Lecture 6 – Linear Regression 2: Semiparametrics and Visualization

Paul Goldsmith-Pinkham

January 29, 2026

This lecture note continues our study of linear regression, focusing on two key themes: understanding what OLS estimates when we include controls, and improving how we visualize and communicate regression results. In the previous lecture, we focused on inference. Today, we turn to interpretation of coefficients and best practices for data visualization.

The goals for this lecture are:

- Understand how OLS weights treatment effects across strata when controls are included
- Learn the Frisch-Waugh-Lovell theorem and its implications for visualization
- Explore semiparametric approaches like binscatter
- Develop principles for effective data visualization in research

Why Linear Regression?

Linear regression remains the workhorse of empirical economics. Why is it so popular? There are several reasons:

1. **Computational efficiency**: OLS has a closed-form analytic solution, and matrix inversion algorithms have become very fast.
2. **Statistical efficiency**: Under classical assumptions (homoskedasticity, no serial correlation), OLS is the Best Linear Unbiased Estimator (BLUE).
3. **Interpretability**: Linear regression provides an intuitive summary of relationships in the data.
4. **Robustness**: While "better" estimators may exist for specific settings, linear regression performs reasonably well across a wide range of situations.
5. **Scalability**: Modern implementations handle high-dimensional fixed effects and large datasets efficiently.

This lecture focuses on how to stay in the world of linear regression while improving our understanding of what it estimates and how we present results.

General Framework for Causal Relationships

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

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

where D_i is the causal variable of interest, W_i represents controls or sources of heterogeneity, and 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 is high-dimensional.

A simpler version separates the error term:

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

Even with this simplification, we face choices about what to report. Should we 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:

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

This can be made more complex through interactions (e.g., $Y_i = D_i\tau + W_i\beta_1 + D_i \times W_i\beta_2 + \epsilon_i$), but even then, there is not always a "single number" to report.

Visualizing Relationships

When plotting relationships between outcomes and causal variables, the data itself often tells a compelling story. However, 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

The challenge becomes more complex when we need to account for control variables.

Residual Regression and the Frisch-Waugh-Lovell Theorem

Consider our basic specification:

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

How does OLS handle the controls W_i? We can understand this through projection matrices.

Definition 1 (Projection and Annihilator Matrices)
Define the projection matrix as:

$$\mathbf{P}_W = \mathbf{W}_n(\mathbf{W}_n'\mathbf{W}_n)^{-1}\mathbf{W}_n'$$

This matrix has the properties that $\mathbf{P}_W\mathbf{W}_n = \mathbf{W}_n$ *and* $\mathbf{P}_W\mathbf{P}_W = \mathbf{P}_W$ *(idempotent). When applied to* $\mathbf{D}_n$*, we get the predicted values from regressing* D_i *on* W_i*.*

The annihilator matrix is:

$$\mathbf{M}_W = \mathbf{I}_n - \mathbf{P}_W$$

This gives us the residuals from the regression on W_i*.*

Theorem 1 (Frisch-Waugh-Lovell)
If we transform $\mathbf{Y}_n^* = \mathbf{M}_W\mathbf{Y}_n$ *and* $\mathbf{D}_n^* = \mathbf{M}_W\mathbf{D}_n$*, then running the regression*

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

yields the same coefficient τ *as the original multivariate regression.*

This theorem is powerful for both computation and visualization. When W is a discrete set of covariates (e.g., fixed effects), the transformation demeans D and Y within each group. The resulting regression estimate gives:

$$\tau = \frac{E(\sigma_D^2(W_i)\tau(W_i))}{E(\sigma_D^2(W_i))}, \qquad \sigma_D^2(W_i) = E((D_i - E(D_i|W_i))^2|W_i) \tag{1}$$

where $\tau(W_i)$ is the conditional treatment effect given W_i. This shows that OLS produces a *variance-weighted* average of the conditional treatment effects.

Binary Treatment with Binary Controls

To build intuition, consider both W_i and D_i binary. Consider the regression:

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

with $D_i, W_i \in \{0,1\}$. By definition, U_i is a mean-zero regression residual uncorrelated with (D_i, W_i).

