The logit model

Derivation, normalization and parameters estimation

Michel Bierlaire

Mathematical Modeling of Behavior

EPFL

Outline

The logit model

Probability for individual n to choose alternative i within the set $\mathcal{C}_n$:

$$P(i|\mathcal{C}_n) = \frac{e^{\mu_n V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu_n V_{jn}}}.$$

where

$$V_{in} = \sum_{k=1}^{K} \beta_k z_{ink},$$

where

- $\mathcal{C}_n$ is the set of alternatives for individual n,
- z_{in} are the attributes of alternative i for individual n, and
- μ_n and β_k, $k = 1, \ldots, K$ are parameters to be estimated from data.

Example: transportation mode choice in Switzerland

Three alternatives

- Car,
- public transportation (PT),
- Slow modes (SM).

Six attributes

- Travel cost (car and PT).
- Travel time (car and PT).
- Waiting time (PT).
- Distance (SM).

Example: n is Priya

Who is Priya?

- German speaking.
- Age: 44.
- Gender: female.
- Subscription: no GA.
- Socio-prof. category: manager.
- Income: high.
- Car availability: yes.

Attributes

- Car cost: $z_{\text{car},\text{Priya},1} = 0.13$ CHF.
- Time by car: $z_{\text{car},\text{Priya},2} = 1.0$ min.
- PT cost: $z_{\text{PT},\text{Priya},1} = 3.0$ CHF.
- Time by PT: $z_{\text{PT},\text{Priya},2} = 10.0$ min.
- Waiting time: $z_{\text{PT},\text{Priya},3} = 0.0$ min.
- Distance: $z_{\text{SM},\text{Priya},4} = 0.1$ km.

Example: n is Priya

Utility: functions of the attributes

$$
\begin{aligned}
V_{\text{car},\text{Priya}} &= 18 - z_{\text{car},\text{Priya},1} - 1.73\, z_{\text{car},\text{Priya},2}^{0.757} \\
&= 18 - 0.13 - 1.73 \cdot 1 \\
&= 16.2, \\
V_{\text{PT},\text{Priya}} &= -8.4 - z_{\text{PT},\text{Priya},1} - 0.48\, z_{\text{PT},\text{Priya},2}^{0.757} - 1.9\, z_{\text{PT},\text{Priya},3} \\
&= -8.4 - 3 - 0.48 \cdot 10^{0.757} - 1.9 \cdot 0 \\
&= -14.1, \\
V_{\text{SM},\text{Priya}} &= -237\, z_{\text{SM},\text{Priya},4} \\
&= -237 \cdot 0.1 \\
&= -23.7.
\end{aligned}
$$

Example: n is Priya

Logit model: probability for Priya to choose the car

$$\mathcal{C}_{\text{Priya}} = \{\text{car}, \text{PT}, \text{SM}\}, \;\; \mu_n = 0.0373.$$

$$\begin{aligned} P(\text{car}|\mathcal{C}_{\text{Priya}}) &= \frac{e^{0.0373 \; V_{\text{car},\text{Priya}}}}{\sum_{j \in \mathcal{C}_{\text{Priya}}} e^{0.0373 \; V_{j,\text{Priya}}}} \\ &= \frac{e^{0.0373 \cdot 16.2}}{e^{0.0373 \cdot 16.2} + e^{0.0373 \cdot (-14.1)} + e^{0.0373 \cdot (-23.7)}} \\ &= \frac{1.83}{2.83} \\ &= 0.646. \end{aligned}$$

Example: n is Priya

Logit model: probability for Priya to choose public transportation

$$\mathcal{C}_{\text{Priya}} = \{\text{car}, \text{PT}, \text{SM}\}, \ \mu_n = 0.0373.$$

$$\begin{aligned}
P(\text{PT}|\mathcal{C}_{\text{Priya}}) &= \frac{e^{0.0373 \ V_{\text{PT},\text{Priya}}}}{\sum_{j \in \mathcal{C}_{\text{Priya}}} e^{0.0373 \ V_{j,\text{Priya}}}} \\
&= \frac{e^{0.0373 \cdot -14.1}}{e^{0.0373 \cdot 16.2} + e^{0.0373 \cdot (-14.1)} + e^{0.0373 \cdot (-23.7)}} \\
&= \frac{0.59}{2.83} \\
&= 0.208.
\end{aligned}$$

Example: n is Priya

Logit model: probability for Priya to choose a slow mode

$$\mathcal{C}_{\text{Priya}} = \{\text{car}, \text{PT}, \text{SM}\}, \;\; \mu_n = 0.0373.$$

$$\begin{aligned}
P(\text{SM}|\mathcal{C}_{\text{Priya}}) &= \frac{e^{0.0373\; V_{\text{SM},\text{Priya}}}}{\sum_{j\in\mathcal{C}_{\text{Priya}}} e^{0.0373\; V_{j,\text{Priya}}}} \\
&= \frac{e^{0.0373\,\cdot\,-23.7}}{e^{0.0373\,\cdot\,16.2} + e^{0.0373\,\cdot\,(-14.1)} + e^{0.0373\,\cdot\,(-23.7)}} \\
&= \frac{0.412}{2.83} \\
&= 0.146.
\end{aligned}$$

Example: n is Mateo

Who is Mateo?

- French speaking.
- Age: 35.
- Gender: male.
- Subscription: GA.
- Socio-prof. category: craftman.
- Income: low.
- Car availability: no.

Attributes (same as Priya, except cost PT)

- Car cost: $z_{\text{car},\text{Mateo},1} = 0.13$ CHF.
- Time by car: $z_{\text{car},\text{Mateo},2} = 1.0$ min.
- PT cost: $z_{\text{PT},\text{Mateo},1} = 0.0$ CHF.
- Time by PT: $z_{\text{PT},\text{Mateo},2} = 10.0$ min.
- Waiting time: $z_{\text{PT},\text{Mateo},3} = 0.0$ min.
- Distance: $z_{\text{SM},\text{Mateo},4} = 0.1$ km.

