Class 2: CEF & OLS

MFE 402

Dan Yavorsky

Last Class

- Introduced this course
- Reviewed key topics from probability and statistics
  - Emphasized random variables and their distributions
  - Discussed statistics
  - Barely mentioned anything about data
- Reviewed key ideas and notation when working with matrices
  - Common mathematical operations
  - The transpose, determinant, and inverse of a matrix
  - Briefly mentioned definiteness, Cholesky decomposition, and the Trace operator
  - Provided a reference for common derivatives with vectors

Topics for Today

1. Joint, Marginal, and Conditional Distributions
2. Conditional Expectation Function (CEF) and CEF Model
3. The Special Case of a Linear CEF
4. A Linear CEF Model
5. Least Squares Estimator

Joint, Marginal, and Conditional Distributions & Densities

Joint Distribution & Density Functions

We begin by generalizing from a univariate random variable to a vector of random variables (aka **multivariate random vector**) using the bivariate case to illustrate:

CDF: The joint **cumulative distribution** function is

$$F_{X,Y}(x, y) = \mathbb{P}_{X,Y}[X \leq x, Y \leq y]$$

PDF:

For vectors of **discrete** random variables, the joint probability **mass** function is

$$p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y]$$

For vectors of **continuous** random variables, the joint probability **density** function is

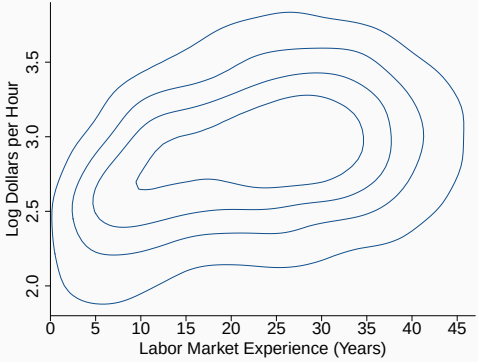
$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y)$$

Example: Joint Probability Mass/Density Functions

Discrete Example

	Y=1	**Y=2**	**Y=3**
X=1	1/60	4/60	9/60
X=2	2/60	6/60	12/60
X=3	3/60	8/60	3/60
X=4	4/60	2/60	6/60

Continuous Example



Marginal (Univariate) Distribution & Density Functions

The marginal univariate cumulative distribution function (or just **marginal distribution**) of X is

$$\begin{aligned} F_X(x) &= \mathbb{P}[X \leq x] \\ &= \mathbb{P}[X \leq x, Y \leq \infty] \\ &= \lim_{y \to \infty} F_{X,Y}(x, y) \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{x} f_{X,Y}(u, v)\, du\, dv \end{aligned}$$

The marginal univariate probability density function (or just **marginal density**) of X is

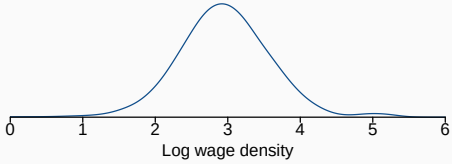
$$\begin{aligned} f_X(x) &= \frac{d}{dx} F_X(x) && \text{derivative of univariate CDF} \\ &= \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy && \text{integrate } Y \text{ out of joint PDF} \end{aligned}$$

Example: Marginal Distributions

Discrete Example

	Y=1	Y=2	Y=3	f(x)
X=1	1/60	4/60	9/60	14/60
X=2	2/60	6/60	12/60	20/60
X=3	3/60	8/60	3/60	14/60
X=4	4/60	2/60	6/60	12/60
f(y)	10/60	20/60	30/60	

Continuous Example



Conditional Distribution & Density Functions

The **conditional distribution** function of Y given $X = x$ is

$$F_{Y|X}(y|x) = \mathbb{P}[Y \leq y|X = x]$$

The **conditional density** function of Y given $X = x$ is

$$\begin{aligned} f_{Y|X}(y|x) &= \frac{\partial}{\partial y} F_{Y|X}(y|x) \\ &= \frac{f_{X,Y}(x,y)}{f_X(x)} \end{aligned}$$

Think of this as the distribution and density of Y for the subpopulation with a specific value of $X = x$

