Advanced Econometrics

Lecture 8: Nonlinearities and Flexible Functional Forms

Eduard Brüll

Fall 2025

Lecture 8

Advanced Econometrics

Literature: Wooldridge Ch. 6 & 8; Greene Ch. 9; Hastie & Tibshirani (1990)

8.1: Nonlinearities within OLS

Review: Linear in Parameters $\neq$ Linear in Variables

- OLS assumes the model is linear in parameters, not necessarily in variables.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

  is still a linear regression model.

- The conditional mean function $E[y|X]$ can be nonlinear in x.

$$\text{“Linear”} \Rightarrow \text{ additive in } \beta, \text{ not necessarily in } x.$$

- Nonlinearities in variables allow marginal effects to vary with x.

$$\frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

Marginal Effects that Depend on x

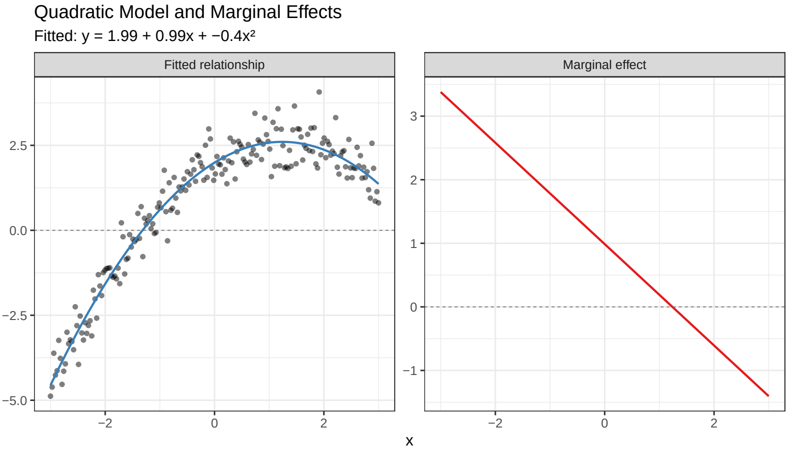
- In a linear model, $\partial y/\partial x = \beta$ is constant.
- In a nonlinear function of x, the slope changes:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

$$\Rightarrow \frac{\partial y}{\partial x} = \beta_1 + 2\beta_2 x$$

- Interpretation:
  - $\beta_2 > 0$: increasing effect of x.
  - $\beta_2 < 0$: diminishing returns.
- Visual check: plot $\hat{y}(x)$ or $\frac{d\hat{y}}{dx}$.

Illustration: Marginal Effect for a Quadratic Function



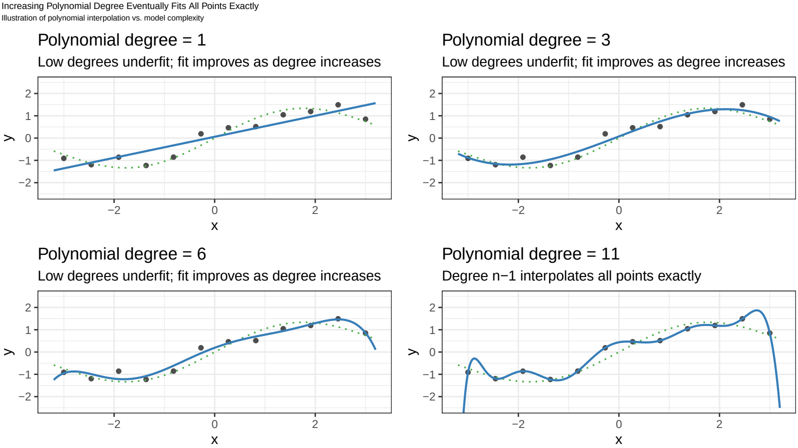
Polynomial Models

- General form:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_r x_i^r + \varepsilon_i$$

- Captures curvature in $E[y|x]$ while remaining linear in β.
- Choose degree r:
  - Sequential F-tests for higher-order terms.
  - Information criteria (AIC, BIC) for fit vs. complexity.
  - Or choose via LASSO-regression (more later)
- Watch out for:
  - Extrapolation instability at high degrees.
  - Multicollinearity among x^j terms.

