Advanced Econometrics
03 The Linear Regression Model

Eduard Brüll

Fall 2025

Lecture 3

Advanced Econometrics

Literature: Greene Chapter 2 and 3

3.1.1: The Conditional Expectation Function

The Conditional Expectation Function

Definition: The **conditional expectation function** for a dependent variable Y_i, given a $K + 1 \times 1$ vector of covariates $\mathbf{X}_i$, describes the average value of Y_i in the population when we hold $\mathbf{X}_i$ fixed.

Written as $\mathbb{E}[Y_i \mid \mathbf{X}_i]$, the CEF is a function of $\mathbf{X}_i$.

Examples:

- $\mathbb{E}[\text{Income}_i \mid \text{Education}_i]$
- $\mathbb{E}[\text{Birth weight}_i \mid \text{Air quality}_i]$

We will generally assume $\mathbf{X}_i$ is a random variable, which implies that $\mathbb{E}[Y_i \mid \mathbf{X}_i]$ is also a random variable.

The Conditional Expectation Function (contd.)

Formally, for continuous Y_i with conditional density $f_Y(t \mid \mathbf{X}_i = x)$,

$$\mathbf{E}[Y_i \mid \mathbf{X}_i = x] \ = \ \int t \, f_Y(t \mid \mathbf{X}_i = x) \, \mathrm{d}t.$$

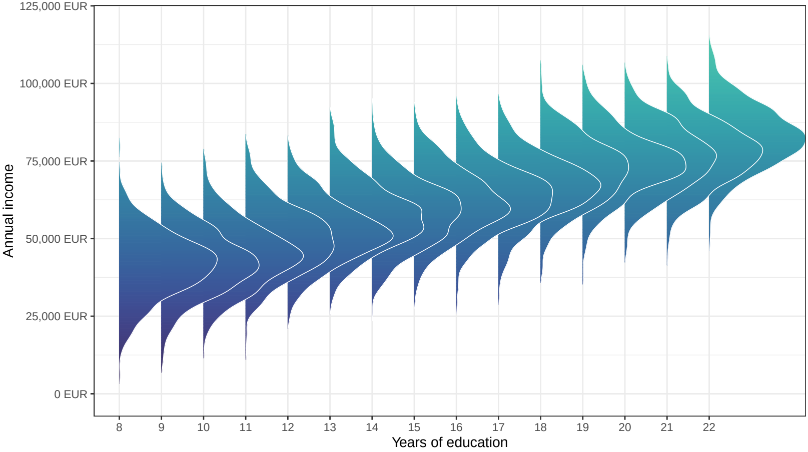
For discrete Y_i with conditional probability mass function $\mathbb{P}(Y_i = t \mid \mathbf{X}_i = x)$,

$$\mathbf{E}[Y_i \mid \mathbf{X}_i = x] \ = \ \sum_t t \, \mathbb{P}(Y_i = t \mid \mathbf{X}_i = x).$$

Notice: We are focusing on the population. The goal is to build intuition about the parameters that we will eventually estimate.

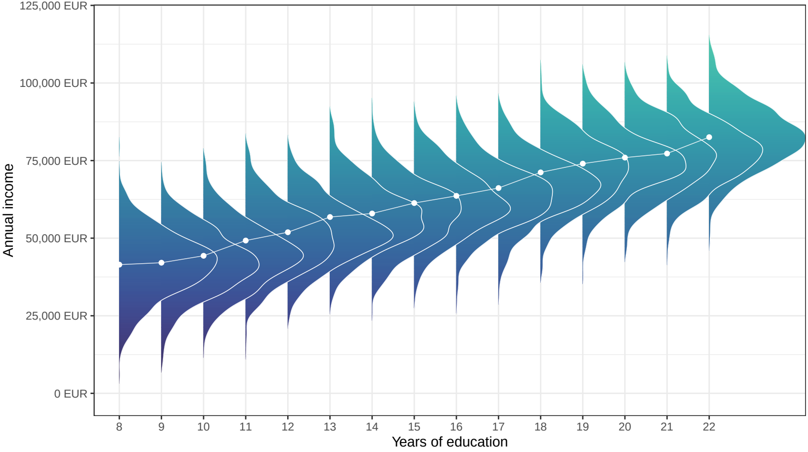
The CEF Graphically

The conditional distributions of Y_i for $X_i \in [8, ..., 22]$:



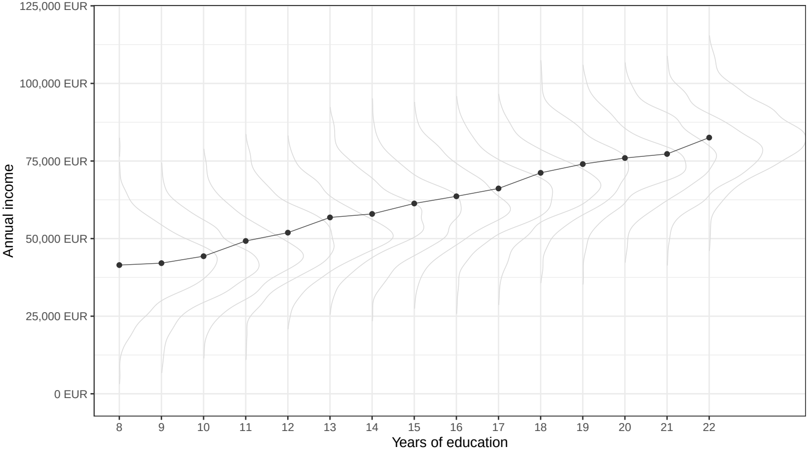
The CEF Graphically (contd.)

