Lecture 6 – Linear Regression 2 — Semiparametrics and Visualization

Paul Goldsmith-Pinkham

January 24, 2025

Linear regression remains one of the most widely used tools in empirical research. This popularity stems from several key advantages:

1. Computational efficiency through analytic solutions and improved matrix inversion 2. Statistical efficiency (OLS is BLUE under classical assumptions) 3. Most importantly: - Provides intuitive summaries of data relationships - Serves as a robust default method compared to alternatives that may only perform well in specific settings - Handles common empirical challenges well (high dimensional fixed effects, large datasets)

General Framework for Causal Relationships

Without imposing structure, we can describe relationships in our data as:

$$Y_i = F(D_i, W_i, \epsilon_i)$$

where: - D_i is our causal variable of interest - W_i represents controls/heterogeneity - ϵ_i captures unobservable noise

This general formulation is very challenging to estimate when ϵ_i enters non-separably or when D_i or W_i are high-dimensional. A simpler version separates out the error term:

$$Y_i = F(D_i, W_i) + \epsilon_i$$

Even with this simplification, we face choices about what features of the function to estimate and report: - Average partial effects: $E\left(\frac{\partial F}{\partial D_i}\middle| W_i = w\right)$ - Population average effects: $E\left(\frac{\partial F}{\partial D_i}\right)$

The linear model further restricts this to:

$$Y_i = D_i\tau + W_i\beta + \epsilon_i$$

We can add complexity through interactions:

$$Y_i = D_i\tau + W_i\beta_1 + D_i \times W_i\beta_2 + \epsilon_i$$

Visualizing Relationships

Comment 1
When plotting relationships between outcomes and causal variables, regression lines provide useful summaries, especially when: 1. The underlying relationship appears approximately linear 2. There are many data points making patterns difficult to discern 3. We want to communicate the average relationship succinctly

Multivariate Relationships and Controls

A key challenge arises when considering control variables. Consider the case where W represents discrete fixed effects. In the potential outcomes framework with propensity scores, we would estimate:

$$\tau(w) = E(Y|D_i = 1, W = w) - E(Y|D_i = 0, W = w)$$

and aggregate using inverse probability weighting (IPW). OLS performs this aggregation automatically through its weighting scheme.

The Residual Regression Approach

To understand how OLS handles controls, consider our basic specification:

$$Y_i = \tau D_i + \beta W_i + \epsilon_i$$

We can use projection matrices to better understand this: - Define projection matrix $\mathbf{P}_W = \mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'$ - Note $\mathbf{P}_W\mathbf{W}_n = \mathbf{W}_n$ and $\mathbf{P}_W\mathbf{P}_W = \mathbf{P}_W$ - $\mathbf{P}_W\mathbf{D}_n$ gives predicted values from regressing D_i on W_i - Define annihilator matrix $\mathbf{M}_W = \mathbf{I}_n - \mathbf{P}_W$ which gives residuals

Comment 2 (Frisch-Waugh-Lovell Theorem)
If we transform $\mathbf{Y}_n^* = \mathbf{M}_W\mathbf{Y}_n$ and $\mathbf{D}_n^* = \mathbf{M}_W\mathbf{D}_n$, running the regression

$$Y_i^* = \tau D_i^* + \tilde{\epsilon}_i$$

yields the same coefficient τ as the original multivariate regression. This provides a powerful tool for both computation and visualization.

