

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

This document gathers the material used in the course, that is the slides of the videos as well as the text files. It makes it easier to search the material of the course and to find some concepts using the search functionality of your pdf reader.

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 1

Introduction to behavior modeling

Motivation

Michel Bierlaire

Introduction to choice models



Motivation

Motivation

Human dimension in

- ▶ engineering
- ▶ business
- ▶ marketing
- ▶ planning
- ▶ policy making

Motivation

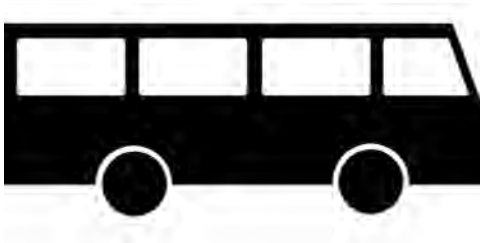
Concept of demand

Willingness and ability to purchase a commodity or service [Merriam-Webster]

Applications

Transportation

- ▶ Choice of destination
- ▶ Choice of transportation mode
- ▶ Choice of itinerary
- ▶ Choice of vehicle



Applications



Marketing

- ▶ Choice of packaging
- ▶ Choice of store
- ▶ Choice of product
- ▶ Choice of brand

Applications

Health

- ▶ Choice of treatment
- ▶ Choice of doctor
- ▶ Choice of training for doctors



Applications



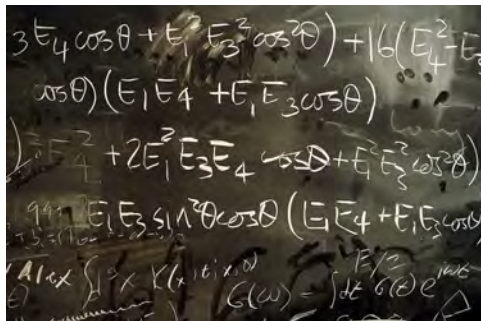
Energy

- ▶ Choice of appliances
- ▶ Choice of energy savings measures
- ▶ Choice of heating equipment

Motivation

Need for

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



Handwritten mathematical equations on a chalkboard, including:

$$3E_4 \cos \theta + E_1 E_3 \cos^2 \theta) + 16(E_4^2 - E_3^2 \cos^2 \theta)(E_1 E_4 + E_1 E_3 \cos \theta)$$
$$E_4^2 + 2E_1^2 E_3 E_4 \cos \theta + E_1^2 E_3^2 \cos^2 \theta)$$
$$94 E_1 E_3 \sin^2 \theta \cos \theta (E_1 E_4 + E_1 E_3 \cos \theta)$$
$$A(x) \int_0^x K(x,t) f(t) dt$$
$$G(\omega) = \int dt G(t) e^{i\omega t}$$

In this course...

Focus

- ▶ Individual / disaggregate 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**

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”

Introduction to behavior modeling

Simple example

Michel Bierlaire

Introduction to choice models



Simple example: introduction

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 electric cars

Choice problem

Choice

Consumer's choice to

- ▶ own an electric car
- ▶ own a car with combustion engine

Research questions

- ▶ what is the current market penetration of electric cars relative to combustion engine cars?
- ▶ how will the penetration change in the future?

Data

Population

- ▶ adults aged 20 and above
- ▶ in Switzerland
- ▶ owning a car

Sample

- ▶ 2500 individuals
- ▶ randomly selected

Questions

Is your car electric?

- ▶ Yes,
- ▶ No.

What is your age range?

- ▶ 20–39
- ▶ 40–64
- ▶ 65+

Data

Contingency table

	Age		
	20–39	40–64	65+
Electric	65	55	5
Not electric	835	1045	495

Data

This slide is not shown. I will write by hand its content on the previous slide.

Contingency table

	Age			Total
	20–39	40–64	65+	
Electric	65	55	5	125
Not electric	835	1045	495	2375
	900	1100	500	2500

Market penetration

- ▶ In the sample
 $125/2500 = 5\%$
- ▶ Currently in the population: by inference: 5%
- ▶ How do we predict?
We need a model.

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Definitions.

A mathematical model involves *variables*. There are two different types of variables. The first is the *dependent* or *endogenous* variable. It is what we are explaining. In the context of discrete choice, the dependent variable is the choice. Actually, the term *discrete choice* emphasizes that the dependent variable (that is, the choice) is discrete, and not continuous like in linear regression.

The other type of variables are called *independent*, *exogenous*, or *explanatory* variables. A model typically involves several independent variables. They can be either discrete or continuous.