Illustration: Beware of High-Degree Polynomials



Dummy Variables

- A **dummy variable** (or indicator) takes values 0 or 1 to represent categories:

$$D_i = \begin{cases} 1 & \text{if observation } i \text{ belongs to group A} \\ 0 & \text{otherwise.} \end{cases}$$

- Model with one dummy:

$$y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$$

- Interpretation:

$$E[y|D = 1] - E[y|D = 0] = \beta_1 \quad \Rightarrow \quad \beta_1 = \text{mean difference between groups.}$$

- You can **one-hot encode** multiple categories this way, but you must **omit one base category** to avoid perfect collinearity (“dummy variable trap”).

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + \cdots + \varepsilon_i$$

- Interactions allow slope differences by group:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 (x_i \times D_i) + \varepsilon_i$$

Interaction Terms

- Allow the effect of one regressor to depend on another.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \beta_3 (x_i \times z_i) + \varepsilon_i$$

- Marginal effect of x:

$$\frac{\partial y}{\partial x} = \beta_1 + \beta_3 z$$

- Examples:
  - Gender differences in wage returns to education.
  - Policy effect only active in treated regions.
- Always include base levels of z_i and x_i for if you are interested in an interaction term!

Interactions with Dummy Variables

Interacting a continuous variable x_i with a dummy D_i allows for **group-specific slopes**.

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 D_i + \beta_3 (x_i \times D_i) + \varepsilon_i$$

The model implies two regression lines:

$$E[y \mid D] = \begin{cases} \beta_0 + \beta_1 x, & \text{if } D = 0, \\ (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x, & \text{if } D = 1. \end{cases}$$

Interpretation:

- β_2: difference in intercepts between groups ($x = 0$).
- β_3: difference in slopes between groups — how the effect of x changes when $D = 1$.

Graphically:

$$\text{Parallel lines if } \beta_3 = 0, \quad \text{different slopes if } \beta_3 \neq 0.$$

Example:

- Wage regression with x = years of education and D = female.
- $\beta_3 < 0$: smaller returns to education for women.

Why Logarithmic Transformations?

- Many economic relationships are multiplicative rather than additive:

$$y = Ax^{\beta}e^{\varepsilon}$$

- Taking logs makes this relationship additive:

$$\ln y = \ln A + \beta \ln x + \varepsilon$$

- Now, β approximates how y changes in **percentage terms** when x changes in percentage terms.
- Intuition:

A 1% increase in $x \Rightarrow$ about a $\beta\%$ change in y

Why the Log Approximation Works

- We want to understand why a **change in the log of** x measures a **percentage change in** x.

$$\Delta \ln x = \ln(x + \Delta x) - \ln(x) = \ln\left(1 + \frac{\Delta x}{x}\right)$$

- Let $z = \frac{\Delta x}{x}$ = the **relative (percentage) change** in x.
- Expand $\ln(1+z)$ around $z = 0$ (using a Taylor series):

$$\ln(1+z) = z - \frac{z^2}{2} + \frac{z^3}{3} - \dots$$

- When z is small (say a few percent), the higher-order terms are negligible:

$$\ln(1+z) \approx z$$

- Therefore:

$$\Delta \ln x = \ln(x + \Delta x) - \ln(x) \approx \frac{\Delta x}{x}$$

- So a small **percentage change** in x produces roughly the same **change in log(**x**)**.

Log Models and Interpretation

- Using the approximation:

  Linear–log: $y = \beta_0 + \beta_1 \ln x + \varepsilon \quad \Rightarrow \quad \dfrac{\Delta y}{\Delta x / x} \approx 0.01\beta_1$ (semi-elasticity)

  Log–linear: $\ln y = \beta_0 + \beta_1 x + \varepsilon \quad \Rightarrow \quad \dfrac{\Delta y / y}{\Delta x} \approx \beta_1$ (semi-elasticity)

  Log–log: $\ln y = \beta_0 + \beta_1 \ln x + \varepsilon \quad \Rightarrow \quad \dfrac{\Delta y / y}{\Delta x / x} \approx \beta_1$ (elasticity)

- The log transformation thus links linear regression coefficients to interpretable economic quantities (percent or proportional effects).

How Accurate is the Log Approximation?

