Chapter 11: Why Model?

Contents

Part I of this book was mostly conceptual, with calculations kept to a minimum. In contrast, Part II requires the use of computers to fit regression models. This chapter describes the differences between the **nonparametric estimators** used in Part I and the **parametric (model-based) estimators** used in Part II. It reviews the concept of smoothing and the bias-variance trade-off in modeling decisions, motivating the need for models in data analysis regardless of whether the goal is causal inference or prediction.

This chapter is based on Hernán and Robins (2020, chap. 11, pp. 147-156).

Important context: Part II uses real data that can be downloaded from the book's website. The analyses require regression techniques (linear and logistic models), and the book provides code in R, SAS, Stata, and Python.

1 11.1 Data Cannot Speak for Themselves (pp. 147-150)

Even the simple task of estimating a population mean requires modeling assumptions when data become sparse.

1.1 Example: HIV Treatment and CD4 Count

Consider a study of 16 HIV-positive individuals randomly sampled from a super-population. Each receives treatment A (antiretroviral therapy), and we measure outcome Y (CD4 cell count, cells/mm^3).

Goal: Estimate the population mean $\mathrm{E}[Y|A = a]$ for each treatment level a.

1.2 Scenario 1: Binary Treatment

Treatment $A \in \{0, 1\}$ with 8 individuals in each group.

Estimator: Sample average within each group

- Estimate for $A = 0$: $\bar{Y}_{A=0} = 67.50$
- Estimate for $A = 1$: $\bar{Y}_{A=1} = 146.25$

This **nonparametric estimator** (sample mean) is consistent and unbiased.

Why this works:

- 8 observations per group provides reasonable precision
- No modeling assumptions needed
- Sample mean is the maximum likelihood estimator under minimal assumptions

Confidence intervals: With 8 observations, we can compute valid confidence intervals around each mean.

1.3 Scenario 2: Four Treatment Levels

Treatment $A \in \{1, 2, 3, 4\}$ (none, low-dose, medium-dose, high-dose) with 4 individuals per group.

Estimates: 70.0, 80.0, 117.5, 195.0 for $A = 1, 2, 3, 4$ respectively.

Issue: With only 4 individuals per category:

- Sample averages are still unbiased
- But estimates are less precise (wider confidence intervals)
- More variability in estimates across categories

1.4 Scenario 3: Continuous Treatment

Treatment A is dose in mg/day, taking integer values from 0 to 100 mg.

Problem: With 16 individuals and 101 possible treatment values:

- Many treatment levels have zero observations
- Cannot compute sample average for unobserved treatment levels
- **The nonparametric estimator is undefined** for A values with no data

Question: How do we estimate $\mathrm{E}[Y|A = 90]$ when no one received dose 90?

This is the fundamental motivation for modeling:

When data are sparse relative to the dimension of covariates, nonparametric estimation fails or becomes imprecise. We need to **borrow strength** across treatment levels by making assumptions about the relationship between treatment and outcome.

The curse of dimensionality returns: As the number of treatment levels (or combinations of confounders) increases, we have fewer observations per cell, requiring parametric assumptions.

2 11.2 Parametric Estimators of the Conditional Mean (pp. 150-152)

Parametric models make assumptions about the functional form relating treatment to outcome, allowing estimation even with sparse data.

2.1 Linear Regression Model

Assume the conditional mean follows a linear function:

$$\mathrm{E}[Y|A=a] = \beta_0 + \beta_1 a$$

Parameters: (β_0, β_1) define the line.

Estimation: Fit the model using least squares to estimate $(\hat{\beta}_0, \hat{\beta}_1)$.

Prediction: For any value a, estimate $\mathrm{E}[Y|A=a] = \hat{\beta}_0 + \hat{\beta}_1 a$.

Example 2.1 (Linear Model for Continuous Treatment)**.** With the HIV data and continuous treatment dose:

- Fit: $\mathrm{E}[Y|A] = \beta_0 + \beta_1 A$
- Obtain estimates: $\hat{\beta}_0 = 70$, $\hat{\beta}_1 = 1.25$
- Predict for $A = 90$: $\hat{E}[Y|A=90] = 70 + 1.25(90) = 182.5$

Even though no one received dose 90, the model provides an estimate by **interpolation** from observed doses.