Example 1 (Project STAR)
Consider a stylized version of Project STAR, where D_i *indicates assignment to a small classroom and* Y_i *is the student's average test score. Randomization was stratified by school, so the probability of assignment to small vs. large classroom depends on school. Let* W_i *denote the school fixed effect (binary for simplicity: only 2 schools).*

Using potential outcomes notation $Y_i(d)$:

- 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: $p_1(W_i) = \Pr(D_i = 1|W_i) = E[D_i|W_i]$

Under conditional random assignment $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D_i|W_i$, we get the key result from **?**:

Theorem 2 (OLS with Binary Treatment and Controls) *Under conditional random assignment,*

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

where

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

Proof 1

Let $\tilde{D}_i = D_i - E[D_i|W_i]$ denote the residual from regressing D_i on W_i. By the FWL theorem:

$$\beta = \frac{E[\tilde{D}_i Y_i]}{E[\tilde{D}_i^2]} = \frac{E[E[\tilde{D}_i Y_i(0)|W_i]]}{E[\tilde{D}_i^2]} + \frac{E[E[\tilde{D}_i D_i \tau_{i1}|W_i]]}{E[\tilde{D}_i^2]}$$

The first term equals zero because $E[\tilde{D}_i|W_i] = 0$ (not just $corr(\tilde{D}_i, W_i) = 0$) and by random assignment. For the second term, note that $E[\tilde{D}_i D_i|W_i] = \text{var}(D_i|W_i)$. Hence:

$$\beta = \frac{E[\text{var}(D_i|W_i)\tau(W_i)]}{E[\text{var}(D_i|W_i)]}$$

which gives the stated result when W_i is binary.

Comment 1 (Key Features of the OLS Estimator with Controls) *Several important properties follow from* **??** *2:*

1. *The weights $\phi \in (0,1)$ are guaranteed to be positive*
2. *No need to explicitly estimate propensity scores*
3. *OLS puts larger weight on strata with higher variation in D_i*
4. *The estimand $\neq$ ATE unless $\tau(w)$ is constant or $p_1(w)$ is constant across strata*
5. *This weighting helps avoid identification problems under overlap failure (e.g., $p_1(0) = 0$) or precision issues under weak overlap ($p_1(0)$ close to o)*

See **?** *for discussion of how this weighting may lead to "unrepresentative" estimands.*

Multiple Treatment Arms

The framework extends to multiple treatments, but with important complications. Consider adding a second treatment arm (e.g., the teaching aide arm 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 $j = 1, 2$.

Define:

- $\tau_{ik} = Y_i(k) - Y_i(0)$ as the treatment effect for arm k
- $\tau_k(W_i) = E[\tau_{ik}|W_i]$ as the conditional effect
- $p_{0k}(w) = E[X_{ik}|W_i = w]$ as the propensity scores

Under conditional random assignment $(Y_i(0), Y_i(1), Y_i(2)) \perp\!\!\!\perp X_i|W_i$, we can derive the causal interpretation of β_1:

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

where $\lambda_{11}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1}X_{i1}|W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \geq 0$ and $\lambda_{12}(W_i) = \frac{E[\tilde{\tilde{X}}_{i1}X_{i2}|W_i]}{E[\tilde{\tilde{X}}_{i1}^2]} \neq 0$ in general.

Comment 2 (Contamination Bias)
The key insight is that $\tilde{\tilde{X}}_{i1}$ is the residual from regressing X_{i1} on W_i, a constant, and *X_{i2}. Since X_{i2} depends non-linearly on X_{i1} (they cannot both equal 1), the coefficient β_1 is "contaminated" by the effects of treatment 2. This is an important consideration when interpreting regression coefficients in settings with multiple treatment arms. See* **?** *for further discussion.*

Visualization with Controls

The FWL theorem provides a powerful approach for visualizing relationships while controlling for covariates. The basic approach is:

1. Residualize both Y and D by regressing each on W
2. Plot the residualized Y^* against residualized D^*
3. The slope of the fitted line equals the coefficient from the multivariate regression

However, there are practical considerations:

- Residualized variables can be hard to interpret (they're centered at zero)

- A simple fix: add back the overall means to make interpretation more intuitive
- Be careful that the visualization reflects the variation identifying your parameter

Comment 3
When adding back means for visualization, ensure the relationship you display corresponds to the estimand you care about. Simply adding raw means may not achieve this if there are compositional differences across strata.

Nonparametric and Semiparametric Approaches

Our interest often lies in conditional expectation functions $E(Y|D)$. The approaches we use can be categorized into three types:

Definition 2 (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 the error distribution $F(\theta_i)$ is infinite-dimensional

It is important to distinguish between *nuisance parameters* (which we don't care about estimating, like θ_i in the robust standard error case) and parameters of interest.