Discrete Treatment Effects with Controls

Consider a simplified example with binary treatment D_i and covariate $W_i \in \{0, 1\}$:

$$Y_i = \alpha + D_i\beta + W_i\gamma + U_i$$

Using potential outcomes notation:

- $Y_i(d)$ represents potential outcomes
- Individual treatment effect $\tau_{i1} = Y_i(1) - Y_i(0)$
- Conditional treatment effect $\tau_1(w) = E[\tau_{i1}|W_i = w]$
- Observed outcome $Y_i = Y_i(0) + \tau_{i1}D_i$
- Propensity score $p_1(W_i) = Pr(D_i = 1|W_i)$

Under conditional random assignment $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|W_i$, we get:

$$\beta = \phi\tau_1(0) + (1 - \phi)\tau_1(1)$$

where

$$\phi = \frac{Var(D_i|W_i = 0)Pr(W_i = 0)}{\sum_{w=0}^{1} Var(D_i|W_i = w)Pr(W_i = w)}$$

Example 1 (Key Properties)
The OLS estimator with controls has several important features: 1. Weights ϕ are bounded between 0 and 1 2. No need to explicitly estimate propensity scores 3. Puts more weight on strata with higher treatment variation 4. May differ from ATE unless treatment effects are constant or propensity scores are constant across strata 5. Weighting structure helps handle lack of overlap

Multiple Treatment Arms

The framework extends to multiple treatments. Consider adding a second treatment arm (e.g., teaching aide in Project STAR):

$$Y_i = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2 + W_i\gamma + U_i$$

where: - $X_{ij} = \mathbb{1}\{D_i = j\}$ for treatments $j = 1, 2$ - $\tau_{ik} = Y_i(k) - Y_i(0)$ are treatment effects - $\tau_k(W_i) = E[\tau_{ik}|W_i]$ are conditional effects - $p_{ok}(w) = E[X_{ik}|W_i = w]$ are propensity scores

The interpretation becomes more complex with multiple treatments. For β_1:

$$\beta_1 = E[\lambda_{11}(W_i)\tau_1(W_i)] + E[\lambda_{12}(W_i)\tau_2(W_i)]$$

where $\lambda_{11}(W_i)$ and $\lambda_{12}(W_i)$ are weights.

Comment 3
A key insight is that β_1 is "contaminated" by τ_2 effects because X_{i1} and X_{i2} have a nonlinear relationship - they cannot occur simultaneously. The residualization of X_{i1} involves both W_i and X_{i2}.

Visualization with Controls

The Frisch-Waugh-Lovell theorem provides a powerful visualization approach:

1. Plot residualized Y^* against residualized D^* 2. Add back overall means to make interpretation more intuitive 3. Consider potential issues: - Residualized variables can be hard to interpret - Need to carefully consider what mean adjustments to make - Should reflect the variation identifying your parameter of interest

Comment 4
When adding back means, be careful about interpretation - the relationship you visualize should correspond to the estimand you care about. Simply adding back raw means may not achieve this.

Nonparametric and Semiparametric Approaches

Let's consider more flexible approaches to modeling relationships. Our interest lies in conditional expectation functions $E(Y|D)$. The framework can be categorized into:

Definition 1 (Model Types)
1. Parametric: Finite dimensional specification

$$Y_i = D_i\beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2)$$

2. Nonparametric: Infinite dimensional specification

$$Y_i = F(D_i, \theta_i)$$

where θ_i is infinite-dimensional
3. Semiparametric: Combination of both

$$Y_i = D_i\beta + \epsilon_i, \quad \epsilon_i \sim F(\theta_i)$$

where β is finite-dimensional but θ_i is infinite-dimensional

Binscatter Analysis

One popular approach to flexible estimation is binscatter analysis. Consider:

$$Y_i = f(D_i, \theta) + \epsilon_i$$

The binscatter approach approximates $f(\cdot)$ by: 1. Dividing observations into equally-sized bins 2. Computing means within each bin 3. Plotting these means against bin centers

Comment 5
Key considerations for binscatter: 1. Choice of bin width affects visualization 2. Tradeoff between smoothness and noise 3. Need to handle controls appropriately 4. Statistical inference remains important

Advances in Binscatter (Cattaneo et al.)

