Linear Models: CEF and Best Linear Predictor

Robert Gulotty

University of Chicago

February 20, 2026

Populations, Projections, and Structure

- Today we study relationships between random variables Y and $X = (X_1, \ldots, X_k)$ in the population.
- The **Conditional Expectation Function (CEF)**: $m(X) = \mathbb{E}[Y \mid X]$.
  A potentially non-linear expectation of Y based on X.
- The **Best Linear Predictor (BLP)**: $\beta^* = \arg\min_b \mathbb{E}[(Y - X'b)^2]$.
  The linear projection of Y onto X in the population.
- Later we will use algebra to fit a line in the sample and use probability theory to estimate the BLP.
- Structural (exogeneity) assumptions determine whether the BLP coincides with the CEF.

Properties of Conditional Expectation

- $m(X) \equiv \mathbb{E}[Y \mid X]$ (Definition)
  - Note: $\mathbb{E}[Y|X]$ is a random variable (a function of X), not a number!
- $\mathbb{E}[\mathbb{E}[Y \mid X] \mid X] = \mathbb{E}[Y \mid X]$ (Idempotence)
- $\mathbb{E}[\mathbb{E}[Y \mid X]] = \mathbb{E}[Y]$ (Law of Iterated Expectations)
- $m(X) = \arg\min_g \mathbb{E}[(Y - g(X))^2]$ (Best predictor of Y using X)

CEF Error and Mean Independence

- Define the CEF error: $e = Y - m(X)$, so $Y = m(X) + e$.
- The CEF error has conditional mean zero (by construction):

$$\mathbb{E}[e|X] = \mathbb{E}[Y - m(X)|X] = m(X) - m(X) = 0$$

- And unconditional mean zero (by LIE):

$$\mathbb{E}[e] = \mathbb{E}[\mathbb{E}[e|X]] = 0$$

- $\mathbb{E}[e|X] = 0$ is called *mean independence*—true by construction, this is **not** full independence.

Hansen Theorem 2.4: Properties of the CEF Error e

1. $\mathbb{E}[e|X] = 0$
2. $\mathbb{E}[e] = 0$
3. If the rth moment of Y exists, the rth moment of e exists. [Technical]
4. For *any measurable function* h(X) where the expected covariance with e exists $[\mathbb{E}[|h(X)e|] < \infty]$, then, $\mathbb{E}[h(X)e] = 0$

Key Point

These are properties, not assumptions; they follow from LIE and the definition of the CEF.

Covariance of Estimator and Disturbance

e is a random variable, so we can calculate its covariance with $m(X)$:

$$\mathbb{E}[em(X)] = \mathbb{E}[\mathbb{E}[em(X)|X]] = \mathbb{E}[m(X)\mathbb{E}[e|X]] = \mathbb{E}[m(X) * 0] = 0$$

because $\mathbb{E}(Y|X) = m(X)$ is a measurable function of X, 2.4.4 applies.

$$Cov(e, m(X)) = \mathbb{E}[em(X)] - \mathbb{E}[e]\mathbb{E}[m(X)] = 0 - 0 * m(X) = 0$$

The disturbance is uncorrelated with the conditional expectation. This is what allows us to separate the signal from the noise.

Covariance of Estimator and Disturbance

e is a random variable, so we can calculate its covariance with $m(X)$:

$$\mathbb{E}[em(X)] = \mathbb{E}[\mathbb{E}[em(X)|X]] = \mathbb{E}[m(X)\mathbb{E}[e|X]] = \mathbb{E}[m(X) * 0] = 0$$

because $\mathbb{E}(Y|X) = m(X)$ is a measurable function of X, 2.4.4 applies.

$$Cov(e, m(X)) = \mathbb{E}[em(X)] - \mathbb{E}[e]\mathbb{E}[m(X)] = 0 - 0 * m(X) = 0$$

The disturbance is uncorrelated with the conditional expectation. This is what allows us to separate the signal from the noise.

Covariance of Estimator and Disturbance

e is a random variable, so we can calculate its covariance with $m(X)$:

$$\mathbb{E}[em(X)] = \mathbb{E}[\mathbb{E}[em(X)|X]] = \mathbb{E}[m(X)\mathbb{E}[e|X]] = \mathbb{E}[m(X) * 0] = 0$$

because $\mathbb{E}(Y|X) = m(X)$ is a measurable function of X, 2.4.4 applies.

$$Cov(e, m(X)) = \mathbb{E}[em(X)] - \mathbb{E}[e]\mathbb{E}[m(X)] = 0 - 0 * m(X) = 0$$

The disturbance is uncorrelated with the conditional expectation. This is what allows us to separate the signal from the noise.

Covariance of Estimator and Disturbance

e is a random variable, so we can calculate its covariance with $m(X)$:

$$\mathbb{E}[em(X)] = \mathbb{E}[\mathbb{E}[em(X)|X]] = \mathbb{E}[m(X)\mathbb{E}[e|X]] = \mathbb{E}[m(X) * 0] = 0$$

because $\mathbb{E}(Y|X) = m(X)$ is a measurable function of X, 2.4.4 applies.

$$Cov(e, m(X)) = \mathbb{E}[em(X)] - \mathbb{E}[e]\mathbb{E}[m(X)] = 0 - 0 * m(X) = 0$$

The disturbance is uncorrelated with the conditional expectation. This is what allows us to separate the signal from the noise.

Mean Independence $\neq$ Independence

Let

$$X \sim \text{Uniform}(-1, 1), \qquad Y = X^2 + X\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ and $\varepsilon \perp X$.

Step 1: Compute the CEF

$$m(X) = \mathbb{E}[Y|X] = X^2 + X\mathbb{E}[\varepsilon|X] = X^2.$$

Step 2: CEF error

$$e = Y - m(X) = X\varepsilon.$$

Step 3: Mean independence holds

$$\mathbb{E}[e|X] = \mathbb{E}[X\varepsilon|X] = X\mathbb{E}[\varepsilon|X] = 0.$$

But independence fails:

$$\text{Var}(e|X) = \text{Var}(X\varepsilon|X) = X^2.$$

Mean Independence $\neq$ Independence

Let

$$X \sim \text{Uniform}(-1, 1), \qquad Y = X^2 + X\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ and $\varepsilon \perp X$.

Step 1: Compute the CEF

$$m(X) = \mathbb{E}[Y|X] = X^2 + X\mathbb{E}[\varepsilon|X] = X^2.$$

Step 2: CEF error

$$e = Y - m(X) = X\varepsilon.$$

Step 3: Mean independence holds

$$\mathbb{E}[e|X] = \mathbb{E}[X\varepsilon|X] = X\mathbb{E}[\varepsilon|X] = 0.$$

But independence fails:

$$\text{Var}(e|X) = \text{Var}(X\varepsilon|X) = X^2.$$

Mean Independence $\neq$ Independence

Let

$$X \sim \text{Uniform}(-1, 1), \qquad Y = X^2 + X\varepsilon,$$

where $\varepsilon \sim \mathcal{N}(0, 1)$ and $\varepsilon \perp X$.

Step 1: Compute the CEF

$$m(X) = \mathbb{E}[Y|X] = X^2 + X\mathbb{E}[\varepsilon|X] = X^2.$$

Step 2: CEF error

$$e = Y - m(X) = X\varepsilon.$$

Step 3: Mean independence holds

$$\mathbb{E}[e|X] = \mathbb{E}[X\varepsilon|X] = X\mathbb{E}[\varepsilon|X] = 0.$$

But independence fails:

$$\text{Var}(e|X) = \text{Var}(X\varepsilon|X) = X^2.$$

Variance of the CEF Error

Unconditional:

$$\sigma^2 \equiv Var[e] = \mathbb{E}[e^2]$$

A measure of variation in Y not explained by $m(X)$.

Conditional:

$$\sigma^2(x) = Var[e|X = x] = \mathbb{E}[e^2|X = x]$$

Averaging: $\sigma^2 = \mathbb{E}[\sigma^2(X)]$

Mean-variance representation:

Define $u = \frac{e}{\sigma(X)}$, with $\mathbb{E}[u|X] = 0$, $Var[u|X] = 1$.

$$Y = m(X) + \sigma(X)u$$

If $\sigma(X)$ is constant: homoskedastic.

X matters in two ways: through the mean and through the variance.

Hansen Theorem 2.6

Call $e_1 = Y - m(X_1)$, $e_{12} = Y - m(X_1, X_2)$,

$$var[Y] \geq var[e_1] \geq var[e_{12}]$$

You will always explain (weakly) more of the variance with more variables.

This is why R^2 is a poor measure of model fit.

Proof of Theorem 2.6: Setup

Goal: Show $\mathbb{E}[m(X_1)^2] \leq \mathbb{E}[m(X_1, X_2)^2]$, where $m(\cdot)$ denotes the relevant CEF.