Recall:

$$\ln(1+z) \approx z \quad \text{for small } z = \frac{\Delta x}{x}.$$

- Compare the exact and approximate values:

Relative change z	$\ln(1+z)$	Approx. z	Error (%)
0.01	0.00995	0.01000	0.5%
0.05	0.04879	0.05000	2.5%
0.10	0.09531	0.10000	4.9%
0.25	0.22314	0.25000	12.1%
0.50	0.40547	0.50000	23.3%
1.00	0.69315	1.00000	44.3%

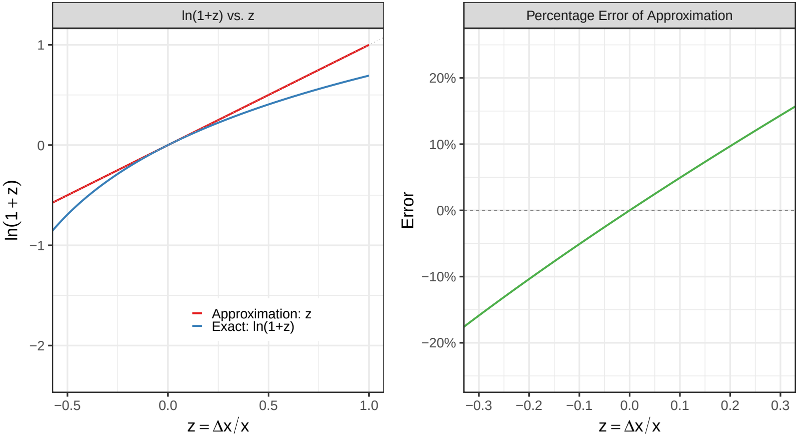
- The approximation is very accurate for small relative changes (say below 10%), but deteriorates for larger ones.
- Visually: $\ln(1+z)$ bends below the 45° line as z grows.

Rule of thumb:

Use the log approximation only for $|\Delta x/x| \lesssim 0.1$.

Alternatively, economists often report **log points** directly instead of percentage points to avoid this approximation issue.

Illustration: Accuracy of the Log-Approximation



8.2: Polynomial Models

Functional Form and Economic Theory

Before relying on statistical tests, start from **economic theory**.

- Theory suggests shape restrictions: monotonicity, concavity, saturation, thresholds, etc.
- Example: diminishing returns $\Rightarrow$ negative second derivative ($\beta_2 < 0$).
- Utility, production, or demand functions often imply specific curvature.

Polynomials can be a **flexible approximation** to such theoretical shapes:

$$f(x) \approx \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r$$

- But without theory, higher-degree terms risk capturing noise, not structure.
- Therefore:
  1. Use theory to motivate the expected shape of $E[y|x]$.
  2. Use statistical tests (e.g., sequential F-tests) only to check adequacy of that shape.

Sequential F-Tests for Polynomial Terms

- To decide whether to include higher-order terms, test:

$$H_0 : \beta_r = 0 \quad \text{vs.} \quad H_1 : \beta_r \neq 0$$

- More generally:

$$H_0 : \beta_{q+1} = \cdots = \beta_r = 0$$

- Compute the **F-statistic** comparing restricted (degree *q*) and unrestricted (degree *r*) models:

$$F = \frac{(SSR_R - SSR_U)/(r - q)}{SSR_U/(n - r - 1)}$$

- If $F > F_{r-q,\, n-r-1;1-\alpha}$, reject $H_0 \longrightarrow$ higher-degree terms improve fit.
- Repeat sequentially: degree $1 \rightarrow 2 \rightarrow 3 \rightarrow \ldots$ until H_0 not rejected.

Example: Choosing Polynomial Degree

Fit models of increasing degree:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

and so on.

- Use the F-test to compare models, e.g.:

$$F_{2\text{ vs. }3} = \frac{(SSR_2 - SSR_3)/1}{SSR_3/(n-4)}$$

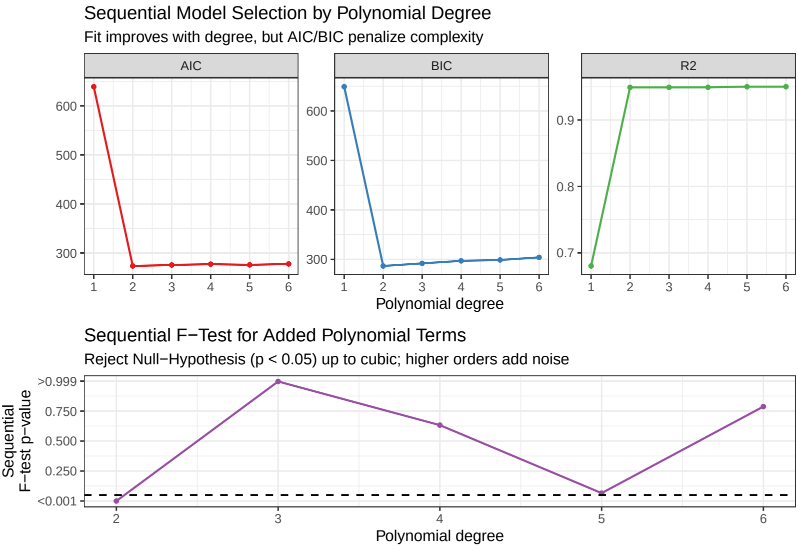
- Stop adding terms when *F*-test is insignificant.

