Causal Regression

Causal Inference using Machine Learning
Master in Economics, UNT

Andres Mena

Spring 2024

Table of Contents

1. Conditional Expectation Function
2. Regression Properties
3. Causal Regression
4. Inference in Regression

Content

1. Conditional Expectation Function
2. Regression Properties
3. Causal Regression
4. Inference in Regression

Definition of CEF

The Conditional Expectation Function (CEF)

- The *Conditional Expectation Function (CEF)* is the population average of y_i given a $k \times 1$ vector of covariates X_i, denoted $E[y_i|X_i]$.

Definition of CEF

The Conditional Expectation Function (CEF)

- The *Conditional Expectation Function (CEF)* is the population average of y_i given a $k \times 1$ vector of covariates X_i, denoted $E[y_i|X_i]$.
- For discrete y_i:

$$E[y_i|X_i = x] = \sum_t tP(y_i = t|X_i = x)$$

where $P(y_i = t|X_i = x)$ is the conditional probability mass function.

Definition of CEF

The Conditional Expectation Function (CEF)

- The *Conditional Expectation Function (CEF)* is the population average of y_i given a $k \times 1$ vector of covariates X_i, denoted $E[y_i|X_i]$.
- For discrete y_i:

$$E[y_i|X_i = x] = \sum_t tP(y_i = t|X_i = x)$$

  where $P(y_i = t|X_i = x)$ is the conditional probability mass function.
- For continuous y_i:

$$E[y_i|X_i = x] = \int tf_y(t|X_i = x)\,dt$$

  where $f_y(t|X_i = x)$ is the conditional density of y_i given $X_i = x$.

The Law of Iterated Expectations

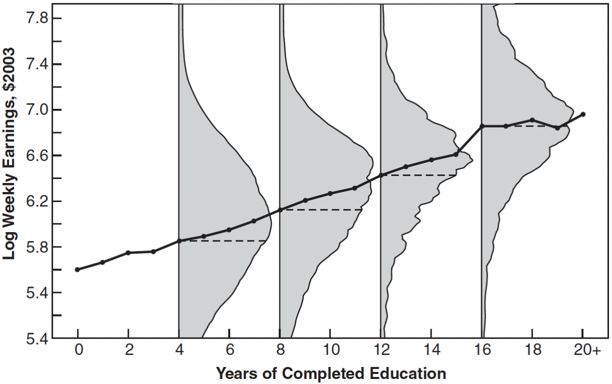

Figure: CEF of Log Wage on Schooling (Mostly Harmless).

LIE:

$$E[y_i] = E\{E[y_i|X_i]\}$$

Theorem: The CEF Decomposition Property

Theorem:

$$y_i = E[y_i|X_i] + \varepsilon_i$$

Theorem: The CEF Decomposition Property

Theorem:

$$y_i = E[y_i|X_i] + \varepsilon_i$$

- ε_i is the mean independent, i.e $E[\varepsilon_i] = E[\varepsilon_i|X_i] = 0$.

Theorem: The CEF Decomposition Property

Theorem:

$$y_i = E[y_i|X_i] + \varepsilon_i$$

- ε_i is the mean independent, i.e $E[\varepsilon_i] = E[\varepsilon_i|X_i] = 0$.
- ε_i is uncorrelated with any function of X_i.

Theorem: The CEF Decomposition Property

Theorem:

$$y_i = E[y_i|X_i] + \varepsilon_i$$

- ε_i is the mean independent, i.e $E[\varepsilon_i] = E[\varepsilon_i|X_i] = 0$.
- ε_i is uncorrelated with any function of X_i.

Proof: i) Replace ϵ, ii) Use LIE to show Orthogonality h(X) and ϵ.

Theorem: The CEF Prediction Property

Theorem:

$$E[y_i|X_i] = \arg\min_{m(X_i)} E[(y_i - m(X_i))^2]$$

- The CEF $E[y_i|X_i]$ solves the minimum mean squared error (MMSE) prediction problem.

Proof: Find Normal equations and take conditional expectation.

Theorem: The ANOVA Theorem

Theorem:

$$V(y_i) = V(E[y_i|X_i]) + E[V(y_i|X_i)]$$

- The total variance equals the variance of the CEF plus the average conditional variance of y_i given X_i.

Theorem: The ANOVA Theorem

Theorem:

$$V(y_i) = V(E[y_i|X_i]) + E[V(y_i|X_i)]$$

- The total variance equals the variance of the CEF plus the average conditional variance of y_i given X_i.