The models developed in this course are *probabilistic*, in the sense that they associate a probability with different values of the variables.

For example, consider a dependent variable i (the choice) and an independent variable k . The probability that i is equal to ℓ and k is equal to j is called the *joint probability* and denoted

$$\Pr(i = \ell, k = j). \quad (1)$$

The probability for a single variable to be equal to a given value is called the *marginal probability*. In the above example, the marginal probability for the variable i is denoted $\Pr(i = \ell)$ and is derived from the joint probabilities by enumerating all the possible values of the other variable(s):

$$\Pr(i = \ell) = \sum_j \Pr(i = \ell, k = j). \quad (2)$$

If the value of one of the variables is known, the probability associated with the other variable is called the *conditional probability*. For instance, the

probability that $i = \ell$, conditional to the fact that $k = j$ is denoted

$$\Pr(i = \ell | k = j). \quad (3)$$

The joint probability can be decomposed into the product of a conditional probability and a marginal probability:

$$\begin{aligned} \Pr(i = \ell, k = j) &= \Pr(i = \ell | k = j) \Pr(k = j) \\ &= \Pr(k = j | i = \ell) \Pr(i = \ell). \end{aligned} \quad (4)$$

Using (4) into (2), we obtain the *law of total probability*:

$$\Pr(i = \ell) = \sum_j \Pr(i = \ell | k = j) \Pr(k = j). \quad (5)$$

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Practice quiz

Consider the simple example about electric cars' ownership discussed in this section. The associated contingency table is the following:

	Age			Total
	20–39	40–64	65+	
Electric	65	55	5	125
Not electric	835	1045	495	2375
	900	1100	500	2500

Perform the following tasks:

1. calculate all possible joint probabilities,
2. calculate all possible marginal probabilities in two different ways:
 - (a) from the joint probabilities calculated in 1,
 - (b) directly from the contingency table,
3. calculate all possible conditional probabilities.

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Solution of the practice quiz

1. We denote by i the choice, which in this case refers to electric car or not, and by k the age category. In order to calculate the joint probability $P(i = \ell, k = j)$, we just need to divide the cell of the contingency table corresponding to $i = \ell$ and $k = j$ by the total size of the sample. For instance, the joint probability of $i = \text{electric}$ and $k = 20\text{--}39$ is calculated as

$$P(i = \text{electric}, k = 20\text{--}39) = \frac{65}{2500} = 2.6\%.$$

The remaining joint probabilities are calculated in an analogous way and included in the following table:

	Age		
	20–39	40–64	65+
Electric	2.6 %	2.2%	0.2%
Not electric	33.4%	41.8%	19.8%

2. The marginal probability for a single variable to be equal to a given value can be derived in two different ways. For example, for $i = \text{electric}$, it can be calculated

(a) by adding the joint probabilities of $i = \text{electric}$ and all possible

age categories:

$$\begin{aligned}
 P(i = \text{electric}) &= P(i = \text{electric}, k = 20-39) + \\
 &\quad P(i = \text{electric}, k = 40-64) + \\
 &\quad P(i = \text{electric}, k = 65+) \\
 &= \frac{65}{2500} + \frac{55}{2500} + \frac{5}{2500} = \frac{125}{2500} = 5\%,
 \end{aligned}$$

- (b) directly from the contingency table by dividing the row total of $i = \text{electric}$ by the total size of the sample:

$$P(i = \text{electric}) = \frac{125}{2500} = 5\%.$$

Analogously, $P(i = \text{not electric}) = \frac{2375}{2500} = 95\%$. The marginal probabilities associated with the different age categories are calculated in a similar way:

- $P(k = 20-39) = 36\%$,
- $P(k = 40-64) = 44\%$,
- $P(k = 65+) = 20\%$.

3. The probability that $i = \ell$ conditional to the fact that $k = j$ can be obtained from the corresponding joint probability and the marginal probability of $k = j$:

$$P(i = \ell | k = j) = \frac{P(i = \ell, k = j)}{P(k = j)}.$$

For instance, if $i = \text{electric}$ and $k = 20 - 39$:

$$\begin{aligned}
 P(i = \text{electric} | k = 20-39) &= \frac{P(i = \text{electric}, k = 20-39)}{P(k = 20-39)} = \frac{\frac{65}{2500}}{\frac{900}{2500}} \\
 &= \frac{65}{900} = 7.2\%.
 \end{aligned}$$

The remaining conditional probabilities of the form $P(i = \ell | k = j)$ are included in the following table:

