Mathematical modeling

Discrete variables

Michel Bierlaire

Introduction to transportation systems

EPFL

Parameter estimation

Models covered so far:

- X and Y are both discrete: contingency table.
- X and Y are both continuous: linear regression.

Easy case

- Y is continuous.
- X is discrete.

More complex case

- Y is discrete.
- X is continuous.

Binary variables

Coding of qualitative variables

- Example: X is level of comfort: $\mathcal{A} = \{$very comfortable, comfortable, rather comfortable, not comfortable$\}$.
- Define binary variables.

	z_{vc}	z_{c}	z_{rc}	z_{nc}
very comfortable	1	0	0	0
comfortable	0	1	0	0
rather comfortable	0	0	1	0
not comfortable	0	0	0	1

Binary variables

Regression

- Now we can write

$$Y = \theta_1 z_{\mathrm{vc}} + \theta_2 z_{\mathrm{c}} + \theta_3 z_{\mathrm{rc}} + \theta_4 z_{\mathrm{nc}} + \sigma\varepsilon$$

- We can rely on the methodology for Y and X continuous.
- Linear regression.

Parameter estimation

Models covered so far:

- X and Y are both discrete.
- X and Y are both continuous.
- Y continuous and X discrete.

More complex case

- Y is discrete.
- X is continuous.

Discrete choice

Choice situation

- Traveler has the choice to take public transportation or not.
- Y: transportation mode. Qualitative with $\mathcal{C}$ ={public transport, others}.
- X_1: travel time. Continuous variables.
- X_2: travel cost. Continuous variables.
- We cannot write

$$Y = \theta_1 X_1 + \theta_2 X_2 + \sigma\varepsilon$$

- We need to go back to utility theory

Utility theory

Attributes

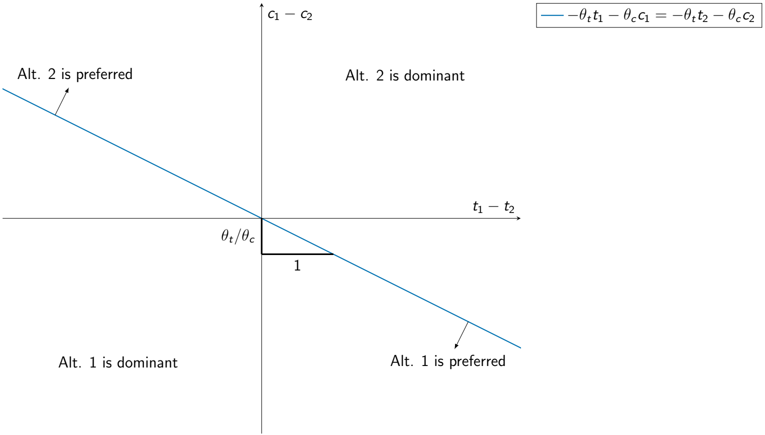
	Attributes	
Alternatives	Travel time (t)	Travel cost (c)
PT (1)	t_1	c_1
not PT (2)	t_2	c_2

Utility functions

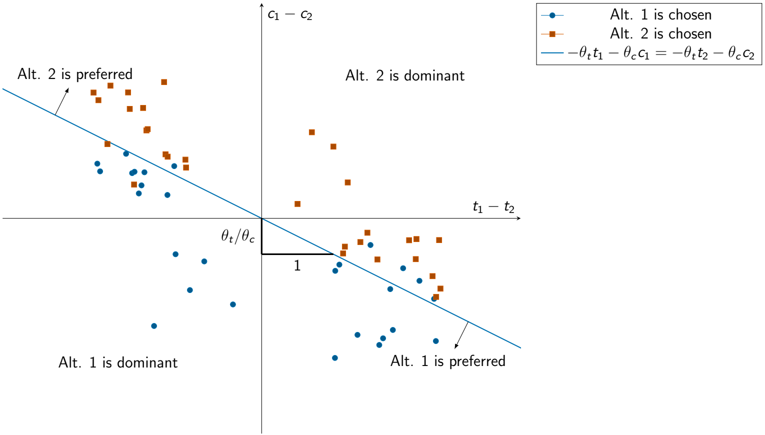
$$u_1 = -\theta_t t_1 - \theta_c c_1$$
$$u_2 = -\theta_t t_2 - \theta_c c_2$$

where $\theta_t > 0$ and $\theta_c > 0$ are parameters.