Important: Always include all lower-order terms when testing a higher-order one.

Model Fit vs. Complexity

- Higher-degree polynomials improve fit in-sample ()$R^2 \uparrow$), but may overfit.
- Sequential F-tests guard against adding unnecessary terms, but:
  - Depend on chosen α (risk of multiple testing).
  - Are not ideal for predictive performance.
- Alternative: use **information criteria** like AIC/BIC to penalize complexity (more on them later).
  $\rightarrow$ Choose model with minimal BIC/AIC

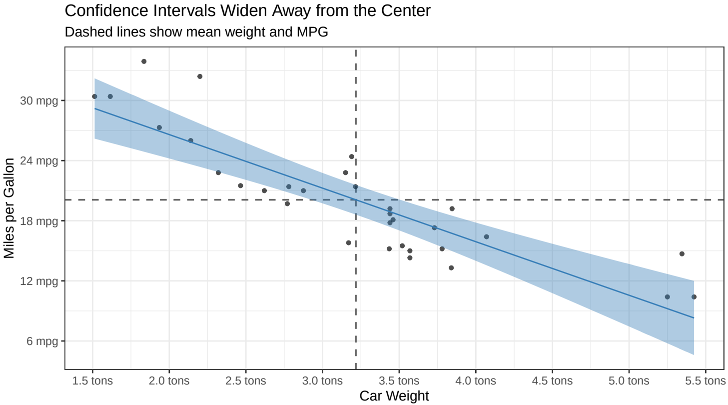
Illustration: Polynomial Choice



Simulation setup: $y = 2 + 1x - 0.4x^2 + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.5^2)$

8.3 Confidence Intervals and Leverage

Confidence Bands in Practice



Example: Fitted line for the `mtcars` data. Confidence bands widen at the edges even though residual variance is constant.

Question: Why?

Variance of $\hat{y}_0$

Start from the linear model:

$$y = X\beta + \varepsilon, \qquad \mathbf{E}[\varepsilon] = 0, \quad \mathbf{var}(\varepsilon) = \sigma^2 I_n.$$

The OLS estimator is:

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

The fitted value at a new point x_0 is:

$$\hat{y}_0 = x_0'\hat{\beta} = x_0'\beta + x_0'(X'X)^{-1}X'\varepsilon.$$

Take expectations, using $\mathbf{E}[\varepsilon|X] = 0$:

$$\begin{aligned}
\mathbf{E}[\hat{y}_0] &= \mathbf{E}\big[\,\mathbf{E}[\hat{y}_0 \mid X]\big] \\
&= \mathbf{E}\Big[\,\mathbf{E}\big[x_0'(X'X)^{-1}X'(X\beta + \varepsilon)\,\big|\,X\big]\Big] \\
&= \mathbf{E}\Big[x_0'\beta + x_0'(X'X)^{-1}X'\,\mathbf{E}[\varepsilon \mid X]\Big] \\
&= \mathbf{E}[x_0'\beta] = x_0'\beta \qquad \text{due to Exogeneity } \mathbf{E}[\varepsilon \mid X] = 0
\end{aligned}$$

Variance of $\hat{y}_0$

Subtract the mean and use $\mathbf{var}(a'Z) = a'\,\mathbf{var}(Z)a$:

$$\mathbf{var}(\hat{y}_0) = \mathbf{var}\left(x_0'(X'X)^{-1}X'\varepsilon\right) = x_0'(X'X)^{-1}X'\,\mathbf{var}(\varepsilon)X(X'X)^{-1}x_0.$$

Substitute $\mathbf{var}(\varepsilon) = \sigma^2 I_n$:

$$\mathbf{var}(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}X'X(X'X)^{-1}x_0.$$

Simplify $X'X$ in the middle:

$$\boxed{\mathbf{var}(\hat{y}_0) = \sigma^2 x_0'(X'X)^{-1}x_0.}$$

Interpretation:

- The term $x_0'(X'X)^{-1}x_0$ measures how far the point x_0 is from the **center of the data cloud**, taking into account how the data are spread and correlated.
- In geometry, this acts like a *stretch-adjusted* squared distance (the **Mahalanobis distance**)
- So, predictions made far from where most data lie have larger distance and therefore larger variance.