	Age		
	20–39	40–64	65+
Electric	7.2 %	5.0%	1.0%
Not electric	92.8%	95.0%	99.0%

Analogously, the conditional probabilities of the form $P(k = j|i = \ell)$ are computed as

$$P(k = j|i = \ell) = \frac{P(i = \ell, k = j)}{P(i = \ell)},$$

and included in the following table:

	Choice	
	Electric	Not electric
20–39	52.0 %	35.2%
40–64	44.0 %	44.0%
65+	4.0 %	20.8%

Introduction to behavior modeling

Simple example

Michel Bierlaire

Introduction to choice models



Simple example: developing a model

Model

Variables

- ▶ i : status of electric car ownership (yes or no)
- ▶ k : age category (20–39, 40–64 or 65+)

Model

Decomposition

$$P(i, k) = P(i|k)P(k) = P(k|i)P(i)$$

Interpretation

- ▶ $P(i|k)$: age explains electric car ownership
- ▶ $P(k|i)$: electric car ownership explains age

Model



Model

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

Model

Specification

$$P(i = \text{yes} \mid k = 20\text{--}39) = \pi_1,$$

$$P(i = \text{yes} \mid k = 40\text{--}64) = \pi_2,$$

$$P(i = \text{yes} \mid k = 65+) = \pi_3.$$

Parameters

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

Model estimation

$$\pi_j = P(i = \text{yes} \mid k = j)$$

Model estimation

Slide not shown. Write by hand on the previous slide

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

Exercise

Calculate the estimates of the parameters π_1 , π_2 and π_3 from the contingency table using the above formula.

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Practice quiz

Consider the simple example about electric cars' ownership discussed in this section. The associated contingency table is the following:

	Age			Total
	20–39	40–64	65+	
Electric	65	55	5	125
Not electric	835	1045	495	2375
	900	1100	500	2500

The model that is considered acceptable from a behavioral point of view involves two variables:

- i , defined as the status of electric car ownership (yes or no), and,
- k , defined as the age category (20–39, 40–64, 65+).

The model is defined as

$$P(i = \text{yes} | k = 20\text{--}39) = \pi_1,$$

$$P(i = \text{yes} | k = 40\text{--}64) = \pi_2,$$

$$P(i = \text{yes} | k = 65+) = \pi_3.$$

1. What is the dependent variable?
2. What is the independent variable?
3. Calculate the estimates of the parameters π_1 , π_2 and π_3 using the contingency table.

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Solution of the practice quiz

1. The behavioral model is $P(i|k)$, where i is defined as the status of electric car ownership (yes or no) and k as the age category (20–39, 40–64, 65+). The status of electric car ownership i is the dependent, or endogenous, variable.
2. The age, k , is the independent, exogenous, or explanatory variable.
3. The parameters specifying the model are the following:

$$P(i = \text{yes}|k = 20\text{--}39) = \pi_1,$$

$$P(i = \text{yes}|k = 40\text{--}64) = \pi_2,$$

$$P(i = \text{yes}|k = 65+) = \pi_3.$$

Since we do not have access to the full population, we infer the value of the parameters from the sample as follows (the required probabilities are calculated in the previous practice quiz):

$$\pi_1 \approx \hat{\pi}_1 = \hat{P}(i = \text{yes}|k = 20\text{--}39) = \frac{\hat{P}(i = \text{yes}, k = 20\text{--}39)}{\hat{P}(k = 20\text{--}39)} = \frac{65}{900} = 0.0722,$$

$$\pi_2 \approx \hat{\pi}_2 = \hat{P}(i = \text{yes}|k = 40\text{--}64) = \frac{\hat{P}(i = \text{yes}, k = 40\text{--}64)}{\hat{P}(k = 40\text{--}64)} = \frac{55}{1100} = 0.0500,$$

$$\pi_3 \approx \hat{\pi}_3 = \hat{P}(i = \text{yes}|k = 65+) = \frac{\hat{P}(i = \text{yes}, k = 65+)}{\hat{P}(k = 65+)} = \frac{5}{500} = 0.0100.$$

Introduction to behavior modeling

Simple example

Michel Bierlaire

Introduction to choice models



Simple example: quality of the estimates

Parameters estimates

$$\begin{aligned}\hat{\pi}_1 &= 65/900 = 0.0722, \\ \hat{\pi}_2 &= 55/1100 = 0.0500, \\ \hat{\pi}_3 &= 5/500 = 0.0100.\end{aligned}$$

