Conditional Expectations (CE)

- Regression analyses are often used to establish possible causal relationships between some variables of interests.

- Some commonly used terms:
  - Explained variable, dependent variable, regressand, response variable (usually denoted by $y$)
  - Explanatory variables, independent variables, regressors, control variables, or covariates (usually denoted by $\mathbf{x} = (x_1, \ldots, x_K)$)

- Conditional Expectation: $E(y|x_1, \ldots, x_K) = \mu(x_1, \ldots, x_K)$. Note that the CE of $y$ here is a function of $(x_1, \ldots, x_K)$.

- Partial effects: how $y$ changes when element of $\mathbf{x}$ change. Assuming that $\mu(\cdot)$ is differentiable and $x_j$ is continuous, the partial effect of $x_j$ is captured by the partial derivative $\partial \mu(\mathbf{x}) / \partial x_j$. If $x_j$ is a discrete variable, partial effects are computed by comparing $E(y|\mathbf{x})$ at different settings of $x_j$.

Example:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2 + u$$

$$E(y|\mathbf{x}) \equiv \mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2^2$$

$$\partial \mu(\mathbf{x}) / \partial x_1 = \beta_1$$

$$\partial \mu(\mathbf{x}) / \partial x_2 = 2\beta_2 x_2.$$
• The error term: $E(u|x) = 0$. This implies

- $E(u) = 0$ (by law of iterated expectations)
- $u$ is uncorrelated with any function of $x$. Note that this is stronger than saying $u$ is uncorrelated with $x$. The latter implies that $\text{cov}(u, x_j) = 0$, but does not rule out cases like $\text{cov}(u, f(x_j)) \neq 0$, where $f(x_j)$ is a non-linear function of $x_j$. On the other hand, $E(u|x) = 0$ is weaker than the independence between $u$ and $x$, which implies $E(u|x) = E(u)$.

• Law of Iterated Expectations (LIE): Let $x = f(w)$, then

$$E(y|x) = E[E(y|x)|w]$$
$$E(y|x) = E[E(y|w)|x]$$

*The smaller information set always dominates.* (As a function of $w$, $x$ has a smaller information set).

- $E(y|x) = E[E(y|x, z)|x]$
- Suppose $E(y|x) = g[f(x)]$, then $E[y,f(x)] = E[y|x] = g[f(x)]$
- $E(y) = E[E(y|x)] = E[\mu(x)]$. Let $p(y|x), p(x)$ and $p(x,y)$ be the conditional density, the marginal density (for $x$) and the joint density respectively. Then

$$\mu(x) = E(y|x) = \int yp(y|x) dy$$
$$E[E(y|x)] = \int \int yp(y|x) dyp(x) dx$$
$$= \int \int yp(x,y) dydx$$
$$= E(y).$$

E.g., suppose $x$ is discrete with $\text{Prob.}(x = 0) = 0.4$ and $\text{Prob.}(x = 1) = 0.6;$
\[ E(y|x = 0) = 1 \] and \[ E(y|x = 1) = 2 \]. We then have

\[
E(y) = 1 \times 0.4 + 2 \times 0.6 \\
= E(y|x = 0) \text{ Prob.}(x = 0) + E(y|x = 1) \text{ Prob.}(x = 1) \\
= E[E(y|x)].
\]

For discrete \( x \), the CE is essentially a weighted average.

- Conditional Jensen’s Inequality: If \( c: \mathcal{R} \to \mathcal{R} \) is a convex function defined on \( R \) and \( E[|y|] < \infty \), then

\[
c[E(y|x)] \leq E[c(y)|x].
\]

Exercise: Show that this leads to unconditional Jensen’s inequality immediately.

