Multinomial Logit / Conditional Logit

Why You Can't Just Run Four Logits

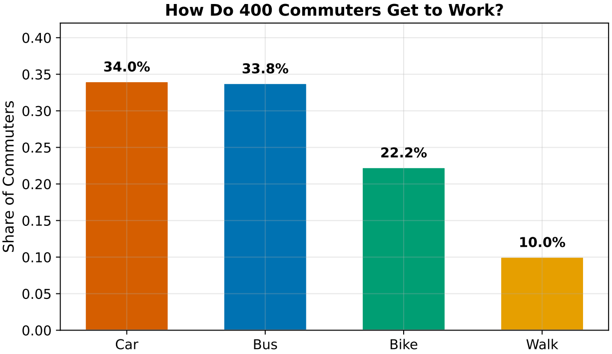
Jake Anderson

May 16, 2026

Outline

The Data

A city surveys **400 commuters** about how they get to work. Each person picks one of four modes: **Car**, **Bus**, **Bike**, or **Walk**.



The outcome is **categorical with 4 levels**. Binary logit handles 2 options. How do we handle 4?

First Instinct: Run Separate Binary Logits

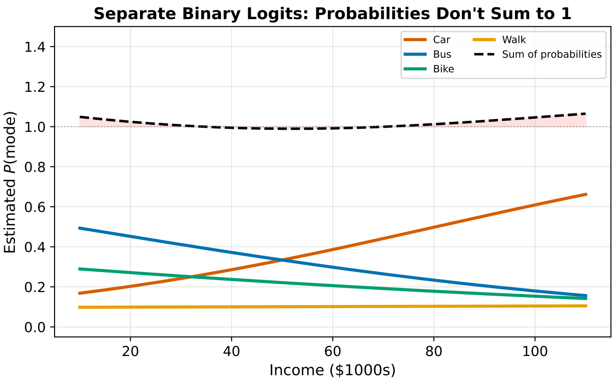
You already know binary logit. Why not run one for each mode?

1. Logit for Car vs. Not Car
2. Logit for Bus vs. Not Bus
3. Logit for Bike vs. Not Bike
4. Logit for Walk vs. Not Walk

Each model gives $\hat{P}(\text{mode})$ as a function of income. Seems straightforward.

But there is a structural problem: four separate models know nothing about each other.

Separate Binary Logits: The Failure



The four probabilities don't sum to 1. At some income levels, the total exceeds 1; at others, it falls short.

$\implies$ Separate binary models violate a basic requirement: a commuter must choose **exactly one** mode, so probabilities must sum to 1 across alternatives.

What Went Wrong, and What We Need

The problem: separate binary logits model each mode in isolation.

- Each model is free to assign any probability between 0 and 1
- Nothing forces the four probabilities to coordinate
- $\Longrightarrow$ They don't sum to 1, so they aren't valid choice probabilities

The requirement: we need a single model that assigns probabilities to *all four modes simultaneously*, with:

- Every $P(y_i = j) \in (0, 1)$
- $\sum_{j=1}^{4} P(y_i = j) = 1$ by construction

$\Longrightarrow$ The multinomial logit model achieves exactly this.

Outline

Random Utility Framework

Each commuter i assigns a **utility** to each mode j:

$$U_{ij} = V_{ij} + \varepsilon_{ij}$$

- V_{ij} = the part we can model (income, travel time, cost)
- ε_{ij} = unobserved taste variation (idiosyncratic preferences)

Decision rule: commuter i chooses mode j if $U_{ij} > U_{ik}$ for every other mode k.

$\Longrightarrow$ We never observe utility directly. We observe the *choice*, which reveals which mode had the highest utility for that person.

The Multinomial Logit Probability

If the ε_{ij} follow an i.i.d. Type I Extreme Value (Gumbel) distribution (chosen for mathematical convenience), the choice probability takes a clean form:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_{k=1}^{J} e^{V_{ik}}}$$

This is the **softmax** function (the same function used in machine learning classifiers). Note:

- Every $P(y_i = j) \in (0, 1)$
- $\sum_{j=1}^{J} P(y_i = j) = 1$ by construction

$\implies$ The softmax forces probabilities to be positive and sum to 1. This is the structural constraint that separate binary logits violate.