Proof: Use decomposition property and show $E[\epsilon_i^2] = E[V(y_i|X_i)]$.

Content

1 Conditional Expectation Function

2 Regression Properties

3 Causal Regression

4 Inference in Regression

Regression Problem

Regression Problem:

$$\beta_{OLS} = \arg\min_b E[(y_i - X_i b)^2]$$

- The vector of population regression coefficients, β_{OLS}, is defined as the solution to the population least squares problem.
- Using the first-order condition:

$$E[X_i(y_i - X_i b)] = 0$$

- Solution:

$$\beta_{OLS} = E[X_i X_i']^{-1} E[X_i y_i]$$

Linear CEF Theorem

Theorem:

- If the CEF is linear, then $X_i'\beta_{OLS}$ is the CEF.

Proof: Assume $E[y_i|X_i] = X_i'\beta^*$ for any and β^*, and show $\beta_{OLS} = \beta^*$.

Best Linear Predictor (BLP) of Y

Theorem:

- The function $X_i'\beta_{OLS}$ is the best linear predictor of y_i given X_i in the MMSE sense.

Proof: $\beta_{BLP} = \beta_{OLS}$

$$\beta_{BLP} = \arg\min_b E[(y_i - X_i b)^2]$$

BLP of the CEF

Theorem:

- The function $X_i'\beta$ provides the MMSE linear approximation to $E[y_i \mid X_i]$:
$$\beta = \arg\min_b E[(E[y_i \mid X_i] - X_i'b)^2]$$
- This motivates regression as an approximation to the CEF.

BLP of the CEF

Theorem:

- The function $X_i'\beta$ provides the MMSE linear approximation to $E[y_i \mid X_i]$:
$$\beta = \arg\min_b E[(E[y_i \mid X_i] - X_i'b)^2]$$
- This motivates regression as an approximation to the CEF.

Proof: Take conditional expectation of the BLP for Y and note that is the same problem as BLP of the CEF.

Figure: Linear CEF Approximation

Illustration:

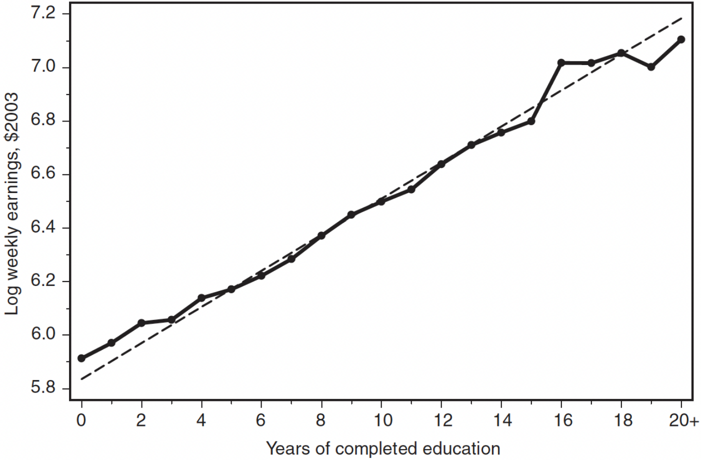

Figure: Regression threads the CEF (dots = CEF, dashes = regression line).

Bivariate Regression

Centering: Subtract the mean of X and Y to define centered variables:

$$\tilde{X}_i = X_i - \bar{X}, \quad \tilde{Y}_i = Y_i - \bar{Y}$$

Bivariate Regression

Centering: Subtract the mean of X and Y to define centered variables:

$$\tilde{X}_i = X_i - \bar{X}, \quad \tilde{Y}_i = Y_i - \bar{Y}$$

Centered Regression:

$$\tilde{Y}_i = \beta_1 \tilde{X}_i + \varepsilon_i$$

Bivariate Regression

Centering: Subtract the mean of X and Y to define centered variables:

$$\tilde{X}_i = X_i - \bar{X}, \quad \tilde{Y}_i = Y_i - \bar{Y}$$

Centered Regression:

$$\tilde{Y}_i = \beta_1 \tilde{X}_i + \varepsilon_i$$

Key Result: The slope coefficient in the centered regression is:

$$\beta_1 = \frac{\sum \tilde{X}_i \tilde{Y}_i}{\sum \tilde{X}_i^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}(X)}$$

Bivariate Regression

Centering: Subtract the mean of X and Y to define centered variables:

$$\tilde{X}_i = X_i - \bar{X}, \quad \tilde{Y}_i = Y_i - \bar{Y}$$

Centered Regression:

$$\tilde{Y}_i = \beta_1 \tilde{X}_i + \varepsilon_i$$

Key Result: The slope coefficient in the centered regression is:

$$\beta_1 = \frac{\sum \tilde{X}_i \tilde{Y}_i}{\sum \tilde{X}_i^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}$$

Why Center?