Step 1. By LIE, the coarser CEF is a conditional expectation of the finer one:

$$m(X_1) \;=\; \mathbb{E}[Y|X_1] \;=\; \mathbb{E}\big[\,\mathbb{E}[Y|X_1, X_2] \,\big|\, X_1\big] \;=\; \mathbb{E}\big[\,m(X_1, X_2) \,\big|\, X_1\big]$$

Step 2. Jensen's inequality: for any convex function φ (such as $z \mapsto z^2$),

$$\varphi\big(\mathbb{E}[Z|X_1]\big) \;\leq\; \mathbb{E}\big[\varphi(Z) \,\big|\, X_1\big]$$

Apply with $Z = m(X_1, X_2)$ and $\varphi(z) = z^2$:

$$\Big(\mathbb{E}\big[m(X_1, X_2) \,\big|\, X_1\big]\Big)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \,\big|\, X_1\big]$$

The left side is $m(X_1)^2$ by Step 1, so:

$$m(X_1)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \,\big|\, X_1\big]$$

Proof of Theorem 2.6: Setup

Goal: Show $\mathbb{E}[m(X_1)^2] \leq \mathbb{E}[m(X_1, X_2)^2]$, where $m(\cdot)$ denotes the relevant CEF.

Step 1. By LIE, the coarser CEF is a conditional expectation of the finer one:

$$m(X_1) \;=\; \mathbb{E}[Y|X_1] \;=\; \mathbb{E}\big[\,\mathbb{E}[Y|X_1, X_2] \,\big|\, X_1\big] \;=\; \mathbb{E}\big[\,m(X_1, X_2) \,\big|\, X_1\big]$$

Step 2. Jensen's inequality: for any convex function φ (such as $z \mapsto z^2$),

$$\varphi\big(\mathbb{E}[Z|X_1]\big) \;\leq\; \mathbb{E}\big[\varphi(Z) \,\big|\, X_1\big]$$

Apply with $Z = m(X_1, X_2)$ and $\varphi(z) = z^2$:

$$\Big(\mathbb{E}\big[m(X_1, X_2) \,\big|\, X_1\big]\Big)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \,\big|\, X_1\big]$$

The left side is $m(X_1)^2$ by Step 1, so:

$$m(X_1)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \,\big|\, X_1\big]$$

Proof of Theorem 2.6: Setup

Goal: Show $\mathbb{E}[m(X_1)^2] \leq \mathbb{E}[m(X_1, X_2)^2]$, where $m(\cdot)$ denotes the relevant CEF.

Step 1. By LIE, the coarser CEF is a conditional expectation of the finer one:

$$m(X_1) \;=\; \mathbb{E}[Y|X_1] \;=\; \mathbb{E}\big[\,\mathbb{E}[Y|X_1, X_2] \,\big|\, X_1\big] \;=\; \mathbb{E}\big[\,m(X_1, X_2) \,\big|\, X_1\big]$$

Step 2. Jensen's inequality: for any convex function φ (such as $z \mapsto z^2$),

$$\varphi\big(\mathbb{E}[Z|X_1]\big) \;\leq\; \mathbb{E}\big[\varphi(Z) \,\big|\, X_1\big]$$

Apply with $Z = m(X_1, X_2)$ and $\varphi(z) = z^2$:

$$\Big(\mathbb{E}\big[m(X_1, X_2) \,\big|\, X_1\big]\Big)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \,\big|\, X_1\big]$$

The left side is $m(X_1)^2$ by Step 1, so:

$$m(X_1)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \,\big|\, X_1\big]$$

Proof of Theorem 2.6: Setup

Goal: Show $\mathbb{E}[m(X_1)^2] \leq \mathbb{E}[m(X_1, X_2)^2]$, where $m(\cdot)$ denotes the relevant CEF.

Step 1. By LIE, the coarser CEF is a conditional expectation of the finer one:

$$m(X_1) \;=\; \mathbb{E}[Y|X_1] \;=\; \mathbb{E}\big[\,\mathbb{E}[Y|X_1, X_2] \;\big|\; X_1\big] \;=\; \mathbb{E}\big[\,m(X_1, X_2) \;\big|\; X_1\big]$$

Step 2. Jensen's inequality: for any convex function φ (such as $z \mapsto z^2$),

$$\varphi\big(\mathbb{E}[Z|X_1]\big) \;\leq\; \mathbb{E}\big[\varphi(Z) \;\big|\; X_1\big]$$

Apply with $Z = m(X_1, X_2)$ and $\varphi(z) = z^2$:

$$\Big(\mathbb{E}\big[m(X_1, X_2) \;\big|\; X_1\big]\Big)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \;\big|\; X_1\big]$$

The left side is $m(X_1)^2$ by Step 1, so:

$$m(X_1)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \;\big|\; X_1\big]$$

Proof of Theorem 2.6: Setup

Goal: Show $\mathbb{E}[m(X_1)^2] \leq \mathbb{E}[m(X_1, X_2)^2]$, where $m(\cdot)$ denotes the relevant CEF.

Step 1. By LIE, the coarser CEF is a conditional expectation of the finer one:

$$m(X_1) \;=\; \mathbb{E}[Y|X_1] \;=\; \mathbb{E}\big[\,\mathbb{E}[Y|X_1, X_2] \;\big|\; X_1\big] \;=\; \mathbb{E}\big[\,m(X_1, X_2) \;\big|\; X_1\big]$$

Step 2. Jensen's inequality: for any convex function φ (such as $z \mapsto z^2$),

$$\varphi\big(\mathbb{E}[Z|X_1]\big) \;\leq\; \mathbb{E}\big[\varphi(Z) \;\big|\; X_1\big]$$

Apply with $Z = m(X_1, X_2)$ and $\varphi(z) = z^2$:

$$\Big(\mathbb{E}\big[m(X_1, X_2) \;\big|\; X_1\big]\Big)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \;\big|\; X_1\big]$$

The left side is $m(X_1)^2$ by Step 1, so:

$$m(X_1)^2 \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2 \;\big|\; X_1\big]$$

Proof of Theorem 2.6: Conclusion

Step 3. Take unconditional expectations of both sides:

$$\mathbb{E}\big[m(X_1)^2\big] \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2\big]$$

Step 4. Convert to variances. Recall $\text{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$, and by LIE:

$$\mathbb{E}[m(X_1)] = \mathbb{E}[m(X_1, X_2)] = \mathbb{E}[Y]$$

so $(\mathbb{E}[Z])^2$ is the same for both. Subtracting it from the inequality above:

$$\text{Var}[m(X_1)] \;\leq\; \text{Var}[m(X_1, X_2)]$$

Step 5. Since $\text{Var}[Y] = \text{Var}[m] + \text{Var}[e]$ (by the CEF decomposition), larger $\text{Var}[m]$ means smaller $\text{Var}[e]$:

$$\text{Var}[e_1] \;\geq\; \text{Var}[e_{12}] \qquad \square$$

Proof of Theorem 2.6: Conclusion

Step 3. Take unconditional expectations of both sides:

$$\mathbb{E}\big[m(X_1)^2\big] \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2\big]$$

Step 4. Convert to variances. Recall $\text{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$, and by LIE:

$$\mathbb{E}[m(X_1)] = \mathbb{E}[m(X_1, X_2)] = \mathbb{E}[Y]$$

so $(\mathbb{E}[Z])^2$ is the same for both. Subtracting it from the inequality above:

$$\text{Var}[m(X_1)] \;\leq\; \text{Var}[m(X_1, X_2)]$$

Step 5. Since $\text{Var}[Y] = \text{Var}[m] + \text{Var}[e]$ (by the CEF decomposition), larger $\text{Var}[m]$ means smaller $\text{Var}[e]$:

$$\text{Var}[e_1] \;\geq\; \text{Var}[e_{12}] \qquad \square$$

Proof of Theorem 2.6: Conclusion

Step 3. Take unconditional expectations of both sides:

$$\mathbb{E}\big[m(X_1)^2\big] \;\leq\; \mathbb{E}\big[m(X_1, X_2)^2\big]$$

Step 4. Convert to variances. Recall $\text{Var}[Z] = \mathbb{E}[Z^2] - (\mathbb{E}[Z])^2$, and by LIE:

$$\mathbb{E}[m(X_1)] = \mathbb{E}[m(X_1, X_2)] = \mathbb{E}[Y]$$

so $(\mathbb{E}[Z])^2$ is the same for both. Subtracting it from the inequality above:

$$\text{Var}[m(X_1)] \;\leq\; \text{Var}[m(X_1, X_2)]$$

Step 5. Since $\text{Var}[Y] = \text{Var}[m] + \text{Var}[e]$ (by the CEF decomposition), larger $\text{Var}[m]$ means smaller $\text{Var}[e]$:

$$\text{Var}[e_1] \;\geq\; \text{Var}[e_{12}] \qquad \square$$

Hansen Theorem 2.7: CEF as Best Predictor

- Goal: Minimize mean squared prediction error: $(Y - g(X))^2$. Assumption: $\mathbb{E}[Y^2] < \infty$.