The CEF $\mathbf{E}[Y_i \mid \mathbf{X}_i]$ connects these conditional distributions' means:



The CEF Graphically (contd.)

Focusing on $\mathbf{E}[Y_i \mid \mathbf{X}_i]$:



Interlude: Law of Iterated Expectations (LIE)

Definition: For any random variables Y and X,

$$\mathbf{E}[Y] \;=\; \mathbf{E}\big[\, \mathbf{E}[Y \mid X] \,\big].$$

Intuition: The overall expectation of Y can be obtained in two steps:

1. First take the conditional expectation of Y given X.
2. Then average this conditional expectation over the distribution of X.

Example: Average income can be computed as

$$\mathbf{E}[\text{Income}] \;=\; \mathbf{E}\big[\,\mathbf{E}[\text{Income} \mid \text{Education}]\,\big]\,.$$

CEF Decomposition

The Law of Iterated Expectations (LIE) tells us that any random variable Y_i can be written as two components:

$$Y_i = \mathbf{E}[Y_i \mid \mathbf{X}_i] + \varepsilon_i$$

Interpretation:

1. **The conditional expectation function (CEF)** captures the systematic part of Y_i explained by $\mathbf{X}_i$.
2. A **residual** ε_i, which has special properties:
   2.1 $\mathbf{E}[\varepsilon_i \mid \mathbf{X}_i] = 0$ (zero mean given $\mathbf{X}_i$),
   2.2 ε_i is uncorrelated with any function of $\mathbf{X}_i$.

Takeaway: The CEF provides the predictable part of Y_i, while the residual is the unpredictable variation.

Proof of Mean Indepence

To show:

$$\mathbf{E}[\varepsilon_i \mid \mathbf{X}_i] = 0 \quad \text{for} \quad Y_i \ = \ \mathbf{E}[Y_i \mid \mathbf{X}_i] \ + \ \varepsilon_i$$

Proof:

$$\begin{aligned}
&\mathbf{E}[\varepsilon_i \mid \mathbf{X}_i] = \\
&\mathbf{E}\big[Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i] \,\big|\, \mathbf{X}_i\big] = \\
&\mathbf{E}[Y_i \mid \mathbf{X}_i] \ - \ \mathbf{E}\big[\,\mathbf{E}[Y_i \mid \mathbf{X}_i] \,\big|\, \mathbf{X}_i\big] = \\
&\mathbf{E}[Y_i \mid \mathbf{X}_i] - \mathbf{E}[Y_i \mid \mathbf{X}_i] = 0
\end{aligned}$$

Proof of Zero Correlation

To show:

$\mathbf{E}[h(\mathbf{X}_i)\varepsilon_i] = 0$ for any measurable h where $Y_i = \mathbf{E}[Y_i \mid \mathbf{X}_i] + \varepsilon_i$

Proof:

$$\begin{aligned}
\mathbf{E}[h(\mathbf{X}_i)\varepsilon_i] &= \mathbf{E}\Big[\ \mathbf{E}\big[h(\mathbf{X}_i)\varepsilon_i \mid \mathbf{X}_i\big]\ \Big] \\
&= \mathbf{E}\Big[h(\mathbf{X}_i)\ \mathbf{E}[\varepsilon_i \mid \mathbf{X}_i]\Big] \\
&= \mathbf{E}\big[h(\mathbf{X}_i) \times 0\big] \\
&= 0\,.
\end{aligned}$$

The Prediction Property of the CEF

Claim: The conditional expectation function $\mathbf{E}[Y_i \mid \mathbf{X}_i]$ is the best predictor of Y_i given $\mathbf{X}_i$, in the sense of minimizing mean squared error (MSE).

Formally: For any measurable function $g(\mathbf{X}_i)$,

$$\mathbf{E}\left[\,(Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i])^2\,\right] \;\leq\; \mathbf{E}\left[\,(Y_i - g(\mathbf{X}_i))^2\,\right].$$

Intuition:

- The CEF captures all predictable variation in Y_i from $\mathbf{X}_i$.
- Any other predictor $g(\mathbf{X}_i)$ can only add noise.

Proof of the Prediction Property

For any $g(\mathbf{X}_i)$, decompose:

$$\begin{aligned}\mathbf{E}\left[(Y_i - g(\mathbf{X}_i))^2\right] &= \mathbf{E}\Big[\,(Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i] + \mathbf{E}[Y_i \mid \mathbf{X}_i] - g(\mathbf{X}_i))^2\,\Big] \\ &= \mathbf{E}\left[(Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i])^2\right] \\ &\quad + \mathbf{E}\left[(\mathbf{E}[Y_i \mid \mathbf{X}_i] - g(\mathbf{X}_i))^2\right] \\ &\quad + 2\ \mathbf{E}\Big[(Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i])(\mathbf{E}[Y_i \mid \mathbf{X}_i] - g(\mathbf{X}_i))\Big].\end{aligned}$$

Key: The cross term vanishes since

$$\mathbf{E}\big[Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i] \,\big|\, \mathbf{X}_i\big] = 0.$$

Thus:

$$\mathbf{E}\left[(Y_i - g(\mathbf{X}_i))^2\right] = \mathbf{E}\left[(Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i])^2\right] + \mathbf{E}\left[(\mathbf{E}[Y_i \mid \mathbf{X}_i] - g(\mathbf{X}_i))^2\right] \;\geq\; \mathbf{E}\left[(Y_i - \mathbf{E}[Y_i \mid \mathbf{X}_i])^2\right].$$

3.1.2: The Population Regression Line

From the CEF to Linear Regression

Recall: The conditional expectation function (CEF) is

$$\mathbf{E}[Y_i \mid \mathbf{X}_i].$$

It fully describes the systematic relationship between Y_i and $\mathbf{X}_i$.

