Best Linear Prediction

Econometrics II

Douglas G. Steigerwald

UC Santa Barbara
How to approximate $\mathbb{E}(y|x)$?

- If $x$ is continuous, $\mathbb{E}(y|x)$ is generally unknown
  - Linear approximation $x^T \beta$
    - $\beta$ is the linear predictor (or projection) coefficient ($\beta_{lpc}$)
    - $\beta$ is not $\nabla_x \mathbb{E}(y|x)$
  - $\beta$ is identified if $\mathbb{E}(xx^T)$ is invertible

- Linear prediction error $u$ is uncorrelated with $x$ by construction
Approximate the CEF

- conditional mean $\mathbb{E}(y|x)$
  - "best" predictor (mean squared prediction error)
  - functional form generally unknown
    ➤ unless $x$ discrete (and low dimension)

- approximate $\mathbb{E}(y|x)$ with $x^T \beta$
  - linear approximation, thus a linear predictor

1. select $\beta$ to form "best" linear predictor of $y$: $\mathcal{P}(y|x)$
2. select $\beta$ to form "best" linear approximation to $\mathbb{E}(y|x)$

1 and 2 yield identical $\beta$
  - either criterion could be used to define $\beta$
  - we use 1 and refer to $x^T \beta$ as the best linear predictor
Best Linear Predictor Coefficient

1. select $\beta$ to minimize mean-square prediction error

$$S(\beta) = \mathbb{E} \left( y - x^T \beta \right)^2$$

$\beta := \beta_{lpc}$ satisfies

$$\mathbb{E} (xx^T) \beta_{lpc} = \mathbb{E} (xy) \quad \text{Solution}$$

2. select $\beta$ to minimize mean-square approximation error

$$d(\beta) = \mathbb{E}_x \left( \mathbb{E} (y|x) - x^T \beta \right)^2$$

solution satisfies

$$\mathbb{E} (xx^T) \beta_{lac} = \mathbb{E} (xy) \quad \text{Solution}$$
Identification (General)
- $\theta$ and $\theta'$ are separately identified iff $P_\theta \neq P_{\theta'} \Rightarrow \theta \neq \theta'$

Identification - Background

Identification (Best Linear Predictor)
- $\beta$ and $\beta'$ are separately identified iff $(E(xx^T))^{-1} E(xy)$ from $P_\beta$ does not equal $(E(xx^T))^{-1} E(xy)$ from $P_{\beta'}$
  - i.e. there is a unique solution to $\beta_{lpc} = (E(xx^T))^{-1} E(xy)$
  - i.e. $E(xx^T)$ is invertible
Can we uniquely determine $\beta_{lpc}$?

$$\mathbb{E} \left( xx^T \right) \beta_{lpc} = \mathbb{E} (xy)$$

- if $\mathbb{E} \left( xx^T \right)$ is invertible
  - there is a unique value of $\beta_{lpc}$ that solves the equation
    - $\beta_{lpc}$ is identified as there is a unique solution
- if $\mathbb{E} \left( xx^T \right)$ is not invertible
  - there are multiple values of $\beta_{lpc}$ that solve the equation
    - $\beta_{lpc}$ is not identified as there is not a unique solution
    - mathematically $\beta_{lpc} = \left( \mathbb{E} \left( xx^T \right) \right)^{-1} \mathbb{E} (xy)$  
      \textit{Generalized Inverse}
  - all solutions yield an equivalent best linear predictor $x^T \beta_{lpc}$
    - best linear predictor is identified
Invertibility

*Required assumption:* \( \mathbb{E} \left( xx^T \right) \) is positive definite

- for any non-zero \( \alpha \in \mathbb{R}^k \):
  \[
  \alpha^T \mathbb{E} \left( xx^T \right) \alpha = \mathbb{E} \left( \alpha^T xx^T \alpha \right) = \mathbb{E} \left( \alpha^T x \right)^2 \geq 0
  \]