Example: Conditional Distributions

Discrete Example

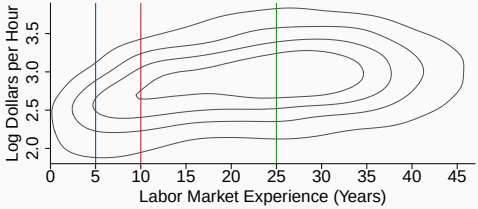
	Y=1	Y=2	Y=3	f(x)
X=1	1/60	4/60	9/60	14/60
X=2	2/60	6/60	12/60	20/60
X=3	3/60	8/60	3/60	14/60
X=4	4/60	2/60	6/60	12/60
f(y)	10/60	20/60	30/60	

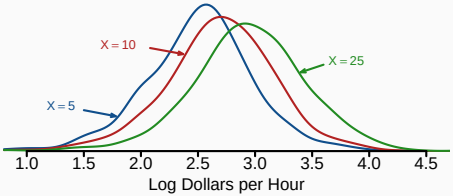
$f_{Y|X}(y = 1|x = 2) = (2/60)/(20/60) = 0.1$

$f_{Y|X}(y = 2|x = 2) = (6/60)/(20/60) = 0.3$

$f_{Y|X}(y = 3|x = 2) = (12/60)/(20/60) = 0.6$

Continuous Example





The CEF

Conditional Expectation Function

The **Conditional Expectation Function** (CEF) of Y given X is $\mathbb{E}[Y|X=x]$:

$$\begin{aligned}\mathbb{E}[Y|X=x] &= \int_{-\infty}^{\infty} y\ f_{Y|X}(y|X=x)\,dy \\ &= \int_{-\infty}^{\infty} y\ \frac{f_{X,Y}(X=x,y)}{f_X(X=x)}\,dy \\ &= \frac{1}{f_X(x)}\int_{-\infty}^{\infty} y\ f_{X,Y}(X=x,y)\,dy\end{aligned}$$

It is a function of x:

$$m(x) = \mathbb{E}[Y|X=x]$$

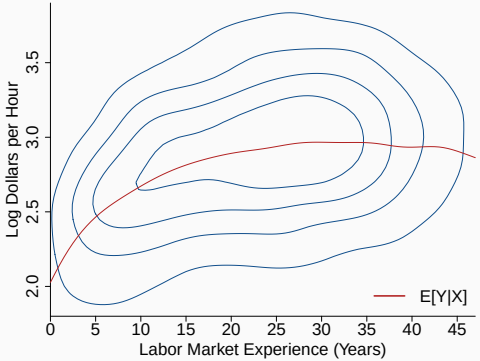
Example: CEF

Discrete Example

	Y=1	Y=2	Y=3	f(x)
X=1	1/60	4/60	9/60	14/60
X=2	2/60	6/60	12/60	20/60
X=3	3/60	8/60	3/60	14/60
X=4	4/60	2/60	6/60	12/60
f(y)	10/60	20/60	30/60	

$$\mathbb{E}[Y|X=1] = 1(1/14) + 2(4/14) + 3(9/14) = 2.571$$
$$\mathbb{E}[Y|X=2] = 1(2/20) + 2(6/20) + 3(12/20) = 2.500$$
$$\mathbb{E}[Y|X=3] = 1(3/14) + 2(8/14) + 3(3/14) = 2.000$$
$$\mathbb{E}[Y|X=4] = 1(4/12) + 2(2/12) + 3(6/12) = 2.167$$

Continuous Example



Law of Iterated Expectations

The expectation of the CEF is the unconditional expectation:

$$\mathbb{E}\left[\mathbb{E}[Y|X]\right] = \mathbb{E}[Y]$$

That is, for each value x, we have $\mathbb{E}[Y|X = x]$, and then we take the probability-weighted average across the x's:

$$\begin{aligned}
\mathbb{E}\left[\mathbb{E}[Y|X]\right] &= \int_{-\infty}^{\infty} m(x)\, f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\infty} y\, f_{Y|X}(y|x)\, dy \right) f_X(x)\, dx \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y\, f_{X,Y}(x,y)\, dy\, dx \\
&= \int_{-\infty}^{\infty} y\, f_Y(y)\, dy \\
&= \mathbb{E}[Y]
\end{aligned}$$