Problem:

- The true population CEF may be unknown.
- We often need a tractable approximation for estimation and inference.

Solution: Approximate the CEF with a linear function of $\mathbf{X}_i$:

$$\mathbf{E}[Y_i \mid \mathbf{X}_i] \;\approx\; \mathbf{X}_i'\beta.$$

The Population Regression Line

Definition: The population regression line is the best linear approximation to the CEF:

$$\mathbf{X}_i'\beta = \arg\min_{g \in \mathcal{G}_{\text{linear}}} \mathbf{E}\left[\,(Y_i - g(\mathbf{X}_i))^2\,\right],$$

where $\mathcal{G}_{\text{linear}}$ is the set of linear functions of $\mathbf{X}_i$.

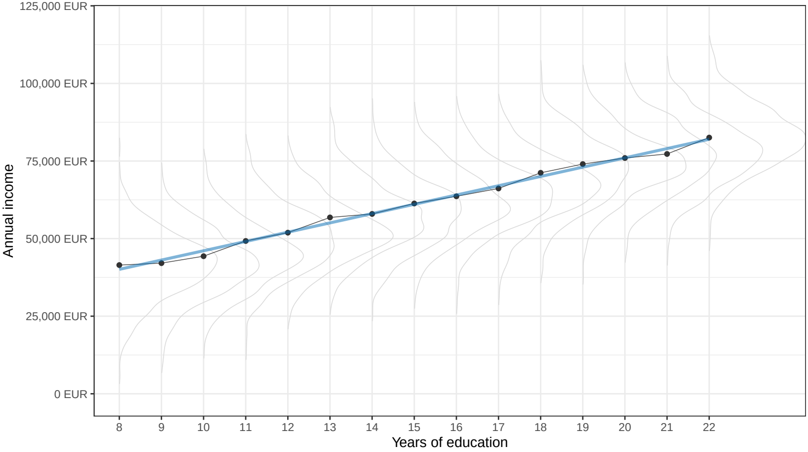
Characterization:

- β are the population OLS coefficients.
- The residual $u_i = Y_i - \mathbf{X}_i'\beta$ satisfies

$$\mathbf{E}[\mathbf{X}_i u_i] = 0 \quad \text{(orthogonality condition).}$$

Takeaway: The population regression line gives the linear predictor of Y_i that comes closest to the (possibly nonlinear) CEF in mean squared error.

The Population Regression Line as linear approximation of the CEF:



From Population to Sample Regression

Population regression:

$$\beta = \arg\min_{b} \ \mathbf{E}\left[\, (Y_i - \mathbf{X}_i' b)^2 \,\right].$$

Problem: The expectation $\mathbf{E}[\cdot]$ is unknown.

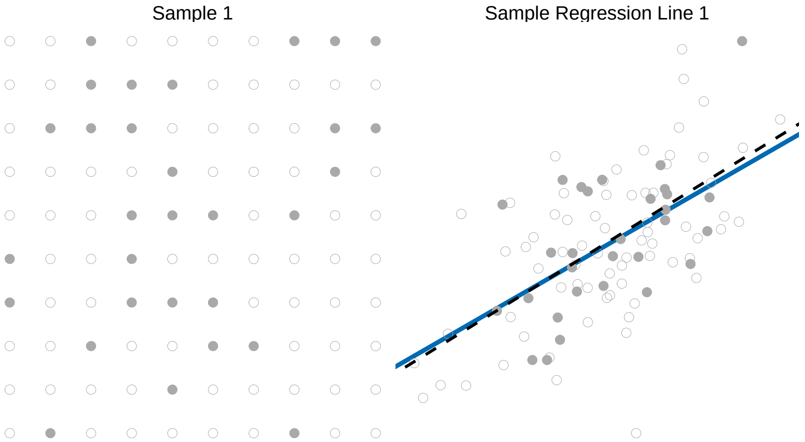
Idea: Replace expectations with sample averages.

$$\hat{\beta} = \arg\min_{b} \ \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i' b)^2.$$

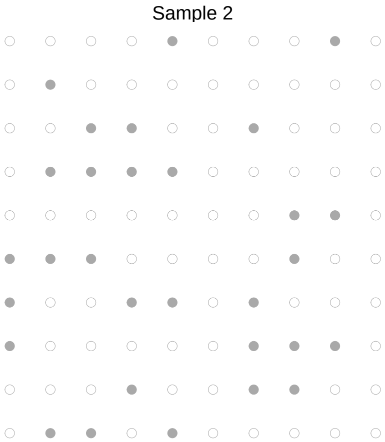
This is the principle of OLS: Estimate the coefficients that minimize the average squared residuals in the sample.