- so \( \mathbb{E} \left( xx^T \right) \) is positive semi-definite by construction
- positive semi-definite matrices are invertible IFF they are positive definite
- if we assume \( \mathbb{E} \left( xx^T \right) \) is positive definite, then
  - \( \mathbb{E} \left( \alpha^T x \right)^2 > 0 \)
  - there is no non-zero \( \alpha \) for which \( \alpha^T x = 0 \)
    - implies there are no redundant variables in \( x \)
    - i.e. all columns are linearly independent
Best Linear Predictor: Error

- best linear predictor (linear projection)

\[ \mathcal{P} (y|x) = x^T \beta_{lpc} \]

- decomposition

\[ y = x^T \beta_{lpc} + u \quad u = e + \left( \mathbb{E} (y|x) - x^T \beta_{lpc} \right) \]

- choice of \( \beta_{lpc} \) implies \( \mathbb{E} (xu) = 0 \)

  - \( \mathbb{E} (xu) = \mathbb{E} \left( x \left( y - x^T \beta_{lpc} \right) \right) = \)
    \[ \mathbb{E} (xy) - \mathbb{E} (xx^T) \left( \mathbb{E} (xx^T) \right)^{-1} \mathbb{E} (xy) = 0 \]
  - error from projection onto \( x \) is orthogonal to \( x \)
Variance of $u$ equals the variance of the error from a linear projection

- **Variance of $u$**
  - $\mathbb{E} u^2 = \mathbb{E} (y - x^T \beta)^2 = \mathbb{E} y^2 - \mathbb{E} (yx^T) \beta$
  - because $\mathbb{E} (x^T \beta)^2 = \mathbb{E} (yx^T) \beta$

- **Variance of projection error**
  - projection error is defined as $\|u\| = \|y\| - \|x^T \beta\|$
  - because $y^2 = \|y\|^2$
  - $Var(u) = Var(\|u\|)$
Best Linear Predictor: Covariate Error Correlation

- $\mathbb{E}(xu) = 0$ is a set of $k$ equations, as

$$\mathbb{E}(x_j u) = 0$$

- if $x$ includes an intercept, $\mathbb{E} u = 0$

- because

$$\text{Cov}(x_j, u) = \mathbb{E}(x_j u) - \mathbb{E}x_j \cdot \mathbb{E}u$$

- covariates are uncorrelated with $u$ by construction

- for $r \geq 2$ if $\mathbb{E}|y|^r < \infty$ and $\mathbb{E}\|x\|^r < \infty$ then $\mathbb{E}|u|^r < \infty$

- if $y$ and $x$ have finite second moments then the variance of $u$ exists

- note: $\mathbb{E}|y|^r < \infty \Rightarrow \mathbb{E}|y|^s < \infty$ for all $s \leq r$ (Liapunov’s Inequality)
Linear Projection Model

linear projection model is

\[ y = x^T \beta + u \quad \mathbb{E}(xu) = 0 \quad \beta = \left( \mathbb{E}(xx^T) \right)^{-1} \mathbb{E}(xy) \]

- \( x^T \beta \) is the best linear predictor
  - not necessarily the conditional mean \( \mathbb{E}(y|x) \)

- \( \beta \) is the linear prediction coefficient
  - not the conditional mean coefficient if \( \mathbb{E}(y|x) \neq x^T \beta \)
  - not a causal (structural) effect if:
    - \( \mathbb{E}(y|x) \neq x^T \beta \)
    - \( \mathbb{E}(y|x) = x^T \beta \) but \( \nabla_x e \neq 0 \)
How Does the Linear Projection Differ from the CEF?