CEF Error

The CEF error e (some textbooks use ε) is defined as the difference between Y and the CEF evaluated at X:

$$e = Y - m(X)$$

By construction, this yields the "breakdown" formula via simple rearrangement:

$$Y = m(X) + e$$

Notice: e is derived from $F_{X,Y}(X, Y)$ because it's a function of Y and X, so it's properties are derived from this construction. Some important properties:

- $\mathbb{E}[e|X] = 0$ (i.e., e is "mean independent" of X, but not necessarily independent of X)
- $\mathbb{E}[e] = 0$
- $\mathbb{E}[h(X)\, e] = 0$ for any function of X, $h(X)$

CEF as the Best Predictor of Y

Why focus on the CEF?

- The CEF $m(X)$ is the best predictor of Y (in a minimum mean-squared error sense).

Proof:

- Suppose you have X and want to predict Y: $\hat{Y} = g(X)$. One way to measure the (ex ante and non-stochastic) magnitude of the prediction error is with the mean-squared error function:

$$
\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(m(X) + e - g(X))^2\right] \\
&= \mathbb{E}\left[e^2\right] + 2 \times \mathbb{E}\left[e\,(m(X) - g(X))\right] + \mathbb{E}\left[(m(X) - g(X))^2\right] \\
&= \mathbb{E}\left[e^2\right] + \mathbb{E}\left[(m(X) - g(X))^2\right] \\
&\geq \mathbb{E}\left[e^2\right] \\
&= \mathbb{E}\left[(Y - m(X))^2\right] \qquad \text{with equality when} \quad g(X) = m(X)
\end{aligned}
$$

CEF Gradient

One way to interpret the CEF is how marginal changes in the regressors x_i for $i = 1, \ldots, k$ imply changes in the conditional expectation of the response variable Y:

$$\nabla_i m(\mathbf{x}) = \frac{\partial}{\partial x_i} m(\mathbf{x}) = \frac{\partial}{\partial x_i} m(x_1, x_2, \ldots, x_k)$$

Notice:

1. The effect of each variable is calculated holding the other variables constant.
   - This interpretation is not "all else held constant" but more precisely "all else **in the model** held constant"
2. The effect of each variable is on $m(\mathbf{x}) = \mathbb{E}[Y|X = \mathbf{x}]$, not necessarily on the value of Y.
   - This is the change in the actual value of Y only if the error e is unaffected by the change in the regressor x_i. We'll come back to this when we discuss endogeneity.

Linear CEF

Linear CEF

An important special case is when the CEF is linear in $\mathbf{x}$:

$$m(\mathbf{x}) = x_1\beta_1 + x_2\beta_2 + \ldots + x_k\beta_k$$

We often use shorthand notation, defining the vectors $\mathbf{x}$ and $\boldsymbol{\beta}$ as

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{k\text{-}1} \\ 1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_{k\text{-}1} \\ \beta_k \end{bmatrix}$$

so that we can write the CEF as $m(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$.

Notice we set $x_k = 1$ so that β_k is the intercept (we'll sometimes call the intercept β_0)

Interpretation of Coefficients whe the CEF is Linear

One of the most appealing features of a Linear CEF is that the coefficients ($\boldsymbol{\beta}$) are the CEF derivatives:

$$\nabla m(\mathbf{x}) = \begin{bmatrix} \nabla_1 m(\mathbf{x}) \\ \nabla_2 m(\mathbf{x}) \\ \vdots \\ \nabla_k m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x_1} m(\mathbf{x}) \\ \frac{\partial}{\partial x_2} m(\mathbf{x}) \\ \vdots \\ \frac{\partial}{\partial x_k} m(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \boldsymbol{\beta}$$

Therefore, the coefficients have simple and natural interpretations as the marginal effects of changing one variable, holding the others constant.