Binscatter Analysis

One popular semiparametric approach is *binscatter*, which approximates the conditional expectation function using binned means. Consider:

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

The binscatter approach works as follows:

1. Divide observations into equally-sized bins based on values of D_i
2. Compute the mean of Y_i within each bin
3. Plot these means against the bin centers (or midpoints)

This is particularly useful when:

- There are too many data points to see patterns in a raw scatter plot
- We want to examine whether a linear relationship is appropriate
- We want a flexible visualization of the conditional expectation function

Example 2 (Choice of Bins)
The choice of bin width affects the visualization substantially. Too few bins may oversimplify the relationship; too many bins produce noisy estimates. For example, plotting income on health insurance coverage with 10, 20, or 50 bins can yield quite different visual impressions of the same underlying relationship.

Advances in Binscatter: Cattaneo et al.

Recent work by **?** provides important methodological advances for binscatter:

Formal statistical framework. The paper recognizes binscatter as a nonparametric estimation problem, providing a basis for inference rather than treating it as purely descriptive.

Optimal bin selection. The paper develops data-driven methods for choosing the number of bins that balance bias (too few bins misses curvature) against variance (too many bins are noisy). The canonical rule suggests approximately $n^{1/3}$ bins.

Correct handling of controls. This is perhaps the most practically important contribution. The traditional approach of residualizing D_i before binning produces incorrect results when f is nonlinear.

Comment 4 (The FWL Problem with Binscatter)
Consider the model:

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

If f is nonlinear, you cannot simply residualize D_i by W_i and recover the function f. The correct approach is:

1. *Create bins based on the raw treatment variable D_i*
2. *Within each bin, estimate the effect controlling for W_i*
3. *Plot these conditional effects*

Unfortunately, the traditional Stata binscatter command residualized first, which could produce misleading results.

Statistical inference. The framework allows for construction of confidence intervals and tests for shape restrictions (e.g., monotonicity).

Code for implementing these methods is available at https://nppackages.github.io/binsreg/. For a simpler fix to the FWL issue in the traditional approach, see https://github.com/mdroste/stata-binscatter2.

Design Principles for Research Communication

Binscatter's success reflects a broader lesson: improving data visualization dramatically improves communication of results. The status quo of large regression tables is often ineffective at conveying findings.

Four Design Goals

1. **Minimize tables**: Tables are important repositories of information but make comparison difficult and tend to include unnecessary information. Control variable coefficients, for example, rarely have a causal interpretation [**?**]. Consider moving detailed tables to online appendices.
2. **Have describable goals for every exhibit**: The purpose of each figure or table should be immediately obvious. If it's not clear, either:
   - Too much information obscures the main message
   - Insufficient emphasis on key elements

3. **Craft not-ugly figures**: There's almost no good reason to have bad figures. Small improvements yield big returns:
   - Fix the color scheme (default schemes are often poor)
   - Label axes clearly
   - Make the color scheme accessible
   - Adjust line weights and point sizes for emphasis
4. **Do not mislead readers**: Present data honestly. For example, in event study plots with discrete time periods, avoid smooth confidence bands that imply continuity that isn't present.

Schwabish's Guidelines for Figures

? provides excellent guidelines:

1. Show the data
2. Reduce clutter
3. Integrate graphics and text
4. Avoid providing extraneous information
5. Start with grey (add color strategically for emphasis)

Practical Tips

Comment 5 (Making Good Figures)
- *Bar graphs should almost always be horizontal for readable labels*
- *Don't put confidence intervals on bar graphs; use point-range plots instead*
- *Directly label on your figure rather than using legends when possible*
- *Fix your units: round numbers, add commas, include currency symbols, use zero padding*
- *Label your y-axis at the top of the graph rather than rotated 90 degrees*
- *Use gestalt principles to highlight key elements: shape, thickness, saturation, color, size, position*

Academic papers differ from media visualizations. Our figures often:

- Present multiple variations of similar analyses
- Support robustness checks
- Build understanding incrementally

The goal is to provide a polished way to convey bite-sized pieces of information, so that once readers understand the main result, subsequent robustness checks are easily processed.

Discussion Questions 1

1. *Consider a paper you've recently read. How could its main figures be improved using the principles discussed?*
2. *When would you choose binscatter over a simple linear regression visualization? What are the tradeoffs?*
3. *In the multiple treatment arm setting, how might you visualize the contamination problem discussed in this lecture?*