Lecture 16: Hierarchical Models and Bayesian Logistic Regression

CS109A: Introduction to Data Science

Harvard University

- **Course:** CS109A: Introduction to Data Science
- **Lecture:** Lecture 16
- **Instructors:** Pavlos Protopapas, Kevin Rader, Chris Tanner
- **Topics:** Logistic Regression Review, Log-Scale Regression, Beta-Binomial Model, Hierarchical Models, Bayesian Logistic Regression, Posterior Predictive Distribution

Key Summary

This lecture applies Bayesian thinking to logistic regression and introduces hierarchical models—one of the most powerful tools for handling structured data with groups.

Key Topics:

- Review: Interpreting logistic regression coefficients (NBA shooting example)
- Log-log regression for multiplicative relationships (housing prices)
- Review of the Beta-Binomial model and conjugacy
- Hierarchical models: Why and how to model grouped data
- The shrinkage effect: Borrowing strength across groups
- Bayesian logistic regression: Priors on β coefficients
- Posterior predictive distributions for forecasting
- Preview: When conjugacy fails, we need simulation (MCMC)

Contents

1 Review: Interpreting Logistic Regression

Before diving into new material, let's solidify our ability to interpret logistic regression models through a concrete example.

1.1 NBA Shooting Example

Data: All NBA field goal attempts from the previous season.

Response: $Y = 1$ if shot is successful, $Y = 0$ if missed.

Predictor: Distance from the hoop (in feet).

After fitting a logistic regression model:

$$\log\left(\frac{p}{1-p}\right) = 0.796 - 0.0474 \times \text{Distance}$$

1.2 Step-by-Step Interpretation

1. Interpreting the Intercept (0.796):

The intercept represents the log-odds when Distance = 0 (a layup right at the basket).

- Log-odds = 0.796
- Odds = $e^{0.796} = 2.22$
- Probability = $\frac{2.22}{1+2.22} = \frac{e^{0.796}}{1+e^{0.796}} \approx 0.69$

Interpretation: A shot from 0 feet has approximately a 69% probability of success.

2. Interpreting the Slope (-0.0474):

The slope represents the change in log-odds for each additional foot of distance.

- For each 1-foot increase in distance, log-odds **decreases** by 0.0474
- Odds ratio = $e^{-0.0474} \approx 0.954$
- Each additional foot **multiplies** the odds by 0.954 (a 4.6% decrease)

Interpretation: Longer shots are harder—makes sense!

3. Finding the Classification Boundary:

Where does the predicted probability equal 0.5?

At $P = 0.5$: Log-odds = 0

$$0 = 0.796 - 0.0474 \times \text{Distance}$$
$$\text{Distance} = \frac{0.796}{0.0474} \approx 16.8 \text{ feet}$$

Interpretation: Shots from less than 17 feet are predicted as makes; shots from beyond 17 feet are predicted as misses.

Key Information

The Complete Interpretation Recipe

1. Write out the model equation
2. Interpret intercept: Log-odds when all predictors = 0
3. Convert to probability: $p = \frac{e^{\text{log-odds}}}{1+e^{\text{log-odds}}}$
4. Interpret slopes: Change in log-odds per unit change
5. Convert to odds ratio: e^{β} = multiplicative change in odds
6. Find decision boundary: Set log-odds = 0, solve for X

2 Log-Log Regression: Multiplicative Models

Sometimes the relationship between X and Y is **multiplicative** rather than additive. This is common in financial and economic data.

2.1 The Problem with Standard Linear Regression

Consider predicting house prices from square footage. Both variables:

- Are strictly positive
- Have right-skewed distributions
- Show heteroscedasticity (variance increases with the mean)

2.2 The Solution: Log-Transform Both Variables

If we take the logarithm of both X and Y:

$$\log_2(Y) = \beta_0 + \beta_1 \log_2(X)$$

Example:

Housing Price Example Model on log-log scale:

$$\log_2(\text{Price}) = 12.46 + 0.722 \times \log_2(\text{SqFt})$$

Interpreting the slope (0.722):
What does a 1-unit change in $\log_2(\text{SqFt})$ mean?

