# Applied Regression (I)

Francisco Villamil

Applied Quantitative Methods II

IC3JM, Spring 2026

# Roadmap

# Review: Key Concepts from AQMSS-I

# The regression model

The most common tool in social science:

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

- $Y$: outcome we want to explain

# The regression model

The most common tool in social science:

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

- $Y$: outcome we want to explain
- $X$: explanatory variable(s)

# The regression model

The most common tool in social science:

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

- $Y$: outcome we want to explain
- $X$: explanatory variable(s)
- $\beta$: coefficients (what we estimate)

# The regression model

The most common tool in social science:

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

- $Y$: outcome we want to explain
- $X$: explanatory variable(s)
- $\beta$: coefficients (what we estimate)
- $\varepsilon$: error term (what we can't explain)

# What regression tells us

- Regression estimates **conditional expectations**

# What regression tells us

- Regression estimates **conditional expectations**
- "What is the average $Y$ for units with a given value of $X$?"

# What regression tells us

- Regression estimates **conditional expectations**
- "What is the average $Y$ for units with a given value of $X$?"

# What regression tells us

- Regression estimates **conditional expectations**
- "What is the average $Y$ for units with a given value of $X$?"

- The slope $\beta_1$ tells us:

# What regression tells us

- Regression estimates **conditional expectations**
- "What is the average $Y$ for units with a given value of $X$?"

- The slope $\beta_1$ tells us:
  - $\rightarrow$ How much $Y$ changes, on average

# What regression tells us

- Regression estimates **conditional expectations**
- "What is the average $Y$ for units with a given value of $X$?"

- The slope $\beta_1$ tells us:
  - $\rightarrow$ How much $Y$ changes, on average
  - $\rightarrow$ When comparing units that differ by 1 in $X$

# Descriptive vs. Causal interpretation

- **Descriptive**: How do units with different $X$ values compare?
  - $\rightarrow$ "People with more education earn more, on average"
- **Causal**: What happens if we change $X$ for a given unit?
  - $\rightarrow$ "If we give someone more education, they will earn more"
- Same coefficient, very different claims!

# The challenge of causal inference

- Causal effects are about **counterfactuals**

# The challenge of causal inference

- Causal effects are about **counterfactuals**
- "What would have happened if things were different?"

# The challenge of causal inference

- Causal effects are about **counterfactuals**
- "What would have happened if things were different?"

# The challenge of causal inference

- Causal effects are about **counterfactuals**
- "What would have happened if things were different?"

- The problem: we can't observe counterfactuals

# The challenge of causal inference

- Causal effects are about **counterfactuals**
- "What would have happened if things were different?"

- The problem: we can't observe counterfactuals
- We need strategies to infer them

# The challenge of causal inference

- Causal effects are about **counterfactuals**
- "What would have happened if things were different?"

- The problem: we can't observe counterfactuals
- We need strategies to infer them
- This will be a recurring theme throughout the course

# Today's goals

- Understand regression as modeling conditional expectations

# Today's goals

- Understand regression as modeling conditional expectations
- Review the logic of OLS

# Today's goals

- Understand regression as modeling conditional expectations
- Review the logic of OLS
- Discuss when regression can tell us about causation

# Today's goals

- Understand regression as modeling conditional expectations
- Review the logic of OLS
- Discuss when regression can tell us about causation
- Learn how to think about control variables

# Roadmap

# Regression as Conditional Expectations

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"
- This is the **conditional expectation function** (CEF)

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"
- This is the **conditional expectation function** (CEF)

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"
- This is the **conditional expectation function** (CEF)
- Written as: $E[Y|X]$

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"
- This is the **conditional expectation function** (CEF)
- Written as: $E[Y|X]$

# What question does regression answer?

- "What is the average value of $Y$ for different values of $X$?"
- This is the **conditional expectation function** (CEF)
- Written as: $E[Y|X]$
- Regression approximates this function

# Example: Income and support for redistribution

- Research question: How does income relate to support for redistribution?