Utility theory



Utility theory



Utility theory

Random utility

- U_i is a continuous random variable.
- For example,

$$U_i = u_i + \varepsilon_i = -\theta_t t_i - \theta_c c_i + \varepsilon_i$$

- Individuals maximize their utility:

$$\Pr(Y = i) = \Pr(U_i \geq U_j)$$

- Causality:

$$Y|U|X$$

Random utility model

Latent variable

- X and Y are observed.
- U is not observed. It is latent.

Logit model

- Consider that Y corresponds to a set $\mathcal{C}$ of alternatives.
- $U_i = u_i + \varepsilon_i$ is the random utility for alternative i.
- If the ε_i are i.i.d. Extreme Value$(0, \mu)$, then

$$\Pr(Y = i) = \Pr(U_i \geq U_j, \forall j \in \mathcal{C}) = \frac{e^{\mu u_i}}{\sum_{j \in \mathcal{C}} e^{\mu u_j}}.$$

Random utility model

Shift invariance

$$\Pr(Y = i) = \Pr(U_i + K \geq U_j + K, \forall j \in \mathcal{C}), \ \forall K \in \mathbb{R}.$$

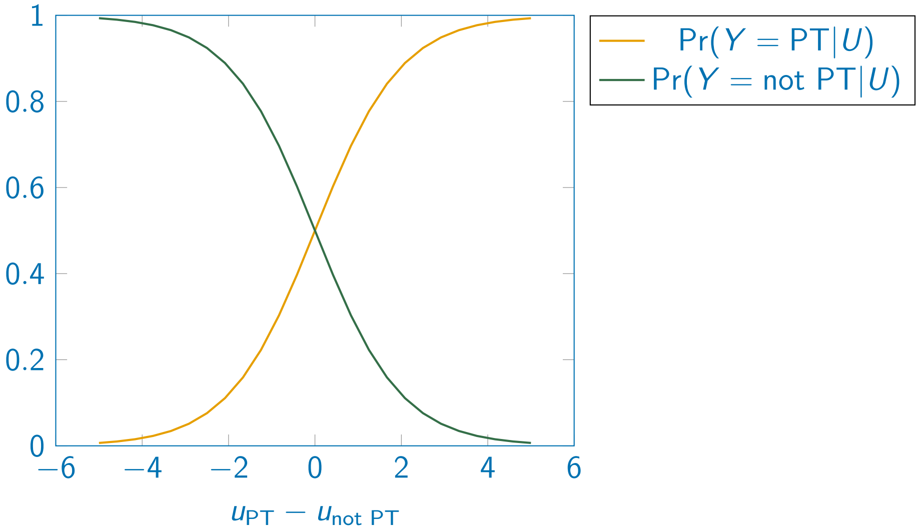
Scale invariance

$$\Pr(Y = i) = \Pr(\mu U_i \geq \mu U_j, \forall j \in \mathcal{C}), \ \forall \mu > 0.$$

Modeling implications for estimation

- Normalization of one intercept to zero.
- Normalization of the scale parameter to one.