$$
\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(e + m(X) - g(X))^2\right] && \text{(b/c } Y = e + m(X)\text{)} \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2] \\
&= \mathbb{E}[e^2] + 2 \times 0 + \mathbb{E}[(m(X) - g(X))^2] && \text{(by thm 2.4.4.)} \\
&\geq \mathbb{E}[e^2] && \text{(b/c squares are positive.)} \\
&= \mathbb{E}[(Y - m(X))^2] \neq \infty && \text{(by thm 2.4.3.)}
\end{aligned}
$$

Hansen Theorem 2.7: CEF as Best Predictor

- Goal: Minimize mean squared prediction error: $(Y - g(X))^2$. Assumption: $\mathbb{E}[Y^2] < \infty$.

$$\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(e + m(X) - g(X))^2\right] && \text{(b/c } Y = e + m(X)\text{)} \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2] \\
&= \mathbb{E}[e^2] + 2 \times 0 + \mathbb{E}[(m(X) - g(X))^2] && \text{(by thm 2.4.4.)} \\
&\geq \mathbb{E}[e^2] && \text{(b/c squares are positive.)} \\
&= \mathbb{E}[(Y - m(X))^2] \neq \infty && \text{(by thm 2.4.3.)}
\end{aligned}$$

Hansen Theorem 2.7: CEF as Best Predictor

- Goal: Minimize mean squared prediction error: $(Y - g(X))^2$. Assumption: $\mathbb{E}[Y^2] < \infty$.

$$\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(e + m(X) - g(X))^2\right] && \text{(b/c } Y = e + m(X)\text{)} \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2] \\
&= \mathbb{E}[e^2] + 2 \times 0 + \mathbb{E}[(m(X) - g(X))^2] && \text{(by thm 2.4.4.)} \\
&\geq \mathbb{E}[e^2] && \text{(b/c squares are positive.)} \\
&= \mathbb{E}[(Y - m(X))^2] \neq \infty && \text{(by thm 2.4.3.)}
\end{aligned}$$

Hansen Theorem 2.7: CEF as Best Predictor

- Goal: Minimize mean squared prediction error: $(Y - g(X))^2$. Assumption: $\mathbb{E}[Y^2] < \infty$.

$$\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(e + m(X) - g(X))^2\right] && \text{(b/c } Y = e + m(X)\text{)} \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2] \\
&= \mathbb{E}[e^2] + 2 \times 0 + \mathbb{E}[(m(X) - g(X))^2] && \text{(by thm 2.4.4.)} \\
&\geq \mathbb{E}[e^2] && \text{(b/c squares are positive.)} \\
&= \mathbb{E}[(Y - m(X))^2] \neq \infty && \text{(by thm 2.4.3.)}
\end{aligned}$$

Hansen Theorem 2.7: CEF as Best Predictor

- Goal: Minimize mean squared prediction error: $(Y - g(X))^2$. Assumption: $\mathbb{E}[Y^2] < \infty$.

$$
\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(e + m(X) - g(X))^2\right] && \text{(b/c } Y = e + m(X)\text{)} \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2] \\
&= \mathbb{E}[e^2] + 2 \times 0 + \mathbb{E}[(m(X) - g(X))^2] && \text{(by thm 2.4.4.)} \\
&\geq \mathbb{E}[e^2] && \text{(b/c squares are positive.)} \\
&= \mathbb{E}[(Y - m(X))^2] \neq \infty && \text{(by thm 2.4.3.)}
\end{aligned}
$$

Hansen Theorem 2.7: CEF as Best Predictor

- Goal: Minimize mean squared prediction error: $(Y - g(X))^2$. Assumption: $\mathbb{E}[Y^2] < \infty$.

$$
\begin{aligned}
\mathbb{E}\left[(Y - g(X))^2\right] &= \mathbb{E}\left[(e + m(X) - g(X))^2\right] && \text{(b/c } Y = e + m(X)\text{)} \\
&= \mathbb{E}[e^2] + 2\mathbb{E}[e(m(X) - g(X))] + \mathbb{E}[(m(X) - g(X))^2] \\
&= \mathbb{E}[e^2] + 2 \times 0 + \mathbb{E}[(m(X) - g(X))^2] && \text{(by thm 2.4.4.)} \\
&\geq \mathbb{E}[e^2] && \text{(b/c squares are positive.)} \\
&= \mathbb{E}[(Y - m(X))^2] \neq \infty && \text{(by thm 2.4.3.)}
\end{aligned}
$$

Implications of Theorem 2.7

- Regardless of the distribution of X and Y, the best predictor is the CEF $m(X)$.
  - For example: Consider the intercept only model $\mu = \mathbb{E}[Y]$.
  - The best predictor for Y, among constants, is μ.
- Since the CEF is the best predictor, anything else must be identical or worse.
- We don't know what the joint distribution is, so we don't generally know the CEF, we have to use an estimator.

Linear CEF and Best Prediction

If m(x) is *linear* in X, we can write $m(X) = X'\beta$. The **linear regression model** is:

$$Y = X'\beta + e, \quad \mathbb{E}[e|X] = 0$$

If homoskedastic: $\mathbb{E}[e^2|X] = \sigma^2$.

Best prediction property:

- If the CEF is linear, the solution to $\beta = \arg\min_b \mathbb{E}[(Y - X'b)^2]$ gives the right formula for β.
- If the CEF is nonlinear, the solution gives the best linear approximation to the true CEF.

Linear CEF and Best Prediction

If m(x) is *linear* in X, we can write $m(X) = X'\beta$. The **linear regression model** is:

$$Y = X'\beta + e, \quad \mathbb{E}[e|X] = 0$$

If homoskedastic: $\mathbb{E}[e^2|X] = \sigma^2$.

Best prediction property:

- If the CEF is linear, the solution to $\beta = \arg\min_b \mathbb{E}[(Y - X'b)^2]$ gives the right formula for β.
- If the CEF is nonlinear, the solution gives the best linear approximation to the true CEF.

From CEF to Linear Predictors

- The CEF provides: summarization, MMSE prediction, and ANOVA decomposition.
- But the CEF requires knowing the conditional distribution $f_{Y|X}(y|x)$.
  - Exception: CEF is always linear if X is discrete and all interactions are included.
- We will instead approximate the CEF using a **Best Linear Predictor** (BLP).
- Later we will show that OLS estimates are optimal **if** the CEF is linear and the residuals are homogeneous.

From CEF to Linear Predictors

- The CEF provides: summarization, MMSE prediction, and ANOVA decomposition.
- But the CEF requires knowing the conditional distribution $f_{Y|X}(y|x)$.
  - Exception: CEF is always linear if X is discrete and all interactions are included.
- We will instead approximate the CEF using a **Best Linear Predictor** (BLP).
- Later we will show that OLS estimates are optimal **if** the CEF is linear and the residuals are homogeneous.

From CEF to Linear Predictors

- The CEF provides: summarization, MMSE prediction, and ANOVA decomposition.
- But the CEF requires knowing the conditional distribution $f_{Y|X}(y|x)$.
  - Exception: CEF is always linear if X is discrete and all interactions are included.
- We will instead approximate the CEF using a **Best Linear Predictor** (BLP).
- Later we will show that OLS estimates are optimal **if** the CEF is linear and the residuals are homogeneous.

The CEF in Political Science

- **Voter turnout and age:** $\mathbb{E}[\text{Turnout}|\text{Age}]$ is nonlinear—turnout rises steeply from 18–30, plateaus in middle age, and rises again among retirees.
- **Income and party ID:** $\mathbb{E}[\text{Income}|\text{Party}]$ differs across categories—the CEF is just group means when X is discrete.
- **Conflict and GDP:** $\mathbb{E}[\text{Conflict}|\text{GDP}]$ may be highly nonlinear, with sharp thresholds at low GDP.

Takeaway: The CEF is the target. But when the relationship is complex, we approximate it with a linear predictor and interpret accordingly.

The CEF in Political Science

- **Voter turnout and age:** $\mathbb{E}[\text{Turnout}|\text{Age}]$ is nonlinear—turnout rises steeply from 18–30, plateaus in middle age, and rises again among retirees.
- **Income and party ID:** $\mathbb{E}[\text{Income}|\text{Party}]$ differs across categories—the CEF is just group means when X is discrete.
- **Conflict and GDP:** $\mathbb{E}[\text{Conflict}|\text{GDP}]$ may be highly nonlinear, with sharp thresholds at low GDP.