Linear CEF with Non-Linear Effects

The linear CEF is less restrictive than it first appears. Take the following CEF as an example:

$$m(\mathbf{x}) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_2^2\beta_3$$

Here, $m(\mathbf{x})$ is non-linear in x_2 but we can define $x_3 = x_2^2$ to re-write the CEF as

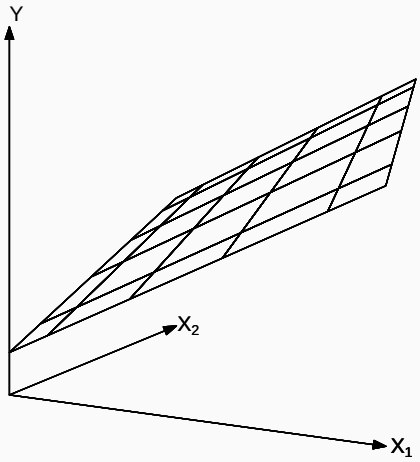
$$m(\mathbf{x}) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_3\beta_3$$

This creates a linear CEF, which is sufficient for most econometric purposes (ie, estimation and inference of the parameters). The one major exception is with the analysis of CEF derivatives, which should be defined with respect to the "original" variables:

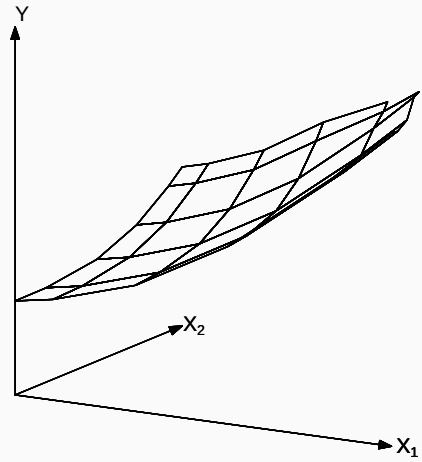
$$\frac{\partial}{\partial x_1} m(\mathbf{x}) = \beta_1$$
$$\frac{\partial}{\partial x_2} m(\mathbf{x}) = \beta_2 + 2x_2\beta_3$$

Example: Linear vs Non-Linear Effects

Linear CEF Surface



Non-Linear CEF Surface



Linear Projection Coefficient

When the CEF is a linear function of X, that is $m(X) = X'\beta$, we call β the linear projection coefficient.

Recall that $\mathbb{E}[Xe] = 0$.Then:

$$
\begin{aligned}
&= \mathbb{E}[Xe] \\
&= \mathbb{E}[X(Y - X'\beta)] \\
&= \mathbb{E}[XY] - \mathbb{E}[XX']\beta \\
\beta &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right] \\
&= \mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} \qquad \text{(in BHE notation)}
\end{aligned}
$$

This derivation requires a couple of technical, mathematical properties:

- The means, variances, and covariance between X and Y must be finite (ie, not infinite)
- That $\mathbb{E}[XX']$ is invertible

Linear CEF Model

Regression Function

A regression function $g(X)$ is a function of X used to predict Y

We saw earlier that the CEF is the best predictor of Y given X
(in the sense that it minimizes the mean-squared error)

- Therefore, the CEF is the best regression function
- But we rarely know the CEF, so we need to pick a functional form to use
- We should pick one that does a good job approximating the CEF
  (which feels like tough criteria: since we don't know the CEF, how can we know if our choice of functional form will do a good job approximating it)

The Linear CEF Model

The Linear CEF **Model** is two assumptions:

$$Y = X'\beta + e \qquad \text{with} \qquad \mathbb{E}[e|X] = 0$$

Comments:

- These modeling assumptions are for simplicity – we rarely know the true shape of the CEF
- The assumption $\mathbb{E}[e|X] = 0$ can be a *major* assumption. It implies a particular relationship between X and e such that $\text{Cov}(X, e) = 0$. We'll return to this when we discuss endogeneity.

Three Reasons in Support of a Linear CEF Model

Define the **population linear regression function** as $X'\beta$ with $\beta = (\mathbb{E}[XX'])^{-1}\mathbb{E}[XY]$ derived exactly as before

1. When the **CEF is linear**, the population linear regression function is it
   - Other than coincidence, there are two situations that lead to a linear CEF:
     1. A fully-saturated model – discrete X's that are fully interacted with each other
     2. The joint distribution of X and Y $F_{X,Y}(x, y)$ is multivariate normal in $\mathbb{R}^{k+1}$
2. The population linear regression function is the **best linear predictor** of Y given X, even when the CEF is non-linear
3. The population linear regression function is the **best linear approximation** to the CEF, even when the CEF is non-linear

1.1 Linear CEF with Dummy Variables

If all regressors (the X's) take a finite set of values, the CEF can be written as a linear function.

