# Lecture 09: Probability and Maximum Likelihood Estimation

CS109A: Introduction to Data Science

Harvard University

|  |  |
|---|---|
| ■ **Course:** | CS109A: Introduction to Data Science |
| ■ **Lecture:** | Lecture 09 |
| ■ **Instructor:** | Pavlos Protopapas, Kevin Rader, Chris Gumb |
| ■ **Objective:** | Understanding the probabilistic foundations of linear regression, connecting OLS to MLE, and comparing formula-based inference with bootstrapping |

## Contents

# 1 Introduction and Motivation

**Lecture Overview**

This lecture bridges the gap between machine learning and statistics. We'll discover that our "loss function" approach to linear regression has deep connections to probability theory—and this connection will help us understand when to trust our model's estimates.

**Key Topics:**

- **Probability Foundations**: Random variables, PMF, PDF
- **Key Distributions**: Normal (Gaussian) and Binomial
- **The Central Limit Theorem**: Why normal distributions are everywhere
- **Likelihood and MLE**: A new perspective on model fitting
- **The Big Connection**: OLS = MLE under normality assumptions
- **Formula-Based vs. Bootstrap Inference**: When each method shines

## 1.1 Today's Question: How Much is a House Worth?

Professor Rader motivates probability theory with a concrete question:

*"What determines the selling price of a home in Cambridge/Somerville?"*

Using data from Redfin.com (592 home sales), we'll explore:

- Response variable ($Y$): Selling price (in thousands of dollars)
- Predictors ($X$): Type (condo, single-family, etc.), bedrooms, bathrooms, square footage, lot size, year built, distance to Harvard Square T-stop

This real-world example will illustrate why probability theory matters for data science.

# 2 Exploratory Data Analysis: The Housing Dataset

## 2.1 Data Cleaning

Before modeling, we need to address data quality issues:

**Data Preprocessing Steps**

1. **Missing values in lot size**: Mostly condos/townhouses that don't own land
   - Solution: Impute with 0 (reasonable assumption)
2. **Missing values in HOA fees**: Mostly single-family homes without HOA
   - Solution: Impute with 0 (no HOA = no fees)
3. **Price scale**: Convert from dollars to thousands of dollars
   - Makes coefficients easier to interpret
4. **Zip code type**: Convert from numeric to categorical

- Zip codes don't have meaningful numerical order!

## 2.2 Key Observations from EDA

### 2.2.1 Heteroscedasticity

**Definition: Heteroscedasticity**

**Heteroscedasticity** occurs when the variance of residuals is not constant across all values of the predictor.

**In this dataset**: Smaller homes have less price variability. Larger homes have much more price variability.

This violates the assumption of constant variance in linear regression!

**Example: Visualizing Heteroscedasticity**

When plotting price vs. square footage:

- Small homes (1000 sqft): Prices cluster tightly, perhaps \$400K-\$600K
- Large homes (3000 sqft): Prices spread widely, perhaps \$800K-\$2M+

The "funnel" or "cone" shape is a classic sign of heteroscedasticity.

### 2.2.2 Collinearity Strikes Again

When fitting a multiple regression model, an interesting result emerges:

```
                 coef    std err          t      P>|t|
Intercept  -1949.0670    745.203     -2.615      0.009
sqft             0.6411      0.044     14.720      0.000
beds           -89.9345     23.532     -3.822      0.000  <-- Negative!
baths          198.4646     31.332      6.334      0.000
```

**The Negative Bedroom Coefficient**

**Why is the coefficient for "beds" negative?**

This seems counterintuitive—shouldn't more bedrooms increase price?

**Interpretation**: "Holding *square footage constant*, each additional bedroom is associated with a \$89,934 *decrease* in price."

**The insight**: If you're cramming another bedroom into the *same* total square footage, you're creating smaller, more cramped rooms. Homes with lots of tiny bedrooms sell for less than homes with fewer, larger rooms (same total sqft).

This is collinearity at work—"beds" and "sqft" are highly correlated, so interpreting one while holding the other constant creates unusual but meaningful interpretations.

# 3 Review: Cross-Validation and Regularization

Before diving into probability, let's solidify some key concepts from previous lectures.

## 3.1 When to Use Cross-Validation

**Cross-Validation is for Model Selection**

Cross-validation can be used whenever you need to **choose between models**:

- **A. Choosing $k$ in k-NN**: Different $k$ values = different models
- **B. Choosing $\lambda$ in Ridge/Lasso**: Different $\lambda$ values = different models
- **C. Choosing predictors**: Different feature sets = different models
- **D. Choosing model families**: k-NN vs. linear regression = different models

**Answer: ALL OF THE ABOVE!**

Whenever you have a choice between models, cross-validation helps you make that choice objectively.