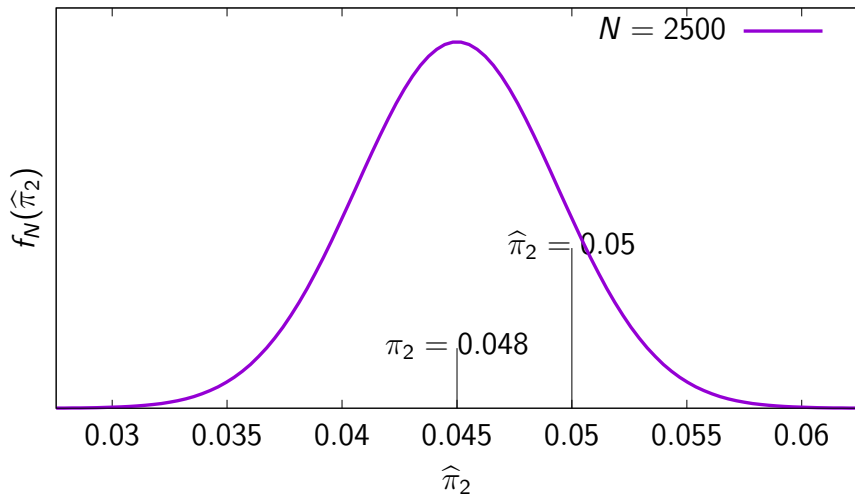
Informal checks

- ▶ Do these estimates make sense?
- ▶ Do they match our a priori expectations?
- ▶ Here: as age increases, the market share of electric cars decreases.

Quality of the estimates

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

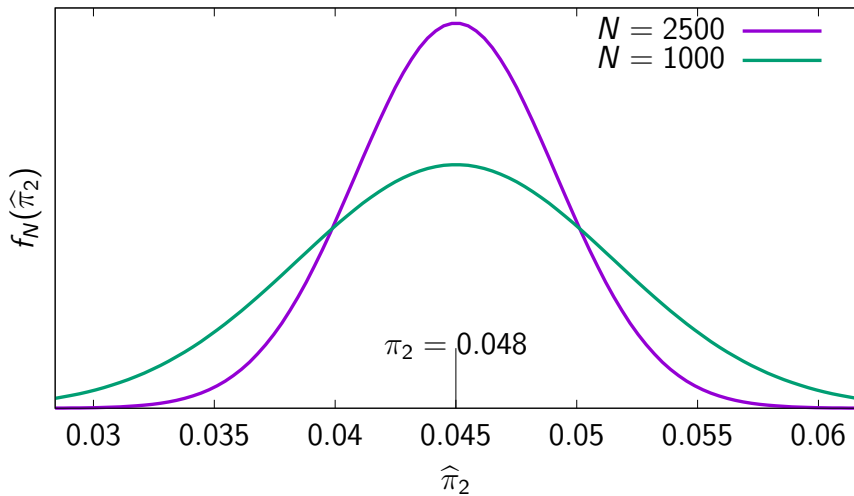
Distribution of $\hat{\pi}_2$



Distribution of $\hat{\pi}_2$

- ▶ 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

Distribution of $\hat{\pi}_2$



Statistical 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:

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

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Practice quiz

Consider the simple example about electric cars' ownership discussed in this section. The associated contingency table is the following:

	Age			Total
	20–39	40–64	65+	
Electric	65	55	5	125
Not electric	835	1045	495	2375
	900	1100	500	2500

Estimate the standard errors of the estimates $\hat{\pi}_1 = 0.0722$, $\hat{\pi}_2 = 0.0500$ and $\hat{\pi}_3 = 0.0100$ calculated in the previous practice quiz.

Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Solution of the practice quiz

The standard errors of the estimates $\hat{\pi}_j$ are calculated as follows:

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

where N_j is the number of observations used for the estimation. In this case, N_j is the number of individuals in age category j :

$$\hat{s}_{\pi_1} = \sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/N_1} = \sqrt{0.0722(1 - 0.0722)/900} = 0.0086,$$

$$\hat{s}_{\pi_2} = \sqrt{\hat{\pi}_2(1 - \hat{\pi}_2)/N_2} = \sqrt{0.0500(1 - 0.0500)/1100} = 0.0066,$$

$$\hat{s}_{\pi_3} = \sqrt{\hat{\pi}_3(1 - \hat{\pi}_3)/N_3} = \sqrt{0.0100(1 - 0.0100)/500} = 0.0044.$$