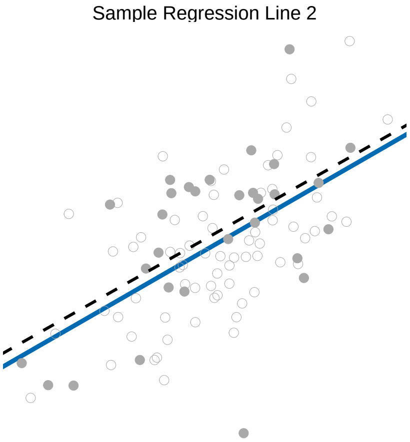
Population vs. Sample Graphically

Population vs. Sample Graphically

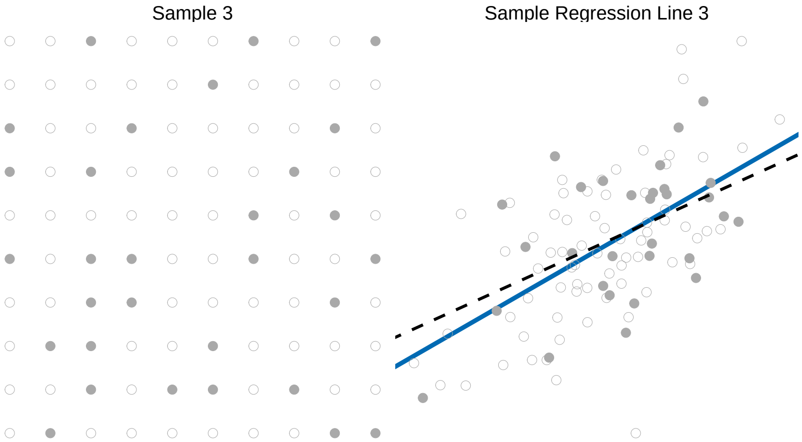


Population vs. Sample Graphically

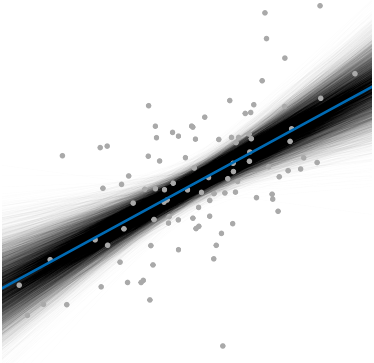


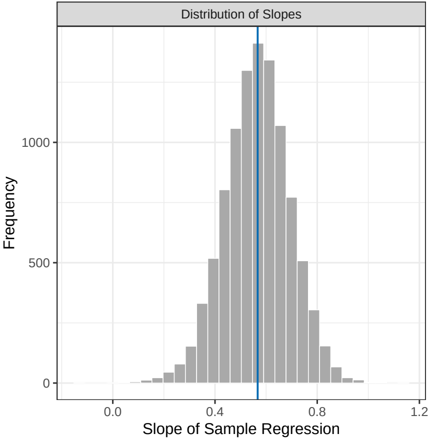


Population vs. Sample Graphically



Drawing 10,000 samples





3.2: The Linear Regression Model

The Linear Regression Model

Model setup:

$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{iK}\beta_K + u_i$ or in compact form: $y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$.

Notation:

- $i = 1, \ldots, n$ observations
- $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{iK})'$ is a $(K+1) \times 1$ regressor vector
- $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$ is a $(K+1) \times 1$ parameter vector
- u_i is the regression error

Matrix form:

$$\mathbf{y} = X\boldsymbol{\beta} + \mathbf{u},$$

where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1K} \\ 1 & x_{21} & x_{22} & \ldots & x_{2K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nK} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}.$$

3.2.1: Classical Linear Regression Assumptions

Assumptions on the Data Generating Process

The Classical Linear Regression Model Assumptions:

A1: **Linearity** The regression model is linear in parameters: $y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{iK}\beta_K + \varepsilon_i$.

A2: **Identifiability** X has full column rank $(K+1)$, so that $(X'X)^{-1}$ exists.

A3: **(Strict) Exogeneity** $\mathbb{E}[u_i \mid \mathbf{x}_i] = 0$.

A4: **Homoskedasticity (and Nonautocorrelation)** $\text{Var}(\varepsilon_i \mid \mathbf{x}_i) = \sigma^2 < \infty \quad \forall i$.

A5: **Data Generating Process** The regressor matrix X may be fixed (conditional analysis) or random (stochastic regressors).

A6: **Normality (for inference)** $\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n)$. This implies that the ε_i are independent and identically distributed.

Note: A6 is only needed for exact small-sample inference. We will later relax this assumption and rely on asymptotics instead.

Data Generating Process: Linearity

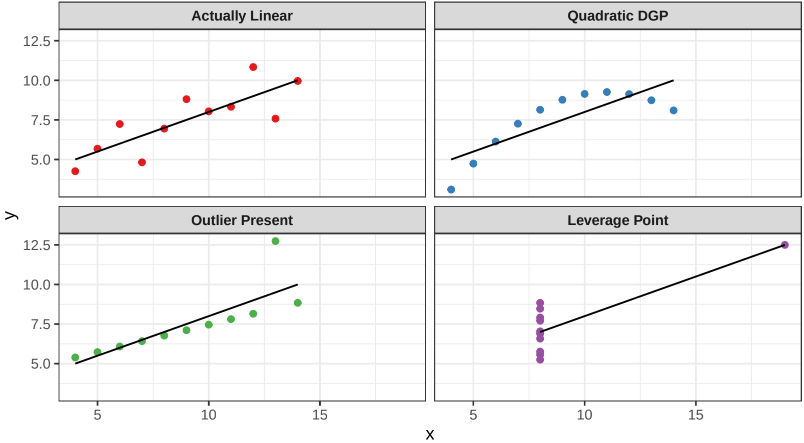
A1: Linearity

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_K x_{iK} + \varepsilon_i \text{ and } E(\varepsilon_i) = 0.$$

A1 assumes that the

- functional relationship is linear in parameters β_k
- error term ε_i enters additively
- parameters β_k are constant across observations i

Anscombe's Quartet



All four sets are identical when examined using linear statistics, but very different when graphed. Correlation between x and y is 0.816. Linear regression $y = 3.00 + 0.50x$.