Example: n is Mateo

Utility: functions of the attributes

$$
\begin{aligned}
V_{\text{car,Mateo}} &= 3.84 - z_{\text{car,Mateo,1}} - 2.85\, z_{\text{car,Mateo,2}}^{0.757} \\
&= 3.84 - 0.13 - 2.85 \cdot 1 \\
&= 0.858, \\
V_{\text{PT,Mateo}} &= 12.1 - z_{\text{PT,Mateo,1}} - 1.02\, z_{\text{PT,Mateo,2}}^{0.757} - 0.17\, z_{\text{PT,Mateo,3}} \\
&= 12.1 - 0 - 1.02 \cdot 10^{0.757} - 0.17 \cdot 0 \\
&= 6.23, \\
V_{\text{SM,Mateo}} &= -167\, z_{\text{SM,Mateo,4}} \\
&= -167 \cdot 0.1 \\
&= -16.7.
\end{aligned}
$$

Example: n is Mateo

Logit model: probability for Mateo to choose the car

$$\mathcal{C}_{\text{Mateo}} = \{\text{PT}, \text{SM}\}, \; \mu_n = 0.0725.$$

$$P(\text{car}|\mathcal{C}_{\text{Mateo}}) = 0.$$

Example: n is Mateo

Logit model: probability for Mateo to choose public transportation

$$\mathcal{C}_{\text{Mateo}} = \{\text{PT}, \text{SM}\}, \ \mu_n = 0.0725.$$

$$\begin{aligned} P(\text{PT}|\mathcal{C}_{\text{Mateo}}) &= \frac{e^{0.0725 \ V_{\text{PT},\text{Mateo}}}}{\sum_{j \in \mathcal{C}_{\text{Mateo}}} e^{0.0725 \ V_{j,\text{Mateo}}}} \\ &= \frac{e^{0.0725 \cdot 6.23}}{e^{0.0725 \cdot (6.23)} + e^{0.0725 \cdot (-16.7)}} \\ &= \frac{1.57}{1.87} \\ &= 0.841. \end{aligned}$$

Example: n is Mateo

Logit model: probability for Mateo to choose a slow mode

$$\mathcal{C}_{\text{Mateo}} = \{\text{PT}, \text{SM}\}, \;\; \mu_n = 0.0725.$$

$$\begin{aligned}
P(\text{SM}|\mathcal{C}_{\text{Mateo}}) &= \frac{e^{0.0725 \; V_{\text{SM},\text{Mateo}}}}{\sum_{j \in \mathcal{C}_{\text{Mateo}}} e^{0.0725 \; V_{j,\text{Mateo}}}} \\
&= \frac{e^{0.0725 \cdot -16.7}}{e^{0.0725 \cdot (6.23)} + e^{0.0725 \cdot (-16.7)}} \\
&= \frac{0.298}{1.87} \\
&= 0.159.
\end{aligned}$$

How does it work?

- Where does the logit model come from?
  - With two alternatives.
  - With multiple alternatives.
- How do we specify the utility functions?
  - What variables can be involved?
  - How do we come up with a fonctional form?
  - How do we derive a different model for different individuals?
- How do we estimate the parameters?

Outline

The binary logit model

Two alternatives: $\mathcal{C}_n = \{i, j\}$

$$\begin{array}{rcl} U_{in} & = & V_{in} + \varepsilon'_{in}, \\ U_{jn} & = & V_{jn} + \varepsilon'_{jn}. \end{array}$$

Main issue

- Utility is latent, not observed.
- Only the choice is observed.
- More complicated than linear regression.
- How do we know the “zero” of utility?
- How do we know the units of utility?

Binary choice model

Choice model

$$P_n(i|\{i,j\}) = \Pr(U_{in} \geq U_{jn}).$$

Invariant to shifts

$$P_n(i|\{i,j\}) = \Pr(U_{in} + \eta \geq U_{jn} + \eta), \ \forall \eta \in \mathbb{R}.$$

Invariant to scale

$$P_n(i|\{i,j\}) = \Pr(\mu U_{in} \geq \mu U_{jn}), \ \forall \mu \in \mathbb{R}, \mu > 0.$$

Binary choice model

Choice model

$$
\begin{aligned}
P_n(i|\{i,j\}) &= \Pr(U_{in} \geq U_{jn}) \\
&= \Pr(V_{in} + \varepsilon'_{in} \geq V_{jn} + \varepsilon'_{jn}) \\
&= \Pr(V_{in} - V_{jn} \geq \varepsilon'_{jn} - \varepsilon'_{in}) \\
&= \Pr\left(\varepsilon'_n \leq V_{in} - V_{jn}\right),
\end{aligned}
$$

where $\varepsilon'_n = \varepsilon'_{jn} - \varepsilon'_{in}$.

Note

- For binary choice, it would be sufficient to make assumptions about $\varepsilon'_n = \varepsilon'_{jn} - \varepsilon'_{in}$.
- But we want to generalize later on.

Error term

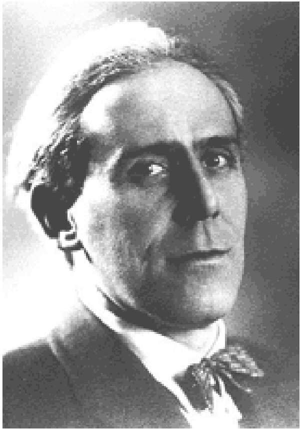
Assumptions about the random variables ε'_{in} and ε'_{jn}

ε'_{in} and ε'_{jn} are the maximum of many r.v. capturing unobserved attributes (e.g. mood, experience), measurement and specification errors.

Gumbel theorem

The maximum of many i.i.d. random variables approximately follows an Extreme Value distribution: EV(η, μ), with $\mu > 0$.

Extreme value distribution



Emil Julius Gumbel (1891–1966)

- father of extreme value theory,
- politically involved left-wing pacifist in Germany,
- strongly against right wing's campaign of organized assassination (1919),
- first German professor to be expelled from university under the pressure of the Nazis,
- 1932: he left Heidelberg to Paris, where he met Borel and Fréchet,
- 1940: he had to escape to New-York, where he continued his fight against Nazism by helping the US secret service.

