Applied Regression

Francisco Villamil

Applied Quantitative Methods II

IC3JM, Spring 2026

Today's goals

- Review regression as modeling conditional expectations
- Understand multiple regression and control variables
- Learn how to model conditional relationships (interactions)
- Present results effectively with `modelsummary`

Roadmap

Regression Review

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

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)
- β: coefficients (what we estimate)
- ε: error term (what we can't explain)

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

Interpreting the slope coefficient

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

- β_1 represents:
  - → The difference in average Y
  - → Between groups that differ by 1 unit in X

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

Descriptive vs. Causal interpretation

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

Running a regression in R

- The basic function: `lm(y ~ x, data = df)`
- Getting tidy output:
  - → `broom::tidy(model)` — coefficients as a data frame
  - → `broom::glance(model)` — model-level statistics (R^2, etc.)
- These are much easier to work with than `summary()`

Roadmap

Multiple Regression

Adding predictors

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

- β_1 now represents:
  - → The difference in average Y
  - → Between groups that differ by 1 in X_1
  - → **Holding X_2 constant**
- This is the "controlled" effect of X_1

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
- The bias formula:

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

- Depends on:
  → How strongly the confounder affects Y
  → How strongly the confounder relates to X

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

Categorical predictors

- What if X is a category (region, party, gender)?
- R automatically creates **dummy variables**
  - → One indicator (0/1) for each category
  - → One category is the **reference** (omitted)
- Coefficients represent the difference from the reference
- Example: `lm(income ~ factor(region), data = df)`
  - → If reference is "North", the "South" coefficient means: average income in South minus average income in North

Roadmap

Interaction Effects

When effects depend on context

- Sometimes, the effect of X on Y depends on another variable Z
- Examples:
  - → Effect of education on income may differ by gender
  - → Effect of campaign spending may differ by incumbency status
  - → Effect of democracy on growth may depend on economic development
- We model this with **interaction terms**

The interaction model

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 (X \times Z) + \varepsilon$$

- β_1: effect of X when $Z = 0$
- β_2: effect of Z when $X = 0$
- β_3: how the effect of X changes as Z increases

The marginal effect of X

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z$$

- The effect of X is no longer a single number
- It's a **function** of Z

- Need to report effects at meaningful values of Z

Continuous × categorical interactions

- When Z is categorical (e.g., gender, regime type)
- The interaction gives a **different slope** for each group

- Example: `lm(income ~ education * gender, data = df)`
  - → One slope for men, a different slope for women

- Equivalent to fitting separate regressions by group
- But estimated jointly (shares the error variance)

Continuous × continuous interactions

- When both X and Z are continuous
- The slope of X varies smoothly with Z (and vice versa)

- Harder to interpret from coefficients alone

- Best communicated through plots:
  - → Predicted values at different combinations of X and Z
  - → Marginal effect of X across values of Z

Common mistakes with interactions

- **Mistake 1**: Interpreting β_1 as "the effect of X"
  - → It's only the effect when $Z = 0$
  - → May not even be meaningful!
- **Mistake 2**: Omitting constitutive terms
  - → Always include X and Z separately, not just $X \times Z$
- **Mistake 3**: Not showing how the effect varies
  - → Plot the marginal effect across values of Z

Visualizing interactions

- Tables of coefficients are hard to interpret
- Better approach:
  - → Plot predicted values of Y for different combinations of X and Z
  - → Plot the marginal effect of X across values of Z
  - → Include confidence intervals
- In R: `marginaleffects::plot_predictions()`

Roadmap

Presenting Results

Why presentation matters

- A regression table is not the end of the analysis
- Readers need to understand the **substance** of your findings
- Good presentation:
  - → Shows what the results **mean**, not just what they are
  - → Communicates **uncertainty** honestly
  - → Helps readers evaluate the **size** of effects

The `modelsummary` package

- Creates publication-quality tables from model objects
- Basic usage:
  - → `modelsummary(model)`
  - → `modelsummary(list(m1, m2, m3))`
- Output formats: LaTeX, HTML, Word, markdown
- Highly customizable: statistics, labels, notes

Coefficient plots

- A visual alternative to tables

- `modelsummary::modelplot(model)`
  - → Each coefficient as a point with confidence interval
  - → Easy to compare multiple models

- Often more effective than tables for communicating results
- Readers immediately see which effects are large vs. small

Building sequential models

- Common strategy: show how results change as you add variables

- Step 1: Bivariate model (just X and Y)
- Step 2: Add control variables
- Step 3: Add interactions

- Present all three in one table:
  - → `modelsummary(list(m1, m2, m3))`

- Shows robustness and what adding controls does to the estimate

Example workflow in R

```
m1 <- lm(y ~ x, data = df)
m2 <- lm(y ~ x + z1 + z2, data = df)
m3 <- lm(y ~ x * z1 + z2, data = df)
modelsummary(list(m1, m2, m3))
modelplot(list(m1, m2, m3))
plot_predictions(m3, condition = c("x", "z1"))
```


Summary: Key takeaways

- Regression estimates conditional expectations
- Multiple regression: "holding constant" interpretation
- Control variables help only if chosen correctly
- Interactions model conditional relationships
- Present results clearly: tables, coefficient plots, marginal effects

For next week

- Read Urdinez & Cruz (2020), chapter 8
- Read Gelman et al., chapters 13–14
- Complete Assignment 2

- Next session: Binary outcomes
  - → Linear probability model vs. logistic regression
  - → Interpreting logit results
  - → Predicted probabilities and marginal effects

Questions?