- **Conditional variance**

  - Definition: \( \text{Var}(y|x) \equiv \sigma^2(x) = E[(y - E(y|x))^2|x] = E[y^2|x] - [E[y|x]]^2. \)

   This is because

\[
E\left\{ y^2 - 2yE[y|x] + (E[y|x])^2 \right\} |x]
\]

\[
= E[y^2|x] - 2E[y|x]E[y|x|x] + E[(E[y|x])^2|x] \\
= E[y^2|x] - 2[E[y|x]]^2 + [E[y|x]]^2.
\]

- \( \text{Var}[a(x)y + b(x)|x] = [a(x)]^2 \text{Var}(y|x). \)

- \( \text{Var}(y) = E[\text{Var}(y|x)] + \text{Var}[E(y|x)] = E[\sigma^2(x)] + \text{Var}(\mu(x)). \)

\[
\text{Var}(y) = E[(y - E(y))^2] \\
= E[(y - E(y|x) + E(y|x) - E(y))^2] \\
= E[(y - E(y|x))^2] + 2E[y - E(y|x)](E[y|x] - E(y)) \\
+ E[(E(y|x) - E(y))^2].
\]
The middle term is zero by LIE. Also by LIE,

\[
E \left[ (y - E(y|x))^2 \right] = E \left\{ E \left[ (y - E(y|x))^2 \right] \right\} = E \left[ \text{Var}(y|x) \right]
\]
\[
E \left[ (E(y|x) - E(y))^2 \right] = E \left[ (E(y|x) - E(E(y|x)))^2 \right] = E \left[ (\mu(x) - E(\mu(x)))^2 \right] = \text{Var}(\mu(x)).
\]

\[\text{Var}(y|x) = E \left[ \text{Var}(y|x, z)|x \right] + \text{Var}(E(y|x, z)|x)\] (again, small information set dominates)

By LIE, we then have

\[
E[\text{Var}(y|x)] = E[E[\text{Var}(y|x, z)|x] + E[\text{Var}(E(y|x, z)|x)]
\]
\[= E[\text{Var}(y|x, z)] + E[\text{Var}(E(y|x, z)|x)] \geq E[\text{Var}(y|x, z)]
\]

Note that

\[
\text{Var}(y) = E[\text{Var}(y|x)] + \text{Var}(E(y|x))
\]
\[= E[\text{Var}(y|x, z)] + \text{Var}(E(y|x, z)).
\]

It follows that

\[
E[\text{Var}(y|x)] \geq E[\text{Var}(y|x, z)] \iff \text{Var}(E(y|x)) \leq \text{Var}(E(y|x, z)).
\]

Define the mean squared error as \(\text{MSE}(y; \mu) = E \left[ (y - \mu(x))^2 \right]\). This result suggests that

\[\text{MSE}[y; E(y|x)] \geq \text{MSE}[y; E(y|x, z)].\]

Namely, one can improve prediction in terms of the MSE by conditioning on more
variables. (This explains why $R^2$ always improves with the number of covariates.)

- Linear projection (OLS)
  
  - Define $L(y|x) = x\beta$ and $y = L(y|x) + u$, where $\beta$ is the OLS coefficient of $y$ on $x$. If $E(x'x)$ is positive-definite, then $\beta$ is unique. Then $E(x'u) = 0$. (orthogonality condition)
  
  - $L\left(\sum_{j=1}^{G} a_j y_j | x \right) = \sum_{j=1}^{G} a_j L(y_j | x)$, where $a_1, \ldots, a_G$ are constants. (Linear projection is a linear operator)
  
  - Law of integrated projections: $L(y|x) = L\left[L(y|x, z) | x \right]$. (Direct application of LIE. Again, small information set dominates.)
  
  - Suppose $L(y|x, z) = x\beta + z\gamma$. Let $r = x - L(x|z)$ and $v = y - L(y|z)$. Then
    
    $$L(v|r) = r\beta, \quad L(y|r) = r\beta$$

**Basic Asymptotic Theory**

- Convergence in probability. A sequence of random variables $\{x_N : N = 1, 2, \ldots\}$ converges in probability to the constant $a$ if for all $\varepsilon > 0$
  
  $$P[|x_N - a| > \varepsilon] \to 0 \text{ as } N \to \infty.$$ 

  We write $x_N \overset{p}{\to} a$ or $\text{plim} x_N = a$. If $a = 0$, we write $x_N = o_p(1)$.

- A sequence of random variables $\{x_N\}$ is bounded in probability if and only if for every $\varepsilon > 0$, there exists a $b_\varepsilon < \infty$ and an integer $N_\varepsilon$ such that
  
  $$P[|x_N| \geq b_\varepsilon] < \varepsilon \text{ for all } N \geq N_\varepsilon.$$
We write $x_N = O_p (1)$.