## 3.2 When to Standardize Predictors

**Standardization Guidance**

Standardize predictors when you want them to be **treated equally**:

- **k-NN**: Without standardization, variables on larger scales (price in dollars) dominate distance calculations over variables on smaller scales (number of rooms)
- **Ridge/Lasso**: Regularization penalizes coefficient magnitude. If one variable is measured in inches vs. feet, its coefficient scale changes artificially

**But not always!** If you have categorical dummies alongside continuous variables, you might *not* want to standardize everything equally. Use judgment!

## 3.3 Reading Trajectory Plots

Trajectory plots show how coefficients change as regularization strength ($\lambda$) increases:

- **Lasso**: Coefficients can become exactly zero (feature selection!)
- **Ridge**: Coefficients approach but never reach zero

**"Textbook" vs. Real-World Trajectory Plots**

**Textbook plots**: Smooth curves where all coefficients steadily shrink toward zero. These assume predictors are **independent**—unrealistic!

**Real-world plots**: Messy curves that may:

- Cross zero (sign changes!)
- Temporarily *increase* before decreasing

These patterns indicate **collinearity**: As one variable gets penalized, a correlated variable "picks up" its predictive power.

# 4 Probability Fundamentals

Now we build the probabilistic foundation needed to understand MLE.

## 4.1 What is Probability?

**Definition: Probability**

**Probability** is the long-run relative frequency of an event occurring.

**Range**: 0 (never happens) to 1 (always happens)

**Why we care in data science**: Our data is a **random realization** from some underlying data-generating process. Probability gives us the language to reason about uncertainty.

## 4.2 Random Variables

**Definition: Random Variable**

A **random variable** assigns numeric values to outcomes of a random phenomenon.

**Example**: Define $X_1 = 1$ if a randomly sampled Harvard student uses a Mac, $X_1 = 0$ otherwise. We don't know $X_1$'s value until we sample—it's "random."

We describe random variables by their **distribution**—the set of possible values and their probabilities.

## 4.3 Discrete vs. Continuous: PMF vs. PDF

Random variables come in two flavors:

| **Aspect** | **Discrete** | **Continuous** |
|---|---|---|
| Values | Countable (0, 1, 2, ...) | Any real number in a range |
| Function | PMF (Probability Mass Function) | PDF (Probability Density Function) |
| $P(X = x)$ | Has a specific probability | **Always equals 0!** |
| Probabilities | Direct from PMF | Area under PDF curve |
| Example | Number of bedrooms | House price |

**Table 1:** *Discrete vs. Continuous random variables*

**Continuous Variables: $P(X$**

For continuous random variables, the probability of any *exact* value is zero.

$P(\text{house price} = \$1,250,000.00) = 0$

Instead, we calculate probabilities over **intervals**:

$P(\$1,200,000 < \text{price} < \$1,300,000) = \text{area under PDF from 1.2M to 1.3M}$

# 5 Key Probability Distributions

## 5.1 The Bernoulli Distribution

The simplest distribution—a single coin flip:

**Definition: Bernoulli Distribution**

$X \sim \text{Bernoulli}(p)$

- $X = 1$ (success) with probability $p$
- $X = 0$ (failure) with probability $1 - p$

**PMF**: $P(X = x) = p^x(1-p)^{1-x}$ for $x \in \{0, 1\}$

**Example**: $X = 1$ if a student uses Mac, 0 otherwise

## 5.2 The Binomial Distribution

Multiple independent Bernoulli trials:

**Definition: Binomial Distribution**

$X \sim \text{Binomial}(n, p)$

Count of successes in $n$ independent trials, each with success probability $p$.

**PMF**: $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ is "n choose k"—the number of ways to arrange $k$ successes among $n$ trials.