The Extreme Value distribution EV(η, μ)

Probability density function (pdf)

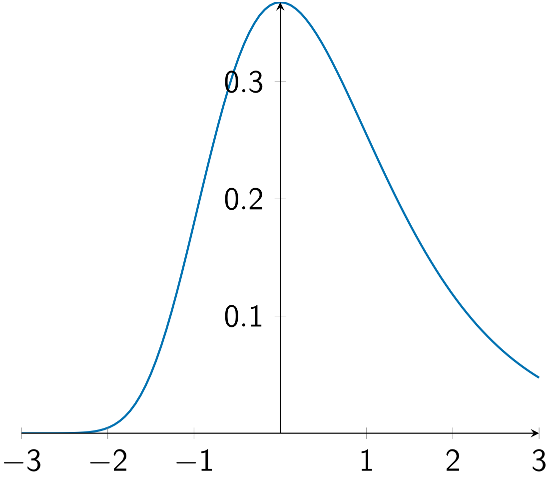
$$f(t) = \mu e^{-\mu(t-\eta)} e^{-e^{-\mu(t-\eta)}}.$$

Cumulative distribution function (CDF)

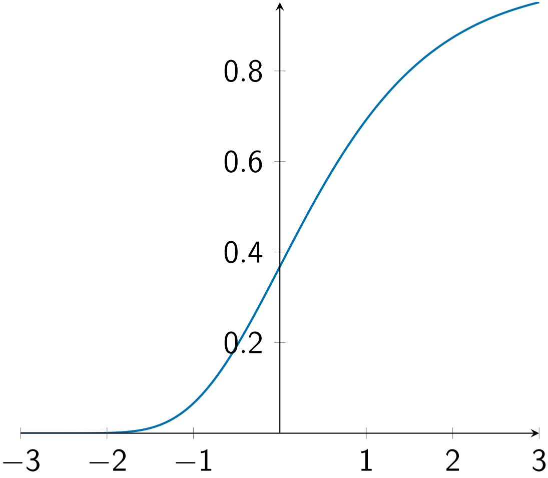
$$\begin{aligned} P(\varepsilon \leq c) = F(c) &= \int_{-\infty}^{c} f(t)dt \\ &= e^{-e^{-\mu(c-\eta)}}. \end{aligned}$$

The Extreme Value distribution

pdf EV(0,1)



CDF EV(0,1)



The Extreme Value distribution

Properties

$$\varepsilon \sim \text{EV}(\eta, \mu).$$

- Mode: η.
- Mean: $\text{E}[\varepsilon] = \eta + \frac{\gamma}{\mu}$ where γ is Euler's constant.
- Variance: $\text{Var}[\varepsilon] = \frac{\pi^2}{6\mu^2}$.

Euler's constant

$$\gamma = -\int_0^\infty e^{-x} \ln x \, dx \approx 0.5772.$$

The Extreme Value distribution

Properties

- Let $\varepsilon \sim EV(\eta, \mu)$, $\alpha > 0$ and $\beta \in \mathbb{R}$. Then

$$\alpha\varepsilon + \beta \sim EV(\alpha\eta + \beta, \mu/\alpha).$$

- In particular, if $\varepsilon \sim \text{EV}(0, 1)$, then, using $\alpha = 1/\mu$ and $\beta = \eta$,

$$\varepsilon' = \eta + \frac{1}{\mu}\varepsilon \sim \text{EV}(\eta, \mu).$$

The Extreme Value distribution

Properties

Let $\varepsilon_1 \sim EV(\eta_1, \mu)$ and $\varepsilon_2 \sim EV(\eta_2, \mu)$

$$\varepsilon = \varepsilon_1 - \varepsilon_2 \sim \text{Logistic}(\eta_1 - \eta_2, \mu),$$

that is

$$F_\varepsilon(x) = \frac{1}{1 + \exp(-\mu(x - (\eta_1 - \eta_2)))}.$$

Note: the two EV distributions must have the same scale μ.

The Extreme Value distribution

Properties

- Let $\varepsilon_1 \sim EV(\eta_1, \mu)$ and $\varepsilon_2 \sim EV(\eta_2, \mu)$ independent. Then,

$$\varepsilon = \max(\varepsilon_1, \varepsilon_2) \sim EV\left(\frac{1}{\mu}\ln(e^{\mu\eta_1} + e^{\mu\eta_2}), \mu\right).$$

- Let $\varepsilon_i \sim EV(\eta_i, \mu)$, $i = 1, \ldots, J$ independent. Then,

$$\varepsilon = \max(\varepsilon_1, \ldots, \varepsilon_J) \sim EV\left(\frac{1}{\mu}\ln\sum_{i=1}^{J} e^{\mu\eta_i}, \mu\right).$$

- The sum of two EV r.v. is not an EV r.v.

Modeling assumptions

Distributions

- ε'_{in} and ε'_{jn} are i.i.d. EV(η, μ).
- $\eta, \mu \in \mathbb{R}$, $\mu > 0$.
- i.i.d. $=$ independent and identically distributed.
- i.i.d. across both i and n.

Modeling assumptions

Change of variables: isolate the parameters

$$
\begin{aligned}
\varepsilon'_{in} &= \eta + \frac{1}{\mu}\varepsilon_{in}, \\
\varepsilon'_{jn} &= \eta + \frac{1}{\mu}\varepsilon_{jn},
\end{aligned}
$$

where ε_{in}, $\varepsilon_{jn} \sim \text{EV}(0, 1)$.

Binary logit model

Specification

If the model is specified as

$$\begin{aligned} U_{in} &= V_{in} + \eta + \frac{1}{\mu}\varepsilon_{in}, \\ U_{jn} &= V_{jn} + \eta + \frac{1}{\mu}\varepsilon_{jn}, \end{aligned}$$

we can assume w.l.o.g. that ε_{in}, $\varepsilon_{jn} \sim \text{EV}(0, 1)$.

