Mathematical Modeling of Behavior

Michel Bierlaire

Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne

Outline

1. Motivation
   - In this course
   - Applications
   - Importance
2. Simple example
   - Choice problem
   - Data
   - Model specification
   - Probabilities
   - Model
   - Estimation
   - Testing
   - Maximum likelihood
   - Hypothesis testing
   - Application

Motivation

Human dimension in

- engineering
- business
- marketing
- planning
- policy making

Need for

- behavioral *theories*
- quantitative *methods*
- operational mathematical *models*

Motivation

Concept of demand

- marketing
- transportation
- energy
- finance

Concept of choice

- brand, product
- mode, destination
- type, usage
- buy/sell, product

In this course...

Focus

- Individual behavior (vs. aggregate behavior)
- Theory of behavior which is
  - descriptive (how people behave) and not normative (how they should behave)
  - general: not too specific
  - operational: can be used in practice for forecasting
- Type of behavior: choice

Applications

Mode choice in the Netherlands

- Context: car vs rail in Nijmegen
- Objective: sensitivity to travel time and cost, inertia.

Mode choice in Switzerland

- Context: Car Postal
- Objective: demand forecasting

Applications

Swissmetro

- Context: new transportation technology
- Objective: demand pattern, pricing

Residential telephone services

- Context: flat rate vs. measured
- Objective: offer the most appropriate service

Airline itinerary choice

- Context: questionnaire about itineraries across the US
- Objective: help airlines and aircraft manufacturer to design a better offer

Importance

Daniel L. McFadden

- UC Berkeley 1963, MIT 1977, UC Berkeley 1991
- Laureate of *The Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel 2000*
- Owns a farm and vineyard in Napa Valley
- “Farm work clears the mind, and the vineyard is a great place to prove theorems”

Outline

Simple example

Objectives

Introduce basic components of choice modeling:

- definition of the problem
- data
- model specification
- parameter estimation
- model application

Application

Analysis of the market for smartphones

Choice problem

Choice

Consumer's choice to

- own a smartphone
- own another ("non-smart") mobile phone.

Questions

- what is the current market penetration of smartphones relative to non-smart phones?
- how will the penetration change in the future?

Data

Population

- adults
- in the US
- owning a mobile phone

Sample

- 2000 adults
- randomly selected

Questions

Is your mobile phone a smartphone

- Yes,
- No.

What is your level of educational attainment?

- No high school diploma,
- High school graduate,
- College graduate.

Data

Contingency table

Smartphone	Low ($k=1$)	Medium ($k=2$)	High ($k=3$)	
	Education			
Yes ($i=1$)	75	500	510	1085
No ($i=2$)	175	500	240	915
	250	1000	750	2000

Market penetration in the sample

- $1085/2000 = 54.3\%$
- How do we predict? We need a model.

Model specification

Variables

Dependent

- or endogenous
- what is explained
- here: choice to use a smartphone
- notation: i
- nature: discrete
- 1 = "yes"; 2= "no"

Independent

- or exogenous
- explanatory
- here: level of education
- notation: k
- nature: discrete
- 1 = "low"; 2= "medium"; 3="high"

Probabilities

Marginal probability

- frequency of smartphone ownership in the population
- $P(i = 1)$
- Inference: use the sample to obtain an estimate
- $P(i = 1) \approx \widehat{P}(i = 1) = 1085/2000 = 0.543$

Joint probability

- frequency of smartphone ownership and medium level of education
- $P(i = 1, k = 2) \approx \widehat{P}(i = 1, k = 2) = 500/2000 = 0.25$

Conditional probability

- frequency of smartphone ownership in the population of people with medium level of education
- $P(i = 1|k = 2) \approx \widehat{P}(i = 1|k = 2) = 500/1000 = 0.50$

Model

$$\begin{aligned} P(i,k) &= P(i|k)P(k) \\ &= P(k|i)P(i) \end{aligned}$$

Interpretation

- $P(i|k)$: level of education explains smartphone ownership
- $P(k|i)$: smartphone ownership explains level of education