What Drives Mode Choice?

We observe two types of variables:

1. **Individual-specific** (same value across all alternatives):
   - Income, age, household size
   - Vary across *people*, not across modes
2. **Alternative-specific** (different value for each mode):
   - Travel time, travel cost
   - The same person faces different times/costs for Car vs. Bus vs. Bike

Higher income $\Longrightarrow$ more likely to drive. Longer bus time $\Longrightarrow$ less likely to take the bus.

MNL with Individual-Specific Variables

When the regressors vary across *people* but not across alternatives (e.g., income), the systematic utility is:

$$V_{ij} = \alpha_j + \beta_j x_i$$

where x_i is a person-level variable (e.g., income in \$10k).

Each alternative gets its **own intercept** α_j and its **own slope** β_j.

- α_j = baseline appeal of mode j ("alternative-specific constant")
- β_j = how income shifts the probability of choosing mode j

$\Longrightarrow$ The same variable (income) has a *different coefficient for each alternative*. This is what makes the probability curves fan out differently.

Numeric Example: Computing Utilities

Suppose Income = \$50k, so $x_i = 5$ (in \$10k units). With Bus as the base ($\alpha_{\text{Bus}} = 0$, $\beta_{\text{Bus}} = 0$):

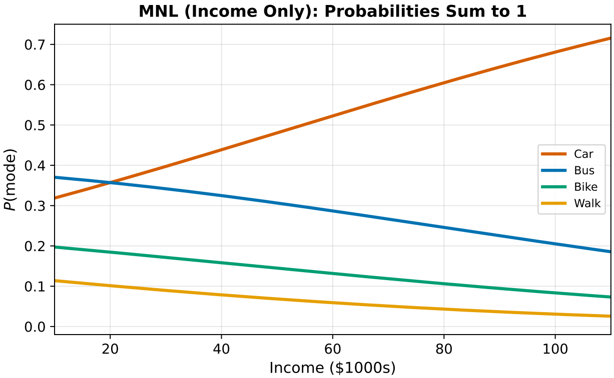
$$
\begin{aligned}
V_{\text{Car}} &= -0.3 + 0.15 \times 5 = 0.45 \\
V_{\text{Bus}} &= 0 + 0 \times 5 = 0 \quad \text{(base)} \\
V_{\text{Bike}} &= -0.6 + (-0.03) \times 5 = -0.75 \\
V_{\text{Walk}} &= -1.1 + (-0.08) \times 5 = -1.50
\end{aligned}
$$

Apply the softmax: $e^{0.45} + e^{0} + e^{-0.75} + e^{-1.50} = 3.26$

$$
P(\text{Car}) = \frac{e^{0.45}}{3.26} = 0.48, \quad P(\text{Bus}) = 0.31, \quad P(\text{Bike}) = 0.14, \quad P(\text{Walk}) = 0.07
$$

$\Longrightarrow$ Car has a lower baseline than Bus ($\alpha_{\text{Car}} < 0$), but the positive income effect ($\beta_{\text{Car}} = 0.15$) makes Car most likely at this income level.

MNL Probabilities: Visualized



As income rises, Car probability increases while Bike and Walk decline. At every income level, the four curves sum to 1.

Normalization: Choosing a Base Category

Absolute utility levels are not identified; only **differences** are identified. Here is why:

If we add the same constant c to every V_{ij}:

$$\frac{e^{V_{ij}+c}}{\sum_k e^{V_{ik}+c}} = \frac{e^c \cdot e^{V_{ij}}}{e^c \cdot \sum_k e^{V_{ik}}} = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

$\implies$ The probabilities are unchanged. We cannot separately identify α_{Car} and α_{Bus}, only their difference.

Fix: set one alternative as the **base category** (e.g., Bus) and normalize:

$$\alpha_{\text{Bus}} = 0, \quad \beta_{\text{Bus}} = 0$$

All estimated coefficients are then interpreted *relative to Bus*:

- $\beta_{\text{Car}} > 0$: higher income increases Car utility *relative to Bus*

Interpreting Coefficients

With Bus as the base category, the log-odds ratio is:

$$\ln\left(\frac{P(y_i = j)}{P(y_i = \text{Bus})}\right) = \alpha_j + \beta_j x_i$$

(This follows directly from dividing the softmax probabilities for mode j and Bus.)