Example 1

- CEF of $\log(wage)$ as a function of $x$ (black and female indicators)
- discrete covariates, small number of values, compute CEF

$$\mathbb{E} (\log (wage) | x) = -.20 \ black - .24 \ female + .10 \ inter + 3.06$$

- $inter = black \cdot female$
  - 20% male race gap (black males 20% below white males)
  - 10% female race gap

- Linear Projection of $\log(wage)$ on $x$ (black and female indicators)

$$\mathcal{P} (\log(wage)|x) = -.15 \ black - .23 \ female + 3.06$$

- 15% race gap
  - average race gap across males and females
  - ignores the role of gender in race gap, even though gender is included
How Does the Linear Projection Differ from the CEF?

Example 2

CEF of white male log(wage) as a function of years of education ($ed$)

- discrete covariate with multiple values
  - could use categorical variables to compute CEF
    - large number of values leads to cumbersome estimation

approximate CEF with linear projections
Approximate CEF of Wage as a Function of Education

Approximation 1

- Linear Projection of $\log(wage)$ on $x = ed$

\[ P(\log(wage) | x) = 0.11 \, ed + 1.50 \]

- 11\% increase in mean wages for every year of education

![Graph showing projections of log(wage) onto education](image)

**Figure 2.8: Projections of log(wage) onto Education**

- works well for $ed \geq 9$, under predicts if education is lower
Approximate CEF of Wage as a Function of Education

Approximation 2: Linear Spline

- Linear Projection of $\log(wage)$ on $x = (ed, spline)$

\[
P(\log(wage) | x) = 0.02 \cdot ed + 0.10 \cdot spline + 2.30
\]

- $spline = (ed - 9) \cdot 1(ed)$
  - 2% increase in mean wages for each year of education below 9
  - 12% increase in mean wages for each year of education above 9

Figure 2.8: Projections of $\log(wage)$ onto Education
How Does the Linear Projection Differ from the CEF?

Example 3

CEF of white male (with 12 years of education) log(wage) as a function of years of experience ($ex$)

- discrete covariate with large number of values
  - approximate CEF with linear projections
- Linear Projection of log(wage) on $x = ex$
  - $P(\log(wage) | x) = 0.011 \cdot ex + 2.50$

over predicts wage for young and old
Approximate CEF of Wage as a Function of Experience

Approximation 2: Quadratic Projection

- Linear Projection of log(wage) on $x = (ex, ex^2)$

$$\mathcal{P}(\log(\text{wage}) | x) = 0.046 ex - 0.001 ex^2 + 2.30$$

- $\nabla \mathcal{P} = 0.046 - 0.001 \cdot ex$

★ captures strong downturn in mean wage for older workers
Properties of the Linear Projection Model

- **Assumption 1**
  - $\mathbb{E}y^2 < \infty \quad \mathbb{E}\|x\|^2 < \infty \quad Q_{xx} = \mathbb{E}(xx^T)$ is positive definite

- **Theorem:** Under Assumption 1
  1. $\mathbb{E}(xx^T)$ and $\mathbb{E}(xy)$ exist with finite elements
  2. The linear projection coefficient exists, is unique, and equals
     \[
     \beta = \left(\mathbb{E}(xx^T)\right)^{-1} \mathbb{E}(xy)
     \]
  3. $\mathcal{P}(y|x) = x^T \left(\mathbb{E}(xx^T)\right)^{-1} \mathbb{E}(xy)$
  4. For $u = y - x^T\beta$, $\mathbb{E}(xu) = 0$ and $\mathbb{E}(u^2) < \infty$
  5. If $x$ contains a constant, $\mathbb{E}u = 0$
  6. If $\mathbb{E}|y|^r < \infty$ and $\mathbb{E}\|x\|^r < \infty$ for $r \geq 2$, then $\mathbb{E}|u|^r < \infty$

**Proof**
Review

- How do we approximate $\mathbb{E} (y|x)$?
- $x^T \beta$

How to do you interpret $\beta$?

- the linear projection coefficient, which is not generally equal to $\nabla_x \mathbb{E} (y|x)$

What is required for identification of $\beta$?

- $\mathbb{E} (xx^T)$ is invertible

What is the correlation between $x$ and $u$?