Example: Variance in a Bivariate Regression

For a bivariate model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
Then

$$X = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \qquad X'X = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}.$$

Invert:

$$(X'X)^{-1} = \frac{1}{n \sum (x_i - \bar{x})^2} \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & n \end{bmatrix}.$$

Plug into $\mathbf{var}(\hat{y}_0) = \sigma^2 x_0' (X'X)^{-1} x_0$, where $x_0 = (1, x_0)'$:

$$\mathbf{var}(\hat{y}_0) = \frac{\sigma^2}{n \sum (x_i - \bar{x})^2} \left[\sum x_i^2 - 2x_0 \sum x_i + n x_0^2 \right]$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2} \right].$$

Interpretation: The variance is smallest at $x_0 = \bar{x}$ (the sample center) and grows quadratically as x_0 moves away. This is why predictions at the edges have high uncertainty.

Leverage

The quantity

$$h_0 = x_0'(X'X)^{-1}x_0$$

is known as the **leverage** of point x_0.

- Leverage measures how far x_0 is from the center of the data in feature space.
- Observations (or prediction points) with high leverage have greater influence on the fitted line.
- The variance of the fitted value is proportional to leverage
- We can compute **leverage for every observation** to gain insights if there are any points that are very influential for our fit.

Influence on Coefficients: DFBETA

- **Leverage** is informative for an observations influence on the fitted line. But this does not mean a high-leverage point necessarily affects our coefficient of interest.

- **DFBETA** measures the actual impact of each observation on each estimated coefficient:

$$\text{DFBETA}_{ij} = \text{change in } \hat{\beta}_j \text{ when observation } i \text{ is removed.}$$

- Intuition:
  - If one data point can noticeably shift a slope or intercept, its DFBETA will be large (positive or negative).
  - A DFBETA close to zero means the observation does not matter much for that coefficient.

- **Rule of thumb:** $|\text{DFBETA}_{ij}| > 2/\sqrt{n}$ indicates influential points.

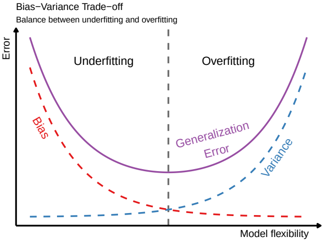
Aggregating DFBETAs for Robustness and Diagnostics

- Software like `R`, `Stata`, or `Python statsmodels` gives a full matrix of DFBETAs: each observation *i* and coefficient *j*.
- These can be **aggregated or filtered** to diagnose robustness:
  - Identify which units, years, or clusters strongly affect a specific coefficient.
  - Compute **average absolute DFBETA** by group (region, industry, firm, etc.) to find influential clusters.
  - **Visual cue:** A **histogram of DFBETAs** for the coefficient of interest shows how influence is distributed across observations whether most points are small and balanced, or a few dominate the estimate.
- **Example application:** In a difference-in-differences regression, highlight units where $\text{DFBETA}_{i,\text{treat}\times\text{post}}$ is large.

8.4: Specification Choice: Information Criteria and Penalized Regression

Model Fit vs. Complexity: The Bias–Variance Tradeoff

- Adding regressors always increases in-sample fit ($R^2 \uparrow$, $SSR \downarrow$).
- But more flexibility $\Rightarrow$ higher estimation variance
- The **expected out-of-sample error** decomposes into:



$$\mathbf{E}[(y - \hat{y})^2] = \text{Bias}^2 + \text{Variance} + \text{Irreducible Noise}$$

As model flexibility increases:

Bias $\downarrow$ but Variance $\uparrow$

The minimum of total (generalization) error gives the **optimal model complexity.**

Penalizing Flexibility: The Idea Behind Information Criteria

- The log-likelihood $\ell = \log L(\hat{\theta})$ measures in-sample fit.
- Adding parameters **always increases** ℓ — even if we only fit noise.
- **Information criteria** correct this by adding a penalty for model complexity:
$$\text{IC} = -2\ell + \text{penalty}(k, n)$$
- Common forms:
$$\text{AIC} = -2\ell + 2k, \qquad \text{BIC} = -2\ell + k \ln n$$
- Choose the model with the **lowest IC**.

Intuition

- **AIC:** smaller penalty $\Rightarrow$ favors better prediction.
- **BIC:** stronger penalty $\Rightarrow$ favors simpler models.