One binary variable example:

- Suppose X represents binary gender with $X = 0$ for males and $X = 1$ for females.
- Let $\mathbb{E}[Y|X = 0] = \mu_0$ and $\mathbb{E}[Y|X = 1] = \mu_1$ and define $\beta_0 = \mu_0$ and $\beta_1 = \mu_1$ - μ_0
- Then $m(x) = \beta_0 + \beta_1 x$

Two binary variables example:

- Suppose X_1 binary gender ($X_1 = 1$ is female) and X_2 marital status ($X_2 = 1$ is married)
- Let $m(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$
- Then:
  - $\mathbb{E}[Y|X_1 = 0, X_2 = 0] = \beta_0$
  - $\mathbb{E}[Y|X_1 = 1, X_2 = 0] = \beta_0 + \beta_1$
  - $\mathbb{E}[Y|X_1 = 0, X_2 = 1] = \beta_0 + \beta_2$
  - $\mathbb{E}[Y|X_1 = 1, X_2 = 1] = \beta_0 + \beta_1 + \beta_2 + \beta_3$

1.2 Linear CEF with Joint Normality

If the joint distribution of Y and X is multivariate normal, then the CEF is linear in X.

To see this, consider the best linear predictor of Y given X: $\quad Y = X'\beta + e$

Properties of the best linear predictor:

- $\mathbb{E}[Xe] = \mathbb{E}[X(Y - X'\beta)] = \mathbb{E}[XY] - \mathbb{E}[XX'] \left(\left(\mathbb{E}[XX'] \right)^{-1} \mathbb{E}[XY] \right) = 0$
- $\mathbb{E}[e] = 0$ if X contains an intercept because then $\mathbb{E}[e]$ is a row of $\mathbb{E}[Xe]$
- Therefore, $\text{Cov}(X, e) = \mathbb{E}[Xe] - \mathbb{E}[X]\mathbb{E}[e] = 0 + 0 = 0$
- Since the vector (e, X) is an affine transformation of the multivariate normal vector (Y, X), it is also multivariate normal.
- Since e and X are normal and uncorrelated, they are independent.

These are the properties of a Linear CEF. Therefore, when (Y, X) are jointly normally distributed, they satisfy a normal linear CEF.

2. Best Linear Predictor

A **linear predictor for** Y is a function $X'\tilde{\beta}$ for some $\tilde{\beta} \in \mathbb{R}^k$.

The mean-squared prediction error is

$$S(\tilde{\beta}) = \mathbb{E}\left[\left(Y - X'\tilde{\beta}\right)^2\right]$$

As a quadratic function of β:

$$S(\tilde{\beta}) = \mathbb{E}\left[Y^2\right] - 2\tilde{\beta}'\mathbb{E}\left[XY\right] + \tilde{\beta}'\mathbb{E}\left[XX'\right]\tilde{\beta}$$

Take a first-order condition and solve for $\tilde{\beta}$:

$$\frac{\partial}{\partial\beta}S(\tilde{\beta}) = -2\mathbb{E}\left[XY\right] + 2\mathbb{E}\left[XX'\right]\tilde{\beta} = 0$$

$$\Rightarrow \tilde{\beta} = \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right] = \mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} = \beta$$

The minimizer $\tilde{\beta} = \arg\min_{\mathbf{b}\in\mathbb{R}^k} S(\mathbf{b})$ is the linear projection coefficient β

3. Best Linear Approximation

A **linear approximation to the CEF** $m(X)$ is a function $X'\bar{\beta}$ for some $\bar{\beta} \in \mathbb{R}^k$.