- Simplifies the formula by removing the intercept (β_0).
- Directly links β_1 to covariance and variance.
- Note that $\bar{X} \to E[X]$ and $\bar{Y} \to E[Y]$ as $n \to \infty$.

The Frisch-Waugh-Lovell Theorem

Theorem: Frisch-Waugh-Lovell (FWL) Theorem

- In a multivariate regression model, the coefficient of a single variable X_k in a multiple regression is equivalent to the coefficient obtained by:

The Frisch-Waugh-Lovell Theorem

Theorem: Frisch-Waugh-Lovell (FWL) Theorem

- In a multivariate regression model, the coefficient of a single variable X_k in a multiple regression is equivalent to the coefficient obtained by:
    1. Regressing Y on all other covariates except X_k, and saving the residuals $\tilde{Y}$.

The Frisch-Waugh-Lovell Theorem

Theorem: Frisch-Waugh-Lovell (FWL) Theorem

- In a multivariate regression model, the coefficient of a single variable X_k in a multiple regression is equivalent to the coefficient obtained by:
  1. Regressing Y on all other covariates except X_k, and saving the residuals $\tilde{Y}$.
  2. Regressing X_k on all other covariates, and saving the residuals $\tilde{X}_k$.

The Frisch-Waugh-Lovell Theorem

Theorem: Frisch-Waugh-Lovell (FWL) Theorem

- In a multivariate regression model, the coefficient of a single variable X_k in a multiple regression is equivalent to the coefficient obtained by:
  1. Regressing Y on all other covariates except X_k, and saving the residuals $\tilde{Y}$.
  2. Regressing X_k on all other covariates, and saving the residuals $\tilde{X}_k$.
  3. Regressing $\tilde{Y}$ on $\tilde{X}_k$.

The Frisch-Waugh-Lovell Theorem

Theorem: Frisch-Waugh-Lovell (FWL) Theorem

- In a multivariate regression model, the coefficient of a single variable X_k in a multiple regression is equivalent to the coefficient obtained by:
  1. Regressing Y on all other covariates except X_k, and saving the residuals $\tilde{Y}$.
  2. Regressing X_k on all other covariates, and saving the residuals $\tilde{X}_k$.
  3. Regressing $\tilde{Y}$ on $\tilde{X}_k$.

Key Insight: The effect of X_k on Y can be isolated by removing the influence of other covariates from both Y and X_k.

Regression Anatomy Formula: Removing estiamted Conditional Expectations under Linearity

Step 1: Residualizing Variables

- Define the residualized variables by removing the estimated conditional expectations $\hat{m}(L|X_{-k})$:

$$\tilde{X}_k = X_k - \hat{m}[X_k \mid X_{-k}], \quad \tilde{Y} = Y - \hat{m}[Y \mid X_{-k}]$$

- This removes the influence of all other covariates (X_{-k}) on both X_k and Y, isolating the unique relationship between X_k and Y.
- All the properties of the *CEF* are preserved, specially the decomposition property.

Regression Anatomy Formula: Removing estiamted Conditional Expectations under Linearity

Step 1: Residualizing Variables

- Define the residualized variables by removing the estimated conditional expectations $\hat{m}(L|X_{-k})$:

$$\tilde{X}_k = X_k - \hat{m}[X_k \mid X_{-k}], \quad \tilde{Y} = Y - \hat{m}[Y \mid X_{-k}]$$

- This removes the influence of all other covariates (X_{-k}) on both X_k and Y, isolating the unique relationship between X_k and Y.
- All the properties of the *CEF* are preserved, specially the decomposition property.

Step 2: Regression on Residualized Variables

$$\tilde{Y} = \beta_k \tilde{X}_k + \tilde{\varepsilon}$$

Regression Anatomy Formula: Removing estiamted Conditional Expectations under Linearity

Step 1: Residualizing Variables

- Define the residualized variables by removing the estimated conditional expectations $\hat{m}(L|X_{-k})$:

$$\tilde{X}_k = X_k - \hat{m}[X_k \mid X_{-k}], \quad \tilde{Y} = Y - \hat{m}[Y \mid X_{-k}]$$

- This removes the influence of all other covariates (X_{-k}) on both X_k and Y, isolating the unique relationship between X_k and Y.
- All the properties of the *CEF* are preserved, specially the decomposition property.