Takeaway: The CEF is the target. But when the relationship is complex, we approximate it with a linear predictor and interpret accordingly.

The CEF in Political Science

- **Voter turnout and age:** $\mathbb{E}[\text{Turnout}|\text{Age}]$ is nonlinear—turnout rises steeply from 18–30, plateaus in middle age, and rises again among retirees.
- **Income and party ID:** $\mathbb{E}[\text{Income}|\text{Party}]$ differs across categories—the CEF is just group means when X is discrete.
- **Conflict and GDP:** $\mathbb{E}[\text{Conflict}|\text{GDP}]$ may be highly nonlinear, with sharp thresholds at low GDP.

Takeaway: The CEF is the target. But when the relationship is complex, we approximate it with a linear predictor and interpret accordingly.

What Have We Learned About the CEF?

CEF Summary

1. $m(X) = \mathbb{E}[Y|X]$ is the **best predictor** of Y given X (Thm 2.7).
2. The CEF error $e = Y - m(X)$ satisfies $\mathbb{E}[e|X] = 0$ **by construction**—not an assumption (Thm 2.4).
3. Signal and noise are uncorrelated: $\text{Cov}(m(X), e) = 0$.
4. More variables $\Rightarrow$ (weakly) smaller error variance (Thm 2.6).
5. $Y = m(X) + \sigma(X)u$: X shapes both the mean *and* the variance.

Next: We approximate $m(X)$ with a *linear* function—the Best Linear Predictor.

(Best) Linear Predictors (bivariate case)

- Suppose we want a predictor $\mathbb{E}^*(Y|X)$ that is **linear**:

$$f(X) = a + bX$$

Our standard for prediction is to minimize the mean-square error (best):

- Whereas the CEF might be infinitely complex, the BLP is characterized just by two numbers, a and b.
  Choose a and b to minimize

$$M = \mathbb{E}[(Y - (a + bX))^2]$$

(Best) Linear Predictors (bivariate case)

- Suppose we want a predictor $\mathbb{E}^*(Y|X)$ that is **linear**:

$$f(X) = a + bX$$

Our standard for prediction is to minimize the mean-square error (best):

- Whereas the CEF might be infinitely complex, the BLP is characterized just by two numbers, a and b.

Choose a and b to minimize

$$M = \mathbb{E}[(Y - (a + bX))^2]$$

(Best) Linear Predictors (bivariate case)

- Suppose we want a predictor $\mathbb{E}^*(Y|X)$ that is **linear**:

$$f(X) = a + bX$$

  Our standard for prediction is to minimize the mean-square error (best):
- Whereas the CEF might be infinitely complex, the BLP is characterized just by two numbers, a and b.
  Choose a and b to minimize

$$M = \mathbb{E}[(Y - (a + bX))^2]$$

Solving for Best Linear Predictor (a and b)

FOC w.r.t. a:

$$0 = -2\mathbb{E}[Y] + 2\mathbb{E}(a + bX) \implies a = \mathbb{E}(Y) - b\mathbb{E}(X)$$

FOC w.r.t. b:

$$\begin{aligned} 0 &= -2\mathbb{E}[YX] + 2\mathbb{E}[(a + bX)(X)] \\ \mathbb{E}(YX) &= [\mathbb{E}(Y) - b\mathbb{E}(X)]\mathbb{E}(X) + b\mathbb{E}(X^2) \\ \mathbb{E}(YX) - \mathbb{E}(Y)E(X) &= b[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \\ b &= \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

Solving for Best Linear Predictor (a and b)

FOC w.r.t. a:

$$0 = -2\mathbb{E}[Y] + 2\mathbb{E}(a + bX) \implies a = \mathbb{E}(Y) - b\mathbb{E}(X)$$

FOC w.r.t. b:

$$\begin{aligned} 0 &= -2\mathbb{E}[YX] + 2\mathbb{E}[(a + bX)(X)] \\ \mathbb{E}(YX) &= [\mathbb{E}(Y) - b\mathbb{E}(X)]\mathbb{E}(X) + b\mathbb{E}(X^2) \\ \mathbb{E}(YX) - \mathbb{E}(Y)E(X) &= b[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \\ b &= \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

Solving for Best Linear Predictor (a and b)

FOC w.r.t. a:

$$0 = -2\mathbb{E}[Y] + 2\mathbb{E}(a + bX) \implies a = \mathbb{E}(Y) - b\mathbb{E}(X)$$

FOC w.r.t. b:

$$\begin{aligned} 0 &= -2\mathbb{E}[YX] + 2\mathbb{E}[(a + bX)(X)] \\ \mathbb{E}(YX) &= [\mathbb{E}(Y) - b\mathbb{E}(X)]\mathbb{E}(X) + b\mathbb{E}(X^2) \\ \mathbb{E}(YX) - \mathbb{E}(Y)E(X) &= b[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \\ b &= \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

Solving for Best Linear Predictor (a and b)

FOC w.r.t. a:

$$0 = -2\mathbb{E}[Y] + 2\mathbb{E}(a + bX) \implies a = \mathbb{E}(Y) - b\mathbb{E}(X)$$

FOC w.r.t. b:

$$0 = -2\mathbb{E}[YX] + 2\mathbb{E}[(a + bX)(X)]$$
$$\mathbb{E}(YX) = [\mathbb{E}(Y) - b\mathbb{E}(X)]\mathbb{E}(X) + b\mathbb{E}(X^2)$$
$$\mathbb{E}(YX) - \mathbb{E}(Y)E(X) = b[\mathbb{E}(X^2) - \mathbb{E}(X)^2]$$
$$b = \frac{Cov(X, Y)}{Var(X)}$$

Solving for Best Linear Predictor (a and b)

FOC w.r.t. a:

$$0 = -2\mathbb{E}[Y] + 2\mathbb{E}(a + bX) \implies a = \mathbb{E}(Y) - b\mathbb{E}(X)$$

FOC w.r.t. b:

$$\begin{aligned} 0 &= -2\mathbb{E}[YX] + 2\mathbb{E}[(a + bX)(X)] \\ \mathbb{E}(YX) &= [\mathbb{E}(Y) - b\mathbb{E}(X)]\mathbb{E}(X) + b\mathbb{E}(X^2) \\ \mathbb{E}(YX) - \mathbb{E}(Y)E(X) &= b[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \\ b &= \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

Solving for Best Linear Predictor (a and b)

FOC w.r.t. a:

$$0 = -2\mathbb{E}[Y] + 2\mathbb{E}(a + bX) \implies a = \mathbb{E}(Y) - b\mathbb{E}(X)$$

FOC w.r.t. b:

$$\begin{aligned} 0 &= -2\mathbb{E}[YX] + 2\mathbb{E}[(a + bX)(X)] \\ \mathbb{E}(YX) &= [\mathbb{E}(Y) - b\mathbb{E}(X)]\mathbb{E}(X) + b\mathbb{E}(X^2) \\ \mathbb{E}(YX) - \mathbb{E}(Y)E(X) &= b[\mathbb{E}(X^2) - \mathbb{E}(X)^2] \\ b &= \frac{Cov(X, Y)}{Var(X)} \end{aligned}$$

Best Linear Predictor

The best linear predictor is the population linear projection:

$$\mathbb{E}^*[Y|X] = \beta_0 + \beta_1 X$$

$$Y = \beta_0 + \beta_1 X + e$$

- Parameters (in Greek)
  - The slope $\beta = \frac{Cov(X,Y)}{Var(X)}$
  - The intercept $\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X)$
- $\mathcal{P}(Y|X) = \mathbb{E}(Y) + \frac{Cov(X,Y)}{Var(X)} * (X - \mathbb{E}(X))$

Best Linear Predictor

The best linear predictor is the population linear projection:

$$\mathbb{E}^*[Y|X] = \beta_0 + \beta_1 X$$
$$Y = \beta_0 + \beta_1 X + e$$

- Parameters (in Greek)
  - The slope $\beta = \frac{Cov(X,Y)}{Var(X)}$
  - The intercept $\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X)$
- $\mathcal{P}(Y|X) = \mathbb{E}(Y) + \frac{Cov(X,Y)}{Var(X)} * (X - \mathbb{E}(X))$