Penalized Regression: Controlling Complexity Directly

- Recall OLS in matrix form:

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- Penalized regression adds a constraint on coefficient magnitude:

$$\hat{\boldsymbol{\beta}}_{\lambda} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left[(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda P(\boldsymbol{\beta}) \right],$$

  where $\lambda \geq 0$ controls how strongly we penalize complexity.

- Examples of penalty functions:

$$P(\boldsymbol{\beta}) = \begin{cases} \sum_j \beta_j^2 & \text{(Ridge)} \\ \sum_j |\beta_j| & \text{(LASSO)} \end{cases}$$

- Larger $\lambda \Rightarrow$ simpler model, smaller coefficients.

Bias-Variance Logic of Penalization

- The penalty **shrinks** coefficients toward zero. This **reduces variance** at the cost of introducing some **bias**.

$$\mathbf{E}[\hat{\boldsymbol{\beta}}_\lambda] \neq \boldsymbol{\beta}_0 \quad \text{but} \quad \mathbf{var}(\hat{\boldsymbol{\beta}}_\lambda) \ll \mathbf{var}(\hat{\boldsymbol{\beta}}_{\text{OLS}})$$

- When prediction is the goal, a small bias can be optimal if it cuts variance substantially.
- Intuitively:

  Shrink noisy slopes slightly $\Rightarrow$ lower mean-squared error overall

- The penalty strength λ determines where we sit on the bias–variance curve.
- In practice, we **choose λ by cross-validation**: fit the model on subsamples, test on held-out data, and pick the λ with the smallest average prediction error.

- Most software (`glmnet`, `sklearn`, `Stata cvlasso`, `tidymodels`) automatically **cross-validate** λ:
  1. Split data into folds (e.g. 10-fold CV),
  2. Estimate the model on training folds,
  3. Compute out-of-sample fit on validation folds,
  4. Pick λ that minimizes average prediction error.
- LASSO can set some coefficients exactly to zero $\Rightarrow$ automatic variable selection.
- But LASSO estimates are **biased** because of the shrinkage term.
- Hence, after variable selection, economists often estimate:

$$\hat{\boldsymbol{\beta}}_{\text{post-LASSO}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} (\mathbf{y} - \mathbf{X}_{\text{selected}}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}_{\text{selected}}\boldsymbol{\beta}).$$

- **Post-LASSO:** Re-estimate OLS on the selected variables to remove shrinkage bias.

Statistical Selection vs. Economic Theory

- Economists are often cautious about purely statistical model selection.
- LASSO is powerful when:
  - we have many potential controls,
  - but the focus is on the main regressor(s), not each control's interpretation.
- Always cross-check results with:
  - domain knowledge
  - theory-based restrictions
  - robustness to alternative control sets

In short:

Use LASSO to narrow down; use economics to decide what makes sense.

Application: Double Selection

- In causal inference, we are often interested in a single regressor of interest d_i:
$$y_i = \alpha d_i + x_i' \beta + \varepsilon_i$$
where x_i are many potential controls.

- A simple LASSO for the outcome regression may omit controls that are weakly related to y_i but strongly related to d_i.
- Omitted variables correlated with $d_i \Rightarrow$ bias in $\hat{\alpha}$.
- **Idea:** Run two selection steps:
  1. Regress y_i on all x_i with LASSO to select controls related to y.
  2. Regress d_i on all x_i with LASSO to select controls related to d.
- Take the **union** of both selected variable sets, and estimate α by OLS controlling for them.

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). *"Inference on Treatment Effects after Selection among High-Dimensional Controls."* **Review of Economic Studies**, 81(2), 608–650.

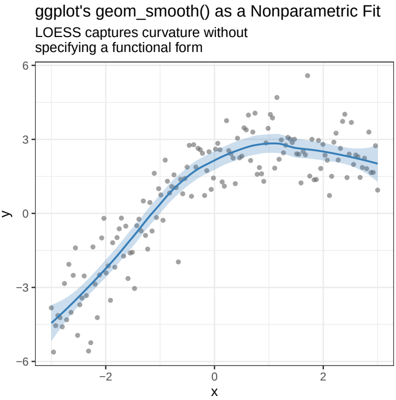
8.5: Local Linear Regressions

Smooth Fits You Already Know: `geom_smooth()`

- `geom_smooth()` in `ggplot2` uses **LOESS** by default, a local regression that adapts to the data's shape.
- It provides a quick, nonparametric way to visualize nonlinear relationships:

$$y_i = f(x_i) + \varepsilon_i,$$

$f(\cdot)$ estimated locally.