Step 2: Regression on Residualized Variables

$$\tilde{Y} = \beta_k \tilde{X}_k + \tilde{\varepsilon}$$

Result: The partial regression coefficient is:

$$\beta_k = \frac{\mathrm{Cov}(Y, \tilde{X}_k)}{\mathrm{Var}(\tilde{X}_k)}$$

Content

1 Conditional Expectation Function

2 Regression Properties

3 Causal Regression

4 Inference in Regression

Causal Regression Model: Constant Treatment Effect

Causal Model $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$

Causal Regression Model: Constant Treatment Effect

Causal Model $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$
Regression Model $Y_i = \alpha + \beta D_i + \varepsilon_i$

Causal Regression Model: Constant Treatment Effect

Causal Model $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$
Regression Model $Y_i = \alpha + \beta D_i + \varepsilon_i$

- β: Constant treatment effect $\beta = Y_i(1) - Y_i(0)$
- α: $E[Y_i(0)]$
- ε_i: $Y_i(0) - E[Y_i(0)]$

Causal Regression Model: Constant Treatment Effect

Causal Model $Y_i = Y_i(0) + D_i(Y_i(1) - Y_i(0))$
Regression Model $Y_i = \alpha + \beta D_i + \varepsilon_i$

- β: Constant treatment effect $\beta = Y_i(1) - Y_i(0)$
- α: $E[Y_i(0)]$
- ε_i: $Y_i(0) - E[Y_i(0)]$

Show difference between $E[Y_i|D_i = 1]$ and $E[Y_i|D_i = 0]$

Regression Coefficient with Covariates (β)

Setup:

$$Y_i = \beta_0 + \beta D_i + \gamma' X_i + \epsilon_i$$

Regression Coefficient with Covariates (β)

Setup:

$$Y_i = \beta_0 + \beta D_i + \gamma' X_i + \epsilon_i$$

Regression Coefficient with Covariates (β)

Setup:

$$Y_i = \beta_0 + \beta D_i + \gamma' X_i + \epsilon_i$$

Using the regression anatomy formula:

$$\beta = \frac{\text{Cov}(Y_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)}, \quad \tilde{D}_i = D_i - \mathbb{E}[D_i | X_i].$$

Regression Coefficient with Covariates (β)

Setup:

$$Y_i = \beta_0 + \beta D_i + \gamma' X_i + \epsilon_i$$

Using the regression anatomy formula:

$$\beta = \frac{\text{Cov}(Y_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)}, \quad \tilde{D}_i = D_i - \mathbb{E}[D_i | X_i].$$

Expanding the Numerator:

$$\text{Cov}(Y_i, \tilde{D}_i) = \mathbb{E}[(D_i - \mathbb{E}[D_i | X_i]) Y_i].$$

Regression Coefficient with Covariates (β)

Setup:

$$Y_i = \beta_0 + \beta D_i + \gamma' X_i + \epsilon_i$$

Using the regression anatomy formula:

$$\beta = \frac{\text{Cov}(Y_i, \tilde{D}_i)}{\text{Var}(\tilde{D}_i)}, \quad \tilde{D}_i = D_i - \mathbb{E}[D_i|X_i].$$

Expanding the Numerator:

$$\text{Cov}(Y_i, \tilde{D}_i) = \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])Y_i].$$

By the Law of Iterated Expectations (LIE), substitute Y_i with $\mathbb{E}[Y_i|D_i, X_i]$.
Define CATE: $\tau(X_i) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i]$.

$$\mathbb{E}[Y_i|D_i, X_i] = \mathbb{E}[Y_i|D_i = 0, X_i] + \tau(X_i)D_i.$$

Regression Coefficient with Covariates

Decomposing the Covariance:

$$\mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i = 0, X_i]\big] + \mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\tau(X_i)D_i\big]$$

Regression Coefficient with Covariates

Decomposing the Covariance:

$$\mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i = 0, X_i]\big] + \mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\tau(X_i)D_i\big]$$

- The first term vanishes (why?)