Binary logit model

Choice model

$$
\begin{aligned}
P_n(i|\{i,j\}) &= \Pr\left(U_{in} \geq U_{jn}\right) \\
&= \Pr(\tfrac{1}{\mu}(\varepsilon_{jn} - \varepsilon_{in}) \leq V_{in} + \not{\eta} - V_{jn} - \not{\eta}), \\
&= \Pr(\varepsilon_{jn} - \varepsilon_{in} \leq \mu V_{in} - \mu V_{jn}).
\end{aligned}
$$

Property of EV

$$
\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in} \sim \text{Logistic}(0, 1).
$$

The Logistic distribution: Logistic(η,μ)

Probability density function (pdf)

$$f(t) = \frac{\mu e^{-\mu(t-\eta)}}{(1 + e^{-\mu(t-\eta)})^2}.$$

Cumulative distribution function (CDF)

$$P(c \geq \varepsilon) = F(c) = \int_{-\infty}^{c} f(t)dt = \frac{1}{1 + e^{-\mu(c-\eta)}}.$$

with $\mu > 0$.

Binary logit model

Choice model

$$P_n(i|\{i,j\}) = \Pr\left(\varepsilon_n \leq \mu V_{in} - \mu V_{jn}\right) = F_\varepsilon(\mu V_{in} - \mu V_{jn}).$$

The binary logit model

$$P_n(i|\{i,j\}) = \frac{1}{1 + e^{-\mu(V_{in} - V_{jn})}} = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{jn}}}.$$

The binary logit model

Key element of the specification

$$\mu V_{in}.$$

Comments

- η does not play any role in the model.
- The units of V_{in} must be fixed. The model must be normalized.
- Before doing it, we extend the model to more than two alternatives.

Outline

Multiple alternatives

Choice set: $\mathcal{C}_n = \{1, \ldots, J_n\}$

$$\begin{aligned} U_{1n} &= V_{1n} + \varepsilon_{1n}, \\ &\vdots \\ U_{J_n n} &= V_{J_n n} + \varepsilon_{J_n n}. \end{aligned}$$

Choice set

Universal choice set: $\mathcal{C}$

- All potential alternatives for the population.
- Alternatives relevant to the analyst.

Mode choice

- driving alone,
- sharing a ride,
- taxi,
- motorcycle,
- bicycle,
- walking,
- transit bus,
- rail rapid transit.

Choice set

Individual's choice set: $\mathcal{C}_n$

- No driver license.
- No auto available.
- Awareness of transit services.
- Transit services unreachable.
- Walking not an option for long distance.

Mode choice

- ~~driving alone~~,
- sharing a ride,
- taxi,
- motorcycle,
- bicycle,
- ~~walking~~,
- ~~transit bus~~,
- rail rapid transit.

Choice set

Choice set generation is tricky

- How to model "awareness"?
- What does "long distance" exactly mean?
- What does "unreachable" exactly mean?

We assume here deterministic rules

- Car is available if n has a driver license and a car is available in the household.
- Walking is available if trip length is shorter than 4km.

Availability conditions

$$\delta_{in} = \begin{cases} 1 & \text{if } i \in \mathcal{C}_n, \\ 0 & \text{otherwise.} \end{cases} \quad \text{or} \quad \ln \delta_{in} = \begin{cases} 0 & \text{if } i \in \mathcal{C}_n, \\ -\infty & \text{otherwise.} \end{cases}$$

Choice model

$$P_n(i|\mathcal{C}_n) = P_n(i|\delta_n, \mathcal{C}) = \Pr(U_{in} + \ln \delta_{in} \geq U_{jn} + \ln \delta_{jn}).$$

Error terms

Logit: same assumptions as for binary logit

ε_{in} are

- independent and
- identically distributed,
- extreme value EV(η, μ).

Comments

i.i.d. across *i* and *n*.

The logit model: derivation

$$P(i|\mathcal{C}_n) = \Pr(U_{in} \geq \max_{j \in \mathcal{C}_n \setminus \{i\}} U_{jn}) = \Pr(V_{in} + \varepsilon_{in} \geq \max_{j \in \mathcal{C}_n \setminus \{i\}} V_{jn} + \varepsilon_{jn}).$$

Best alternative different from i

$$U_{-in} = \max_{j \in \mathcal{C}_n \setminus \{i\}} U_{jn} = \max_{j \in \mathcal{C}_n \setminus \{i\}} (V_{jn} + \varepsilon_{jn}).$$

Binary choice model

$$P(i|\mathcal{C}_n) = \Pr(U_{in} \geq U_{-in}).$$

The logit model

Property of Extreme Value distribution

$$U_{-in} = V_{-in} + \varepsilon_{-in}$$

where

$$V_{-in} = \frac{1}{\mu} \ln \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}},$$

and

$$\varepsilon_{-in} \sim \mathrm{EV}(0, \mu).$$

The logit model

Binary logit

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_{-in}}}$$

Therefore...

$$\begin{aligned}
V_{-in} &= \tfrac{1}{\mu} \ln \textstyle\sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}, \\
e^{\mu V_{-in}} &= e^{\ln \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}} = \textstyle\sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}, \\
P(i|\mathcal{C}_n) &= \tfrac{e^{\mu V_{in}}}{e^{\mu V_{in}} + \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}} = \tfrac{e^{\mu V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}}}.
\end{aligned}$$

The logit model

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{\sum_{j\in\mathcal{C}_n} e^{\mu V_{jn}}}.$$

where

$$V_{in} = \sum_k \beta_k z_{ink},$$

where z_{in} is the vector of attributes of alternative i for individual n.

Outline

Choosing the units

Issue

- As utility is latent, the units are arbitrary.
- We need to choose the units.
- We first introduce a specification that is convenient to interpret.

Context

- Utility contains a cost/price variable (in CHF, say).
- We constrain its coefficient to be -1.
- Utility = opposite of generalized cost.
- Units: CHF.

Example

Setting $\beta_c = -1$

$$V_{in} = -\text{cost}_{in} + \beta_t \text{time}_{in} + \beta_h \text{direct}_{in}.$$

Interpretation of the coefficients

- Willingness to pay for an increase of the variable.
- β_t: transforms minutes into CHF: value of time (opposite).
- β_h: transforms the feature of direct service into CHF.