$\implies$ β_j measures how income changes the **log-odds of mode j relative to Bus**.

Mode	$\hat{\alpha}_j$	$\hat{\beta}_j$	**Interpretation**
Bus (base)	0	0	(normalized)
Car	$-$	$+$	Lower baseline than Bus, but income pulls toward Car
Bike	$-$	$-$	Lower baseline, income pushes further away
Walk	$-$	$-$	Lowest baseline, income pushes further away

The sign of β_j tells us the direction *relative to Bus*. It does not directly give the marginal effect on probability (same nonlinearity issue as binary logit).

Marginal Effects in the Multinomial Logit

For an individual-specific variable x_i, the marginal effect on $P(y_i = j)$ is:

$$\frac{\partial P(y_i = j)}{\partial x_i} = P(y_i = j)\left[\beta_j - \sum_{k=1}^{J} P(y_i = k)\,\beta_k\right]$$

This looks complicated, but the intuition is simple:

- A variable can increase $P(j)$ even if $\beta_j = 0$, as long as it *decreases* the probability of other alternatives
- The effect depends on *all* alternatives' probabilities, not just mode j

Example: suppose $\beta_{\text{Car}} = 0.5$, $\beta_{\text{Bike}} = 0$, $\beta_{\text{Walk}} = -0.3$, and income rises by \$1k. Even though $\beta_{\text{Bike}} = 0$, income may *increase* $P(\text{Bike})$ if the probability-weighted average $\bar{\beta} = \sum_k P_k \beta_k$ is negative. In practice, $\bar{\beta}$ is usually dominated by the positive Car coefficient, so $P(\text{Bike})$ falls.

$\Longrightarrow$ Report Average Marginal Effects (AMEs) for each alternative, just as in binary logit.

Estimation: Maximum Likelihood

Define an indicator $d_{ij} = 1$ if person i chose mode j, and 0 otherwise.

Each person contributes one term to the likelihood: the probability of the mode they actually chose.

The log-likelihood:

$$\ell = \sum_{i=1}^{N} \sum_{j=1}^{J} d_{ij} \ln P(y_i = j)$$

where $P(y_i = j) = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$.

No closed-form solution $\Longrightarrow$ solved numerically, just like binary logit. Software handles this automatically.

Outline

The Limitation Multinomial Logit Cannot Overcome

MNL models **who** chooses what. But it ignores the attributes of the alternatives themselves.

Concrete example: two people with identical incomes face different bus routes.

- Person A: bus takes 20 minutes
- Person B: bus takes 50 minutes

MNL (income only) gives them the **same** probability of taking the bus:

$$V_{\text{Bus},A} = \alpha_{\text{Bus}} + \beta_{\text{Bus}} \cdot \text{Income} = V_{\text{Bus},B}$$

Travel time does not appear in the model. Person B, facing a 50-minute bus ride, gets the same $P(\text{Bus})$ as Person A with a 20-minute ride.

$\Longrightarrow$ We need to incorporate **alternative-specific variables**: attributes that vary across both people *and* modes (travel time, travel cost).

Conditional Logit: Alternative-Specific Variables

Conditional logit (McFadden, 1974) enters alternative-specific variables with a **single coefficient** shared across all modes:

$$V_{ij} = \beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}$$

- Time_{ij} = travel time person i faces for mode j
- β_{time} = the same for all modes

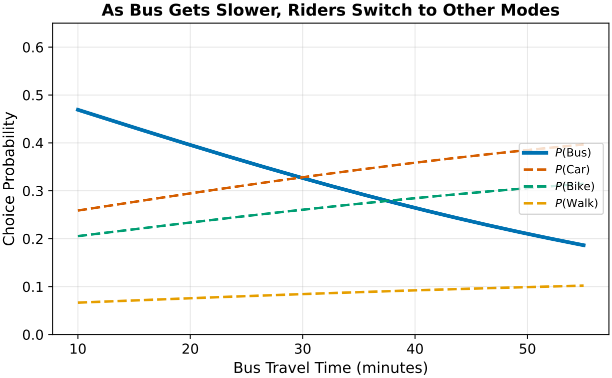
(In practice, we also include alternative-specific constants α_j. Here we focus on the distinctive feature: alternative-varying regressors with shared coefficients.)