**Mean**: $\mu = np$

**Standard Deviation**: $\sigma = \sqrt{np(1-p)}$

**Example: Binomial Example**

20% of Harvard students are varsity athletes. In a random sample of 200 students:

**Expected number of athletes**: $np = 200 \times 0.2 = 40$

**Probability of exactly 50 athletes**:

$$P(X = 50) = \binom{200}{50}(0.2)^{50}(0.8)^{150}$$

This can be computed in Python: `stats.binom.pmf(50, 200, 0.2)`

## 5.3 The Normal (Gaussian) Distribution

The most important distribution in statistics:

**Definition: Normal Distribution**

$X \sim N(\mu, \sigma^2)$

A continuous, bell-shaped distribution parameterized by:

- $\mu$ (mu): The mean (center of the bell)
- $\sigma^2$ (sigma squared): The variance (spread of the bell)

**PDF**: $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
**Standard Normal**: $Z \sim N(0, 1)$ (mean 0, variance 1)
**Standardization**: Any $X \sim N(\mu, \sigma^2)$ can be converted to standard normal:

$$Z = \frac{X - \mu}{\sigma}$$

$Z$ tells you "how many standard deviations away from the mean."

### 5.4 The Central Limit Theorem (CLT)

**Very Important: Why Normal Distributions Are Everywhere**

**Central Limit Theorem (CLT)**:
Regardless of the original population distribution, the **sample mean** $\bar{X}$ of $n$ independent observations approaches a normal distribution as $n$ gets large:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

**Key implications**:

- The mean of $\bar{X}$ equals the population mean $\mu$
- The variance of $\bar{X}$ **shrinks** as $n$ increases (by factor of $1/n$)
- More data → more precise estimates!

**Example: Why Human Height is Normally Distributed**

Adult height results from the accumulation of many small factors:

- Genetics (many genes, each with small effect)
- Nutrition during childhood
- Sleep quality
- Exercise patterns
- And countless other factors...

By CLT, the "sum" of many small independent factors tends toward a normal distribution. This is why height histograms look bell-shaped!

# 6 From Probability to Inference: The Likelihood Function

## 6.1 Probability vs. Inference: Two Directions

**The Two Directions of Statistical Reasoning**

**Probability (Deduction)**: Model → Data

- *Question*: "Given a fair coin ($p = 0.5$), what's the probability of getting 8 heads in 10 flips?"
- We know the model, we predict the data

**Inference (Induction)**: Data → Model

- *Question*: "I flipped a coin 10 times and got 8 heads. Is this coin fair ($p = 0.5$) or biased?"
- We have data, we infer the model

**Data science is mostly inference!** We observe data and try to figure out what process generated it.

## 6.2 The Likelihood Function

**Definition: Likelihood Function**

The **likelihood function** is mathematically the same as the PMF/PDF, but with a different perspective:

- **PMF/PDF** $f(x|\theta)$: Fix parameters $\theta$, vary data $x$
  - "Given this model, how probable is this data?"
- **Likelihood** $L(\theta|x)$: Fix data $x$, vary parameters $\theta$
  - "Given this data, how plausible is this model?"

**For independent observations**:

$$L(\theta|x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i|\theta)$$

The likelihood is a **product** of individual likelihoods (under independence).

## 6.3 Log-Likelihood: Making Life Easier

Products are mathematically inconvenient. Taking logs converts products to sums:

**Definition: Log-Likelihood**

The **log-likelihood** function:

$$\ell(\theta|x_1, \ldots, x_n) = \log L(\theta) = \sum_{i=1}^{n} \log f(x_i|\theta)$$

**Why use log-likelihood?**

1. Products become sums (much easier to work with!)
2. Since log is monotonically increasing, maximizing $L$ is equivalent to maximizing $\ell$

3. Numerical stability (avoids very small numbers from many multiplications)

### 6.4 Maximum Likelihood Estimation (MLE)

**Definition: Maximum Likelihood Estimation**

**MLE** finds the parameter value(s) that maximize the likelihood function:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} L(\theta|\text{data}) = \arg\max_{\theta} \ell(\theta|\text{data})$$

**Intuition**: "Find the parameter that makes the observed data most probable."

**Example: MLE for Normal Distribution**

You observe three values: 3, 5, and 10. Assume they come from $N(\mu, 4)$ (variance = 4, unknown mean).

**What's the MLE for $\mu$?**

Intuitively: The sample mean! $\hat{\mu}_{MLE} = \bar{x} = \frac{3+5+10}{3} = 6$

**Why?** The likelihood function (plotted against $\mu$) is maximized when $\mu = \bar{x}$.

This isn't a coincidence—for normal distributions, the MLE of the mean is always the sample mean.

### 6.5 How to Find the MLE

Two approaches:

1. **Analytical (calculus)**: Take derivative of $\ell(\theta)$ with respect to $\theta$, set equal to zero, solve
   - Works when closed-form solution exists
   - Example: Linear regression
2. **Numerical (optimization)**: Use algorithms like gradient descent to minimize **negative** log-likelihood
   - Works when no closed-form solution
   - Example: Logistic regression (coming soon!)