- 0 by construction!
Best Linear Predictor Coefficient Solution

- $\beta_{lpc}$ is the value of $\beta$ that minimizes

$$S(\beta) = \mathbb{E} y^2 - 2\beta^T \mathbb{E} (xy) + \beta^T \mathbb{E} (xx^T) \beta \quad \text{Vector Calculus}$$

  - first derivative $-2\mathbb{E} (xy) + 2\mathbb{E} (xx^T) \beta$

- solution (linear projection coefficient)

$$\mathbb{E} (xx^T) \beta_{lpc} = \mathbb{E} (xy)$$

- required assumption

  - $\mathbb{E} y^2 < \infty \quad \mathbb{E} \|x\|^2 < \infty \quad \text{Euclidean Length}$

Return to Best Linear Predictor Coefficient
Best Linear Approximation Coefficient Solution

let $m(x) := \mathbb{E}(y|x)$

- $\beta_{lac}$ is the value of $\beta$ that minimizes

$$d(\beta) = \int_{\mathbb{R}^k} \left( m(x) - x^T \beta \right)^2 f_x(x) \, dx$$

- $d(\beta) = \mathbb{E}m(x)^2 - 2\beta^T \mathbb{E}(xm(x)) + \beta^T \mathbb{E}(xx^T) \beta$
  
  ▶ first derivative
  
  $$-2\mathbb{E}(xm(x)) + 2\mathbb{E}(xx^T) \beta$$

  ▶ $\mathbb{E}(xm(x)) = \mathbb{E}(x\mathbb{E}(y|x)) = \mathbb{E}(\mathbb{E}(xy|x)) = \mathbb{E}(xy)$

- solution (linear approximation coefficient)

$$\mathbb{E}(xx^T) \beta_{lac} = \mathbb{E}(xy)$$
Vector Calculus

- vector derivative: inner product
  - $(2 \times 1)$ vectors: $B$ and $C$
  - $B^T C = B_1 C_1 + B_2 C_2$
  - $\frac{\partial B^T C}{\partial B} = \begin{bmatrix} \frac{\partial B^T C}{\partial B_1} \\ \frac{\partial B^T C}{\partial B_2} \end{bmatrix} = \begin{bmatrix} C_1 \\ C_2 \end{bmatrix} = C$

- vector derivative: quadratic form
  - $(2 \times 2)$ matrix: $D$
  - $B^T DB = B_1^2 D_{11} + B_1 B_2 D_{12} + B_1 B_2 D_{21} + B_2^2 D_{22}$
  - $\frac{\partial B^T DB}{\partial B} = \begin{bmatrix} (D_{11} + D_{11}) B_1 + (D_{12} + D_{21}) B_2 \\ (D_{21} + D_{12}) B_1 + (D_{22} + D_{22}) B_2 \end{bmatrix} = (D + D^T) B$

Return to Solution
Euclidean Length

- **Pythagorean Theorem**
  - \( a^2 + b^2 = c^2 \) so the length of the hypotenuse is \( c = (a^2 + b^2)^{1/2} \)

- \( c \) is a vector of dimension 2, so for \( x \) a vector of dimension \( n \)
  - the Euclidean length (norm) is \( \|x\| = (x_1^2 + x_2^2 + \cdots + x_n^2)^{1/2} \)

- therefore
  - \( E \|x\|^2 = E (x_1^2 + x_2^2 + \cdots + x_n^2) \)
  - \( B^TDB = B_1^2D_{11} + B_1B_2D_{12} + B_1B_2D_{21} + B_2^2D_{22} \)

Return Again to Solution
Identification - Background

- identification is important in structural econometric modeling
  - $F$ distribution of observed data (for example $(y, x)$)
  - $\mathcal{F}$ a collection of distributions $F$
  - $\theta$ a parameter of interest (for example $\mathbb{E}y$)

  * identification means that a parameter is uniquely determined by the distribution of the observed variables

Definition

* A parameter $\theta \in \mathbb{R}$ is identified on $\mathcal{F}$ if for all $F \in \mathcal{F}$ there is a uniquely determined value of $\theta$.

