# Advanced Econometrics

University of Warsaw

Jerzy Mycielski

Doctoral School of Social Sciences, 2024

# Linear CEF model

- Assume that CEF is the following function

$$m(x) = x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_{k-1} + \beta_k = x'\beta.$$

- $\beta_k$ is known as constant term or intercept
- $x = \begin{bmatrix} X_1 \\ \vdots \\ X_k \end{bmatrix}, \beta = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}$

$$\begin{aligned} Y =& X'\beta + e \\ E\left( e | X \right) =& 0 \end{aligned}$$

- Additionaly homoscedasticity assumption $var\left( e | X \right) = \sigma^2$ is sametimes added

# Linear CEF with nonlinear effects, with dummy variables

- Quadratic regression

$$m\left(x1, x2\right) = x_1\beta_2 + x_2\beta_2 + x_1^2\beta_3 + x_2^2\beta_4 + x_1x_2\beta_5 + \beta_6$$

- It is nonlinear in the regressors $x_1, x_2$ but it is linear in the coefficients.
- Regression derivative w.r.t. $x_1$

$$\frac{\partial m\left(x1, x2\right)}{\partial x_1} = \beta_2 + 2x_1\beta_3 + x_2\beta_5$$

- $\beta_5$ is related to so called interaction effect
- Qualitative variables are coded using binary (dummy) variables

$$\begin{cases} X_1 = 1 & \text{if } gender = man \\ X_1 = 0 & \text{if } gender = woman. \end{cases}$$

- If qualitative variable has more than $s > 2$ possible values we use $s - 1$ dummy variables to represent it

# Best linear predictor

- Assume $\mathbb{E}\left[Y^2\right] < \infty$, $\mathbb{E}\left\|X\right\|^2 < \infty$, $\mathbb{E}\left[XX'\right]$ is positive semidefinite (invertible)
- Linear Predictor of $Y$ given $X$ is $\mathscr{P}\left[Y|X\right] = X'\beta$ where $\beta$ minimizes the mean squared prediction error:

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

- Solving this minimization problem gives the following formula for $\beta$ (Linear Projection Coefficient)

$$\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[XY\right] = Q_{XX}^{-1} Q_{XY}$$

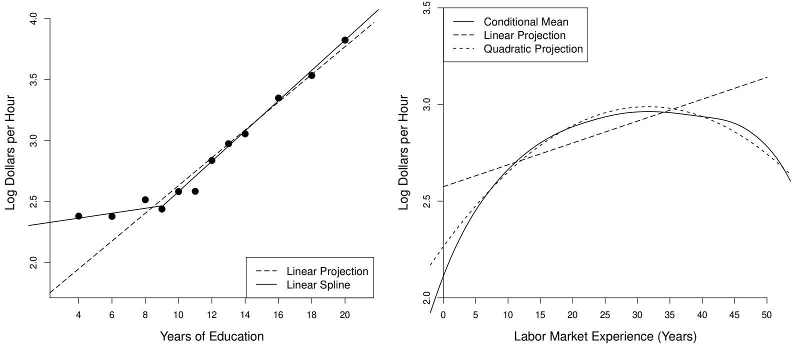
- Then the Best Linear Predictor (Linear Projection) is given by

$$\mathscr{P}\left[Y|X\right] = X'\left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[XY\right]$$

- In the linear projection model $Y = X'\beta + \alpha + e$, $\mu_Y = \mathbb{E}\left(Y\right)$, $\mu_X = \mathbb{E}\left(x\right)$

$$\alpha = \mu_Y - \mu_X \beta$$
$$\beta = var\left[X\right]^{-1} cov\left(X, Y\right).$$

# Wage, spline and polynomial projections




(a) Projections onto *education*

(b) Projections onto *experience*

Figure 2.6: Projections of log(*wage*) onto *education* and *experience*

Source: Hansen (2022)