Advantages of parametric models:

1. **Handle sparse data**: Estimate for all treatment levels, not just observed ones
2. **Precision**: Borrow strength across observations, smaller standard errors
3. **Smoothness**: Avoid choppiness from sample means in small groups

Disadvantages:

1. **Model misspecification**: If true relationship isn't linear, estimates are biased
2. **Bias-variance trade-off**: Reduce variance at the cost of potential bias

Key assumption: The linear functional form is correct (or close enough).

2.2 Other Parametric Models

Quadratic model:

$$\mathrm{E}[Y|A=a] = \beta_0 + \beta_1 a + \beta_2 a^2$$

Logarithmic model:

$$\mathrm{E}[Y|A=a] = \beta_0 + \beta_1 \log(a)$$

Piecewise linear (splines): Different linear relationships in different ranges of A.

Each model makes different assumptions about the shape of the dose-response curve.

3 11.3 Smoothing (pp. 152-153)

Smoothing refers to techniques that estimate the conditional mean as a smooth function, balancing between nonparametric flexibility and parametric smoothness.

3.1 The Smoothing Spectrum

Nonparametric (no smoothing): - Sample means within groups - No assumptions about functional form - High variance when data are sparse

Parametric (maximum smoothing): - Linear, quadratic, etc. models - Strong assumptions about functional form - Low variance but potential bias

Semiparametric (intermediate smoothing): - Methods that smooth but make weaker assumptions - Examples: kernel smoothing, local regression, splines - Balance bias and variance

Smoothing intuition:

When estimating $\mathrm{E}[Y|A = 90]$ with no observations at $A = 90$:

- **No smoothing**: Cannot estimate (undefined)
- **Weak smoothing**: Use nearby observations (e.g., $A = 85, 95$) with weights decreasing by distance
- **Strong smoothing**: Use all data, assuming a global functional form (e.g., linear)

The bias-variance trade-off:

- More smoothing → Less variance, more potential bias
- Less smoothing → More variance, less bias
- Optimal smoothing depends on the true (unknown) relationship and sample size

3.2 Kernel Smoothing

Idea: Estimate $\mathrm{E}[Y|A = a]$ using a weighted average of nearby observations.

$$\hat{E}[Y|A = a] = \frac{\sum_i K\left(\frac{A_i - a}{h}\right) Y_i}{\sum_i K\left(\frac{A_i - a}{h}\right)}$$

where:

- $K(\cdot)$ is a **kernel function** (e.g., Gaussian)
- h is the **bandwidth** (controls amount of smoothing)

Bandwidth selection:

- Large h: More smoothing, use distant observations
- Small h: Less smoothing, use only nearby observations

4 11.4 The Bias-Variance Trade-Off (pp. 153-155)

Every statistical estimator involves a trade-off between **bias** and **variance**.

4.1 Definitions

Bias: Systematic error, the difference between the expected value of the estimator and the true parameter.

$$\mathrm{Bias}(\hat{\theta}) = \mathrm{E}\left[\hat{\theta}\right] - \theta$$

Variance: Random error, the variability of the estimator across repeated samples.

$$\mathrm{Var}(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta} - \mathrm{E}\left[\hat{\theta}\right])^2\right]$$

Mean squared error (MSE): Combines both sources of error.

$$MSE(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$$

4.2 The Trade-Off

Nonparametric estimators (e.g., sample means):

- Bias: Low (if sufficient data in each cell)
- Variance: High (when data are sparse)
- Works well with dense data, fails with sparse data

Parametric estimators (e.g., linear regression):

- Bias: Potentially high (if model is misspecified)
- Variance: Low (uses all data efficiently)
- Works with sparse data, but relies on correct specification

Example: Estimating $\mathbf{E}[Y|A = 90]$

Nonparametric (sample mean of individuals with $A = 90$):