- A 1-unit increase in $\log_2(X)$ means X **doubles**
- This produces a 0.722-unit increase in $\log_2(Y)$
- A 0.722 increase in $\log_2(Y)$ means Y is multiplied by $2^{0.722} \approx 1.65$

Final interpretation: When you **double** the square footage, the price increases by approximately **65%**.

Warning

Why Base 2?
Using $\log_2$ makes interpretation intuitive:

- 1 unit change = **doubling**
- Easy to conceptualize

If you use ln (natural log), a 1-unit change means multiplying by $e \approx 2.718$, which is harder to interpret.

Alternative interpretation (for natural log): For small β_1, approximately $\beta_1 \times 100\%$ change in Y per 1% change in X.

3 Review: Beta-Binomial Model

The Beta-Binomial model is the foundation for understanding Bayesian approaches to classification.

3.1 The Setup

Data: n independent trials, $\sum y_i$ successes, $n - \sum y_i$ failures.

Likelihood: Bernoulli (or Binomial)

$$Y_i|p \sim \text{Bernoulli}(p)$$

Prior: Beta distribution on p

$$p \sim \text{Beta}(a_0, b_0)$$

Hyperparameters: a_0 and b_0 encode our prior belief about p.

3.2 Conjugacy: Why Beta is Special

When we multiply the prior and likelihood:

$$\begin{aligned}\text{Posterior} &\propto \text{Likelihood} \times \text{Prior} \\ f(p|Y) &\propto p^{\sum y_i}(1-p)^{n-\sum y_i} \times p^{a_0-1}(1-p)^{b_0-1} \\ &= p^{(a_0+\sum y_i)-1}(1-p)^{(b_0+n-\sum y_i)-1}\end{aligned}$$

This is exactly a Beta distribution!

Important:

Beta-Binomial Conjugacy

$$\text{Prior: } p \sim \text{Beta}(a_0, b_0)$$

$$\text{Posterior: } p|Y \sim \text{Beta}(a_0 + \sum y_i, b_0 + n - \sum y_i)$$

The posterior is just the prior with successes and failures "added in."

3.3 Posterior Mean: A Weighted Average

The posterior mean is:

$$E[p|Y] = \frac{a_0 + \sum y_i}{a_0 + b_0 + n}$$

This is a **weighted average** of:

- The prior mean: $\frac{a_0}{a_0+b_0}$
- The MLE (sample proportion): $\frac{\sum y_i}{n}$

Key insight: As n (data) increases, the posterior mean approaches the MLE. The prior matters less

with more data.

4 Why Hierarchical Models?

Hierarchical models address a fundamental problem: data often has **natural grouping structure**.

4.1 Examples of Hierarchically Structured Data

- **Government**: Individual voters within counties within states within regions
- **Education**: Students within classrooms within schools within districts
- **Medicine**: Repeated measurements within patients within hospitals
- **Biology**: Cells within tissues within organs within organisms
- **Sports**: Shots within players within teams within leagues

4.2 The NBA Shooting Problem

Goal: Predict shot success based on distance, accounting for player ability.

Data structure:

- Response: Y_{ij} = success/failure of shot i by player j
- Predictor: X_{ij} = distance of shot i by player j
- Grouping: 600 different players, varying number of shots each

4.3 Three Approaches

Approach 1: Complete Pooling (Ignore Players)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \times \text{Distance}_{ij}$$

Problem: Assumes all players are identical. Ignores obvious differences (LeBron vs a rookie).

Approach 2: No Pooling (One-Hot Encoding)

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_0 + \beta_1 \times \text{Distance}_{ij} + \sum_j \gamma_j \times \mathbf{1}[\text{Player} = j]$$

Problem: 600+ parameters! Severe overfitting, especially for players with few shots.

Approach 3: Hierarchical Model (Partial Pooling)

The best of both worlds!

5 Hierarchical Models: The Setup

5.1 The Model

Level 1 (Shots within Players):

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j + \beta_1 \times \text{Distance}_{ij}$$

Each player j has their own intercept α_j (baseline shooting ability).

