOC we have

$$\mathbb{E}_n X_i (Y_i - X_i' \hat{\beta}) = 0 \tag{6}$$

Best Linear Predictor: In Sample

1. defining the in-sample regression error as

$$\hat{\varepsilon}_i := (Y_i - \hat{\beta}' X_i) \tag{7}$$

   we have the simple decomposition of Y:

$$Y_i = X_i' \hat{\beta} + \hat{\varepsilon}_i, \;\; \mathbb{E}_n X_i \hat{\varepsilon}_i = 0 \tag{8}$$

2. $X_i' \hat{\beta}$ is the part of Y that can be predicted and
3. $\hat{\varepsilon}_i$ is the unexplained or residual part.

Analysis of Variance (ANOVA)

POPULATION	SAMPLE
$$\begin{aligned} Y &= \beta' X + \varepsilon, \quad E\varepsilon X = 0 \\ EY^2 &= E(\beta' X)^2 + E\varepsilon^2 \\ MSE_{pop} &= E\varepsilon^2 \\ R^2_{pop} &:= \frac{E(\beta' X)^2}{EY^2} = \\ 1 - \frac{E\varepsilon^2}{EY^2} &\in [0,1] \end{aligned} \quad (9)$$	$$\begin{aligned} Y_i &= \hat\beta' X_i + \hat\varepsilon_i \\ \mathbb{E}_n Y_i^2 &= \mathbb{E}_n(\hat\beta' X_i)^2 + \mathbb{E}_n \hat\varepsilon_i^2 \\ MSE_{sample} &= \mathbb{E}_n \hat\varepsilon_i^2 \\ R^2_{sample} &:= \frac{\mathbb{E}_n(\hat\beta' X_i)^2}{\mathbb{E}_n Y_i^2} = \\ 1 - \frac{\mathbb{E}_n \hat\varepsilon_i^2}{\mathbb{E}_n Y_i^2} &\in [0,1] \end{aligned} \quad (10)$$

By law of large numbers when p/n is small and n is large:

$$\begin{gathered} \mathbb{E}_n Y_i^2 \approx EY^2, \ \ \mathbb{E}_n(\hat\beta' X_i)^2 \approx E(\beta' X)^2, \ \ \mathbb{E}_n \hat\varepsilon_i^2 \approx E\varepsilon^2 \\ R^2_{sample} \approx R^2_{pop} \ \ and \ \ MSE_{sample} \approx MSE_{pop} \end{gathered} \tag{11}$$

Overfitting: What happens when p/n is not small

When p/n is not small, the discrepancy between the in-sample and out-of-sample measures of fit can be substantial. Let´s check the next example :

-

$$\begin{gathered} X \sim N(0, I_p) \;\; and \;\; Y \sim N(0,1), \;\; \beta' X = 0, \;\; R^2_{pop} = 0 \\ if \;\; p = n, \;\; then \;\; R^2_{sample} \;\; is \;\; 1 \gg 0 \\ if \;\; p = \frac{n}{2}, \;\; then \;\; R^2_{sample} \;\; is \;\; about \;\; 0.5 \gg 0 \\ if \;\; p = \frac{n}{20}, \;\; then \;\; R^2_{sample} \;\; is \;\; about \;\; 0.05 \end{gathered} \tag{12}$$

Better measures of out-of-sample predictive ability are the "adjusted" R^2 and MSE.

$$MSE_{adjusted} = \frac{n}{n-p}\mathbb{E}_n\hat{\varepsilon}_i^2, \;\; R^2_{adjusted} := 1 - \frac{n}{n-p}\frac{\mathbb{E}_n\hat{\varepsilon}_i^2}{\mathbb{E}_n Y_i^2} \tag{13}$$

Measuring Predictive Ability by Sample Splitting

To measure out-of-sample performance: **Data splitting**. The idea can be summarized in two parts:

1. Use a random part of data, called the **training sample**, for estimating/training the prediction rule.
2. Use the other part, called the **testing sample**, to evaluate the quality of the prediction rule, recording out-of-sample mean squared error and R^2.

Generic Evaluation of Prediction Rules by Sample-Splitting

1. Randomly partition the data into training and testing samples. Suppose we use n observations for training and m for testing/validation.
2. Use the training sample to compute a prediction rule $\hat{f}(X)$, for example, $\hat{f}(X) = \beta' X$.
3. Let V denote the indexes of the observations in the test sample. Then the out-of-sample/test mean squared error is

$$MES_{test} = \frac{1}{m} \sum_{k \in V} (Y_k - \hat{f}(X_k))^2 \tag{14}$$

and the out-of-sample/test R^2 is

$$R^2_{test} = 1 - \frac{MSE_{test}}{\frac{1}{m} \sum_{k \in V} Y_k^2} \tag{15}$$

Measuring Predictive Ability by Sample Splitting

1. The Linear Model Overfitting R Notebook
2. The Linear Model Overfitting Python Notebook