The mean-squared approximation error is

$$\begin{aligned} d(\bar{\beta}) &= \mathbb{E}\left[\left(m(X) - X'\bar{\beta}\right)^2\right] \\ &= \mathbb{E}\left[m(X)^2\right] - 2\bar{\beta}'\mathbb{E}\left[Xm(X)\right] + \bar{\beta}'\mathbb{E}\left[XX'\right]\bar{\beta} \end{aligned}$$

Take a first-order condition and solve for $\bar{\beta}$:

$$\begin{aligned} \frac{\partial}{\partial \bar{\beta}} d(\bar{\beta}) &= -2\mathbb{E}\left[Xm(X)\right] + 2\mathbb{E}\left[XX'\right]\bar{\beta} = 0 \\ \Rightarrow \bar{\beta} &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[Xm(X)\right] \\ &= \left(\mathbb{E}\left[XX'\right]\right)^{-1}\mathbb{E}\left[XY\right] = \mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} = \beta \end{aligned}$$

The minimizer $\bar{\beta} = \arg\min_{\mathbf{b} \in \mathbb{R}^k} d(\mathbf{b})$ is also the linear projection coefficient β

Why the term "Regression"

Galton studied children height (Y) as a function of parent height (X).

This situation has 2 special properties:

1. height is approximately normally distributed
2. inter-generational height distributions are stable across generations

So Galton assumes $F_{X,Y}(x, y)$ is MVN (with linear CEF) and $\mu_X = \mu_Y$, $\sigma_X^2 = \sigma_Y^2$.

Then (partly based on the upcoming derivation on slide 35):

$$\beta = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \rho \qquad \text{and} \qquad \alpha = \mathbb{E}[Y] - \beta\mathbb{E}[X] = \mu - \rho\mu = \mu(1 - \rho)$$

Putting this together, we have the assumed CEF is:

$$\mathbb{E}[Y|X] = \alpha + \beta X = \mu(1 - \rho) + \rho X$$

Thus, the height of a child is a weighted average of the parent height (X) and the population average (μ), hence the term "regression to the mean"

Ordinary Least Squares Estimator: $\hat{\beta}$

Samples

We have been considering a pair of random variables $(Y, X) \in \mathbb{R} \times \mathbb{R}^k$, from which we discussed $\beta = \left(\mathbb{E}\left[XX'\right]\right)^{-1} \mathbb{E}\left[XY\right]$ as:

- the coefficient vector of the CEF when it is linear
- the linear projection coefficient vector, shown to be the best linear predictor of Y given X and the best linear approximation to the CEF

We are now interested in **estimating β from samples**: joint measurements of (Y, X).

Notice that

- the random variables are Y and X (X may be a vector of length k)
- the observations in the sample are Y_i and X_i
- n is the sample size, such that the dataset is $\{(Y_i, X_i) : i = 1, \ldots, n\}$

We'll assume our sample is iid or "random" – that is, the random vectors of length $k+1$ (Y_i, X_i) for $i = 1, \ldots, n$ are independent and identically distributed; they are draws from a common distribution $F_{X,Y}(x, y)$.

Moment Estimators

The Law(s) of Large Numbers show that the average approaches the expectation as the sample size grows: $\bar{X}_n \longrightarrow \mathbb{E}[X]$

We use this idea to develop "analog" or "plug in" or "moment" estimators.

For example,

- Suppose $\mu = \mathbb{E}[Y]$. Estimate μ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$
- Suppose $\mu = \mathbb{E}[h(Y)]$. Estimate μ with $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} h(Y_i)$

There's a general method of estimation and inference in econometrics called the **Generalized Method of Moments (GMM)** devoted to this idea.

MM Estimator for Linear CEF

Suppose the CEF is linear in X. We can use moment estimators of the expectations:

$$\mathbf{Q}_{XX} = \mathbb{E}[XX'] \qquad \Rightarrow \hat{\mathbf{Q}}_{XX} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i'$$

$$\mathbf{Q}_{XY} = \mathbb{E}[XY] \qquad \Rightarrow \hat{\mathbf{Q}}_{XY} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i$$

and so

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

$$\Rightarrow \hat{\beta} = \left(\hat{\mathbf{Q}}_{XX}\right)^{-1}\hat{\mathbf{Q}}_{XY} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

OLS Estimator as Best Linear Approximation

While the CEF is the best predictor of Y among all functions of X, it's functional form is typically unknown, and is thought to be linear only in special cases. Usually, then, it is more realistic to view the linear model specification $m(X) = X'\beta$ as an *approximation*.