# Example: Income and support for redistribution

- Research question: How does income relate to support for redistribution?

# Example: Income and support for redistribution

- Research question: How does income relate to support for redistribution?
- CEF: "What is the average support for redistribution among people earning \$50k? Among those earning \$100k?"

# Example: Income and support for redistribution

- Research question: How does income relate to support for redistribution?

- CEF: “What is the average support for redistribution among people earning $50k? Among those earning $100k?”

# Example: Income and support for redistribution

- Research question: How does income relate to support for redistribution?
- CEF: “What is the average support for redistribution among people earning \$50k? Among those earning \$100k?”
- We can estimate this with regression

# Linear regression as approximation

- The true CEF might be complicated

# Linear regression as approximation

- The true CEF might be complicated
- Linear regression fits the **best linear approximation**

# Linear regression as approximation

- The true CEF might be complicated
- Linear regression fits the **best linear approximation**

# Linear regression as approximation

- The true CEF might be complicated
- Linear regression fits the **best linear approximation**

- Even if the true relationship is non-linear

# Linear regression as approximation

- The true CEF might be complicated
- Linear regression fits the **best linear approximation**

- Even if the true relationship is non-linear
- The linear fit is still the best predictor among linear functions

# Linear regression as approximation

- The true CEF might be complicated
- Linear regression fits the **best linear approximation**

- Even if the true relationship is non-linear
- The linear fit is still the best predictor among linear functions

# Linear regression as approximation

- The true CEF might be complicated
- Linear regression fits the **best linear approximation**

- Even if the true relationship is non-linear
- The linear fit is still the best predictor among linear functions

- Why linear? Simple, interpretable, often good enough

# The OLS formula

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

- This gives us the slope that minimizes squared errors

# The OLS formula

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

- This gives us the slope that minimizes squared errors
- Intuition: how much does $Y$ move when $X$ moves?

# The OLS formula

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

- This gives us the slope that minimizes squared errors
- Intuition: how much does $Y$ move when $X$ moves?
- Scaled by how much $X$ varies

# Interpreting the slope coefficient

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

- $\beta_1$ represents:

# Interpreting the slope coefficient

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

- $\beta_1$ represents:
  - $\rightarrow$ The difference in average $Y$

# Interpreting the slope coefficient

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

- $\beta_1$ represents:
  - $\rightarrow$ The difference in average $Y$
  - $\rightarrow$ Between groups that differ by 1 unit in $X$

# Interpreting the slope coefficient

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

- $\beta_1$ represents:
  - $\rightarrow$ The difference in average $Y$
  - $\rightarrow$ Between groups that differ by 1 unit in $X$

# Interpreting the slope coefficient

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

- $\beta_1$ represents:
  - $\rightarrow$ The difference in average $Y$
  - $\rightarrow$ Between groups that differ by 1 unit in $X$

- This is a **comparison**, not necessarily a causal effect

# Roadmap

# From Description to Causation

# When can we interpret regression causally?

- Descriptive interpretation: always valid

# When can we interpret regression causally?

- Descriptive interpretation: always valid
  - → "Higher income is associated with less support for redistribution"

# When can we interpret regression causally?

- Descriptive interpretation: always valid
  - → "Higher income is associated with less support for redistribution"

# When can we interpret regression causally?

- Descriptive interpretation: always valid
  - → "Higher income is associated with less support for redistribution"
- Causal interpretation: requires additional assumptions

# When can we interpret regression causally?

- Descriptive interpretation: always valid
  - → "Higher income is associated with less support for redistribution"
- Causal interpretation: requires additional assumptions
  - → "Increasing someone's income would decrease their support"

# When can we interpret regression causally?

- Descriptive interpretation: always valid
  - → "Higher income is associated with less support for redistribution"

- Causal interpretation: requires additional assumptions
  - → "Increasing someone's income would decrease their support"

# When can we interpret regression causally?

- Descriptive interpretation: always valid
  - → "Higher income is associated with less support for redistribution"
- Causal interpretation: requires additional assumptions
  - → "Increasing someone's income would decrease their support"
- The difference is crucial!