Data Generating Process: Identifiability

A2: Identifiability (Full Rank)

$$\text{rank}(X) = K+1 \iff (X'X)^{-1} \text{ exists}$$

$$(x_{i0}, x_{i1}, \ldots, x_{iK}) \text{ are not linearly dependent}$$

A2 implies (see Greene, A−46): $\text{rank}(X'X) = \text{rank}(X) = K+1.$

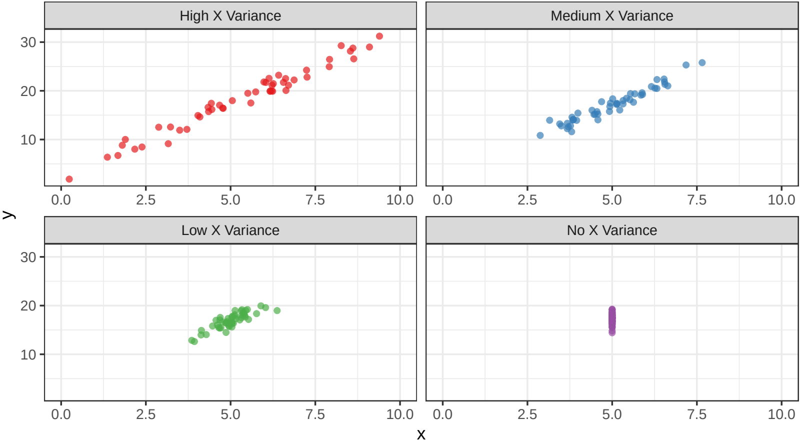
Interpretation / practice:

- *No perfect multicollinearity*: no column of X (including the constant) is an exact linear combination of the others.
- Regressors (except the constant) must have nonzero variation: $0 < \text{Var}(x_{ik})$.
- Watch out for the dummy variable trap: intercept + full set of category dummies $\Rightarrow$ drop one category.
- Avoid exact linear transforms (e.g. include either x and $x - \bar{x}$, not both; or avoid x, $2x$ together).

Every explanatory variable should add independent information to the model.

The Identifying Variation from x_{ik}

Using which dataset would you get a more accurate regression line?



Data Generating Process: (Strict) Exogeneity

A3: (Strict) Exogeneity

$$\mathbf{E}[\varepsilon_i \mid X] = \begin{pmatrix} \mathbf{E}[\varepsilon_1 \mid X] \\ \vdots \\ \mathbf{E}[\varepsilon_n \mid X] \end{pmatrix} = \mathbf{0} \qquad \Longleftrightarrow \qquad \mathbf{E}[\varepsilon_i \mid \mathbf{x}_i] = 0 \; \forall i$$

Implications:

$$\mathbf{E}[\varepsilon_i] = 0, \qquad \text{Cov}(\varepsilon_i, x_{ik}) = 0 \; \forall k, \qquad \mathbf{E}[X'\varepsilon_i] = \mathbf{0} \text{ (orthogonality)}.$$

How this connects to the CEF (earlier in 3.1):

- From the CEF decomposition $Y_i = \mathbf{E}[Y_i \mid \mathbf{X}_i] + \varepsilon_i$ we proved $\mathbf{E}[\varepsilon_i \mid \mathbf{X}_i] = 0$ and $\mathbf{E}[h(\mathbf{X}_i)\varepsilon_i] = 0$. *This is exactly the content of A3 for the regression error.*
- In the linear projection (population regression function), $\varepsilon_i = Y_i - \mathbf{x}_i'\beta$ are residuals orthogonal to regressors: $\mathbf{E}[X'\varepsilon_i] = 0$.

Implication of Linearity and Exogeneity

Key result: Combining A1 (Linearity) and A3 (Exogeneity) implies

$$\mathbf{E}[\mathbf{y} \mid X] = X\beta.$$

Why?

- **A1:** The regression model is linear in parameters: $\mathbf{y} = X\beta + \varepsilon_i$.
- **A3:** Exogeneity ensures $\mathbf{E}[\varepsilon_i \mid X] = 0$.

Interpretation:

- The systematic part of $\mathbf{y}$ given X is exactly $X\beta$.
- The regression line coincides with the conditional expectation function under these assumptions.

Exogeneity as key assumption for causal claims

Last Slide:
The systematic part of $\mathbf{y}$ given X is exactly $X\beta$.

Causal interpretation: If A3 holds, a one-unit increase in x_{ik} shifts $\mathbf{E}[y_i \mid X]$ by β_k (ceteris paribus). Without A3, $\hat{\beta}$ is generally biased for causal effects.

But: A3 is not testable; justify it with design, institutional detail, and diagnostics.