Recall, the linear projection coefficient β is defined as the minimizer of the expected squared error $S(\beta)$:

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

The moment estimator of $S(\beta)$ is the sample average:

$$\hat{S}(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2$$

$\hat{\beta} = \arg\min \hat{S}(\beta)$ is the **ordinary least squares (OLS) estimator** because

- it minimizes the sum of squared errors
- it solves k equations with k unknowns using "ordinary" techniques

Solving for the OLS Estimator with One Regressor

Consider the case where $k = 1$ so that there is a scalar regressor X and two coefficients: $Y = \beta_1 + \beta_2 X + e$

$$S(\beta_0, \beta_1) = \mathbb{E}\left[(Y - \beta_0 - \beta_1 X)^2\right] \quad \Rightarrow \quad \hat{S}(\beta_0, \beta_1) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

First Order Condition for β_0:

$$\begin{aligned}\frac{\partial}{\partial \beta_0}\hat{S}(\hat{\beta}_0, \hat{\beta}_1) &= -2n^{-1} \times \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \\ &= \bar{Y} - \hat{\beta}_0 - \hat{\beta}_1 \bar{X} \\ &= 0 \\ \Rightarrow \hat{\beta}_0 &= \bar{Y} - \hat{\beta}_1 \bar{X}\end{aligned}$$

First Order Condition for β_1:

$$\begin{aligned}\frac{\partial}{\partial \beta_1}\hat{S}(\hat{\beta}_0, \hat{\beta}_1) &= -2n^{-1} \times \sum_{i=1}^{n}\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i\right) \times X_i \\ &= n^{-1}\sum_{i=1}^{n} Y_i X_i - \left(\bar{Y} - \hat{\beta}_1 \bar{X}\right)\sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 \\ &= \left(n^{-1}\sum_{i=1}^{n} Y_i X_i - \bar{Y}\bar{X}\right) - \hat{\beta}_1\left(n^{-1}\sum_{i=1}^{n} X_i^2 - \bar{X}^2\right) \\ &= 0 \\ \Rightarrow \hat{\beta}_1 &= \hat{\mathrm{Cov}}(X_i, Y_i)/\hat{\mathrm{Var}}(X_i)\end{aligned}$$

Solving for the OLS Estimator with Multiple Regressors

Now consider the general case with k regressors (including $X_1 = 1$): $Y = X'\beta + e$

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

$$\hat{S}(\beta) = \sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)^2 = \sum_{i=1}^{n}\left(Y_i - X_i'\beta\right)'\left(Y_i - X_i'\beta\right) = \sum_{i=1}^{n} Y_i^2 - 2\beta'\sum_{i=1}^{n} X_i Y_i + \beta'\left(\sum_{i=1}^{n} X_i X_i'\right)\beta$$

k First Order Conditions

$$\frac{\partial}{\partial\beta}\hat{S}(\hat{\beta}) = -2\sum_{i=1}^{n} X_i Y_i + 2\sum_{i=1}^{n} X_i X_i'\hat{\beta} = 0$$

$$\Rightarrow \hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1}\left(\sum_{i=1}^{n} X_i Y_i\right)$$

BHE uses the notation $\hat{\beta} = \hat{\mathbf{Q}}_{XX}^{-1}\hat{\mathbf{Q}}_{XY}$. This is identical to the Method of Moment estimator from 3 slides ago.

Model in Matrix Notation

It is notationally and computationally **convenient** to write the model and statistics in matrix notation.

The n linear equations $Y_i = X_i'\beta + e_i$ make a system of n equations, which we stack:

$$\begin{aligned} Y_1 &= X_1'\beta + e_1 \\ Y_2 &= X_2'\beta + e_2 \\ &\vdots \\ Y_n &= X_n'\beta + e_n \end{aligned}$$

Then the system of n equations can by written compactly as $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ where:

$$\underset{n\times 1}{\mathbf{y}} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\times k}{\mathbf{X}} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}, \quad \underset{k\times 1}{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \underset{n\times 1}{\mathbf{e}} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Notational Simplifications in Matrix Form

With the matrix notation...