This is the opposite pattern from multinomial logit:

	Regressors vary by. . .	**Coefficients. . .**
Multinomial logit	Individual only	Differ by alternative
Conditional logit	Individual *and* alternative	Same across alternatives

$\implies$ One extra minute of travel time reduces utility by β_{time}, regardless of whether it is a minute on the bus or a minute in the car.

Conditional Logit: Travel Time Effect



As bus travel time increases, $P(\text{Bus})$ falls and the other modes absorb the lost share. The rate of substitution depends on each mode's current probability.

The Full Model: Combining Both Variable Types

In practice, we combine both types of variables in one model. We add a superscript to distinguish income coefficients from the travel-time coefficient:

$$V_{ij} = \underbrace{\alpha_j + \beta_j^{\text{inc}} \cdot \text{Income}_i}_{\text{individual-specific}} + \underbrace{\beta_{\text{time}} \cdot \text{Time}_{ij} + \beta_{\text{cost}} \cdot \text{Cost}_{ij}}_{\text{alternative-specific}}$$

- $\alpha_j, \beta_j^{\text{inc}}$: vary by alternative (one per mode, base category normalized)
- $\beta_{\text{time}}, \beta_{\text{cost}}$: shared across all modes (one coefficient each)

The probability formula is the same softmax:

$$P(y_i = j) = \frac{e^{V_{ij}}}{\sum_k e^{V_{ik}}}$$

$\Longrightarrow$ This combined model is often called the "conditional logit" or "McFadden's choice model" in applied work, though terminology varies across textbooks. It handles both types of variation simultaneously.

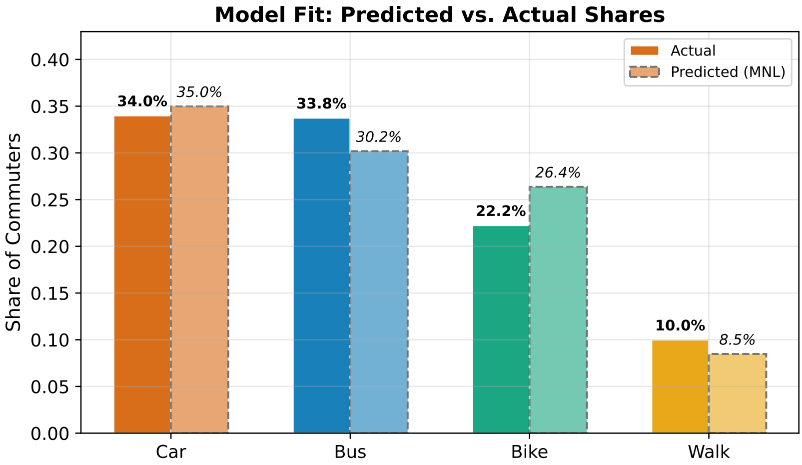
Model Comparison: What Each Specification Handles

	MNL	CL	Combined
Individual-specific variables (income)	✓		✓
Alternative-specific variables (time, cost)		✓	✓
Alternative-specific constants (α_j)	✓		✓
Alternative-varying slopes (β_j)	✓		✓
Shared slopes (β)		✓	✓

$\Longrightarrow$ The combined model subsumes both. It reduces to pure MNL when there are no alternative-specific variables, and to pure CL when there are no individual-specific variables.

All three use the same softmax probability and the same MLE estimator. The only difference is what goes into V_{ij}.

Model Fit: Predicted vs. Actual



The multinomial/conditional logit closely matches the observed mode shares. The model aggregates well even though individual predictions are probabilistic.

Outline

The IIA Assumption

The multinomial logit probability has a special structure. Take the ratio of any two alternatives:

$$\frac{P(y_i = j)}{P(y_i = k)} = \frac{e^{V_{ij}}}{e^{V_{ik}}} = e^{V_{ij} - V_{ik}}$$

This ratio depends **only on j and k**. Adding or removing a third alternative does not change it.