- If $n_{A=90} = 0$: Undefined (infinite variance)
- If $n_{A=90} = 1$: Unbiased but very high variance
- If $n_{A=90} = 100$: Unbiased and low variance

Linear model ($\text{E}[Y|A] = \beta_0 + \beta_1 A$):

- Uses all 16 observations to estimate 2 parameters
- Low variance estimate for any A
- Biased if true relationship is not linear

Practical implications:

1. With small samples and sparse data, we must accept some bias to reduce variance
2. The "best" estimator minimizes MSE, balancing bias and variance
3. Model selection involves choosing the level of smoothing that optimizes this trade-off

5 11.5 The Bias-Variance Trade-Off in Action (pp. 155-156)

Simulation studies can illustrate the bias-variance trade-off across different sample sizes.

5.1 Simulation Setup

1. Specify a true data-generating process (e.g., $\text{E}[Y|A] = \beta_0 + \beta_1 A + \beta_2 A^2$)
2. Generate many datasets of size n from this process
3. Fit different models to each dataset
4. Compute bias, variance, and MSE of each estimator

5.2 Typical Results

Small sample size ($n = 16$):

- Nonparametric: High variance, low bias (where defined)
- Simple parametric (linear): Low variance, moderate bias
- Flexible parametric (quadratic): Moderate variance, low bias
- **Winner**: Simple or moderately flexible model (minimizes MSE)

Large sample size ($n = 1000$):

- Nonparametric: Low variance, low bias
- Simple parametric (linear): Low variance, high bias (if misspecified)
- Flexible parametric: Low variance, low bias

- **Winner**: Flexible models or nonparametric (both work well)

Key lessons:

1. **Small samples require strong assumptions**: With limited data, we need parametric models
2. **Large samples allow flexibility**: With abundant data, we can use flexible or nonparametric approaches
3. **No free lunch**: Every estimator makes trade-offs; the "best" depends on:
    - Sample size
    - True underlying relationship
    - Sparsity of data
4. **Model selection matters**: Choose models that balance the bias-variance trade-off appropriately for your data

Practical strategy:

- Start with simple models
- Add complexity as sample size permits
- Use cross-validation or information criteria to select model complexity
- Conduct sensitivity analyses with different models

6 Summary

This chapter motivated the need for **statistical models** in data analysis.

Key concepts:

1. **Data sparsity problem**:
    - Nonparametric estimation fails when data are sparse
    - Need to borrow strength across observations
2. **Parametric models**:
    - Make assumptions about functional form
    - Allow estimation even with sparse data
    - Examples: Linear, quadratic, logarithmic models
3. **Smoothing**:
    - Spectrum from nonparametric to parametric
    - Semiparametric methods balance flexibility and smoothness
    - Bandwidth/complexity controls amount of smoothing
4. **Bias-variance trade-off**:
    - Bias: Systematic error from modeling assumptions
    - Variance: Random error from finite samples
    - MSE = Bias^2 + Variance
    - Optimal estimator minimizes MSE
5. **Sample size matters**:
    - Small samples: Need strong assumptions (parametric models)
    - Large samples: Can use flexible/nonparametric approaches

Implications for Part II:

The remainder of Part II describes methods for causal inference that use parametric and semiparametric models:

- **Chapter 12**: Standardization with outcome regression models
- **Chapter 13**: IP weighting with propensity score models
- **Chapter 14**: Instrumental variables with parametric models
- **Chapter 15**: G-estimation with structural nested models

All these methods face the bias-variance trade-off. Understanding this trade-off is crucial for:

- Choosing appropriate models
- Assessing model adequacy
- Interpreting results cautiously

The central challenge: We must make modeling assumptions to estimate causal effects with finite data, but wrong assumptions introduce bias. The art of causal inference involves making reasonable assumptions and assessing their plausibility.

7 References

Hernán, Miguel A, and James M Robins. 2020. *Causal Inference: What If*. Chapman & Hall/CRC. https://miguelhernan.org/whatifbook.