Useful for exploration and pattern recognition, but:

- The fitted shape depends on a bandwidth
- **No interpretable parameters!**
- Unstable at data boundaries

Interactive Illustration: Local Smooths in Action

Shiny-App on how LOESS fits work

Inside a Nonparametric Smoother: The Big Picture

Goal: Estimate a smooth function $f(x)$ without imposing a specific parametric form.

For each target point x_0:

1. Assign weights $w_i = K\left(\frac{x_i - x_0}{h}\right)$ to nearby observations.
2. Fit a simple model (often linear) using these weighted observations.
3. Move x_0 across the range of x and repeat to obtain $\hat{f}(x)$.

Key ingredients:

- The **kernel** $K(\cdot)$ decides how fast weights decline with distance.
- The **bandwidth** h controls how wide the local neighborhood is.

Result: A smooth, flexible fit that adapts to the local structure of the data.

Kernels: How Local Weights Are Assigned

The **kernel function** $K(u)$ determines how much weight each observation receives based on distance

$$u = \frac{x_i - x_0}{h}$$

where h is the bandwidth (smoothing parameter).

$$K(u) \geq 0, \quad K(u) = K(-u), \quad \int K(u)\, du = 1$$

Intuition: Nearby points get high weights; distant points get low or zero weight.

Common kernel shapes:

Name	**Kernel function** $K(u)$
Uniform	$\frac{1}{2}\,\mathbb{1}(\lvert u\rvert \leq 1)$ (equal weights within window)
Triangular	$(1 - \lvert u\rvert)\,\mathbb{1}(\lvert u\rvert \leq 1)$ (linearly decreasing weights)
Epanechnikov	$\frac{3}{4}(1 - u^2)\,\mathbb{1}(\lvert u\rvert \leq 1)$ (optimal in MSE sense)
Gaussian	$\frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}$ (smooth, infinite support)

In practice: The kernel shape matters little. Most of the smoothing behavior is driven by the bandwidth h.

Bandwidth: The Smoothing Parameter

- The **bandwidth** $h > 0$ defines the size of the local neighborhood:

$$w_i = K\left(\frac{x_i - x_0}{h}\right)$$

- Smaller $h \Rightarrow$ more local fit:
  - captures fine detail (low bias),
  - but higher variance (less data per fit).
- Larger $h \Rightarrow$ smoother fit:
  - lower variance,
  - but higher bias (averages distant points).

Bias–Variance Tradeoff in Local Regression

- The bandwidth h controls how smooth the local fit is.

$$\hat{f}(x) = \sum_i w_i(x)\, y_i, \quad w_i(x) \propto K\left(\frac{x_i - x}{h}\right)$$

- Small $h \Rightarrow$ low bias, high variance (wiggly fit)
- Large $h \Rightarrow$ high bias, low variance (over-smoothed)
- **Exactly the same bias–variance tradeoff as in prediction:** choosing h balances flexibility and stability.

Why Nonparametrics Are Rare in Economics

In theory: Nonparametric methods make minimal assumptions about functional form.

$$y_i = f(x_i) + \varepsilon_i, \qquad f(\cdot) \text{ estimated flexibly.}$$

In practice: Economists rarely use fully nonparametric estimators because:

- **Curse of dimensionality:** Precision declines exponentially with the number of regressors.

$$n_{\text{effective}} \approx n \cdot h^k \quad \Rightarrow \quad \text{requires huge samples if } k > 2$$

- **No structural interpretation:** Nonparametric fits show patterns, not mechanisms or parameters.
- **Difficult inference:** Confidence intervals and hypothesis testing are less straightforward.
- **Economists prefer interpretable, theory-consistent parameters.**

Therefore: Nonparametrics are mainly used for visualization, validation, or specific designs (e.g. RDD).

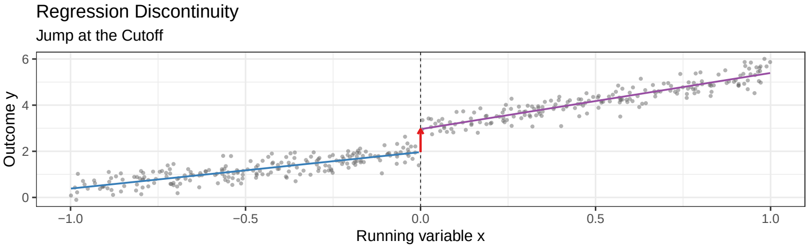
Application: RDD

Regression Discontinuity Designs: A treatment switches on when running variable x crosses a known cutoff c (Here $c = 0$).

$$D_i = \mathbb{1}(x_i \geq c)$$

If potential outcomes are smooth in x, any jump in y at c identifies the treatment effect:

$$\tau = \lim_{x \downarrow c} E[y|x] - \lim_{x \uparrow c} E[y|x]$$



Idea: Compare observations just left vs. right of the cutoff. Similar units, different treatment.