- equivalently, $\theta$ is identified if we can write out a mapping $\theta = g(F)$ on the set $\mathcal{F}$
  - restriction to $\mathcal{F}$ is important

  * most parameters are identified only on a strict subset of the space of all distributions
consider identification of the mean $\mu = \mathbb{E} y$

$\mu$ is uniquely determined if $\mathbb{E} y < \infty$

$\star \mu$ is identified for the set $\mathcal{F} = \left\{ F : \int_{-\infty}^{\infty} |y| \, dF(y) < \infty \right\}$

identification of the conditional mean

**Theorem:** If $\mathbb{E} y < \infty$, the conditional mean $m(x) = \mathbb{E} (y|x)$ is identified almost everywhere.

generally, moments of observed data are identified as long as we exclude degenerate cases
Identification - More Complicated Models

- consider the context of censoring
  - $y$ is a random variable with distribution $F$
  - we observe $y^*$ defined by the censoring rule

$$y^* = \begin{cases} 
  y & \text{if } y \leq \tau \\
  \tau & \text{if } y > \tau
\end{cases}$$

- applies to income surveys, where incomes above the top code are recorded as equal to the top code ("top coded" data)

- observed variable $y^*$ has distribution

$$F^* (u) = \begin{cases} 
  F (u) & \text{if } u < \tau \\
  1 & \text{if } u \geq \tau
\end{cases}$$

- we are interested in the features of $F$ not the censored distribution $F^*$
  - we cannot calculate $\mu = \mathbb{E} y$ from $F^*$ except in the trivial case where there is no censoring $\mathbb{P} (y \geq \tau) = 0$
    - $\mu$ is not generically identified from $F^*$
Assumptions to Restore Identification

- **parametric identification**
  - assume a parametric distribution \((y \sim \mathcal{N}(\mu, \sigma^2))\)
    - so \(\mathcal{F}\) is the set of normal distributions
    - can show that \((\mu, \sigma^2)\) are identified for all \(F \in \mathcal{F}\)
  - not ideal - identification achieved only through use of an arbitrary and unverifiable parametric assumption

- **nonparametric identification**
  - quantiles \(q_\alpha\) of \(F\), for \(\alpha \leq \mathbb{P}(y \leq \tau)\) are identified
    - if 20% of the distribution is censored, can identify all quantiles for \(\alpha \in (0, 0.8)\)

- study of identification focuses attention on what can be learned from the data distributions available

*Return to General Identification*
Generalized Inverse

- for any matrix $A$
  - $A^\dagger$ (Moore-Penrose generalized inverse) exists and is unique
- $A^\dagger$ satisfies
  - $AA^\dagger A = A$
  - $A^\dagger AA^\dagger = A$
  - $AA^\dagger$ and $A^\dagger A$ are symmetric
- example, if $A_{11}^{-1}$ exists and $A = \begin{bmatrix} A_{11} & 0 \\ 0 & 0 \end{bmatrix}$
- then $A^\dagger = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$

Return to Identification
Proof of Theorem 1

4 \|E(xx^T)\| \leq E \|xx^T\| \quad (\text{Expectation Inequality})

E \|xx^T\| = E \|x\|^2 < \infty \quad (\text{Assumption 1})

- $A^{-}$ satisfies
  - $AA^{-}A = A$
  - $A^{-}AA^{-} = A$
  - $AA^{-}$ and $A^{-}A$ are symmetric

- example, if $A_{11}^{-1}$ exists and $A = \begin{bmatrix} A_{11} & 0 \\ 0 & 0 \end{bmatrix}$

- then $A^{-} = \begin{bmatrix} A_{11}^{-1} & 0 \\ 0 & 0 \end{bmatrix}$

Return to Properties of the LPM