$\Longrightarrow$ **Independence of Irrelevant Alternatives (IIA):** the relative odds between any two modes are unaffected by what other modes exist.

This is both a strength (clean, tractable) and a weakness (unrealistic in some settings). Note: this property holds given fixed parameters, which is the standard presentation of IIA.

The Red Bus / Blue Bus Problem

A classic thought experiment exposes when IIA fails.

Before: two options with equal market shares:

- Car: 50% Bus: 50%

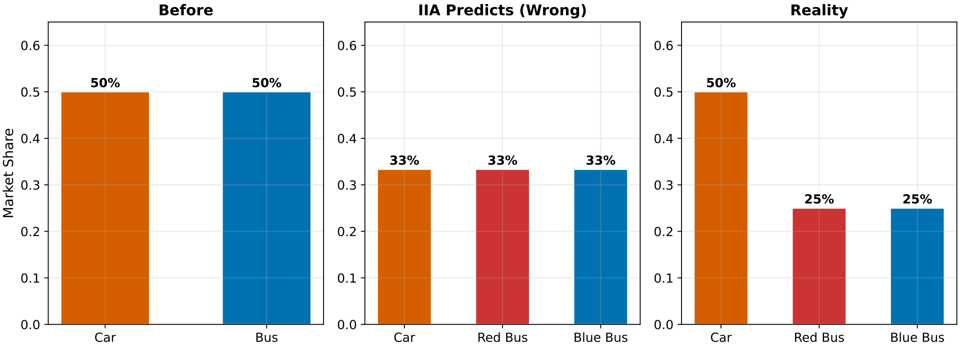
Now the city introduces a **Red Bus** that is identical to the existing (Blue) Bus. Under IIA, the multinomial logit predicts:

- Car: 33%, Red Bus: 33%, Blue Bus: 33%

But common sense says bus riders split between Red and Blue, while car drivers are unaffected:

- Car: 50%, Red Bus: 25%, Blue Bus: 25%

Red Bus / Blue Bus: Visualized



IIA says the relative odds of Car to Blue Bus stay the same after Red Bus enters. In reality, Red Bus steals from Blue Bus (a close substitute), not proportionally from all modes.

When Does IIA Fail?

IIA is problematic when alternatives are **close substitutes** for some people but not for others.

IIA fails when:

- Two alternatives share unobserved attributes (Red Bus $\approx$ Blue Bus)
- The error terms ε_{ij} are **correlated** across alternatives

IIA is reasonable when:

- Alternatives are genuinely distinct (car, bus, bike, walk are quite different)
- The observed variables capture the similarities between alternatives

Hausman test for IIA: estimate the model on a subset of alternatives. If IIA holds, the coefficients should not change significantly when you drop one alternative.

Beyond Multinomial Logit: Relaxing IIA

You don't need to estimate these models for this course, but you should know they exist.

Model	**How it relaxes IIA**
Nested logit	Groups similar alternatives into "nests" (e.g., Motorized: {Car, Bus} vs. Non-motorized: {Bike, Walk}). Allows correlation within nests
Mixed logit	Coefficients vary randomly across individuals. Generates flexible substitution patterns
Multinomial probit	Normal errors with a full covariance structure. Most general, but computationally expensive

$\Longrightarrow$ Start with multinomial logit and test IIA. If it fails, move to nested or mixed logit.

Decision Framework: Which Model to Use

1. **Only individual-specific variables?** (income, age, etc.)
   $\Longrightarrow$ Multinomial logit (MNL)
2. **Only alternative-specific variables?** (time, cost, etc.)
   $\Longrightarrow$ Conditional logit (CL)
3. **Both types of variables?**
   $\Longrightarrow$ Combined model (the usual case in practice)
4. **Alternatives are close substitutes?** (Red Bus / Blue Bus concern)
   $\Longrightarrow$ Test IIA. If it fails, consider nested logit or mixed logit

$\Longrightarrow$ In most applied work, start with the combined model. It nests the other two as special cases.

Thank you!

jakeanderson@g.ucla.edu