Level 2 (Players within League):

$$\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$$

The player intercepts are drawn from a common distribution:

- α_0 = league-average baseline ability
- σ_α^2 = variance of abilities across players

Definition:

Hierarchical Model In a hierarchical model, **parameters** at one level are treated as **random variables** drawn from a distribution defined at a higher level.
This creates "partial pooling"—each group's estimate is informed by:

1. Its own data
2. The overall distribution of all groups

5.2 What Are We Estimating?

Fixed effects (same for everyone):

- β_1: Effect of distance on log-odds (assumed same for all players)

Hyperparameters (describe the population):

- α_0: Mean player ability
- σ_α^2: Variance of player abilities

Random effects (vary by group):

- α_j for each player j: Individual player abilities

6 The Shrinkage Effect

The magic of hierarchical models lies in **shrinkage**—pulling extreme estimates toward the group mean.

6.1 The Problem with OLS

Consider two players:

- **Alandis Williams**: 1 shot, 1 made (100% in sample)
- **Mac McClung**: 2 shots, 0 made (0% in sample)

With standard OLS (no pooling):

- Williams gets coefficient $\to +\infty$ (predicted 100% from any distance!)
- McClung gets coefficient $\to -\infty$ (predicted 0% from any distance!)

These are terrible estimates based on almost no data.

6.2 The Hierarchical Solution

With hierarchical modeling:

- Players with many shots: Estimates based mostly on their own data
- Players with few shots: Estimates **shrunk** toward the league average

Example:

Shrinkage in Action **Top 5 players (OLS estimates):**

- Alandis Williams (never heard of): Coefficient $\approx +10$ (100% predicted!)

n very few shots

Top 5 players (Hierarchical estimates):

- Jarrett Allen, SGA, Damian Lillard... (famous, high-volume shooters!)

The hierarchical model **automatically discounts** estimates based on small samples and gives appropriate credit to players with substantial evidence.

Key Information

How Does Sample Size Enter the Model?

The α_j distribution connects each player's shots to the overall population. A player with:

- **Many shots**: Strong evidence from their Bernoulli trials pulls α_j toward their sample mean
- **Few shots**: Weak evidence, so α_j is dominated by the prior (population mean α_0)

The math automatically handles this trade-off!

7 Extending Hierarchical Models

7.1 Random Slopes

In our current model, β_1 (distance effect) is the same for all players. But maybe:

- Steph Curry is **better** at long-range shots
- Other players are **worse** from distance

We can let the slope vary too:

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \alpha_j + \beta_{1j} \times \text{Distance}_{ij}$$

where both α_j and β_{1j} are drawn from distributions.

7.2 Hyperpriors: Priors on Hyperparameters

What if we're uncertain about α_0 and σ_α^2?

We can add another layer:

- $\alpha_0 \sim N(\mu_0, \tau^2)$ (hyperprior on mean)
- $\sigma_\alpha^2 \sim$ Inverse-Gamma$(\ldots)$ (hyperprior on variance)

It's turtles all the way down!

In practice, we usually stop at 2-3 levels.

7.3 Fully Bayesian vs Empirical Bayes

Fully Bayesian: Put priors on ALL unknown parameters $(\alpha_0, \sigma_\alpha^2, \beta_1)$

Empirical Bayes: Estimate hyperparameters from the data (like treating $\alpha_0, \sigma_\alpha^2$ as fixed unknowns)

The hierarchical model we discussed is often fit in an "empirical Bayes" spirit, which is a blend of frequentist and Bayesian thinking.

8 Bayesian Logistic Regression

Now let's apply full Bayesian thinking to logistic regression parameters.

8.1 Why Not Use Beta Priors?

In the Beta-Binomial model, we put a Beta prior on p.

Can we do the same for logistic regression?

NO! Here's why:

- The Beta distribution has support $[0, 1]$
- In logistic regression, our parameters are $\beta_0, \beta_1, \ldots$ (the coefficients)
- These coefficients are on the **log-odds scale**, which is $(-\infty, +\infty)$
- A Beta prior would be inappropriate!

8.2 Normal Priors on β

Instead, we use **Normal priors**:

$$\beta_j \sim N(\mu_j, \sigma_j^2)$$

Common choices:

- $\mu_j = 0$: We expect coefficients to be near zero (conservative)
- σ_j^2 controls how strongly we believe this

Key Information

Connection to Ridge Regression

A Normal prior centered at 0 is the Bayesian interpretation of Ridge (L2) regularization!