Best Linear Predictor

The best linear predictor is the population linear projection:

$$\mathbb{E}^*[Y|X] = \beta_0 + \beta_1 X$$
$$Y = \beta_0 + \beta_1 X + e$$

- Parameters (in Greek)
  - The slope $\beta = \frac{Cov(X,Y)}{Var(X)}$
  - The intercept $\alpha = \mathbb{E}(Y) - \beta\mathbb{E}(X)$
- $\mathcal{P}(Y|X) = \mathbb{E}(Y) + \frac{Cov(X,Y)}{Var(X)} * (X - \mathbb{E}(X))$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X,Y)}{Var(X)} \qquad \beta Var(X) = Cov(X,Y)$$

$$\begin{aligned}
\mathbb{E}[(Y-(\alpha+\beta X))^2] - \mathbb{E}[Y-(\alpha+\beta X)]^2 &= Var(Y-(\alpha+\beta X)) \\
&= Var(Y-\beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X,Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X, Y)}{Var(X)} \qquad \beta Var(X) = Cov(X, Y)$$

$$\begin{aligned}
\mathbb{E}[(Y - (\alpha + \beta X))^2] - \mathbb{E}[Y - (\alpha + \beta X)]^2 &= Var(Y - (\alpha + \beta X)) \\
&= Var(Y - \beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X, Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X, Y)}{Var(X)} \qquad \beta Var(X) = Cov(X, Y)$$

$$\begin{aligned}
\mathbb{E}[(Y - (\alpha + \beta X))^2] - \mathbb{E}[Y - (\alpha + \beta X)]^2 &= Var(Y - (\alpha + \beta X)) \\
&= Var(Y - \beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X, Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X,Y)}{Var(X)} \qquad \beta Var(X) = Cov(X,Y)$$

$$\begin{aligned}
\mathbb{E}[(Y-(\alpha+\beta X))^2] - \mathbb{E}[Y-(\alpha+\beta X)]^2 &= Var(Y-(\alpha+\beta X)) \\
&= Var(Y-\beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X,Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X,Y)}{Var(X)} \qquad \beta Var(X) = Cov(X,Y)$$

$$\begin{aligned}
\mathbb{E}[(Y-(\alpha+\beta X))^2] - \mathbb{E}[Y-(\alpha+\beta X)]^2 &= Var(Y-(\alpha+\beta X)) \\
&= Var(Y-\beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X,Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X,Y)}{Var(X)} \qquad \beta Var(X) = Cov(X,Y)$$

$$\begin{aligned}
\mathbb{E}[(Y-(\alpha+\beta X))^2] - \mathbb{E}[Y-(\alpha+\beta X)]^2 &= Var(Y-(\alpha+\beta X)) \\
&= Var(Y-\beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X,Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Deriving Variance of BLP at Minimized Values

$$\beta = \frac{Cov(X,Y)}{Var(X)} \qquad \beta Var(X) = Cov(X,Y)$$

$$\begin{aligned}
\mathbb{E}[(Y-(\alpha+\beta X))^2] - \mathbb{E}[Y-(\alpha+\beta X)]^2 &= Var(Y-(\alpha+\beta X)) \\
&= Var(Y-\beta X) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta Cov(X,Y) \\
&= Var(Y) + \beta^2 Var(X) - 2\beta^2 Var(X) \\
&= Var(Y) - \beta^2 Var(X) \\
\sigma_\epsilon^2 &= \sigma_Y^2 - \beta^2 \sigma_X^2
\end{aligned}$$

Example: Population Linear Projection is not the CEF

Example: $Y = X + X^2$, $X \sim N(0,1)$, $\mathbb{E}[X] = \mathbb{E}[X^3] = 0$, $\mathbb{E}[X^2] = 1$
True CEF: $m(X) = X + X^2$
Population Linear Projection: $Y = \alpha + \beta X + e$.

$$\alpha = E(Y) - \beta E(X) = E(X + X^2) - \beta * 0 = 0 + 1 - 0$$
$$\beta = Cov(X, Y)/Var(X) = Cov(X, X + X^2)/1 = 1/1$$
$$\mathcal{P}(Y|X) = 1 + 1 * X$$
$$e = m(X) - \mathcal{P}(Y|X) = X^2 - 1$$
$$\mathbb{E}[eX] = \mathbb{E}[(X^2 - 1)X] = \mathbb{E}[X^3 - X] = 0$$

The projection error e is a function of X, and $\mathbb{E}[e|X] \neq 0$.

Example: Population Linear Projection is not the CEF

Example: $Y = X + X^2$, $X \sim N(0,1)$, $\mathbb{E}[X] = \mathbb{E}[X^3] = 0$, $\mathbb{E}[X^2] = 1$
True CEF: $m(X) = X + X^2$
Population Linear Projection: $Y = \alpha + \beta X + e$.

$$\begin{aligned}
\alpha &= E(Y) - \beta E(X) = E(X + X^2) - \beta * 0 = 0 + 1 - 0 \\
\beta &= Cov(X, Y)/Var(X) = Cov(X, X + X^2)/1 = 1/1 \\
\mathcal{P}(Y|X) &= 1 + 1 * X \\
e &= m(X) - \mathcal{P}(Y|X) = X^2 - 1 \\
\mathbb{E}[eX] &= \mathbb{E}[(X^2 - 1)X] = \mathbb{E}[X^3 - X] = 0
\end{aligned}$$

The projection error e is a function of X, and $\mathbb{E}[e|X] \neq 0$.

CEF vs Linear Predictors / Projection

- CEF ($m(X)$)
  - Best prediction, the prediction that minimizes squared error.
  - Can be non-linear.
  - Requires knowing joint distribution.
  - Prediction Errors are uncorrelated with any function of X. ($E(e|X) = 0.$)
- Linear Predictor/ Projection ($\mathcal{P}(Y|X)$)
  - Draws a line (plane/hyperplane)
  - Requires knowing variances and covariances.
  - Prediction Errors are uncorrelated with X ($E(Xe) = 0.$)

CEF vs BLP: The Big Picture

Why Settle for Linear?

- The CEF is the *best possible* predictor—but we rarely know it.
- The BLP is the best *linear* predictor—and we only need $\text{Cov}(X, Y)$ and $\text{Var}(X)$.
- When the CEF is actually linear (e.g., jointly normal X, Y), BLP = CEF.
- When the CEF is nonlinear, BLP gives the best linear approximation.

Key insight: $\mathbb{E}[Xe] = 0$ (BLP errors uncorrelated with X) is *weaker* than $\mathbb{E}[e|X] = 0$ (CEF errors mean-independent of X). The BLP projection "uses up" the linear information in X, but may leave nonlinear patterns in the residual.

CEF vs BLP: The Big Picture

Why Settle for Linear?

- The CEF is the *best possible* predictor—but we rarely know it.
- The BLP is the best *linear* predictor—and we only need $\text{Cov}(X, Y)$ and $\text{Var}(X)$.
- When the CEF is actually linear (e.g., jointly normal X, Y), BLP = CEF.
- When the CEF is nonlinear, BLP gives the best linear approximation.

Key insight: $\mathbb{E}[Xe] = 0$ (BLP errors uncorrelated with X) is *weaker* than $\mathbb{E}[e|X] = 0$ (CEF errors mean-independent of X). The BLP projection "uses up" the linear information in X, but may leave nonlinear patterns in the residual.

BLP in R: From Formula to Code

```
OLS estimates the same thing from a sample mod <- lm(y
~ x) coef(mod)  compare to c(alpha_formula, beta_formula)
With multiple regressors:  beta =
```

$solve(Q_X X)X < -cbind(1, x1, x2)beta_matrix < -solve(t(X)coef(lm(y \tilde{} x1 + x2))sameanswer$

Indicator (Dummy) Variables

- Our event space Ω is often binary or categorical (or coded to be so).
- Examples: Age bins (18-34), Party (R, D), Country (USA, Japan).

Definition

Suppose A is an event. Define $\mathbb{I} = 1$ if outcome in A, $\mathbb{I} = 0$ otherwise. Then $E[\mathbb{I}(\omega \in A)] = P(A)$.

CEF with a dummy variable:

$$\mathbb{E}[Y|A] = \alpha + \delta D_i + e$$

Indicator Variables: Interpretation

$$\begin{aligned} \mathbb{E}[y_i|\omega_i \in A] &= \beta_0 + \beta_1 \\ \mathbb{E}[y_i|\omega_i \notin A] &= \beta_0 \\ \mathbb{E}[y_i|\omega_i \in A] - E[y_i|\omega_i \notin A] &= \beta_1 \end{aligned}$$

Parallel Slopes

- With a continuous independent variable, adding an indicator control variable is called the "parallel slopes" model.

$$y_i = \beta_0 + \delta d_i + \beta_1 x_i + e_i$$

- The effect of x_i is the same for each group.

Unordered Categorical Variables

- Suppose we have a model of multi-category data.
- For example, ethnic identification $x_i \in \{\text{Kazakh, Russian, Other}\}$

$$d_{2i} = \begin{cases} 1 \text{ if } x_i = \text{Kazakh} \\ 0 \text{ otherwise} \end{cases} \qquad d_3 = \begin{cases} 1 \text{ if } x_i = \text{Russian} \\ 0 \text{ otherwise} \end{cases}$$

- Our estimation becomes

$$y_i = b_1 + b_2 d_{2i} + b_3 d_{3i} + e_i$$

- We can interpret $b_1,\ b_2,\ b_3$ as the difference in means relative to baseline.