Application: Local Linear Regression in RDD

Implementation: Fit separate local linear regressions on each side of the cutoff c:

$$y_i = \alpha_{\text{below}} + \beta_{\text{below}}(x_i - c) + \varepsilon_i, \quad x_i < c,$$
$$y_i = \alpha_{\text{above}} + \beta_{\text{above}}(x_i - c) + \varepsilon_i, \quad x_i \geq c.$$

Each weighted by a kernel $K(\frac{x_i - c}{h})$ emphasizing observations near c:

$$\min_{\alpha_{\text{below}}, \beta_{\text{below}}} \sum_{i: x_i < c} K\left(\frac{x_i - c}{h}\right) (y_i - \alpha_{\text{below}} - \beta_{\text{below}}(x_i - c))^2,$$

$$\min_{\alpha_{\text{above}}, \beta_{\text{above}}} \sum_{i: x_i \geq c} K\left(\frac{x_i - c}{h}\right) (y_i - \alpha_{\text{above}} - \beta_{\text{above}}(x_i - c))^2.$$

The estimated discontinuity:

$$\boxed{\hat{\tau} = \hat{\alpha}_{\text{above}} - \hat{\alpha}_{\text{below}}}$$

Interpretation: $\hat{\tau}$ measures the difference in the fitted lines at $x = c$.

8.6: Beyond the Mean: Quantile and RIF Regressions

Why Go Beyond the Mean?

- OLS estimates the effect of x on the **conditional mean** $E[y|x]$.
- But economic effects can differ across the outcome distribution:

  Wage returns to education, treatment effects, inequality changes.

- Quantile regression allows heterogeneity:

$$Q_\tau(y|x) = x'\beta_\tau, \quad \text{for } \tau \in (0,1)$$

- Each β_τ describes the marginal effect of x at quantile τ, e.g. "effect on the 10th vs. 90th percentile".
- Insight: Policies may compress or stretch the distribution, not just shift its mean.

Quantile Regression: Intuition

- OLS finds the line that makes the **average residual** zero:

$$\mathbf{E}[\varepsilon_i | x_i] = 0 \quad \Rightarrow \quad \text{best fit for the mean of } y|x.$$

- Quantile regression instead finds the line that makes, say, **half the residuals positive and half negative**:

$$\mathbf{E}[\mathbb{1}\{\varepsilon_i < 0\} | x_i] = \tau \quad \Rightarrow \quad \text{best fit for the } \tau\text{-quantile of } y|x.$$

- For $\tau = 0.5$ this gives the conditional median; for $\tau = 0.9$ it fits the 90th percentile, and so on.
- Same idea as OLS, but instead of “best fit for the mean,” it’s the “best fit for a chosen part of the distribution.”

Interpreting Quantile Regressions

- Each β_τ shows how x shifts the τ-quantile of $y|x$:

$$Q_\tau(y|x+1) - Q_\tau(y|x)$$

- Differences across τ reveal **heterogeneous effects**:
  - Education may raise wages mainly at the top quantiles.
  - Minimum wages affect the lower tail more strongly.
- Plotting β_τ against τ shows how effects vary across the outcome distribution.
- Note that, these are **conditional** quantiles. They describe how x affects the distribution **given** covariates!

RIF-Regression: Effects on Unconditional Quantiles

- Quantile regression: effect on **conditional** quantiles $Q_\tau(y|x)$.
- Often we care about how x shifts the **unconditional** distribution, e.g. the overall 10th or 90th percentile

Idea (Firpo, Fortin & Lemieux, 2009): Use the **Recentered Influence Function (RIF)** of a statistic v (such as a quantile).

$$\text{RIF}(y_i; v) = v + \text{IF}(y_i; v)$$

- Each observation's RIF shows how it influences the statistic v.
- Key property: $\mathbf{E}[\text{RIF}(y_i; v)] = v$
- **Common Use:** Regress the RIF on covariates:

$$\mathbf{E}[\text{RIF}(y_i; v) \mid x_i] = x_i' \beta_v$$

- β_v shows how x_i affects the **unconditional quantile** (or other statistic)