# Omitted variable bias

- Consider the following regression model

$$Y = X_1'\beta_1 + X_2'\beta_2 + e$$

- Make Linear Projection of $\dot{Y}$on $X_1$ only
- In such a case Linear Projection Coefficient is

$$\begin{aligned}\gamma_1 =& \mathbb{E}\left[X_1X_1'\right]^{-1}\mathbb{E}\left[X_1Y\right] \\ =& \mathbb{E}\left[X_1X_1'\right]^{-1}\mathbb{E}\left[X_1\left(X_1'\beta_1 + X_2'\beta_2 + e\right)\right] \\ =& \beta_1 + \mathbb{E}\left[X_1X_1'\right]^{-1}\mathbb{E}\left[X_1X_2'\right]\beta_2 = \beta_1 + \Gamma_{12}\beta_2\end{aligned}$$

- Generally speaking $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 \neq \beta_1$, the coefficient is biased estimate of $\beta_1$!

# Misspecified model

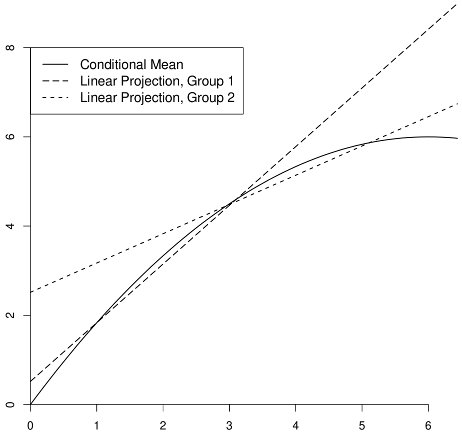



Figure 2.7: Conditional Mean and Two Linear Projections

Source: Hansen (2022)

# CEF and Causal effect (Potential Outcome/Rubin model)

- Model
$$Y = h(D, X, U)$$
- $X$ are observable factors, $U$ unobservable factors
- Causal effect of $D$ on $Y$ is
$$C(X, U) = Y(1) - Y(0) = h(1, X, U) - h(0, X, U)$$
interpreted as change in $Y$ due to treatment while holding $U$ constant
- The conditional average causal effect o of $D$ on $Y$ is
$$ACE(x) = \mathbb{E}[C(X, U)| X = x] = \int_{\mathbb{R}^l} C(x, u) f(u| x) du$$
where $f(u)$ is the density of $U$.
- The unconditional average causal effect of $D$ on $Y$ is
$$ACE = \mathbb{E}[C(X, U)] = \int ACE(x) f(x) dx$$

# Conditional Independence Assumption (CIA)

- We say that variables $U$ and $D$ are conditionally independent if conditional on $X$ the random variables $D$ and $U$ are statistically independent

$$f\left(u|\, D, X\right) = f(u|X)$$

- In such a case

$$\begin{aligned} m\left(d,x\right) =& \mathbb{E}\left[Y|\, D=d, X=x\right] = \mathbb{E}\left[h\left(d,x,U\right)|\, D=d, X=x\right] \\ =& \mathbb{E}\left[h\left(d,x,U\right)|\, X=x\right] \end{aligned}$$

- Therefore

$$\begin{aligned} \nabla m\left(d,x\right) =& m\left(1,x\right) - m\left(0,x\right) \\ =& \mathbb{E}\left[h\left(d,x,U\right)|\, X=x\right] - \mathbb{E}\left[h\left(0,x,U\right)|\, X=x\right] \\ =& \mathbb{E}\left[C\left(X,U\right)|\, X=x\right] = ACE\left(x\right) \end{aligned}$$

- CIA implies $\nabla m\left(d,x\right) = ACE(x)$ !

# The concept of a random sample.

- Sample is the set $\{(Y_i, X_i) : i = 1, ..., n\}$ of $n$ realisations of the random variables $(Y, X)$
- The variables $(Y_i, X_i)$ are a **sample** from the distribution $F$ if they are identically distributed with distribution $F$
- The variables $(Y_i, X_i)$ are a **random sample** if they are mutually independent and identically distributed (i.i.d.) across $i = 1, ..., n$.
- The sample have this properties if the process of selecting the sample from the population satisfies the following conditions:
  - every member of the population have the same probability of being drawn to the sample
  - the probabilities of being drawn from the population are independent: observations are independent

# Bibliography

Hansen, B. (2022). *Econometrics*. Princeton University Press. ISBN: 9780691235899.