Dummies and Interactions in R

```
Interaction:  does party moderate the economy effect?  mod2 <- lm(approval
~ party * economy, data = df)  party * economy expands to:  party + economy + party:economy
Marginal effect of economy for Democrats:  coef(mod2)["economy"] +
    coef(mod2)["partyD:economy"]
```

Interactions with Dummy Variables

- Suppose now we have a model with different averages and different responses to x

$$\text{Group 1: } y_i = \mu + \beta_1 x_i + e_i,$$
$$\text{Group 2: } y_i = \mu + \delta + (\beta_1 + \gamma) x_i + e_i,$$

- Again, we can use a dummy variable $d_i = 0$ if i is in group 1, $d_i = 1$ if i is in group 2.

$$y_i = \mu + \delta d_i + \beta_1 x_i + \gamma d_i x_i + e_i$$

- $d_i x_i$ is called an interaction term.

Interactions Generally and Interpretation

- With continuous moderator z:

$$y_i = \mu + \delta z_i + \beta_1 x_i + \gamma z_i * x_i + e_i$$

- An interaction conditions the effect of x with z:

$$\frac{\partial E[y_i|X]}{\partial x_i} = \beta_1 + \gamma z_i$$

- A one unit increase in x_i produces a $\beta_1 + \gamma z_i$ unit increase in y_i.
- β_1 is the effect when $z_i = 0$.

Interactions Generally and Interpretation

- With continuous moderator z:

$$y_i = \mu + \delta z_i + \beta_1 x_i + \gamma z_i * x_i + e_i$$

- An interaction conditions the effect of x with z:

$$\frac{\partial E[y_i|X]}{\partial x_i} = \beta_1 + \gamma z_i$$

- A one unit increase in x_i produces a $\beta_1 + \gamma z_i$ unit increase in y_i.
- β_1 is the effect when $z_i = 0$.

Data Transformations to Ease Interpretation

- You can demean continuous variables.

$$\begin{aligned}\frac{\partial E[y_i|X]}{\partial x_i} &= \frac{\partial}{\partial x_i}[\mu + \delta(z_i - \bar{z}) + \beta_1(x_i - \bar{x}) + \gamma(z_i - \bar{z}) * (x_i - \bar{x})] \\ &= \beta_1 + \gamma(z_i - \bar{z})\end{aligned}$$

- A one unit increase in x_i produces a $\beta_1 + \gamma(z_i - \bar{z})$ unit increase in y_i.
- But now β_1 is the average marginal effect (when $z_i = \bar{z}_i$)

Regularity Conditions and MSPE

- Given Y and $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ 1 \end{pmatrix}$, we need:
  1. $\mathbb{E}[Y^2] < \infty$
  2. $\mathbb{E}||X||^2 < \infty$
  3. $\mathbf{Q}_{XX} = \mathbb{E}[XX']$ is positive definite.
- A linear predictor: $X'\beta = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \beta_4$
- The mean square prediction error is $S(\beta) = \mathbb{E}[(Y - X'\beta)^2]$.
- The **best linear predictor** $\mathcal{P}[Y|X]$ minimizes $S(\beta)$.

Regularity Conditions and MSPE

- Given Y and $X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ 1 \end{pmatrix}$, we need:
  1. $\mathbb{E}[Y^2] < \infty$
  2. $\mathbb{E}||X||^2 < \infty$
  3. $\mathbf{Q}_{XX} = \mathbb{E}[XX']$ is positive definite.
- A linear predictor: $X'\beta = X_1\beta_1 + X_2\beta_2 + X_3\beta_3 + \beta_4$
- The mean square prediction error is $S(\beta) = \mathbb{E}[(Y - X'\beta)^2]$.
- The **best linear predictor** $\mathcal{P}[Y|X]$ minimizes $S(\beta)$.

Deriving the Linear Projection Coefficient β

$$
\begin{aligned}
S(\beta) &= \mathbb{E}[(Y - X'\beta)(Y - X'\beta)] \\
&= \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta \\
0 &= \frac{\partial}{\partial \beta} S(\beta) = -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\
\mathbf{Q}_{XY} &= \mathbf{Q}_{XX}\beta \\
\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} &= \beta \\
\mathbb{E}[XX']^{-1}\mathbb{E}[XY] &= \beta
\end{aligned}
$$

Deriving the Linear Projection Coefficient β

$$
\begin{aligned}
S(\beta) &= \mathbb{E}[(Y - X'\beta)(Y - X'\beta)] \\
&= \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta \\
0 &= \frac{\partial}{\partial \beta} S(\beta) = -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\
\mathbf{Q}_{XY} &= \mathbf{Q}_{XX}\beta \\
\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} &= \beta \\
\mathbb{E}[XX']^{-1}\mathbb{E}[XY] &= \beta
\end{aligned}
$$

Deriving the Linear Projection Coefficient β

$$
\begin{aligned}
S(\beta) &= \mathbb{E}[(Y - X'\beta)(Y - X'\beta)] \\
&= \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta \\
0 = \frac{\partial}{\partial \beta} S(\beta) &= -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\
\mathbf{Q}_{XY} &= \mathbf{Q}_{XX}\beta \\
\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} &= \beta \\
\mathbb{E}[XX']^{-1}\mathbb{E}[XY] &= \beta
\end{aligned}
$$

Deriving the Linear Projection Coefficient β

$$
\begin{aligned}
S(\beta) &= \mathbb{E}[(Y - X'\beta)(Y - X'\beta)] \\
&= \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta \\
0 = \frac{\partial}{\partial \beta} S(\beta) &= -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\
\mathbf{Q}_{XY} &= \mathbf{Q}_{XX}\beta \\
\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} &= \beta \\
\mathbb{E}[XX']^{-1}\mathbb{E}[XY] &= \beta
\end{aligned}
$$

Deriving the Linear Projection Coefficient β

$$
\begin{aligned}
S(\beta) &= \mathbb{E}[(Y - X'\beta)(Y - X'\beta)] \\
&= \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta \\
0 &= \frac{\partial}{\partial \beta} S(\beta) = -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\
\mathbf{Q}_{XY} &= \mathbf{Q}_{XX}\beta \\
\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} &= \beta \\
\mathbb{E}[XX']^{-1}\mathbb{E}[XY] &= \beta
\end{aligned}
$$

Deriving the Linear Projection Coefficient β

$$
\begin{aligned}
S(\beta) &= \mathbb{E}[(Y - X'\beta)(Y - X'\beta)] \\
&= \mathbb{E}[Y^2] - 2\beta'\mathbb{E}[XY] + \beta'\mathbb{E}[XX']\beta \\
0 = \frac{\partial}{\partial \beta} S(\beta) &= -2\mathbb{E}[XY] + 2\mathbb{E}[XX']\beta \\
\mathbf{Q}_{XY} &= \mathbf{Q}_{XX}\beta \\
\mathbf{Q}_{XX}^{-1}\mathbf{Q}_{XY} &= \beta \\
\mathbb{E}[XX']^{-1}\mathbb{E}[XY] &= \beta
\end{aligned}
$$

Linear Projection and Projection Error

$$\mathcal{P}[Y|X] = X'\beta = X'\mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

The projection error is $e = Y - X'\beta$, so $Y = X'\beta + e$.

The error is uncorrelated with X:

$$\mathbb{E}[Xe] = \mathbb{E}[XY] - \mathbb{E}[XX']\mathbb{E}[XX']^{-1}\mathbb{E}[XY] = 0$$

$$\begin{pmatrix} \mathbb{E}[X_1 e] \\ \mathbb{E}[X_2 e] \\ \vdots \\ \mathbb{E}[1 * e] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

This is k equations, the last of which is $\mathbb{E}[e] = 0$ when there is a constant.

Linear Projection and Projection Error

$$\mathcal{P}[Y|X] = X'\beta = X'\mathbb{E}[XX']^{-1}\mathbb{E}[XY]$$

The projection error is $e = Y - X'\beta$, so $Y = X'\beta + e$.

The error is uncorrelated with X:

$$\mathbb{E}[Xe] = \mathbb{E}[XY] - \mathbb{E}[XX']\mathbb{E}[XX']^{-1}\mathbb{E}[XY] = 0$$

$$\begin{pmatrix} \mathbb{E}[X_1 e] \\ \mathbb{E}[X_2 e] \\ \vdots \\ \mathbb{E}[1 * e] \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

This is k equations, the last of which is $\mathbb{E}[e] = 0$ when there is a constant.

The Design Matrix

$\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$ exists and is unique *if* $\mathbf{Q}_{XX} = \mathbb{E}[XX']$ is invertible.

$\mathbf{Q}_{XX}$ is called the design matrix.