Model

- identify stable causal relationships between the variable
- here: we select $P(i|k)$ as an acceptable behavioral model
- stability over time necessary to forecast

Model

Specification

$$
\begin{aligned}
P(i=1|k=1) &= \pi_1,\\
P(i=1|k=2) &= \pi_2,\\
P(i=1|k=3) &= \pi_3.
\end{aligned}
$$

Parameters

- π_1, π_2, π_3
- unknown
- must be estimated from data

Model estimation

$$\pi_j = P(i=1|k=j) \approx \widehat{\pi}_j = \widehat{P}(i=1|k=j) = \frac{\widehat{P}(i=1,k=j)}{\widehat{P}(k=j)}$$

Using the contingency table:

$$\begin{array}{rcccl} \widehat{\pi}_1 & = & 75/250 & = & 0.300, \\ \widehat{\pi}_2 & = & 500/1000 & = & 0.500, \\ \widehat{\pi}_3 & = & 510/750 & = & 0.680. \end{array}$$

Quality of the estimates

Informal checks

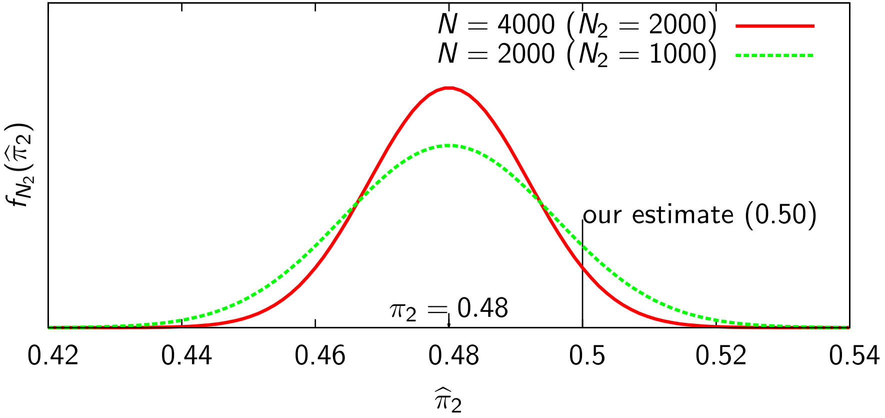
- Do these estimates make sense?
- Do they match our a priori expectations?
- Here: as years of education increases, there is a higher penetration of smartphones.

Quality of the estimates

- How is $\widehat{\pi}_j$ different from π_j?
- We have no access to π_j
- For each sample, we would obtain a different value of $\widehat{\pi}_j$
- $\widehat{\pi}_j$ is distributed.

Quality of the estimates

Distribution of π_2



Quality of the estimates

Distribution of π_2

- Smaller samples are associated with wider spread
- The larger the sample, the better the estimate
- In practice, impossible to repeat the sampling multiple times
- Distributions derived from theoretical results or simulation

Properties

- Bernoulli (0/1) random variables
- Variance: $\sigma_j^2 = \pi_j(1 - \pi_j)$
- Sample average: unbiased estimator
- Standard error of the estimator: $\sqrt{\sigma^2/N}$
- Estimated standard error:

$$\widehat{s}_{\pi_j} = \sqrt{\widehat{\pi}_j(1 - \widehat{\pi}_j)/N_j}$$

Testing

Estimates and standard errors

parameter	$\widehat{\pi}_j$	$\widehat{s}_{\pi_j}$
π_1	0.300	0.029
π_2	0.500	0.016
π_3	0.680	0.017

Maximum likelihood estimation

Likelihood function

$$\mathcal{L}^* = \prod_{n=1}^{N} P(i_n|k_n)$$

- Probability that our model reproduces exactly the observations
- For our example:

$$\mathcal{L}^* = (\pi_1)^{75}(1-\pi_1)^{175}(\pi_2)^{500}(1-\pi_2)^{500}(\pi_3)^{510}(1-\pi_3)^{240}$$