Logit model

Moneymetric utility function

$$V_{in} = -\text{cost}_{in} + \sum_k \beta_k z_{ink}.$$

Choice model

$$P_n(i|\mathcal{C}) = \frac{e^{\mu V_{in}}}{\sum_{j\in\mathcal{C}} e^{\mu V_{in}}} = \frac{e^{-\mu \text{cost}_{in} + \sum_k \mu\beta_k z_{ink}}}{\sum_{j\in\mathcal{C}} e^{-\mu \text{cost}_{jn} + \sum_k \mu\beta_k z_{jnk}}}.$$

Outline

Maximum likelihood estimation

Motivation

- The model involves unknown parameters: μ, β_k.
- Their value must be inferred from a sample of observations.
- We use maximum likelihood to estimate their value.

Example: specification table of the model

	Alternative i	Alternative j
β_c	cost of trip (CHF)	cost of trip (CHF)
β_1	car (0/1)	car (0/1)
β_2	travel time (hours)	travel time (hours)
β_3	headway if train (min.)	headway if train (min.)

Observed variables

1. An indicator variable defined as

$$y_{in} = \begin{cases} 1 & \text{if individual } n \text{ chose alternative } i, \\ 0 & \text{if individual } n \text{ chose alternative } j. \end{cases}$$

   For notational convenience, we also define $y_{jn} = 1 - y_{in}$.
2. Two vectors of explanatory variables z_{in} and z_{jn}, each containing $K = 4$ values.

Example: raw data

	Individual 1	Individual 2	Individual 3
Train cost	40.00	7.80	40.00
Car cost	5.00	8.33	3.20
Train travel time	2.50	1.75	2.67
Car travel time	1.17	2.00	2.55
Headway	60	60	30
Choice	Car	Train	Train

Example: formatted data

n	cost_{in}	car_{in}	time_{in}	headway_{in}	cost_{jn}	car_{jn}	time_{jn}	headway_{jn}
1	5	1	1.17	0	40	0	2.5	60
2	7.8	0	1.75	60	8.33	1	2	0
3	40	0	2.67	30	3.2	1	2.55	0

Chosen alternative: i.

Example: observed variables

$$y_{i1} = y_{i2} = y_{i3} = 1, y_{j1} = y_{j2} = y_{j3} = 0.$$

$$\begin{aligned}
z_{i1} &= (5 \quad 1 \quad 1.17 \quad 0)^T \\
z_{j1} &= (40 \quad 0 \quad 2.5 \quad 60)^T \\
z_{i2} &= (7.8 \quad 0 \quad 1.75 \quad 60)^T \\
z_{j2} &= (8.33 \quad 1 \quad 2 \quad 0)^T \\
z_{i3} &= (40 \quad 0 \quad 2.67 \quad 30)^T \\
z_{j3} &= (3.2 \quad 1 \quad 2.55 \quad 0)^T
\end{aligned}$$

Choice model

$$\beta = \begin{pmatrix} -1 \\ \beta_{\text{car}} \\ \beta_{\text{time}} \\ \beta_{\text{headway}} \end{pmatrix}$$

$$P_n(i; \beta, \mu) = \frac{e^{\mu \beta^T z_{in}}}{e^{\mu \beta^T z_{in}} + e^{\mu \beta^T z_{jn}}}.$$

Likelihood

Probability that the model replicates all the observations.

Example: likelihood

Individuals

- Each individual n has chosen alternative i.
- This choice is predicted by the model with probability $P_n(i; \beta, \mu)$.

Likelihood

$$\mathcal{L}^*(\beta, \mu) = P_1(i; \beta, \mu) P_2(i; \beta, \mu) P_3(i; \beta, \mu).$$

where $\beta \in \mathbb{R}^{K=4}$ and $\mu \in \mathbb{R}$.

Example: likelihood

Assume that

$$\beta = \begin{pmatrix} -1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \; \mu = 10^{-8},$$

we have

n	V_{in}	V_{jn}	$P_n(i)$	$P_n(j)$
1	-5.0	-40.00	0.5	0.5
2	-7.8	-8.33	0.5	0.5
3	-40.0	-3.20	0.5	0.5

$$\mathcal{L}^* = 0.5 \cdot 0.5 \cdot 0.5 = 0.125. \tag{1}$$

Example: likelihood

Assume that

$$\beta = \begin{pmatrix} -1 \\ -1 \\ -15 \\ -0.3 \end{pmatrix}, \; \mu = 0.1$$

we have

n	V_{in}	V_{jn}	$P_n(i)$	$P_n(j)$
1	-23.55	-95.50	0.999	0.001
2	-52.05	-39.33	0.219	0.781
3	-89.05	-42.45	0.009	0.991

$$\mathcal{L}^* = 0.999 \cdot 0.219 \cdot 0.009 = 0.00197.$$

Definitions

Likelihood

$$\mathcal{L}^*(\beta, \mu) = \prod_{n=1}^{N} P_n(i; \beta, \mu)^{y_{in}} P_n(j; \beta, \mu)^{y_{jn}},$$

where $\beta \in \mathbb{R}^K$ and $\mu \in \mathbb{R}$.

Log likelihood

$$\mathcal{L}(\beta, \mu) = \sum_{n=1}^{N} (y_{in} \ln P_n(i; \beta, \mu) + y_{jn} \ln P_n(j; \beta, \mu)).$$

Maximum likelihood estimation

Optimization problem

$$\widehat{\beta}, \widehat{\mu} = \text{argmax}_{\beta \in \mathbb{R}^K, \mu \in \mathbb{R}} \mathcal{L}(\beta, \mu) = \mathcal{L}(\beta_1, \beta_2, \ldots, \beta_K, \mu).$$

Software

`biogeme.epfl.ch`

Estimation of the parameters

Unknown parameters

$$\mu, \beta_k, k = 1, \ldots.$$

Contribution to the likelihood of observation n

$$P_n(i|\mathcal{C}) = \frac{e^{-\mu \text{cost}_{in} + \sum_k \mu \beta_k z_{ink}}}{e^{-\mu \text{cost}_{in} + \sum_k \mu \beta_k z_{ink}} + e^{-\mu \text{cost}_{jn} + \sum_k \mu \beta_k z_{jnk}}}.$$

Issue: non linearity

- Non-concave formulation.
- Algorithms may converge to local maxima.
- A concave formulation is desirable.