## 7 The Big Connection: OLS = MLE

This is the central insight of the lecture!

### 7.1 Setting Up the Probabilistic Model

Recall our linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

**Key assumption**: The residuals $\epsilon_i$ are normally distributed:

$$\epsilon_i \sim N(0, \sigma^2)$$

This implies:

$$Y_i | X_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

Each $Y$ value, given its $X$, comes from a normal distribution centered at the regression line.

### 7.2 Building the Likelihood

For $n$ independent observations:

$$L(\beta_0, \beta_1, \sigma^2 | \text{data}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{2\sigma^2}\right)$$

### 7.3 The Log-Likelihood

Taking logs:

$$\ell(\beta_0, \beta_1, \sigma^2) = \underbrace{-\frac{n}{2}\log(2\pi\sigma^2)}_{\text{doesn't depend on } \beta} \underbrace{- \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2}_{\text{depends on } \beta}$$

### 7.4 Maximizing the Log-Likelihood

To find $\hat{\beta}_0, \hat{\beta}_1$ that maximize $\ell$:

- The first term doesn't involve $\beta$, so it's irrelevant for optimization
- The second term has a negative sign, so maximizing $\ell$ means **minimizing** the sum

**Very Important: The Central Result: OLS**

Maximizing the likelihood (MLE) with respect to $\beta_0, \beta_1$ is equivalent to minimizing:

$$\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2$$

**This is exactly the Sum of Squared Errors (SSE)!**
**Conclusion**: If residuals are normally distributed, then:

**Ordinary Least Squares (OLS) = Maximum Likelihood Estimation (MLE)**

This provides the **probabilistic justification** for why we use MSE as our loss function!

## 8 Statistical Inference: Quantifying Uncertainty

Now that we have the probabilistic framework, we can quantify uncertainty in our estimates.

### 8.1 Point Estimates Are Not Enough

We computed $\hat{\beta}_1 = 0.5898$ for the square footage coefficient. But:

- This is based on one sample of 592 homes
- A different sample would give a different $\hat{\beta}_1$
- How confident should we be in this specific value?

We need to quantify this uncertainty.

## 8.2 Two Approaches to Inference

1. **Bootstrap** (from previous lecture): Resample data, compute estimates, look at distribution
2. **Formula-based** (this lecture): Use mathematical formulas derived from probability theory

## 8.3 Formula-Based Confidence Intervals

Under the linear regression assumptions, we have closed-form formulas for the **standard error** of $\hat{\beta}_1$:

**Definition: Standard Error of the Slope**

$$\hat{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

where $\hat{\sigma}^2 = \frac{\sum(Y_i - \hat{Y}_i)^2}{n-p-1}$ (corrected MSE)

**Interpretation**:

- More data ($n$ larger) → smaller SE (more precise!)
- More spread in $X$ (larger $\sum(X_i - \bar{X})^2$) → smaller SE
- Better model fit (smaller $\hat{\sigma}^2$) → smaller SE

**Definition: Confidence Interval Formula**

$$95\% \text{ CI for } \beta_1 = \hat{\beta}_1 \pm t^* \cdot \hat{SE}(\hat{\beta}_1)$$

where $t^* \approx 2$ for 95% confidence (from the t-distribution).

**Interpretation**: If we repeated the study many times, approximately 95% of the intervals we construct would contain the true $\beta_1$.

## 8.4 Hypothesis Testing

**Hypothesis Testing for Regression Coefficients**

**Hypotheses**:

- $H_0$: $\beta_1 = 0$ (no association between $X$ and $Y$)
- $H_A$: $\beta_1 \neq 0$ (there is an association)

**Test statistic**:

$$t = \frac{\hat{\beta}_1 - 0}{\hat{SE}(\hat{\beta}_1)} = \frac{\hat{\beta}_1}{\hat{SE}(\hat{\beta}_1)}$$

**Interpretation**: How many standard errors is our estimate from zero?

**p-value**: Probability of seeing a $|t|$ this large or larger if $H_0$ were true.

**Decision**: If p-value $< 0.05$, reject $H_0$. Conclude the association is statistically significant.

# 9 Bootstrap vs. Formula-Based Inference

Let's compare the two approaches on our housing data:

## 9.1 The Comparison

For the square footage coefficient ($\hat{\beta}_1 = 0.5898$):

| Method | 95% CI | Width |
|---|---|---|
| Bootstrap | [0.487, 0.705] | 0.218 |
| Formula (statsmodels) | [0.544, 0.636] | 0.092 |

**Table 2:** *Confidence intervals: Bootstrap vs. Formula-based*

**Why the Difference?**

The formula-based CI is much **narrower**—it suggests we're more confident than we should be!