Maximum likelihood estimation

Estimates

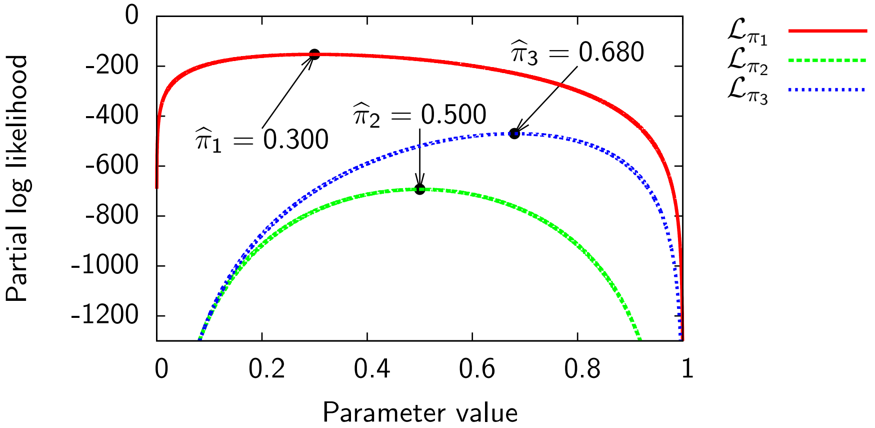
- Values of the parameters that maximize $\mathcal{L}^*$.
- In practice, the logarithm is maximized

$$\mathcal{L} = \ln \mathcal{L}^* = \sum_{n=1}^{N} \ln P(i_n|k_n).$$

Properties

- Consistency
- Asymptotic efficiency

Maximum likelihood



Hypothesis testing

Null hypothesis

- Default hypothesis
- Is accepted except if the data tells otherwise
- Example: education has no effect on smartphone ownership
- Under the null hypothesis, we have a restricted model

$$\pi = \pi_1 = \pi_2 = \pi_3.$$

- We compare the unrestricted and the restricted model

Hypothesis testing

Unrestricted model

- Log likelihood function:

$$\mathcal{L} = 75 \ln(\pi_1) + 175 \ln(1-\pi_1) + 500 \ln(\pi_2) + 500 \ln(1-\pi_2) \\ +510 \ln(\pi_3) + 240 \ln(1-\pi_3)$$

- Estimates: $\widehat{\pi}_1 = 0.300$, $\widehat{\pi}_2 = 0.500$, $\widehat{\pi}_3 = 0.680$.
- Maximum likelihood: -1316.0

Restricted model

- Log likelihood function:

$$\mathcal{L} = 1085 \ln(\pi) + 915 \ln(1-\pi).$$

- Estimate: $\widehat{\pi} = 0.543$
- Maximum likelihood: -1379.1

Hypothesis testing

Property

- If the null hypothesis is true
- the statistic

$$-2(\mathcal{L}^R - \mathcal{L}^U) = -2(-1379.0 + 1316.0) = 126.1$$

- is asymptotically distributed as χ^2 with degrees of freedom equal to the number of restrictions (2 here).

Applying the test

- the critical value of the χ^2 distribution with 2 degrees of freedom at 99% significance is $9.210 < 126.1$.
- The null hypothesis is rejected with at least 99% confidence.
- Education *does* influence smartphone ownership.

Model application

Present scenario

- Level of education: low (12.5%), medium (50%), high (37.5%)
- Penetration rate:
  $0.300 \times 12.5\% + 0.500 \times 50\% + 0.680 \times 37.5\% = 54.3\%$

Future scenario

- Level of education will change in the future
- Level of education: low (10%), medium (40%), high (50%)
- Penetration rate: $0.300 \times 10\% + 0.500 \times 40\% + 0.680 \times 50\% = 57\%$

Note

- Causal relationship does not vary over time
- Values of the explanatory variables evolve over time

Outline

1. Motivation
   - In this course
   - Applications
   - Importance
2. Simple example
   - Choice problem
   - Data
   - Model specification
   - Probabilities
   - Model
   - Estimation
   - Testing
   - Maximum likelihood
   - Hypothesis testing
   - Application