For any non-zero $\alpha \in \mathbb{R}^k$

$$\alpha' \boldsymbol{Q}_{XX} \alpha = \mathbb{E}[\alpha' XX' \alpha] = \mathbb{E}[(\alpha' X)^2] \geq 0$$

This must be strictly > 0 to be invertible, if not, there is no unique solution to $\mathbf{Q}_{XY} = \mathbf{Q}_{XX}\beta$, and we say β is not identified.

Application: Linear Probability Model (Binary Y)

$$P(Y = 1|X) = \beta_0 + \beta_1 X$$

- Suppose we have a binary outcome: $Y = \mathbb{I}(\omega \in B)$
- β_0 is the probability that $Y = 1$ if $X = 0$
- β_1 is the the change in probability that $Y = 1$ given a one unit change in X.
- The BLP is still the best linear predictor: $\beta = \mathbb{E}[XX']^{-1}\mathbb{E}[XY]$.
- However, it is an approximation to a probability, not itself a probability.

Linear Predictor Error Variance

- Hansen defines $\sigma^2 = \mathbb{E}[e^2]$, $Q_{YY} = \mathbb{E}[Y^2]$ and $\mathbf{Q}_{YX} = \mathbb{E}[YX']$, a $1 \times k$ vector.
- Recall, X and β are $k \times 1$ vectors.

$$
\begin{aligned}
\sigma^2 &= \mathbb{E}[(Y - X'\beta)^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\equiv Q_{YY \cdot X}
\end{aligned}
$$

Linear Predictor Error Variance

- Hansen defines $\sigma^2 = \mathbb{E}[e^2]$, $Q_{YY} = \mathbb{E}[Y^2]$ and $\mathbf{Q}_{YX} = \mathbb{E}[YX']$, a $1 \times k$ vector.
- Recall, X and β are $k \times 1$ vectors.

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[(Y - X'\beta)^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\equiv Q_{YY \cdot X}
\end{aligned}$$

Linear Predictor Error Variance

- Hansen defines $\sigma^2 = \mathbb{E}[e^2]$, $Q_{YY} = \mathbb{E}[Y^2]$ and $\mathbf{Q}_{YX} = \mathbb{E}[YX']$, a $1 \times k$ vector.
- Recall, X and β are $k \times 1$ vectors.

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[(Y - X'\beta)^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\equiv Q_{YY \cdot X}
\end{aligned}$$

Linear Predictor Error Variance

- Hansen defines $\sigma^2 = \mathbb{E}[e^2]$, $Q_{YY} = \mathbb{E}[Y^2]$ and $\mathbf{Q}_{YX} = \mathbb{E}[YX']$, a $1 \times k$ vector.
- Recall, X and β are $k \times 1$ vectors.

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[(Y - X'\beta)^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\equiv Q_{YY \cdot X}
\end{aligned}$$

Linear Predictor Error Variance

- Hansen defines $\sigma^2 = \mathbb{E}[e^2]$, $Q_{YY} = \mathbb{E}[Y^2]$ and $\mathbf{Q}_{YX} = \mathbb{E}[YX']$, a $1 \times k$ vector.
- Recall, X and β are $k \times 1$ vectors.

$$\begin{aligned}
\sigma^2 &= \mathbb{E}[(Y - X'\beta)^2] \\
&= \mathbb{E}[Y^2] - 2\mathbb{E}[YX']\beta + \beta'\mathbb{E}[XX']\beta \\
&= Q_{YY} - 2\boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} + \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&= Q_{YY} - \boldsymbol{Q}_{YX}\boldsymbol{Q}_{XX}^{-1}\boldsymbol{Q}_{XY} \\
&\equiv Q_{YY \cdot X}
\end{aligned}$$

Intercepts and Slopes

- We will often exclude the 1 from X and write the linear projection as:

$$Y = X'\beta + \alpha + e$$

- Demeaning we have a useful reinterpretation of the projection coefficient:

$$\begin{aligned} Y - \mu_Y &= (X - \mu_X)'\beta + e \\ \beta &= (\mathbb{E}[(X - \mu_X)(X - \mu_X)'])^{-1}\mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\ \beta &= var[X]^{-1} cov(X, Y) \end{aligned}$$

Partitions and Regression Sub-Vectors

- Divide our regressors into $X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}$, so the linear projection is:

$$Y = X_1'\beta_1 + X_2'\beta_2 + e, \quad \mathbb{E}[Xe] = 0$$

- We can partition $\mathbf{Q}_{XX}$ and $\mathbf{Q}_{XY}$:

$$\mathbf{Q}_{XX} = \begin{bmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{bmatrix} \quad \mathbf{Q}_{XY} = \begin{bmatrix} \mathbf{Q}_{1Y} \\ \mathbf{Q}_{2Y} \end{bmatrix}$$

A Useful Formula

- Recall the error variance was: $Q_{YY \cdot X} \equiv Q_{YY} - \boldsymbol{Q}_{YX} \boldsymbol{Q}_{XX}^{-1} \boldsymbol{Q}_{XY}$
- Analogously $\mathbf{Q}_{11 \cdot 2} \equiv \mathbf{Q}_{11} - \boldsymbol{Q}_{12} \boldsymbol{Q}_{22}^{-1} \boldsymbol{Q}_{21}$,and $\mathbf{Q}_{22 \cdot 1} \equiv \mathbf{Q}_{22} - \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12}$, etc.
- In each case, $\mathbf{Q}_{a \cdot b}$ is the variation in a not predicted by the linear projection of b.

Partitioned Matrix Inversion and Regression

- By similar reasoning:

$$\boldsymbol{Q}_{XX}^{-1} = \begin{bmatrix} \mathbf{Q}_{11\cdot 2}^{-1} & -\mathbf{Q}_{11\cdot 2}^{-1}\mathbf{Q}_{12}\mathbf{Q}_{22}^{-1} \\ -\mathbf{Q}_{22\cdot 1}^{-1}\mathbf{Q}_{21}\mathbf{Q}_{11}^{-1} & \mathbf{Q}_{22\cdot 1}^{-1} \end{bmatrix}$$

- As a result:

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \mathbf{Q}_{11\cdot 2}^{-1}\mathbf{Q}_{1Y\cdot 2} \\ \mathbf{Q}_{22\cdot 1}^{-1}\mathbf{Q}_{2Y\cdot 1} \end{bmatrix}$$

- We can use $\beta_1 = \mathbf{Q}_{11\cdot 2}^{-1}\mathbf{Q}_{1Y\cdot 2}$ to understand multivariate regression.

Interpreting Multivariate Regression

- The multivariate projection equation is $Y = X_1'\beta_1 + X_2'\beta_2 + e$
- X_1 consists of a part explainable by a linear projection of X_2 and a part that is not. Call the latter u_1.
- β_1 is the projection of Y on u_1
- This is the idea behind the Ballantine diagram.

Demonstration: Iterated Projection

Divide X into a single variable in X_1, with all the other variables in X_2, so that

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

Regress X_1 on X_2:

$$\begin{aligned} X_1 &= X_2'\gamma_2 + u_1 \\ \mathbb{E}[X_2 u_1] &= 0 \\ \gamma_2 &= \mathbf{Q}_{22}^{-1}\mathbf{Q}_{21} \\ \mathbb{E}[u_1^2] &= \mathbf{Q}_{11\cdot 2} \\ \mathbb{E}[u_1 Y] &= \mathbf{Q}_{1Y} - \mathbf{Q}_{12}\mathbf{Q}_{22}^{-1}\mathbf{Q}_{2Y} = \mathbf{Q}_{1Y\cdot 2} \\ \beta_1 &= \mathbf{Q}_{11\cdot 2}^{-1}\mathbf{Q}_{1Y\cdot 2} = \frac{\mathbb{E}[u_1 Y]}{\mathbb{E}[u_1^2]} \end{aligned}$$

Omitted Variable Bias

- Begin with $Y = X'\beta + e = X_1'\beta_1 + X_2'\beta_2 + e$
- Suppose that we failed to include X_2 in our model, so we actually estimate a **short regression**:

$$Y = X_1'\gamma_1 + u, \quad E[X_1 u] = 0$$

- We can study the components of γ_1.

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

OVB is Not Solved by Adding Variables

- Suppose $Y = X'\beta + e = X_1'\beta_1 + X_2'\beta_2 + X_3'\beta_3 + e$
- Suppose we cannot measure X_3. Instead, we can choose either:

$$Y = X_1'\gamma_1 + u_1$$

$$Y = X_1'\delta_1 + X_2'\delta_2 + u_2$$

- $\gamma_1 = \beta_1 + \Gamma_{12}\beta_2 + \Gamma_{13}\beta_3$.
- $\delta_1 = \beta_1 + \Gamma_{13\cdot 2}\beta_3$.
- Which is better depends on the signs and sizes of these terms!