**The problem**: Formula-based inference assumes all linear regression assumptions hold, including **constant variance (homoscedasticity)**.

But we already identified **heteroscedasticity** in this data—larger homes have more variable prices!

When assumptions are violated, the formulas give **incorrect standard errors**, leading to **overly optimistic** (too narrow) confidence intervals.

## 9.2 When to Use Each Method

**Very Important: Choosing Your Inference Method**

**Use formula-based inference when**:

- All regression assumptions are reasonably met
- You want fast computation
- Results are easy to report (standard output in statsmodels)

**Use bootstrap when**:

- Assumptions may be violated (especially heteroscedasticity)
- You want robust inference without strong distributional assumptions
- The bootstrap captures the "true" variability in your data

**Bottom line**: Bootstrap is **safer**—it makes fewer assumptions and reflects the actual data distribution.

# 10 Using statsmodels for Inference

```python
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

```python

# Using formula interface (like R!)
model = smf.ols('price ~ sqft + beds + baths + type', data=df)
results = model.fit()

# Get full summary with standard errors, t-stats, p-values, CIs
print(results.summary())

# Extract specific values
print(f"Coefficient for sqft: {results.params['sqft']:.4f}")
print(f"Std Error: {results.bse['sqft']:.4f}")
print(f"p-value: {results.pvalues['sqft']:.4f}")
print(f"95% CI: {results.conf_int().loc['sqft'].values}")
```

Listing 1: Fitting regression with statsmodels

### statsmodels vs. sklearn

**sklearn**: Great for prediction, cross-validation, pipelines

- Doesn't provide standard errors, p-values, or confidence intervals
- Focus on predictive performance

**statsmodels**: Great for inference and interpretation

- Provides full statistical output
- R-like formula interface for easy model specification
- Automatically creates dummy variables for categorical predictors

**Use both!** sklearn for prediction tasks, statsmodels for understanding relationships.

## 11 Quick Reference Summary

Lecture 09 Quick Reference Card

### 1. Key Distributions

- **Bernoulli**: Single binary outcome ($p$)
- **Binomial**: Count of successes in $n$ trials ($n, p$)
- **Normal**: Bell-shaped continuous ($\mu, \sigma^2$)

### 2. Central Limit Theorem

Sample means approach normal distribution as $n$ increases:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

### 3. The Big Connection

Under assumption of normal residuals:

$$\text{OLS (minimize MSE)} \equiv \text{MLE}$$

This justifies using MSE as our loss function!

### 4. Inference Methods

- **Formula-based**: Fast, but assumes all assumptions hold
- **Bootstrap**: Robust, makes fewer assumptions
- When assumptions violated: **Bootstrap wins!**

### 5. Confidence Interval Formula

$$\hat{\beta}_1 \pm t^* \cdot \hat{SE}(\hat{\beta}_1)$$

where $t^* \approx 2$ for 95% CI

## 12 Common Questions and Answers

**Q: Why do we care about the probabilistic interpretation of OLS?**

A: It gives us:

1. Theoretical justification for using MSE
2. A framework for statistical inference (standard errors, p-values, CIs)
3. Understanding of what assumptions we're making

**Q: If bootstrap is more robust, why ever use formulas?**

A: Formulas are much faster (instant computation vs. 1000+ resamples). When assumptions hold, they give the same answer. For quick checks or when you're confident in assumptions, formulas are fine.

**Q: How do I know if heteroscedasticity is a problem?**

A: Plot residuals vs. fitted values (or vs. $X$). Look for "funnel" shapes or patterns. If you see non-constant spread, heteroscedasticity is present. Use bootstrap for inference!

**Q: Why is the coefficient for "beds" negative?**

A: Collinearity with square footage. "Holding sqft constant, more bedrooms means cramming smaller rooms into the same space"—which decreases value. Always interpret coefficients in the context of "holding other variables constant."

**Q: What percentage of Harvard students use Macs?**

A: Professor Rader's informal estimate: around 75-90%. This is used as a fun example for binomial distributions!

## 13 Looking Ahead

In the next lectures, we'll:

- Continue with statistical inference in more detail
- Move to **classification** problems (predicting categories, not numbers)
- Introduce **logistic regression**—where MLE doesn't have a closed-form solution

The probability foundation we built today will be essential for understanding logistic regression's likelihood function!