Introduction to behavior modeling

Simple example

Michel Bierlaire

Introduction to choice models



Simple example: maximum likelihood estimation

Maximum likelihood estimation

Likelihood

Probability that the model correctly predicts all the observations

Likelihood function

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

For our example

$$\mathcal{L}^* = (\pi_1)^{65} (1 - \pi_1)^{835} (\pi_2)^{55} (1 - \pi_2)^{1045} (\pi_3)^5 (1 - \pi_3)^{495}$$

Maximum likelihood estimation

Estimates

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

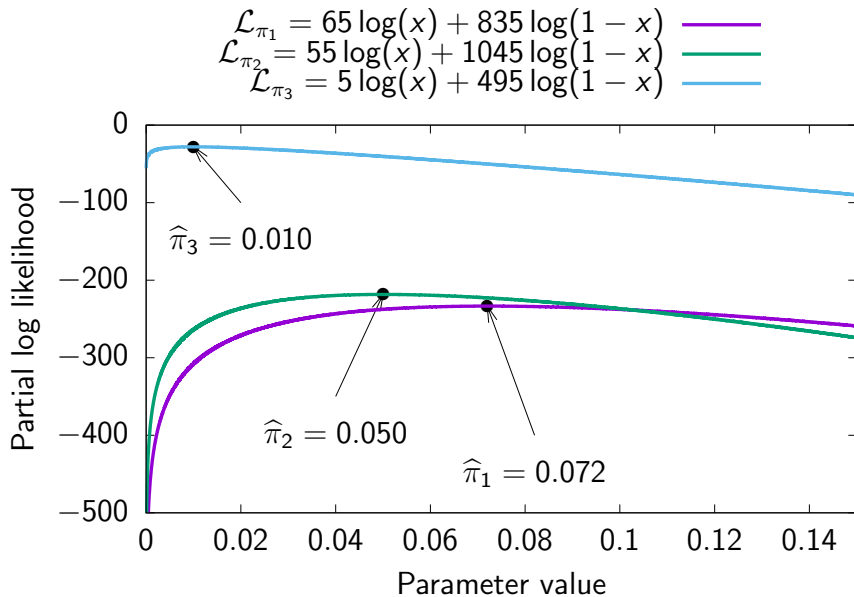
$$\mathcal{L} = \ln \mathcal{L}^* = \sum_{n=1}^N \ln P(i_n | k_n).$$

As $0 \leq \mathcal{L}^* \leq 1$, we have $\mathcal{L} \leq 0$.

Properties

- ▶ Consistency
- ▶ Asymptotic efficiency

Maximum likelihood



Introduction to behavior modeling – 1.2

Simple example

Michel Bierlaire

Hypothesis testing.

The main hypothesis that motivates the model that we are developing in this simple example is that the age is explaining the ownership of an electric car. We need to test this hypothesis against the data.

Assume that our hypothesis is wrong, that age does not explain the ownership of electric car. The terminology of hypothesis testing refers to this as the *null hypothesis*, denoted by H_0 . In that case, the true value of the parameters of our model must be equal:

$$\pi_1 = \pi_2 = \pi_3. \tag{1}$$

Indeed, the proportion of people in the population who own an electric car must be the same for each age category. Now, for a specific sample, there is no guarantee for these values to be equal. Indeed, we may have selected electric car owners among young people, just by chance. This is the key question that we are analyzing here: is the fact that the estimated values of the parameters are different due to a structural reason (electric car ownership indeed varies with age in the population) or purely due to random variations in the sampling procedure?

The null hypothesis is a restricted version of our model of interest, where the restriction is given by (1). Remember that we have used the maximum likelihood procedure to obtain the estimates of the parameters. We have solved the problem:

$$\begin{aligned} \max_{\pi_1, \pi_2, \pi_3} \mathcal{L} &= 65 \log(\pi_1) + 835 \log(1 - \pi_1) \\ &+ 55 \log(\pi_2) + 1045 \log(1 - \pi_2) \\ &+ 5 \log(\pi_3) + 495 \log(1 - \pi_3). \end{aligned} \tag{2}$$

In this context, here is a fairly simple test that we can apply to test our null hypothesis. We estimate two different models: a restricted model and an unrestricted model. Conceptually, if imposing the restriction does *not* lead to a large loss of fit (as measured by a decrease in log likelihood), then we do not reject the null hypothesis.

The unrestricted model is obtained by solving (2), where we found $\hat{\pi}_1 = 0.0722$, $\hat{\pi}_2 = 0.0500$, and $\hat{\pi}_3 = 0.0100$ and the log likelihood for this model is -479.782 , as you can verify by substituting the estimated values of the parameters into equation (2). Now we need to estimate the restricted model.