...sample sums can be written as

$$\sum_{i=1}^{n} X_i X_i' = \mathbf{X}'\mathbf{X}$$

$$\sum_{i=1}^{n} X_i Y_i' = \mathbf{X}'\mathbf{y}$$

...and the least squares (MM or OLS) estimator is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Estimand, Estimator, Estimate

It is vital to distinguish between these three:

- **Estimand**: The parameter we want to estimate (e.g., $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$)
- **Estimator**: A function of the data to estimate of the estimand (e.g., $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$)
- **Estimate**: The value of the estimator on a particular set of data (e.g., $\hat{\beta} = [7, 0.5]'$)

Econometric theory does not focus on the estimate, but rather on the "recipe" (the procedure, algorithm, or function – ie, the estimator). We next focus on what makes a "good" estimator.

- The least squares estimator gets its name because it minimizes the sum of squared residuals – which seems like a reasonable thing to do – but that criteria alone does not say anything specific about the relationship between the estimator ($\hat{\beta}$) and the estimand (β).
- The great popularity of the OLS estimator is that in some estimating problems (but not all!) it scores well on certain criteria.

Unbiasedness of the OLS Estimator

$\hat{\beta}$ is a function of random variables X and Y and so it is a random variable.

This means that it has a distribution, which we call the **sampling distribution** of $\hat{\beta}$.

If the mean of the sampling distribution is centered over the value we seek to estimate, then the estimator is said to be **unbiased**.

$$\begin{aligned}\mathbb{E}[\hat{\beta}|X] &= \mathbb{E}[(X'X)^{-1}X'y|X] \\ &= \mathbb{E}[(X'X)^{-1}X'(X\beta + e)|X] \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\mathbb{E}[e|X] \\ &= \beta + 0\end{aligned}$$

Notice this requires our assumption about the error term: $\mathbb{E}[e|X] = 0$

Use LIE to find that $\mathbb{E}[\hat{\beta}] = \mathbb{E}[\mathbb{E}[\hat{\beta}|X]] = \mathbb{E}[\beta] = \beta$

$\Rightarrow \hat{\beta}$ is an **unbiased** estimator for β.

Computing the OLS Estimator in R

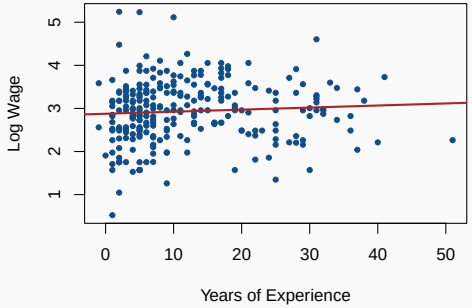
```
# import data
dat <- read.table("support/cps09mar.txt")

# x is experience, y is log(wage)
# subsample of interest is single asian males
exper <- dat[,1] - dat[,4] - 6
lwage <- log( dat[,5]/(dat[,6]*dat[,7]) )
sam <- dat[,11]==4 & dat[,12]==7 & dat[,2]==0

lmcoefs <- coef(lm(lwage[sam] ~ exper[sam]))
print(lmcoefs)
```

```
(Intercept)  exper[sam]
2.876515044 0.004776039
```

```
plot(x=exper[sam], y=lwage[sam], pch=20,
     col="dodgerblue4", ylab="Log Wage",
     xlab="Years of Experience", main="")
abline(a=lmcoefs[1], b=lmcoefs[2],
       col="firebrick", lwd=2)
```



Computing the OLS Estimator in R “by hand”

```
y <- lwage[sam]
x <- cbind(1, exper[sam])

xx <- t(x) %*% x
xy <- t(x) %*% y
betahat <- solve(xx) %*% xy
print(betahat)
```

```
            [,1]
[1,] 2.876515044
[2,] 0.004776039
```

```
# alt way to find x'x
xx <- split(x, 1:nrow(x)) |>
          lapply(\(z) z %*% t(z)) |>
          Reduce(`+`,x=_)
```

```
ssq <- function(beta,x,y) {
    sum((y - x %*% beta)^2)
    }
out <- optim(par=c(0,0), fn=ssq,
             x=x, y=y,
             control=list(reltol=1e-12))
optimcoefs <- out$par
print(optimcoefs)
```

```
[1] 2.876516802 0.004776023
```

Next Time

- Residuals
- Projections
- R Squared
- CEF Error Variance
- Variance of the OLS Estimator