# The potential outcomes framework

- Every unit has two potential outcomes:

# The potential outcomes framework

- Every unit has two potential outcomes:
  - → $Y(1)$: outcome if treated

# The potential outcomes framework

- Every unit has two potential outcomes:
  - → $Y(1)$: outcome if treated
  - → $Y(0)$: outcome if not treated

# The potential outcomes framework

- Every unit has two potential outcomes:
  - → $Y(1)$: outcome if treated
  - → $Y(0)$: outcome if not treated

# The potential outcomes framework

- Every unit has two potential outcomes:
  - → $Y(1)$: outcome if treated
  - → $Y(0)$: outcome if not treated

- Causal effect for unit $i$: $\tau_i = Y_i(1) - Y_i(0)$

# The potential outcomes framework

- Every unit has two potential outcomes:
  - → $Y(1)$: outcome if treated
  - → $Y(0)$: outcome if not treated

- Causal effect for unit $i$: $\tau_i = Y_i(1) - Y_i(0)$

# The potential outcomes framework

- Every unit has two potential outcomes:
  - → $Y(1)$: outcome if treated
  - → $Y(0)$: outcome if not treated

- Causal effect for unit $i$: $\tau_i = Y_i(1) - Y_i(0)$

- The fundamental problem: we only observe one of these

# Why experiments work

- In an experiment, treatment is randomly assigned

# Why experiments work

- In an experiment, treatment is randomly assigned

# Why experiments work

- In an experiment, treatment is randomly assigned
- This means treated and control groups are comparable

# Why experiments work

- In an experiment, treatment is randomly assigned
- This means treated and control groups are comparable

# Why experiments work

- In an experiment, treatment is randomly assigned
- This means treated and control groups are comparable
- We can use the control group's outcomes as counterfactual

# Why experiments work

- In an experiment, treatment is randomly assigned
- This means treated and control groups are comparable
- We can use the control group's outcomes as counterfactual

# Why experiments work

- In an experiment, treatment is randomly assigned
- This means treated and control groups are comparable
- We can use the control group's outcomes as counterfactual
- The simple difference in means estimates the causal effect

# The challenge with observational data

- Most social science data is observational

# The challenge with observational data

- Most social science data is observational
- Treatment is not randomly assigned

# The challenge with observational data

- Most social science data is observational
- Treatment is not randomly assigned

# The challenge with observational data

- Most social science data is observational
- Treatment is not randomly assigned

- Problem: treated and control groups may differ

# The challenge with observational data

- Most social science data is observational
- Treatment is not randomly assigned

- Problem: treated and control groups may differ
- Not just in treatment, but in other ways too

# The challenge with observational data

- Most social science data is observational
- Treatment is not randomly assigned

- Problem: treated and control groups may differ
- Not just in treatment, but in other ways too

# The challenge with observational data

- Most social science data is observational
- Treatment is not randomly assigned

- Problem: treated and control groups may differ
- Not just in treatment, but in other ways too

- These differences can bias our estimates

# Confounding

A **confounder** is a variable that:

- Affects both the treatment and the outcome
- Creates a spurious association between them

- Example: Education, income, and political preferences

# Confounding

A **confounder** is a variable that:

- Affects both the treatment and the outcome
- Creates a spurious association between them

- Example: Education, income, and political preferences
- Education affects both income and political views

# Confounding

A **confounder** is a variable that:

- Affects both the treatment and the outcome
- Creates a spurious association between them

- Example: Education, income, and political preferences
- Education affects both income and political views
- Income-politics relationship may be partly spurious

# The logic of controlling

- If we can identify the confounders...