Regression Coefficient with Covariates

Decomposing the Covariance:

$$\mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i = 0, X_i]\big] + \mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\tau(X_i)D_i\big]$$

- The first term vanishes (why?)
- The second term simplifies:

$$\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])D_i\tau(X_i)] = \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2\tau(X_i)].$$

Regression Coefficient with Covariates

Decomposing the Covariance:

$$\mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i = 0, X_i]\big] + \mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\tau(X_i)D_i\big]$$

- The first term vanishes (why?)
- The second term simplifies:

$$\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])D_i\tau(X_i)] = \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2\tau(X_i)].$$

Final Expression:

$$\beta = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2\tau(X_i)]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]}.$$

Regression Coefficient with Covariates

Decomposing the Covariance:

$$\mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\mathbb{E}[Y_i|D_i = 0, X_i]\big] + \mathbb{E}\big[(D_i - \mathbb{E}[D_i|X_i])\tau(X_i)D_i\big]$$

- The first term vanishes (why?)
- The second term simplifies:

$$\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])D_i\tau(X_i)] = \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2\tau(X_i)].$$

Final Expression:

$$\beta = \frac{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2\tau(X_i)]}{\mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2]}.$$

Weighted Average: Define $\sigma_D^2(X_i) = \mathbb{E}[(D_i - \mathbb{E}[D_i|X_i])^2|X_i]$. Then:

$$\beta = \frac{\mathbb{E}[\sigma_D^2(X_i)\tau(X_i)]}{\mathbb{E}[\sigma_D^2(X_i)]}.$$

Conclusion: - β is a variance-weighted average of CATE $\tau(X_i)$. - The weights depend on the conditional variance of D_i given X_i.

Content

1 Conditional Expectation Function

2 Regression Properties

3 Causal Regression

4 Inference in Regression

Classical Additive Approach - Population Setup

Population Linear Projection:

$$Y = D\beta + X'\gamma + \epsilon, \quad \epsilon \perp (D, X),$$

where:

- D is the treatment indicator,
- $X = (1, W)$ includes an intercept and covariates with $\mathbb{E}[W] = 0$,
- $D \perp W$ (randomized controlled trial).

Key Results: - Decompose $\gamma' X = \gamma_1 + \gamma_2' W$. - For $U := \gamma_2' W + \epsilon$:

$$Y = D\beta + \gamma_1 + U, \quad U \perp (1, D).$$

Interpretation of Coefficients: - $\beta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$ (ATE), - $\gamma_1 = \mathbb{E}[Y(0)]$ (average untreated outcome).

Projection Properties

Projection Setup:

$$Y = D\beta + \gamma_1 + U, \quad U \perp (1, D).$$

Key Implications: - The population projection of Y onto $(1, D)$ yields:

$$Y = \mathbb{E}[Y|D] = D\beta + \gamma_1.$$

- The coefficients (β, γ_1) are the same as those obtained by the two-sample approach in the population:

$$\beta = \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)], \quad \gamma_1 = \mathbb{E}[Y(0)].$$

RCT Setting: - Randomization ensures:

$$D \perp W, \quad \epsilon \perp (D, W).$$

Approximate Normality of $\hat{\beta}$

OLS Estimator for β:

$$\sqrt{n}(\hat{\beta} - \beta) \approx \sqrt{n}\frac{\mathbb{E}_n[\epsilon \tilde{D}]}{\mathbb{E}_n[\tilde{D}^2]} \sim \mathcal{N}(0, V_{11}),$$

where:

- $\tilde{D} = D - \mathbb{E}[D|X]$ (residual after partialling out X),
- $V_{11} = \frac{\mathbb{E}[\epsilon^2 \tilde{D}^2]}{(\mathbb{E}[\tilde{D}^2])^2}$.

Key Derivation Steps: 1. Partial out $X = (1, W)$ from D, 2. Use OLS theory to approximate the distribution of $\hat{\beta}$.

Approximate Normality of $\hat{\gamma}_1$ and Joint Properties

OLS Estimator for γ_1:

$$\sqrt{n}(\hat{\gamma}_1 - \gamma_1) \approx \sqrt{n}\frac{\mathbb{E}_n[\epsilon\tilde{1}]}{\mathbb{E}_n[\tilde{1}^2]} \sim \mathcal{N}(0, V_{22}),$$

where:

- $\tilde{1} = 1 - \mathbb{E}[1|D, X]$ (residual after partialling out D and X),
- $V_{22} = \frac{\mathbb{E}[\epsilon^2\tilde{1}^2]}{(\mathbb{E}[\tilde{1}^2])^2}$.

Joint Normality: The estimators $\hat{\beta}$ and $\hat{\gamma}_1$ are jointly normal:

$$\text{Cov}(\hat{\beta}, \hat{\gamma}_1) \sim V_{12},$$

where:

$$V_{12} = \frac{\mathbb{E}[\epsilon^2\tilde{D}\tilde{1}]}{\mathbb{E}[\tilde{D}^2]\mathbb{E}[\tilde{1}^2]}.$$