Discrete choice



Example

Choice

between Car and PT

Data

#	Time car	Time PT	Choice	#	Time car	Time PT	Choice
1	52.9	4.4	T	11	99.1	8.4	T
2	4.1	28.5	T	12	18.5	84.0	C
3	4.1	86.9	C	13	82.0	38.0	C
4	56.2	31.6	T	14	8.6	1.6	T
5	51.8	20.2	T	15	22.5	74.1	C
6	0.2	91.2	C	16	51.4	83.8	C
7	27.6	79.7	C	17	81.0	19.2	T
8	89.9	2.2	T	18	51.0	85.0	C
9	41.5	24.5	T	19	62.2	90.1	C
10	95.0	43.5	T	20	95.1	22.2	T
				21	41.6	91.5	C

The model

Utility functions

$$
\begin{aligned}
u_{Cn} &= \theta_1 t_{Cn} \\
u_{Tn} &= \theta_1 t_{Tn} + \theta_T
\end{aligned}
$$

Parameters

Let's assume that $\theta_T = 0.5$ and $\theta_1 = -0.1$

First individual

Variables

Let's consider the first observation:

- $t_{C1} = 52.9$
- $t_{T1} = 4.4$
- Choice = PT: $y_{\text{car},1} = 0$, $y_{\text{PT},1} = 1$

Likelihood

What's the probability given by the model that this individual indeed chooses PT?

First individual

Utility functions

$$
\begin{aligned}
u_{C1} &= \theta_1 t_{C1} &= -5.29 \\
u_{T1} &= \theta_1 t_{T1} + \theta_T &= 0.06
\end{aligned}
$$

Contribution of individual 1 to the likelihood

$$
P_1(\mathrm{PT}) = \frac{e^{u_{T1}}}{e^{u_{T1}} + e^{u_{C1}}} = \frac{e^{0.06}}{e^{0.06} + e^{-5.29}} \cong 1
$$

Second individual

Variables

- $t_{C2} = 4.1$
- $t_{T2} = 28.5$
- Choice = PT: $y_{\text{car},2} = 0$, $y_{\text{PT},2} = 1$

Likelihood

What's the probability given by the model that this individual indeed chooses PT?

Second individual

Utility functions

$$
\begin{aligned}
u_{C2} &= \theta_1 t_{C2} &= -0.41 \\
u_{T2} &= \theta_1 t_{T2} + \theta_T &= -2.35
\end{aligned}
$$

Contribution of individual 2 to the likelihood

$$
P_2(\text{PT}) = \frac{e^{u_{T2}}}{e^{u_{T2}} + e^{u_{C2}}} = \frac{e^{-2.35}}{e^{-2.35} + e^{-0.41}} \cong 0.13
$$

Likelihood

Two observations

The probability that the model reproduces both observations is

$$P_1(\text{PT})P_2(\text{PT}) = 0.13$$

All observations

The probability that the model reproduces all observations is

$$P_1(\text{PT})P_2(\text{PT})\ldots P_{21}(\text{car}) = 4.62\ 10^{-4}$$

Likelihood

Likelihood of the sample

$$\mathcal{L}^* = \prod_n \left(P_n(\text{car})^{y_{\text{car},n}} P_n(\text{PT})^{y_{\text{PT},n}} \right)$$

where $y_{j,n}$ is 1 if individual n has chosen alternative j, 0 otherwise

Log likelihood of the sample

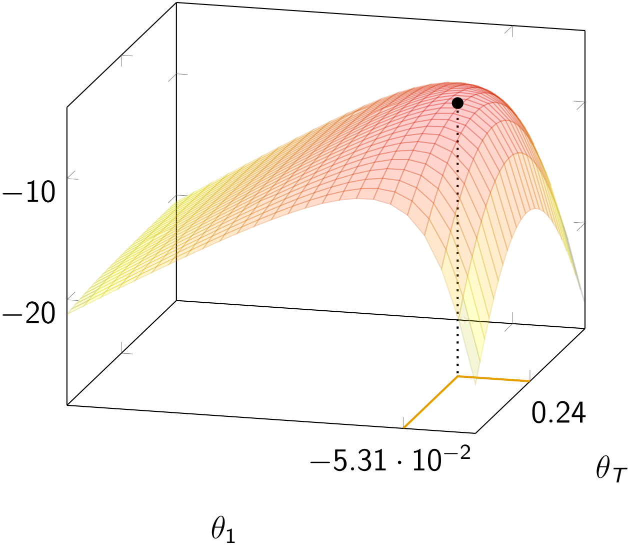
$$\mathcal{L} = \log \mathcal{L}^* = \sum_n \left(y_{\text{car},n} \log P_n(\text{car}) + y_{\text{PT},n} \log P_n(\text{PT}) \right)$$