When A3 fails (why causal designs are necessary)

- Simultaneity / reverse causality ($y \leftrightarrow x$)
- Omitted unobservables (z affects both x and y)
- Measurement error in x (classical or nonclassical)

How to *argue* A3 (make *x* plausibly exogenous)

- **Randomization / encouragement** (lotteries, nudges)
- **Instrumental variables (IV):** relevance ($\mathbf{cov}(z, x) \neq 0$) & exclusion ($z \perp u$)
- **Difference-in-Differences:** parallel trends $\Rightarrow$ timing exogenous
- **Regression Discontinuity:** local randomization at cutoff
- **Panel / FE:** difference out time-invariant unobservables
- **Careful controls / DAG logic:** block backdoor paths; avoid bad controls

Data Generating Process: Homoskedasticity

A4: Homoskedasticity (and No Autocorrelation)

$$\text{Var}[\varepsilon \mid X] = \mathbf{E}[\varepsilon\varepsilon' \mid X] = \begin{pmatrix} \mathbf{E}[\varepsilon_1^2 \mid X] & \mathbf{E}[\varepsilon_1\varepsilon_2 \mid X] & \cdots & \mathbf{E}[\varepsilon_1\varepsilon_n \mid X] \\ \mathbf{E}[\varepsilon_2\varepsilon_1 \mid X] & \mathbf{E}[\varepsilon_2^2 \mid X] & \cdots & \mathbf{E}[\varepsilon_2\varepsilon_n \mid X] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{E}[\varepsilon_n\varepsilon_1 \mid X] & \mathbf{E}[\varepsilon_n\varepsilon_2 \mid X] & \cdots & \mathbf{E}[\varepsilon_n^2 \mid X] \end{pmatrix}$$

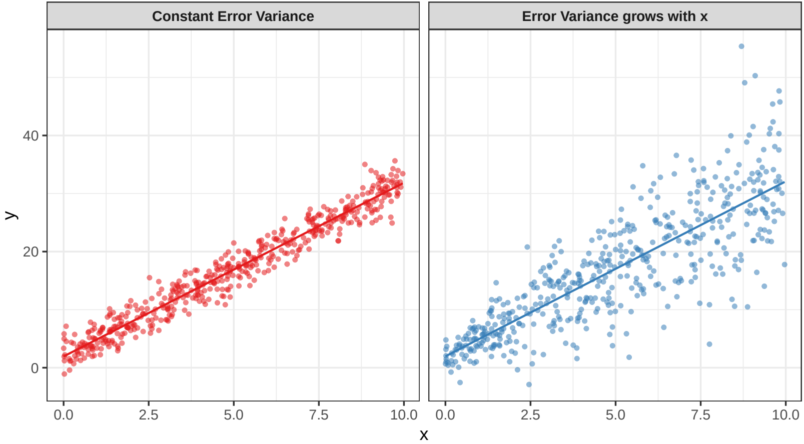
$$= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I_n$$

Implication:

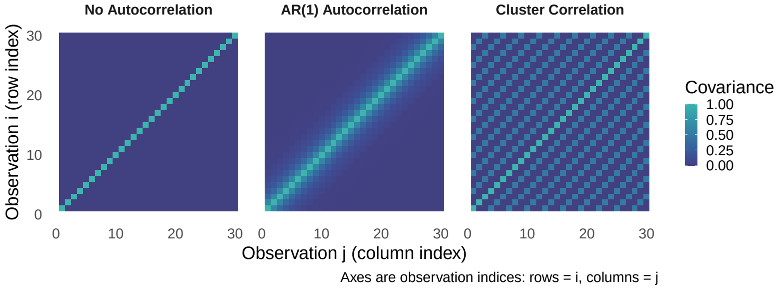
$$\text{Var}[\varepsilon] = \mathbf{E}[\text{Var}(\varepsilon \mid X)] + \text{Var}(\mathbf{E}[\varepsilon \mid X]) = \sigma^2 I_n$$

Homoskedasticity (graphically)

Which line fits our homoskedasticity assumption?



Off-Diagonal Structure of Error Covariance



- **No Autocorrelation:** Errors are independent $\Rightarrow$ only diagonal entries (variances), off-diagonals are zero.
- **AR(1) Autocorrelation:** Nearby errors move together $\Rightarrow$ strong correlation close to the diagonal, fading with distance.
- **Cluster Correlation:** Errors within groups are correlated $\Rightarrow$ block structures along the diagonal.

Data Generating Process: Regressors

A5: Properties of the Regressors

- The regressor matrix X may be treated as
  1. **Nonstochastic** (fixed in repeated samples) – classical textbook case.
  2. **Stochastic** (random) – more realistic in practice.
- In either case, X must be independent of the error process (exogeneity assumption already ensures this).
- Requires that regressors are observed without error.

Interpretation: Whether we treat X as fixed or random does not affect consistency of OLS, but it matters for how we formalize expectations and variances.

Fixed vs. Random Regressors: Why It Matters

⇒ **Fixed X:** Treat X as nonrandom. All uncertainty in $\hat{\beta}$ comes from the random errors ε.

$$E[\hat{\beta} \,|\, X] = \beta, \quad \mathbf{var}(\hat{\beta} \,|\, X) = \sigma^2 (X'X)^{-1}$$

⇒ **Random X:** Both X and ε are random, but $E[\varepsilon|X] = 0$ still ensures unbiasedness. Expectations are now taken over the joint distribution of (X, ε):

$$E[\hat{\beta}] = \beta, \quad \mathbf{var}(\hat{\beta}) = E\left[\sigma^2 (X'X)^{-1}\right]$$

- In large samples, the difference fades:

$$\hat{\beta} \xrightarrow{p} \beta$$

as long as $E[\varepsilon|X] = 0$ and X has full column rank.

Data Generating Process: Normality

A6: Normality (for inference)

$$\varepsilon \mid X \sim \mathcal{N}(0, \sigma^2 I_n)$$

which implies that the errors are

- independent,
- identically distributed,
- Gaussian with mean zero and variance σ^2.