The restricted model has only a single parameter π . The likelihood of the restricted model is

$$\mathcal{L}^* = (\pi)^{125}(1 - \pi)^{2375}, \quad (3)$$

and the log likelihood is

$$\mathcal{L} = 125 \log(\pi) + 2375 \log(1 - \pi). \quad (4)$$

The maximum of this function is attained at $\hat{\pi} = 0.050$. Note that it is the same value as the market share of the electric car. (Can you explain why?) We obtain the maximum value of the restricted log likelihood function by plugging this estimate back into the log likelihood function: $\mathcal{L} = 125 \log(0.050) + 2375 \log(1 - 0.050) = -496.288$. So, clearly there has been a loss of fit from $\mathcal{L}^U = -479.782$ (the log likelihood of the unrestricted model) to $\mathcal{L}^R = -496.288$ (the log likelihood of the restricted model). The loss of fit is expected as we restrict parameters, but the question is whether this loss is statistically significant. For this we need a test statistic.

It can be shown (See Theil, 1971, p. 396 for a derivation) that, under the null hypothesis, the test statistic

$$-2(\mathcal{L}^R - \mathcal{L}^U) \quad (5)$$

is asymptotically distributed as χ^2 with degrees of freedom equal to the number of restrictions (in our case, 2). If this statistic is “large” in the statistical sense, we reject the null hypothesis that the restrictions are true. In our case, the value of the statistic is $-2(-496.288.0 + 479.782.0) = 33.01$. We need to use the χ^2 distribution to determine whether this is large enough to reject the null hypothesis. Using a level of significance of 1% (that is, the probability of rejecting the null hypothesis when it is true), the critical value of the χ^2 distribution with 2 degrees of freedom is 9.210 (see Table 1). As our

test statistic is well above this critical value, we reject the null hypothesis with at least 99% confidence and conclude that age *does* influence the ownership of electric cars.

K	90%	95%	99%	K	90%	95%	99%
1	2.706	3.841	6.635	21	29.615	32.671	38.932
2	4.605	5.991	9.210	22	30.813	33.924	40.289
3	6.251	7.815	11.345	23	32.007	35.172	41.638
4	7.779	9.488	13.277	24	33.196	36.415	42.980
5	9.236	11.070	15.086	25	34.382	37.652	44.314
6	10.645	12.592	16.812	26	35.563	38.885	45.642
7	12.017	14.067	18.475	27	36.741	40.113	46.963
8	13.362	15.507	20.090	28	37.916	41.337	48.278
9	14.684	16.919	21.666	29	39.087	42.557	49.588
10	15.987	18.307	23.209	30	40.256	43.773	50.892
11	17.275	19.675	24.725	31	41.422	44.985	52.191
12	18.549	21.026	26.217	32	42.585	46.194	53.486
13	19.812	22.362	27.688	33	43.745	47.400	54.776
14	21.064	23.685	29.141	34	44.903	48.602	56.061
15	22.307	24.996	30.578	35	46.059	49.802	57.342
16	23.542	26.296	32.000	36	47.212	50.998	58.619
17	24.769	27.587	33.409	37	48.363	52.192	59.893
18	25.989	28.869	34.805	38	49.513	53.384	61.162
19	27.204	30.144	36.191	39	50.660	54.572	62.428
20	28.412	31.410	37.566	40	51.805	55.758	63.691

Table 1: 90%, 95% and 99% of the χ^2 distribution with K degrees of freedom

References

Theil, H. (1971). *Principles of econometrics*, John Wiley and Sons.

Introduction to behavior modeling

Simple example

Michel Bierlaire

Introduction to choice models



Simple example: forecasting

Present situation

Age group	20–39	40–64	65+
Current share	36 %	44%	20%
Market penetration	7.2%	5%	1%

Total market penetration = $36\% \cdot 7.2\% + 44\% \cdot 5\% + 20\% \cdot 1\% = 5\%$

Future scenario

Age structure will change in the future

Age group	20–39	40–64	65+
Current share	36 %	44%	20%
Future share	25 %	50%	25%
Market penetration	7.2%	5%	1%

Future total market penetration = $25\% \cdot 7.2\% + 50\% \cdot 5\% + 25\% \cdot 1\% = 4.55\%$

Forecasting

- ▶ Causal relationship does not vary over time.
- ▶ Characterized by the model specification, including the values of its parameters.
- ▶ Values of the explanatory variables evolve over time.