# The logic of controlling

- If we can identify the confounders...
- ...we can “control” for them in regression

# The logic of controlling

- If we can identify the confounders...
- ...we can “control” for them in regression

# The logic of controlling

- If we can identify the confounders...
- ...we can “control” for them in regression

- The idea: compare units with same confounder values

# The logic of controlling

- If we can identify the confounders...
- ...we can "control" for them in regression

- The idea: compare units with same confounder values
- This eliminates the spurious part of the association

# The logic of controlling

- If we can identify the confounders...
- ...we can "control" for them in regression

- The idea: compare units with same confounder values
- This eliminates the spurious part of the association

# The logic of controlling

- If we can identify the confounders...
- ...we can "control" for them in regression

- The idea: compare units with same confounder values
- This eliminates the spurious part of the association

- But: this requires knowing what the confounders are

# Roadmap

# Control Variables in Practice

# Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- $\beta_1$ now represents:

# Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- $\beta_1$ now represents:
  - $\rightarrow$ The difference in average $Y$

# Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- $\beta_1$ now represents:
  - $\rightarrow$ The difference in average $Y$
  - $\rightarrow$ Between groups that differ by 1 in $X_1$

# Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- $\beta_1$ now represents:
  - → The difference in average $Y$
  - → Between groups that differ by 1 in $X_1$
  - → **Holding $X_2$ constant**

# Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- $\beta_1$ now represents:
  - → The difference in average $Y$
  - → Between groups that differ by 1 in $X_1$
  - → **Holding $X_2$ constant**

# Multiple regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

- $\beta_1$ now represents:
  - $\rightarrow$ The difference in average $Y$
  - $\rightarrow$ Between groups that differ by 1 in $X_1$
  - $\rightarrow$ **Holding $X_2$ constant**

- This is the "controlled" effect of $X_1$

# How controlling works

- OLS with multiple variables "partials out" the controls

# How controlling works

- OLS with multiple variables "partials out" the controls

# How controlling works

- OLS with multiple variables "partials out" the controls
- Technically: we look at variation in $X_1$ that is unrelated to $X_2$

# How controlling works

- OLS with multiple variables “partials out” the controls
- Technically: we look at variation in $X_1$ that is unrelated to $X_2$

# How controlling works

- OLS with multiple variables "partials out" the controls
- Technically: we look at variation in $X_1$ that is unrelated to $X_2$
- This isolates the unique contribution of $X_1$

# Omitted variable bias

- If we omit a confounder, our estimate will be biased

# Omitted variable bias

- If we omit a confounder, our estimate will be biased

# Omitted variable bias

- If we omit a confounder, our estimate will be biased
- The bias formula:

$$\text{Bias} = \beta_{\text{confounder}} \times \delta_{X,\text{confounder}}$$

# Omitted variable bias

- If we omit a confounder, our estimate will be biased
- The bias formula:

$$\text{Bias} = \beta_{\text{confounder}} \times \delta_{X,\text{confounder}}$$

# Omitted variable bias

- If we omit a confounder, our estimate will be biased
- The bias formula:

$$\text{Bias} = \beta_{\text{confounder}} \times \delta_{X,\text{confounder}}$$

- Depends on:

# Omitted variable bias

- If we omit a confounder, our estimate will be biased
- The bias formula:

$$\text{Bias} = \beta_{\text{confounder}} \times \delta_{X,\text{confounder}}$$

- Depends on:
  - $\rightarrow$ How strongly the confounder affects $Y$

# Omitted variable bias

- If we omit a confounder, our estimate will be biased
- The bias formula:

$$\text{Bias} = \beta_{\text{confounder}} \times \delta_{X,\text{confounder}}$$

- Depends on:
  - $\rightarrow$ How strongly the confounder affects $Y$
  - $\rightarrow$ How strongly the confounder relates to $X$

# What makes a good control?

Good controls are variables that:

- Affect both the treatment and the outcome

# What makes a good control?

Good controls are variables that:

- Affect both the treatment and the outcome
- Are determined **before** the treatment

# What makes a good control?

Good controls are variables that:

- Affect both the treatment and the outcome
- Are determined **before** the treatment
- Are not affected by the treatment

# What makes a good control?

Good controls are variables that:

- Affect both the treatment and the outcome
- Are determined **before** the treatment
- Are not affected by the treatment

# What makes a good control?

Good controls are variables that:

- Affect both the treatment and the outcome
- Are determined **before** the treatment
- Are not affected by the treatment

**Pre-treatment confounders** are the key!

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment
- Example: Studying effect of job training on wages

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment

- Example: Studying effect of job training on wages
  $\rightarrow$ Don't control for job type (affected by training)

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment

- Example: Studying effect of job training on wages
  - $\rightarrow$ Don't control for job type (affected by training)
  - $\rightarrow$ Do control for education (determined before training)

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment

- Example: Studying effect of job training on wages
  - → Don't control for job type (affected by training)
  - → Do control for education (determined before training)

# Bad controls: Post-treatment variables

- Never control for variables caused by the treatment
- Example: Studying effect of job training on wages
  - $\rightarrow$ Don't control for job type (affected by training)
  - $\rightarrow$ Do control for education (determined before training)
- Controlling for post-treatment variables can *introduce* bias

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$
- Controlling for it creates a spurious association

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$
- Controlling for it creates a spurious association

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$
- Controlling for it creates a spurious association

- Example: NBA players

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$
- Controlling for it creates a spurious association

- Example: NBA players
  - $\rightarrow$ Height and skill both affect being in NBA

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$
- Controlling for it creates a spurious association

- Example: NBA players
  - $\rightarrow$ Height and skill both affect being in NBA
  - $\rightarrow$ Among NBA players, height and skill are negatively correlated

# Bad controls: Colliders

- A **collider** is caused by both $X$ and $Y$
- Controlling for it creates a spurious association

- Example: NBA players
  - → Height and skill both affect being in NBA
  - → Among NBA players, height and skill are negatively correlated
  - → But not in the general population!

# The limitations of controlling

- We can only control for what we observe and measure

# The limitations of controlling

- We can only control for what we observe and measure

# The limitations of controlling

- We can only control for what we observe and measure
- Unobserved confounders will still bias our estimates

# The limitations of controlling

- We can only control for what we observe and measure
- Unobserved confounders will still bias our estimates

# The limitations of controlling

- We can only control for what we observe and measure
- Unobserved confounders will still bias our estimates
- There's no purely statistical solution to this

# The limitations of controlling

- We can only control for what we observe and measure
- Unobserved confounders will still bias our estimates
- There's no purely statistical solution to this

# The limitations of controlling

- We can only control for what we observe and measure
- Unobserved confounders will still bias our estimates
- There's no purely statistical solution to this
- Need theory + research design, not just more controls

# Summary: Key takeaways

- Regression estimates conditional expectations

# Summary: Key takeaways

- Regression estimates conditional expectations
- Causal interpretation requires additional assumptions

# Summary: Key takeaways

- Regression estimates conditional expectations
- Causal interpretation requires additional assumptions
- Control variables help only if chosen correctly

# Summary: Key takeaways

- Regression estimates conditional expectations
- Causal interpretation requires additional assumptions
- Control variables help only if chosen correctly
- Controlling for the wrong variables can make things worse

# Summary: Key takeaways

- Regression estimates conditional expectations
- Causal interpretation requires additional assumptions
- Control variables help only if chosen correctly
- Controlling for the wrong variables can make things worse
- Always think about what you're comparing

# For next week

- Read Angrist & Pischke (2008), chapters 1-3
- Read Urdinez & Cruz (2020), chapter 5
- Work on Problem Set 1

- Next session: More on regression in practice
  - → Interactions
  - → Non-linear relationships
  - → Standard errors and inference

Questions?