Recent work by Cattaneo et al. provides important methodological advances:

1. Formal statistical framework - Recognizes binscatter as nonparametric estimation - Provides basis for inference

2. Optimal bin selection - Balances bias and variance - Data-driven approach using $\approx n^{1/3}$ rule

3. Correct handling of controls - Cannot simply residualize when f is nonlinear - Need to bin first, then partial out controls

4. Statistical inference - Confidence intervals - Tests for shape restrictions (e.g., monotonicity)

Comment 6
The traditional approach of residualizing before binning can produce misleading results. The correct approach is to: 1. Create bins based on raw treatment variable 2. Estimate bin-specific treatment effects controlling for covariates 3. Plot these conditional effects

I'll continue converting the slides into lecture notes, being careful to use proper LaTeX formatting:

Design Principles for Research Communication

Key design principles for academic work presentation include:

1. Minimize tables
2. Establish clear exhibit goals

3. Create visually appealing figures
4. Present data honestly and clearly

For figure design specifically, Schwabish provides guidelines:

1. Present data clearly
2. Eliminate unnecessary elements
3. Integrate graphics with text effectively
4. Avoid extraneous information
5. Use grey as base color

Table Minimization

Tables serve as important repositories of information but have limitations:

- Difficult to interpret comparisons across specifications
- Often contain unnecessary information
- Better suited for online appendices than main text
- Control variable coefficients rarely have causal interpretation (Hunermund and Louw, 2020)

Example 2
Consider regression output presentation:

- *Traditional table format makes comparison difficult*
- *Coefficient plots show patterns more clearly*
- *Can compress multiple specifications into single visualization*
- *Particularly effective for presentation settings*

Clear Exhibit Goals

Research paper exhibits should:

- Have immediately obvious purpose
- Focus reader attention effectively
- Avoid information overload
- Highlight key findings

Comment 7
If an exhibit's purpose isn't clear, usually either:

- *Too much information obscures the main message*
- *Insufficient emphasis on key elements*

Visual Design Principles

To improve figure quality:

1. Choose appropriate color schemes
2. Label axes clearly
3. Use consistent visual elements
4. Apply appropriate line weights and point sizes
5. Consider readability of all elements

Comment 8
Even small improvements in visual design can significantly enhance communication:

- *Replace default color schemes*
- *Add clear labels*
- *Adjust element sizes for emphasis*
- *Ensure consistent formatting*

Honest Data Representation

Key principles for ethical visualization:

- Avoid misleading visual tricks
- Present uncertainty appropriately
- Use consistent scales
- Show discrete data appropriately (e.g., avoid implying continuity)

Making Good Figures: Practical Guidelines

Best practices for creating effective figures include:

1. Use horizontal bar graphs for readability

2. Avoid confidence intervals on bar graphs - use point range plots
3. Label directly on figures when possible
4. Format units appropriately:
   - Round numbers sensibly
   - Include commas for readability
   - Add appropriate currency symbols
   - Use consistent zero padding
5. Place y-axis labels at top rather than rotated on side

Comment 9
Use gestalt principles to highlight key elements:

- *Shape variations*
- *Line thickness*
- *Color saturation*
- *Size differences*
- *Strategic positioning*
- *Varying sharpness*

Academic Visualization Context

Academic papers differ from media visualizations:

- Focus on communicating research findings clearly
- Present multiple variations of similar analyses
- Break complex information into digestible pieces
- Build understanding incrementally

Comment 10
Effective academic figures should:

- *Present key findings clearly*
- *Support robustness checks*
- *Enable comparison across specifications*
- *Guide readers through complex analyses*

This approach allows readers to:

1. Understand main results quickly
2. Digest supporting evidence systematically
3. Evaluate robustness thoroughly
4. Follow complex analytical arguments

Conclusion

Effective visualization in academic work requires:

- Balance between simplicity and completeness
- Clear communication of key findings
- Systematic presentation of supporting evidence
- Thoughtful design choices that enhance understanding

The effort invested in creating clear, effective visualizations pays dividends in improved communication and understanding of research findings.