OVB in Practice: Does Democracy Cause Growth?

- Barro (1996) regresses GDP growth on a democracy index:

$$\text{Growth}_i = \beta_1 \text{Democracy}_i + e_i$$

- But countries that democratize also tend to have stronger property rights, higher education, and colonial histories that affect growth.
- The OVB formula tells us:

$$\gamma_1 = \beta_1 + \underbrace{\Gamma_{12}}_{\substack{\text{correlation of democracy with institutions}}} \underbrace{\beta_2}_{\substack{\text{effect of institutions on growth}}}$$

- If institutions are positively correlated with democracy ($\Gamma_{12} > 0$) and positively affect growth ($\beta_2 > 0$), then γ_1 overstates the effect of democracy.
- Adding controls can help—but only if they reduce the bias term, not increase it.

OVB in Practice: Does Democracy Cause Growth?

- Barro (1996) regresses GDP growth on a democracy index:

$$\text{Growth}_i = \beta_1 \text{Democracy}_i + e_i$$

- But countries that democratize also tend to have stronger property rights, higher education, and colonial histories that affect growth.
- The OVB formula tells us:

$$\gamma_1 = \beta_1 + \underbrace{\Gamma_{12}}_{\substack{\text{correlation of democracy with institutions}}} \underbrace{\beta_2}_{\substack{\text{effect of institutions on growth}}}$$

- If institutions are positively correlated with democracy ($\Gamma_{12} > 0$) and positively affect growth ($\beta_2 > 0$), then γ_1 overstates the effect of democracy.
- Adding controls can help—but only if they reduce the bias term, not increase it.

OVB in Practice: Does Democracy Cause Growth?

- Barro (1996) regresses GDP growth on a democracy index:

$$\text{Growth}_i = \beta_1 \text{Democracy}_i + e_i$$

- But countries that democratize also tend to have stronger property rights, higher education, and colonial histories that affect growth.
- The OVB formula tells us:

$$\gamma_1 = \beta_1 + \underbrace{\Gamma_{12}}_{\substack{\text{correlation of} \\ \text{democracy with} \\ \text{institutions}}} \underbrace{\beta_2}_{\substack{\text{effect of} \\ \text{institutions} \\ \text{on growth}}}$$

- If institutions are positively correlated with democracy ($\Gamma_{12} > 0$) and positively affect growth ($\beta_2 > 0$), then γ_1 overstates the effect of democracy.
- Adding controls can help—but only if they reduce the bias term, not increase it.

Where We Stand

Recap: From Population to Practice

1. **CEF** ($m(X) = \mathbb{E}[Y|X]$): Best predictor, but requires the full joint distribution.
2. **BLP** ($\mathcal{P}[Y|X] = X'\beta$): Best *linear* predictor, requiring only means, variances, and covariances.
3. **Partitioned regression**: Each coefficient captures the effect of a variable *after removing* what other variables predict.
4. **OVB**: Omitting a correlated variable biases coefficients; adding variables doesn't always help.

Next: All of this is in the *population*. Starting next lecture, we move to *samples*—using OLS to estimate β and studying when OLS does a good job.

What Causal Effect Does Regression Recover?

- Causal inference is interested in:

$$\delta_X \equiv E[Y|X, D = 1] - E[Y|X, D = 0]$$

  That is, given some value of X, what is the expected difference in outcomes between treatment and control?

- Suppose we have the following regression:

$$Y = \delta_R D + X'\beta + e$$

- It turns out δ_R is a weighted average of δ_X, where the weights depend on the variance of X.
- We can use partitions to explain those weights.

Applying Partitions to Treatment Dummy

$$\begin{aligned}
\delta_R &= \frac{Cov(Y, \tilde{D})}{Var(\tilde{D})} \\
&= \frac{E[(D - E[D|X])Y]}{E[(D - E[D|X])^2]} \\
&= \frac{E\{(D - E[D|X])E[Y|D, X]\}}{E[(D - E[D|X])^2]} \qquad \text{(By LIE)}
\end{aligned}$$

Applying Partitions to Treatment Dummy

$$
\begin{aligned}
\delta_R &= \frac{Cov(Y, \tilde{D})}{Var(\tilde{D})} \\
&= \frac{E[(D - E[D|X])Y]}{E[(D - E[D|X])^2]} \\
&= \frac{E\{(D - E[D|X])E[Y|D, X]\}}{E[(D - E[D|X])^2]} \qquad \text{(By LIE)}
\end{aligned}
$$

Applying Partitions to Treatment Dummy

$$
\begin{aligned}
\delta_R &= \frac{Cov(Y, \tilde{D})}{Var(\tilde{D})} \\
&= \frac{E[(D - E[D|X])Y]}{E[(D - E[D|X])^2]} \\
&= \frac{E\{(D - E[D|X])E[Y|D, X]\}}{E[(D - E[D|X])^2]} \qquad \text{(By LIE)}
\end{aligned}
$$

Notational Trick

$$
\begin{aligned}
E[Y|X,D] &= E[Y|X,D=0](1-D) + E[Y|X,D=1]D \\
&= E[Y|X,D=0] + (E[Y|X,D=1] - E[Y|X,D=0])D \\
&= E[Y|X,D=0] + \delta_X D
\end{aligned}
$$

Notational Trick

$$
\begin{aligned}
E[Y|X,D] &= E[Y|X,D=0](1-D) + E[Y|X,D=1]D \\
&= E[Y|X,D=0] + (E[Y|X,D=1] - E[Y|X,D=0])D \\
&= E[Y|X,D=0] + \delta_X D
\end{aligned}
$$

Notational Trick

$$
\begin{aligned}
E[Y|X,D] &= E[Y|X,D=0](1-D) + E[Y|X,D=1]D \\
&= E[Y|X,D=0] + (E[Y|X,D=1] - E[Y|X,D=0])D \\
&= E[Y|X,D=0] + \delta_X D
\end{aligned}
$$

Regression as Weighted Average of Causal Effects

$$= \frac{E\{(D - E[D|X])E[Y|D, X]\}}{E[(D - E[D|X])^2]}$$

$$= \frac{E\{(D - E[D|X])E[Y|X, D = 0]\} + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]}$$

$$= \frac{0 + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]}$$

(X and $\tilde{D}$ are uncorrelated.)

$$= \frac{E[\sigma^2_{D|X}\delta_X]}{E[\sigma^2_{D|X}]}$$

(Where $\sigma^2_{D|X}$ is the conditional variance of D given X.)

Regression puts the most weight on values of X with the highest variance in D. Those have the closest to an even split between treatment and control and are most precisely estimated.

Regression as Weighted Average of Causal Effects

$$= \frac{E\{(D - E[D|X])E[Y|D, X]\}}{E[(D - E[D|X])^2]}$$

$$= \frac{E\{(D - E[D|X])E[Y|X, D = 0]\} + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]}$$

$$= \frac{0 + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]}$$

(X and $\tilde{D}$ are uncorrelated.)

$$= \frac{E[\sigma^2_{D|X}\delta_X]}{E[\sigma^2_{D|X}]}$$

(Where $\sigma^2_{D|X}$ is the conditional variance of D given X.)

Regression puts the most weight on values of X with the highest variance in D. Those have the closest to an even split between treatment and control and are most precisely estimated.

Regression as Weighted Average of Causal Effects

$$= \frac{E\{(D - E[D|X])E[Y|D, X]\}}{E[(D - E[D|X])^2]}$$

$$= \frac{E\{(D - E[D|X])E[Y|X, D = 0]\} + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]}$$

$$= \frac{0 + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]}$$

(X and $\tilde{D}$ are uncorrelated.)

$$= \frac{E[\sigma^2_{D|X}\delta_X]}{E[\sigma^2_{D|X}]}$$

(Where $\sigma^2_{D|X}$ is the conditional variance of D given X.)

Regression puts the most weight on values of X with the highest variance in D. Those have the closest to an even split between treatment and control and are most precisely estimated.

Regression as Weighted Average of Causal Effects

$$
\begin{aligned}
&= \frac{E\{(D - E[D|X])E[Y|D,X]\}}{E[(D - E[D|X])^2]} \\
&= \frac{E\{(D - E[D|X])E[Y|X,D=0]\} + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]} \\
&= \frac{0 + E\{(D - E[D|X])D\delta_X\}}{E[(D - E[D|X])^2]} && \text{(X and } \tilde{D} \text{ are uncorrelated.)} \\
&= \frac{E[\sigma^2_{D|X}\delta_X]}{E[\sigma^2_{D|X}]} && \text{(Where } \sigma^2_{D|X} \text{ is the conditional variance of D given X.)}
\end{aligned}
$$

Regression puts the most weight on values of X with the highest variance in D. Those have the closest to an even split between treatment and control and are most precisely estimated.