- Slutsky’s theorem. Let $g : \mathcal{R}^K \to \mathcal{R}^J$ be a function continuous at some point $c \in \mathcal{R}^K$. Let $\{x_N : N = 1, 2, \ldots \}$ be sequence of $K \times 1$ random vectors such that $x_N \overset{p}{\to} c$. Then $g (x_N) \to g (c)$ as $N \to \infty$. Namely,

$$\text{plim} \ g (x_N) = g (\text{plim} x_N),$$

if $g (\cdot)$ is continuous at $\text{plim} x_N$.

- Convergence in distribution. A sequence of random variables $x_N$ converges in distribution to the continuous random variable $x$ if and only if

$$F_N (\zeta) \to F (\zeta) \quad \text{as} \quad N \to \infty \quad \text{for all} \quad \zeta \in \mathcal{R}$$

where $F_N$ and $F$ are the CDF for $x_N$ and $x$ respectively. We write $x_N \overset{d}{\to} x$.

- If $x \sim N (\mu, \sigma^2)$ and $x_N \overset{d}{\to} x$, then $x_N \overset{d}{\to} N (\mu, \sigma^2)$ or $x_N \sim N (\mu, \sigma^2)$. We say $x_N$ is asymptotically normal.

- Continuous mapping theorem. Let $\{x_N\}$ be a sequence of $K \times 1$ random vectors such that $x_N \overset{d}{\to} x$. If $g : \mathcal{R}^K \to \mathcal{R}^J$ is a continuous functions, then $g (x_N) \overset{d}{\to} g (x)$.

- If $z_n \overset{d}{\to} z$ and $x_n - z_n \overset{p}{\to} 0$, then $x_n \overset{d}{\to} z$.

- Weak Law of Large Numbers (WLLN). Let $\{w_i\}$ be a sequence of iid random variables such that $E (|w_i|) < \infty$. Then $N^{-1} \sum w_i \overset{p}{\to} \mu_w$, where $\mu_w = E w_i$.

- Central Limit Theorem (CLT). Let $\{w_i\}$ be a sequence of iid random vector such that $E (w_i^2) < \infty$ and $E (w_i) = 0$. Then $N^{-1} \sum w_i \overset{d}{\to} N (0, B)$, where $B = \text{Var} (w_i) = E (w_i w_i')$.

- Consistency. If $\hat{\theta}_N \overset{p}{\to} \theta$ for any value of $\theta$, then $\hat{\theta}_N$ is a consistent estimator of $\theta$. 

6
- Asymptotic normality. \( \sqrt{N} \left( \hat{\theta}_N - \theta_N \right) \overset{d}{\rightarrow} N(0, V) \), where \( V \) is positive semidefinite, then \( \hat{\theta}_N \) is asymptotically normal and \( V \) is the asymptotic variance of \( \sqrt{N} \left( \hat{\theta}_N - \theta_N \right) \).

If \( V \) is positive definite and \( \hat{V}_N \overset{p}{\rightarrow} V \), then the asymptotic standard error of \( \hat{\theta}_{Nj} \) is \((\hat{v}_{Njj}/N)^{1/2}\), where \( \hat{v}_{Njj} \) is the \( j \)th diagonal of \( \hat{V}_N \).

- Let \( \sqrt{N} \left( \hat{\theta}_N - \theta_N \right) \overset{d}{\rightarrow} N(0, V) \) and \( \sqrt{N} \left( \tilde{\theta}_N - \theta_N \right) \overset{d}{\rightarrow} N(0, D) \). If \( D - V \) is positive definite, then \( \hat{\theta}_N \) is asymptotically efficient relative to \( \tilde{\theta}_N \); if \( \sqrt{N} \left( \hat{\theta}_N - \tilde{\theta}_N \right) \overset{p}{\rightarrow} 0 \), then \( \hat{\theta}_N \) and \( \tilde{\theta}_N \) are \( \sqrt{N} \)-equivalent.

- Let \( \hat{\theta}_N = \left[ \hat{\theta}_{N1}, \hat{\theta}_{N2} \right] \) with asymptotic variances \( V_1 \) and \( V_2 \) respectively. If

\[
V = \begin{bmatrix}
V_1 & 0 \\
0 & V_2
\end{bmatrix},
\]

then \( \hat{\theta}_{N1} \) and \( \hat{\theta}_{N2} \) are asymptotically independent.