Implications:

- The ε_i are not only uncorrelated but also **independent**.
- OLS estimators $\hat{\beta}$ are normally distributed in finite samples.
- Enables exact *t*- and *F*-tests in small samples.
- Not required for consistency or asymptotic normality of OLS.

In practice, this assumption is often unrealistic; we will later rely on asymptotic approximations instead.

3.2.2: The Least Squares Estimator

The Least Squares Estimator

Setup:

- Observations: $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$
- Population regression model:

$$\mathbf{E}[y_i \mid \mathbf{x}_i] = \mathbf{x}_i' \beta$$

- Disturbance term:

$$\varepsilon_i = y_i - \mathbf{x}_i' \beta$$

Estimation:

- OLS estimates β by $\hat{\beta}$.
- Predicted values and residuals:

$$\hat{y}_i = \mathbf{x}_i' \hat{\beta}, \qquad e_i = y_i - \hat{y}_i.$$

Estimate approximates Population Regression Line:

$$y_i = \mathbf{x}_i' \beta + \varepsilon_i \quad \approx \quad \hat{y}_i + e_i.$$

OLS as Minimization Problem

We minimize the sum of squared residuals:

$$S(\hat{\beta}) = (y - X\hat{\beta})'(y - X\hat{\beta}).$$

Expanding gives:

$$S(\hat{\beta}) = y'y - 2y'X\hat{\beta} + \hat{\beta}'X'X\hat{\beta}.$$

Next step: take the derivative of $S(\hat{\beta})$ with respect to $\hat{\beta}$ to find the minimum.

Deriving the OLS Estimator

FOC: Take the derivative and set equal to zero:

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta} = 0.$$

This gives the normal equations:

$$X'X\hat{\beta} = X'y.$$

Key point: To solve uniquely, $X'X$ must be invertible (A2).

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

Why OLS is a Minimum (Second-Order Condition)

Recall the sum of squared residuals:

$$S(\hat{\beta}) = y'y - 2y'X\hat{\beta} + b'X'X\hat{\beta}.$$

First derivative:

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2X'y + 2X'X\hat{\beta}.$$

Second derivative (Hessian):

$$\frac{\partial^2 S(\hat{\beta})}{\partial \hat{\beta} \partial \hat{\beta}'} = 2X'X.$$

Conclusion: If X has full column rank, $X'X$ is positive definite $\Rightarrow S(\hat{\beta})$ is strictly convex $\Rightarrow$ the OLS solution $\hat{\beta}$ is unique and minimizes $S(\hat{\beta})$.

Key Properties of OLS Residuals

Let $e_i = y_i - \hat{y}_i$ be the residuals.

Two important facts:

- Residuals are uncorrelated with every regressor:

$$\sum_{i=1}^{n} x_{ik} e_i = 0 \quad \text{for each regressor } k.$$

- If a constant is included (which we did!), residuals sum to zero:

$$\sum_{i=1}^{n} e_i = 0.$$

Intuition: The regression line has been chosen so that no systematic pattern is left in the residuals. What remains is “pure noise.”

Projection Interpretation of OLS

OLS can be expressed using projection matrices:

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

where

$$P = X(X'X)^{-1}X'$$

is the **projection matrix**. It projects y onto the part that can be explained by linear combinations of the regressors in X.

Residuals can be written as

$$e = y - \hat{y} = (I - P)y = My,$$

where

$$M = I - P$$

is the **residual maker**.

Properties of P and M

Important properties:

- P and M are **symmetric** and **idempotent**:

$$P^2 = P, \quad M^2 = M.$$

- P keeps any linear combination of regressors unchanged:

$$PX = X.$$

- M removes any linear combination of regressors:

$$MX = 0.$$

- Fitted values and residuals are orthogonal:

$$\hat{y}'e = 0.$$

Example for projection matrix

Example

Show $\boldsymbol{PX} = \boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'X} = \boldsymbol{X}$.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}; \mathbf{X'X} = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ 1 & 1 \end{bmatrix}; \mathbf{X'X}^{-1} = \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 1.5 \end{bmatrix};$$

$$\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1/2 & -1/2 \\ -1/2 & 3/2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix}$$

$$\boldsymbol{PX} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}. \tag{1}$$

Project $\boldsymbol{y}$ on the column space of $\boldsymbol{X}$, i.e. regress $\boldsymbol{y}$ on $\boldsymbol{x}$ and predict $E[\boldsymbol{y}] = \hat{\boldsymbol{y}}$.

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}; \boldsymbol{Py} = \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \hat{\boldsymbol{y}} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}. \tag{2}$$

Example for residual maker matrix

Example

Show $\boldsymbol{M}\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{X} - \boldsymbol{X} = \mathbf{0}$.

$$\boldsymbol{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}; \boldsymbol{X} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix};$$

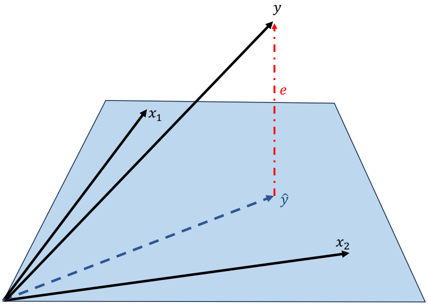
$$\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{P}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} - \begin{bmatrix} 1/2 & 0 & 1/2 \\ 0 & 1 & 0 \\ 1/2 & 0 & 1/2 \end{bmatrix} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix}$$

$$\boldsymbol{M}\boldsymbol{X} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}. \tag{3}$$