Likelihood

Likelihood as a function of the parameters

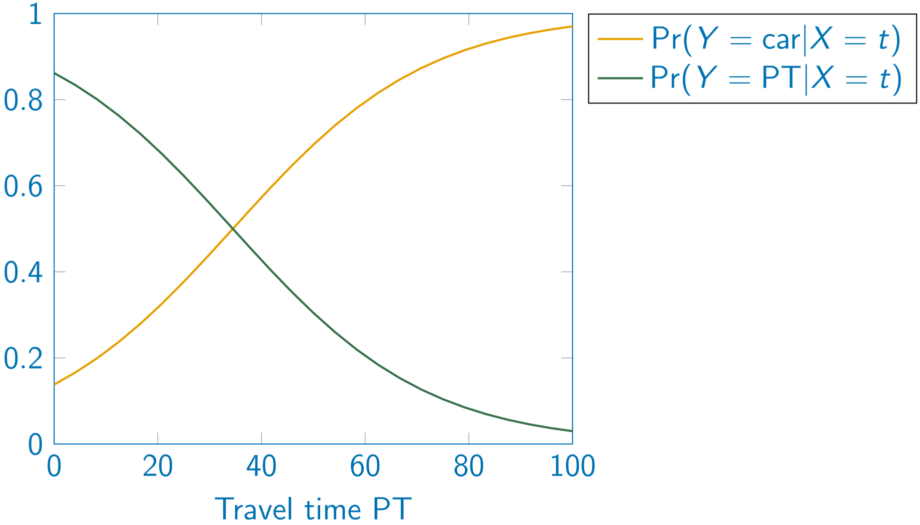
θ_T	θ_1	$\mathcal{L}^*$
0	0	$4.57\ 10^{-07}$
0	-1	$1.97\ 10^{-30}$
0	-0.1	$4.1\ 10^{-04}$
0.5	-0.1	$4.62\ 10^{-04}$

Log likelihood function



Estimated choice model

Assume travel time by car = 30 minutes



Back to the contingency table

	Work	Leisure	Others
PT	172	191	150
Not PT	345	648	494

Use binary variables

$$u_{\text{PT}} = \theta_1 z_{\text{work}} + \theta_2 z_{\text{leis}} + \theta_3 z_{\text{others}}$$

$$u_{\text{not PT}} = 0$$

Logit model

$$\Pr(\text{PT}) = \frac{e^{u_{\text{PT}}}}{e^{u_{\text{PT}}} + e^{u_{\text{not PT}}}}$$

Back to the contingency table

Maximum likelihood estimation

	Work	Leisure	Others
θ_i^*	-0.696	-1.22	-1.19
u_{PT}	-0.696	-1.22	-1.19
Pr(PT)	0.333	0.228	0.233

Conclusion

- Model equivalent to the simple model.
- We can always use logit.

Summary

Dependent variable

- Y continuous: linear regression

$$Y|(X = x_n) = \sum_{k=1}^{K-1} \theta_k x_{nk} + \theta_0 + \sigma\varepsilon$$

- Y discrete: random utility model (logit)

$$\Pr(Y = i|X = x_n) = \frac{e^{u_i(x_n)}}{\sum_{j\in\mathcal{C}} e^{u_j(x_n)}}$$

where

$$u_i(x_n) = \sum_{k=1}^{K-1} \theta_k x_{ink} + \theta_0$$

Summary

Independent variable

When discrete, can be modeled as a set of binary variables.

Estimation of the parameters

Maximum likelihood