Introduction to behavior modeling – 1.3

Summary

Michel Bierlaire

Using a simple example, we went through all the stages of modeling:

- definition of the problem,
- data collection,
- model specification,
- model estimation,
- hypothesis testing,
- model application.

The rest of the course elaborates on these concepts in the general case.

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 2

Theoretical foundations

Ingredients of choice theory

Michel Bierlaire

Introduction to choice models



Ingredients of choice theory

Choice theory

Theory of behavior that is

- ▶ **descriptive**: how people behave and not how they should
- ▶ **abstract**: not too specific
- ▶ **operational**: can be used in practice for forecasting

Building the theory

Define

1. who (or what) is the decision maker,
2. what are the characteristics of the decision maker,
3. what are the alternatives available for the choice,
4. what are the attributes of the alternatives, and
5. what is the decision rule that the decision maker uses to make a choice.

Decision maker

Individual

- ▶ a person
- ▶ a group of persons (internal interactions are ignored)
 - ▶ household, family
 - ▶ firm
 - ▶ government agency
- ▶ notation: n

Characteristics of the decision maker

Disaggregate models

Individuals

- ▶ face different choice situations
- ▶ have different tastes

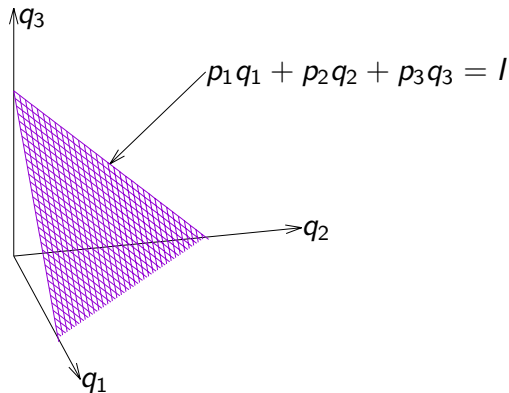
Characteristics

- ▶ income
- ▶ sex
- ▶ age
- ▶ level of education
- ▶ household/firm size
- ▶ etc.

Alternatives: continuous choice set

Commodity bundle

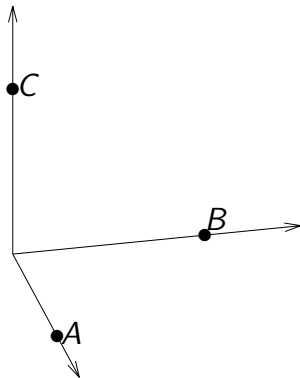
- ▶ q_1 : quantity of milk
- ▶ q_2 : quantity of bread
- ▶ q_3 : quantity of butter
- ▶ Unit price: p_i
- ▶ Budget: I



Alternatives: discrete choice set

List of alternatives

- ▶ Brand *A*
- ▶ Brand *B*
- ▶ Brand *C*



Alternatives: discrete choice set

Choice set

- ▶ Non empty finite and countable set of alternatives
- ▶ Universal: \mathcal{C}
- ▶ Individual specific: $\mathcal{C}_n \subseteq \mathcal{C}$
- ▶ Availability, awareness

Example

Choice of a transportation mode

- ▶ $\mathcal{C} = \{\text{car, bus, metro, walking}\}$
- ▶ If decision maker n has no driver license, and the trip is 12km long

$$\mathcal{C}_n = \{\text{bus, metro}\}$$

Alternative attributes

Characterize each alternative i
for each individual n

- ▶ price
- ▶ travel time
- ▶ frequency
- ▶ comfort
- ▶ color
- ▶ size
- ▶ etc.

Nature of the variables

- ▶ Discrete and continuous
- ▶ Generic and specific

Decision rule

Homo economicus

Rational and narrowly self-interested economic actor who is optimizing her outcome

Preferences

- ▶ $i \succ j$: i is preferred to j ,
- ▶ $i \sim j$: indifference between i and j ,
- ▶ $i \succsim j$: i is at least as preferred as j .

Decision rule

Rationality

- ▶ Completeness: for all alternatives i and j ,

$$i \succ j \text{ or } i \prec j \text{ or } i \sim j.$$

- ▶ Transitivity: for all bundles i, j and k ,

$$\text{if } i \succsim j \text{ and } j \succsim k \text{ then } i \succsim k.$$

- ▶ “Continuity”: if i is preferred to j and k is arbitrarily “close” to i , then k is preferred to j .