Estimation of the parameters

Rename the parameters

$$\beta'_c = -\mu \text{ and } \beta'_k = \mu\beta_k, \ \forall k.$$

$$P_n(i|\mathcal{C}) = \frac{e^{\beta'_c \text{cost}_{in} + \sum_k \beta'_k z_{ink}}}{e^{\beta'_c \text{cost}_{in} + \sum_k \beta'_k z_{ink}} + e^{\beta_c \text{cost}_{jn} + \sum_k \beta_k z_{ijk}}}.$$

Notes

- It is equivalent to the original specification, if μ is normalized to 1.
- Logit with this specification has a concave log-likelihood function.
- Once the parameters are estimated, the inverse transform must be applied to obtain the willingness to pay parameters

$$\beta_t = \frac{\beta'_t}{\mu} = -\frac{\beta'_t}{\beta'_c}.$$

Moneymetric specification

Unnormalized version: includes all β's and μ

$$\mu V_{in} = \mu\beta_c \text{cost}_{in} + \sum_k \mu\beta_k z_{ink}.$$

Normalization: $\beta_c = -1$

$$\mu V_{in} = -\mu \text{cost}_{in} + \sum_k \mu\beta_k z_{ink}.$$

Moneymetric specification

Normalization: $\beta_c = -1$

$$\mu V_{in} = -\mu \text{cost}_{in} + \sum_k \mu \beta_k z_{ink}.$$

Advantages

- Convenient unit.
- Easy interpretation.
- Explicit representation of μ.

Drawbacks

- Not linear in the parameters.
- More complicated to estimate.
- Possibility to be caught in local maxima.

Linear-in-parameters specification

Unnormalized version: includes all β's and μ

$$\mu V_{in} = \mu \beta_c \text{cost}_{in} + \sum_k \mu \beta_k z_{ink}.$$

Normalization: $\mu = 1$

$$\mu V_{in} = \beta_c \text{cost}_{in} + \sum_k \beta_k z_{ink}.$$

Linear-in-parameters specification

Normalization: $\mu = 1$

$$\mu V_{in} = V_{in} = \beta_c \text{cost}_{in} + \sum_k \beta_k z_{ink}.$$

Advantages

- Linear in the parameters.
- Simple to estimate.
- With logit, concave log-likelihood function.

Drawbacks

- Unitless.
- Coefficients difficult to interpret.
- No explicit representation of μ.

Normalization

Notes

- The choice of a specific normalization is arbitrary, as both lead to the exact same choice model.
- The linear-in-parameters normalization has been widely adopted in the literature, for historical reasons.
- The moneymetric normalization provides a better interpretation.
- Warning: if some parameters are assumed to be distributed (see the lecture on mixtures), the choice of the distribution is conditional on the type of normalization.

Comparison with linear regression

Linear regression

$$y_n = \sum_k \beta_k z_{nk} + \varepsilon_n$$

- $\varepsilon_n \sim N(\eta, \sigma^2)$.
- ε_n independent from x.
- y_n is observable.
- All parameters are identified.

Choice model

$$U_{in} = \sum_k \beta_k z_{ink} + \varepsilon_{in}$$

- $\varepsilon_{in} \sim EV(\eta, \mu)$.
- ε_{in} independent from x.
- U_{in} is latent, not observable.
- Location: η does not play any role.
- Units: normalization is needed.

Summary

- ε_{in} i.i.d. EV(η, μ).
- Derivation: from binary logit to multiple alternatives.
- Identification issues due to the latent nature of utility.
- Normalization: η does not play any role.
- Normalization: $\beta_c = -1$: moneymetric specification.
- Alternative normalization: $\mu = 1$.
- Estimation of the parameters: maximum likelihood.

Appendices

- Output of the estimation.
- The binary probit model.
- Gumbel's theorem.

Appendix I: Output of the estimation

Main outputs

- the parameter estimates $\widehat{\beta}$,
- the value of the log likelihood function at the parameter estimates $\mathcal{L}(\widehat{\beta})$.

Other output

- variance-covariance matrix of the estimates,
- standard errors,
- t-statistics,
- p-values,
- goodness of fit.

Variance-covariance: Cramer-Rao bound

Definition

$$-\operatorname{E}\left[\nabla^2\mathcal{L}(\beta)\right]^{-1} = \left\{-\operatorname{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta\partial\beta^T}\right]\right\}^{-1}.$$

Estimator

$$A = \operatorname{E}\left[\frac{\partial^2\mathcal{L}(\beta)}{\partial\beta_k\partial\beta_m}\right] \approx \sum_{n=1}^{N}\left[\frac{\partial^2\left(y_{in}\ln P_n(i) + y_{jn}\ln P_n(j)\right)}{\partial\beta_k\partial\beta_m}\right]_{\beta=\widehat{\beta}},$$

$$\widehat{\Sigma}_{\beta}^{\text{CR}} = -\widehat{A}^{-1}.$$

Variance-covariance: robust estimator

BHHH matrix

$$-E\left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial\beta\partial\beta^T}\right] \approx \sum_{n=1}^{N} \nabla L_n(\widehat{\beta}) \nabla L_n(\widehat{\beta})^T = \widehat{B},$$

where

$$\nabla L_n(\widehat{\beta}) = \nabla(y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)).$$

Robust or sandwich estimator

$$\widehat{\Sigma}_\beta^{\text{R}} = (-\widehat{A})^{-1}\ \widehat{B}\ (-\widehat{A})^{-1} = \widehat{\Sigma}_\beta^{\text{CR}}\ (\widehat{\Sigma}_\beta^{\text{BHHH}})^{-1}\ \widehat{\Sigma}_\beta^{\text{CR}}.$$

Variance-covariance matrix

Notes

- When the true likelihood function is maximized, these estimators are asymptotically equivalent.
- When other consistent estimators are used, different from the maximum likelihood, the robust estimator must be used.