Obtain residuals from a projection of $\boldsymbol{y}$ on the column space of $\boldsymbol{X}$, i.e. regress $\boldsymbol{y}$ on $\boldsymbol{x}$ and predict $\boldsymbol{y} - E[\boldsymbol{y}] = \boldsymbol{y} - \hat{\boldsymbol{y}}$.

$$\boldsymbol{y} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}; \boldsymbol{M}\boldsymbol{y} = \begin{bmatrix} 1/2 & 0 & -1/2 \\ 0 & 0 & 0 \\ -1/2 & 0 & 1/2 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}. \tag{4}$$

Projection



$$y = \hat{y} + e, \qquad \hat{y} = Py$$
$$e = (I - P)y, \qquad P = X(X'X)^{-1}X'$$

Intuition

- The shaded plane is the set of all linear combinations of the regressors in X *(column space)*.
- $\hat{y}$ is the point on this plane that lies closest to the observed y.
- The vector $e = y - \hat{y}$ is the vertical "drop" from y to the plane; $\hat{y}$ and e are orthogonal ($\hat{y}'e = 0$).
- Consequence: along the direction of the regressors, there is no systematic pattern left in the residuals.

Goodness of Fit and the Decomposition of Variation

$$y = \hat{y} + e = Py + My$$

$$\underbrace{y'y}_{\text{Total sum of squares (TSS)}} = \underbrace{y'Py}_{\text{Explained}} + \underbrace{y'My}_{\text{Unexplained}}$$

$$= \hat{y}'\hat{y} + e'e$$

$$(y'y - n\bar{y}^2) = (\hat{y}'\hat{y} - n\bar{y}^2) + e'e$$

Total variation in y

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2$$

Note: $\bar{\hat{y}} = \bar{y}$ only if X contains a constant.

Coefficient of Determination

The share of explained variation is measured by R^2:

$$R^2 = \frac{\text{Explained variation}}{\text{Total variation}} = 1 - \frac{\text{Unexplained variation}}{\text{Total variation}}.$$

Properties:

- $0 \leq R^2 \leq 1$.
- $R^2 = 1$: all outcomes are exactly fitted, residuals equal zero.
- $R^2 = 0$: model does no better than predicting the sample mean $\bar{y}$.
- R^2 always increases when additional regressors are included.

Adjusted R^2

Because R^2 never decreases when adding regressors, we often use the **adjusted** R^2:

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-K}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Key idea: Adjusted R^2 penalizes adding regressors that do not improve fit.

- $\bar{R}^2$ may fall if a new regressor contributes little.
- Helps compare models with different numbers of regressors.

3.2.3: Weighted Least Squares (WLS)

Weighted Least Squares (WLS)

Motivation: Ordinary Least Squares minimizes

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i' \hat{\beta})^2,$$

which gives all observations the same weight.

But in many applications:

- Observations have **different reliability** (e.g., group means from different sample sizes),
- or we wish to reflect a **sampling design** with observation-specific probabilities.

Idea: Assign each observation a nonnegative weight w_i, and minimize

$$S_W(b) = \sum_{i=1}^{n} w_i (y_i - \mathbf{x}_i' \hat{\beta})^2.$$

When w_i reflects sampling probability or precision, larger weights make an observation count more in the fit.

Deriving the WLS Estimator

Write the criterion in matrix form:

$$S_W(\hat{\beta}) = (y - X\hat{\beta})' W (y - X\hat{\beta}), \qquad W = \mathrm{diag}(w_1, \ldots, w_n).$$

First-order condition:

$$\frac{\partial S_W(\hat{\beta})}{\partial \hat{\beta}} = -2X'Wy + 2X'WX\hat{\beta} = 0.$$

Normal equations:

$$X'WX\hat{\beta}_{WLS} = X'Wy.$$

Solution:

$$\hat{\beta}_{WLS} = (X'WX)^{-1}X'Wy.$$

Interpretation of WLS

Equivalent transformation: If $W^{1/2}$ denotes the diagonal matrix of $\sqrt{w_i}$, then WLS is simply OLS on the transformed model:

$$W^{1/2}y = W^{1/2}X\beta + W^{1/2}u.$$

Interpretations:

- Observations with large w_i are given more influence in fitting the regression line.
- When w_i are proportional to the inverse of the sampling variance, this yields an estimator that reflects the relative precision of each observation.
- When w_i correspond to inverse sampling probabilities, the regression estimates are representative of the population defined by that design.

Special case: $w_i = 1$ for all i gives OLS.

When to Use Weighted Least Squares

Common situations:

- Survey data with sampling weights.
- Grouped data where each observation is an average of different sample sizes.
- Heteroskedasticity with known or estimable variance pattern $\sigma_i^2 \propto 1/w_i$.

Practical notes:

- The choice of weights changes the estimand—WLS estimates the linear relationship in the weighted population.
- Always check whether weights are due to sampling design or model assumptions; interpretation differs.

References and Further Ressources

References and Further Resources

- **Greene, W. H.** (2018). Econometric Analysis. Pearson. Chapters 2–3.
- **Rubin, E.** Introduction to Econometrics (EC421). Lecture materials and visual intuition for the Conditional Expectation Function, population vs. sample regression, and Monte Carlo simulations. github.com/edrubin/EC421W22

Several graphical illustrations in this lecture are inspired by and adapted from Ed Rubin's EC421 course materials. Highly recommended as a complementary resource.