- Strong prior (σ^2 small) = Strong regularization
- Weak prior (σ^2 large) = Weak regularization

8.3 The Loss of Conjugacy

Problem: Normal prior $\times$ Bernoulli likelihood $\neq$ Nice closed form!

The posterior distribution doesn't have a recognizable form. We can't write down:

$$E[\beta_j|\text{data}] = \text{simple formula}$$

Solution: Simulation! (MCMC, covered in next lecture)

9 Posterior Predictive Distribution

Once we have a model, we want to predict future observations.

9.1 Frequentist Prediction

In standard regression, prediction is simple:

$$\hat{y}_{\text{new}} = \hat{\beta}_0 + \hat{\beta}_1 x_{\text{new}}$$

Plug in point estimates and get a point prediction.

9.2 Bayesian Prediction

In Bayesian inference, parameters are random variables! We can't just "plug in" single values.

The posterior predictive distribution accounts for:

1. Uncertainty in the parameters (posterior distribution)
2. Inherent randomness in the outcome (likelihood)

Definition:

Posterior Predictive Distribution

$$f(\tilde{y}|\text{data}) = \int f(\tilde{y}|\theta) f(\theta|\text{data}) d\theta$$

This marginalizes out the uncertainty in θ by integrating over all possible parameter values, weighted by their posterior probability.

9.3 Intuition

To predict a new shot for an NBA player:

1. Draw a value of α_j, β_1 from their posterior distribution
2. Use those values to compute p for the new shot
3. Draw a success/failure from Bernoulli(p)
4. Repeat many times to get the full distribution of predictions

This gives us not just a point prediction, but a **distribution** capturing all our uncertainty.

10 Preview: When We Can't Solve Analytically

10.1 The Challenge

For many realistic Bayesian models:

- No conjugacy (posterior doesn't have nice form)
- High-dimensional parameter space
- Complex hierarchical structure

We cannot:

- Write down $f(\theta|\text{data})$ in closed form
- Compute $E[\theta|\text{data}]$ analytically
- Integrate to get posterior predictive distributions

10.2 The Solution: MCMC

Markov Chain Monte Carlo (MCMC) is a family of algorithms that:

1. Generate samples from the posterior distribution
2. Without needing to know its exact form
3. Use those samples to estimate quantities of interest

Key insight: If we can evaluate the posterior *up to a constant* (i.e., we know likelihood $\times$ prior), we can still sample from it!

10.3 Coming Up Next

In the next lecture, we'll cover:

- Monte Carlo integration
- The Metropolis-Hastings algorithm
- Practical considerations for MCMC

11 Summary

Logistic Regression Interpretation

1. Write out: $\log(p/(1-p)) = \beta_0 + \beta_1 X$
2. Intercept: Log-odds when $X = 0$; convert to probability with $\frac{e^{\beta_0}}{1+e^{\beta_0}}$
3. Slope: Change in log-odds per unit X; odds ratio $= e^{\beta_1}$
4. Decision boundary: Solve $\beta_0 + \beta_1 X = 0$ for X

Log-Log Regression

$$\log(Y) = \beta_0 + \beta_1 \log(X)$$

- Doubling X multiplies Y by 2^{β_1}
- Useful for multiplicative/financial relationships
- Often fixes heteroscedasticity

Hierarchical Models

- **Problem**: Grouped data (players, students, patients)
- **Solution**: Parameters vary by group, drawn from common distribution
- **Effect**: Shrinkage—extreme estimates pulled toward mean
- **Benefit**: Better estimates for groups with little data

Model: $y_{ij} \sim f(\alpha_j + \beta X_{ij})$, where $\alpha_j \sim N(\alpha_0, \sigma_\alpha^2)$

Bayesian Logistic Regression

- Put Normal priors on β coefficients (not Beta—wrong support!)
- Normal prior centered at 0 $\Leftrightarrow$ Ridge regularization
- No conjugacy: posterior doesn't have nice form
- Solution: MCMC simulation

Key Concepts

- **Conjugacy**: Prior + Likelihood = Same family posterior
- **Shrinkage**: Pulling estimates toward the mean
- **Hyperparameters**: Parameters of the prior distribution
- **Posterior predictive**: $\int f(\tilde{y}|\theta) f(\theta|\text{data}) d\theta$