Standard errors

Definition

$$\sigma_k = \sqrt{\widehat{\Sigma}_\beta(k,k)},$$

where $\widehat{\Sigma}_\beta(k,k)$ is the kth entry of the diagonal of the matrix $\widehat{\Sigma}_\beta$.

t statistics

Definition

$$t_k = \frac{\widehat{\beta}_k - \beta_0}{\sigma_k},$$

where β_0 is the value associated wth the null hypothesis (usually 0).

Role

Typically used to test the null hypothesis that the true value of the parameter is zero. This hypothesis can be rejected with 95% of confidence if

$$|t_k| \geq 1.96. \tag{2}$$

p values

Definition

- It is the probability to get a t statistic at least as large (in absolute value) as the one reported, under the null hypothesis that $\beta_k = 0$.
- Consider an estimate $\widehat{\beta}_k$ of the parameter β_k, and t_k its t statistic. The p value is calculated as

$$p_k = 2(1 - \Phi(t_k)),$$

where $\Phi(\cdot)$ is the cumulative density function of the univariate standard normal distribution.

Role

- Exact same role as the t statistics.
- The null hypothesis can be rejected at the confidence level p_k.

Goodness of fit

Preliminary remarks

- There are several measures of goodness of fit.
- None of them can be used in an absolute way.
- They can only be used to compare two models, estimated on the same data set, with the same dependent variable.

Goodness of fit

Log likelihood

$$\mathcal{L}(\widehat{\beta}).$$

Normalized log likelihood

$$\rho^2 = 1 - \frac{\mathcal{L}(\widehat{\beta})}{\mathcal{L}(0)}.$$

Comments on ρ^2

- It is not the square of anything. It mimics R^2 in linear regression.
- In general, value strictly between 0 (null model) and 1 (perfect fit).
- But the value is meaningless as such.

Goodness of fit: accounting for the number of parameters

Akaike Information Criterion (AIC)

$$2K - 2\mathcal{L}(\widehat{\beta}).$$

Note: the lower, the better.

Normalized AIC

$$\bar{\rho}^2 = 1 + \frac{\text{AIC}}{2\mathcal{L}(0)} = 1 - \frac{\mathcal{L}(\widehat{\beta}) - K}{\mathcal{L}(0)}.$$

Note: the higher, the better.

Goodness of fit: accounting for sample size

Bayesian Information Criterion (BIC)

$$K \ln(N) - 2\mathcal{L}(\widehat{\beta}).$$

Note: the lower, the better.

Goodness of fit: benchmark models

Benchmark model with 0 parameter

$$P_n(i) = \frac{1}{J_n}.$$

$$\mathcal{L}(0) = -\sum_{n=1}^{N} \log(J_n),$$

where N is the number of observations.

Goodness of fit: benchmark models

Benchmark model with $J - 1$ parameters

We assume that $J_n = J$, $\forall n$:

$$P_n(i) = p_i = \frac{N_i}{N}.$$

There are J parameters $p_1, \ldots, p_J$. They must sum up to one, removing one degree of freedom.

$$\mathcal{L}(c) = \sum_{i=1}^{J} N_i(\ln N_i - \ln N) = \sum_{i=1}^{J} N_i \ln N_i - N \ln N.$$

where N_i is the number of observations choosing alternative i.

Likelihood ratio test

Null hypothesis

Two models are equivalent.

Statistic

$$-2(\mathcal{L}(0) - \mathcal{L}(\widehat{\beta}))$$

is asymptotically distributed as χ^2 with K degrees of freedom.

Statistic

$$-2(\mathcal{L}(c) - \mathcal{L}(\widehat{\beta}))$$

is asymptotically distributed as χ^2 with $K-1$ degrees of freedom.

Appendix II: the probit model

Assumption: similar to linear regression

ε_{in} and ε_{jn} are the sum of many r.v. capturing unobserved attributes (e.g. mood, experience), measurement and specification errors.

Central limit theorem

The sum of many i.i.d. random variables approximately follows a normal distribution: $N(\eta, \sigma^2)$.

The normal distribution $\mathrm{N}(\eta, \sigma^2)$

Probability density function (pdf)

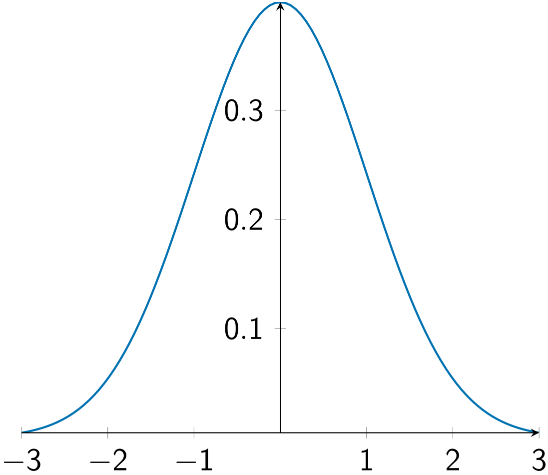
$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{t-\eta}{\sigma}\right)^2}.$$

Cumulative distribution function (CDF)

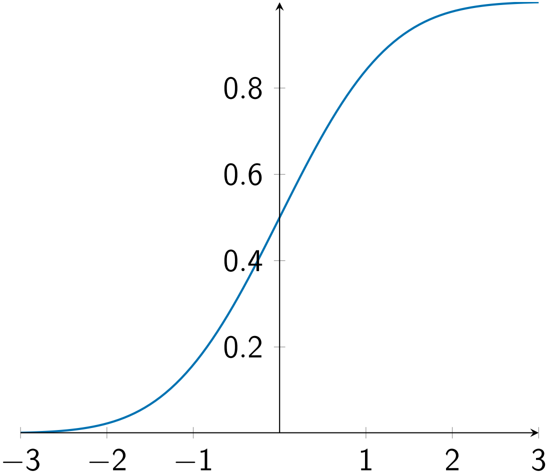
$$P(c \geq \varepsilon) = F(c) \; = \; \int_{-\infty}^{c} f(t)dt.$$

The normal distribution

pdf N(0,1)



CDF N(0,1)



The distribution

Assumptions

- ε_{in} and ε_{jn} are normally distributed, with variance σ_i^2 and σ_j^2, respectively, and covariance σ_{ij}.
- Note: identical distribution across n.
- If an alternative specific constant is in the model, their mean can be assumed to be any constant.
- $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ is also normally distributed, with variance

$$\sigma^2 = \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}.$$

The binary probit model

Choice model

$$P_n(i|\{i,j\}) = \Pr(\varepsilon_n \leq V_{in} - V_{jn}) = F_\varepsilon(V_{in} - V_{jn}).$$

The binary probit model

$$P_n(i|\{i,j\}) = \Phi\left(\frac{V_{in} - V_{jn}}{\sigma}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{(V_{in}-V_{jn})/\sigma} \exp\left(-\frac{1}{2}u^2\right) du.$$

Appendix III: Gumbel theorem

Motivation

- $X_1, \ldots, X_n$ i.i.d.
- $f_{X_i}(x) = f(x)$, $F_{X_i}(x) = F(x)$, $i = 1, \ldots, n$
- $X_n' = \max(X_1, \ldots, X_n)$.
- Applications:
  - rainfall,
  - floods,
  - earthquakes,
  - air pollution,
  - ...

Extreme value distribution

- $X'_n = \max(X_1, \ldots, X_n)$.
- $F_{X'_n} = F(x)^n$. Indeed

$$P(X'_n \leq x) = P(X_1 \leq x)P(X_2 \leq x) \ldots P(X_n \leq x).$$

- Warning: if $n \to \infty$

$$\lim_{n\to\infty} F_{X'_n}(x) = \begin{cases} 1 & \text{if } F(x) = 1, \\ 0 & \text{if } F(x) < 1. \end{cases}$$

Degenerate distribution (if you a die sufficiently many times, the maximum score will always be 6).

Extreme value distribution

- We want a limiting distribution which is non degenerate.
- Limiting distribution of some sequence of transformed “reduced” values.
- For instance $a_n X'_n + b_n$.
- a_n, b_n do not depend on x.
- CDF of limiting distribution: $G(x)$.
- Let’s identify desired properties.

Extreme value distribution

$$
\begin{array}{ccc|l}
X_1 & \dots & X_n & \max(X_1, \dots, X_n) \\
X_{n+1} & \dots & X_{2n} & \max(X_{n+1}, \dots, X_{2n}) \\
\vdots & & \vdots & \\
X_{(i-1)n+1} & \dots & X_{in} & \max(X_{(i-1)n+1}, \dots, X_{in}) \\
\vdots & & \vdots & \\
X_{(N-1)n+1} & \dots & X_{Nn} & \max(X_{(N-1)n+1}, \dots, X_{Nn})
\end{array}
$$

Two ways of seeing $\max(X_1, \dots, X_{Nn})$ when $n \to \infty$.

1. As a max of many X_i, the CDF should look like $G(a_N x + b_N)$.
2. The CDF of the max of each row is $G(x)$.
3. So the CDF of the max of all rows is $G(x)^N$.

Extreme value distribution

Stability postulate (Fréchet, 1927):

$$G(x)^N = G(a_N x + b_N).$$

We consider here the case $a_N = 1$ to obtain the so-called "type I extreme value distribution"

$$G(x)^N = G(x + b_N).$$

We have also

$$\begin{aligned} G(x)^{MN} &= G(x+b_N)^M &= G(x+b_N+b_M), \\ G(x)^{MN} &= G(x+b_{MN}). \end{aligned}$$

Extreme value distribution

Therefore

$$G(x + b_N + b_M) = G(x + b_{MN}),$$

that is

$$b_N + b_M = b_{MN},$$

so that b_N must be of the form

$$b_N = -\mu' \ln N,$$

and the stability postulate becomes

$$G(x)^N = G(x - \mu' \ln N).$$

Let's take the logarithm twice...

Extreme value distribution

$$G(x)^N = G(x - \mu' \ln N).$$

$$N \ln G(x) = \ln G(x - \mu' \ln N).$$

Warning: G is a CDF, so $G(x) \le 1$ and $\ln G(x) \le 0$, $\forall x$.

$$-N \ln G(x) = -\ln G(x - \mu' \ln N).$$

$$\ln N + \ln(-\ln G(x)) = \ln(-\ln G(x - \mu' \ln N)).$$

Define $h(x) = \ln(-\ln G(x))$ to obtain

$$\ln N + h(x) = h(x - \mu' \ln N).$$

h is affine.

Extreme value distribution

$$\begin{aligned}
\ln N + h(x) &= h(x - \mu' \ln N), \\
h(x) &= \alpha x + \beta, \\
h(0) &= \beta, \\
\ln N + \alpha x + \beta &= \alpha(x - \mu' \ln N) + \beta, \\
\alpha &= -\tfrac{1}{\mu'}.
\end{aligned}$$

Therefore

$$h(x) = h(0) - \frac{x}{\mu'}.$$

G is increasing in x (CDF), so h is decreasing in x. Therefore, $\mu' > 0$.

Extreme value distribution

$$h(x) = \ln(-\ln G(x)) = h(0) - \frac{x}{\mu'}.$$

$$-\ln G(x) = \exp\left(h(0) - \frac{x}{\mu'}\right) = \exp\left(-\frac{x - \mu' h(0)}{\mu'}\right).$$

$$G(x) = \exp\left(-\exp\left(-\frac{x - \mu' h(0)}{\mu'}\right)\right).$$

Let $\mu = 1/\mu'$ and $\eta = \mu' h(0) = \ln(-\ln G(0))/\mu$

$$G(x) = \exp\left(-\exp\left(-\mu(x - \eta)\right)\right).$$