Utility

$$U_n : \mathcal{C}_n \longrightarrow \mathbb{R} : i \rightsquigarrow U_n(i)$$

Consistent with the preferences

$$U_n(i) \geq U_n(j) \iff i \succsim j.$$

- ▶ Unique up to an order-preserving transformation.
- ▶ Captures the attractiveness of an alternative.
- ▶ Measure that the decision maker wants to optimize.

Behavioral assumptions

- ▶ the preference structure of the decision maker is fully characterized by a utility associated with each alternative
- ▶ the decision maker is a perfect optimizer
- ▶ the alternative with the highest utility is chosen

Theoretical foundations – 2.2 Microeconomic consumer theory

Michel Bierlaire

Preferences and utility.

This document is a short extract from Bierlaire and Lurkin (2017).

Disaggregate demand models are rooted in microeconomics, the branch of economics that focuses on the decision-making behavior of economic actors. We refer to these actors as *individuals*, although they can also be households or firms, for instance.

Consider a set X of goods, bundles, or actions. The objective is to determine what element(s) of X will be chosen/purchased by a given individual. The preferences of the individual are assumed to be characterized by a preference-indifference operator \succeq .

Consider two goods a and b . Then $a \succeq b$ means that the individual either prefers a to b , or is indifferent between a and b . Other operators can be derived from the preference-indifference operator:

- $a \sim b$ is defined as $(a \succeq b \text{ and } b \succeq a)$ and means that the individual is indifferent between a and b ,
- $a \prec b$ is defined as $(\text{not } a \succeq b)$ and means that the individual strictly prefers b to a .

Operators \preceq and \succ can be defined similarly.

A fundamental assumption is that each individual is *rational*. Formally, it means that her preferences must satisfy completeness and transitivity over the set X . *Completeness* means that, for each $a, b \in X$, it is possible to decide if $a \succeq b$ is true or false. *Transitivity* means that, for any $a, b, c \in X$, if $a \succeq b$ and $b \succeq c$, then $a \succeq c$.

It is possible to represent the preference structure of individuals using a *utility function* (see Debreu, 1954). Let $u : X \rightarrow \mathbb{R}$ be a function mapping the set of goods to the real numbers. We say that u represents \succeq on X if

$$a \succeq b \iff u(a) \geq u(b).$$

The transitivity and completeness of the preferences guarantee the existence of a utility function. It can also be shown that it is unique, up to order preserving transformations¹. Appropriate assumptions can also be made on the preference structure in order to obtain desirable properties of the utility function, such as continuity or differentiability. We refer the reader to textbooks in microeconomics such as Nicholson and Snyder (2007), Pindyck and Rubinfeld (2008), or Varian and Repcheck (2010) for more details.

References

- Bierlaire, M. and Lurkin, V. (2017). Introduction to disaggregate demand models, in R. Batta and J. Peng (eds), *Tutorials in Operations Research Leading Developments from INFORMS Communities*, Tutorials in Operations Research, Institute for Operations Research and the Management Sciences (INFORMS), pp. 48–67. ISBN:978-0-9906153-0-9.
- Debreu, G. (1954). Representation of a preference ordering by a numerical function, *Decision Processes*, Wiley.
- Nicholson, W. and Snyder, C. M. (2007). *Microeconomic Theory: Basic Principles and Extensions*, South Western/Thomson.
- Pindyck, R. S. and Rubinfeld, D. L. (2008). *Microeconomics*, 7th edn, Prentice Hall.
- Varian, H. R. and Repcheck, J. (2010). *Intermediate microeconomics: a modern approach*, Vol. 6, WW Norton & Company New York.

¹An equivalence relation on utility functions can be defined, where u and v are equivalent if v is the composition of u and a strictly increasing function g , that is $v(a) = g(u(a))$ for each $a \in X$. The uniqueness applies across equivalence classes.

Theoretical foundations

Microeconomic consumer theory

Michel Bierlaire

Introduction to choice models



The case of continuous goods

Continuous choice set

- ▶ Consumption bundle

$$Q = \begin{pmatrix} q_1 \\ \vdots \\ q_J \end{pmatrix} \quad p = \begin{pmatrix} p_1 \\ \vdots \\ p_J \end{pmatrix}$$

- ▶ Budget constraint

$$p^T Q = \sum_{\ell=1}^J p_{\ell} q_{\ell} \leq I.$$

- ▶ No attributes, just quantities and prices

Choice

Solution of an optimization problem

$$\max_Q \tilde{U}(Q)$$

subject to

$$p^T Q \leq I, \quad Q \geq 0.$$

Demand function

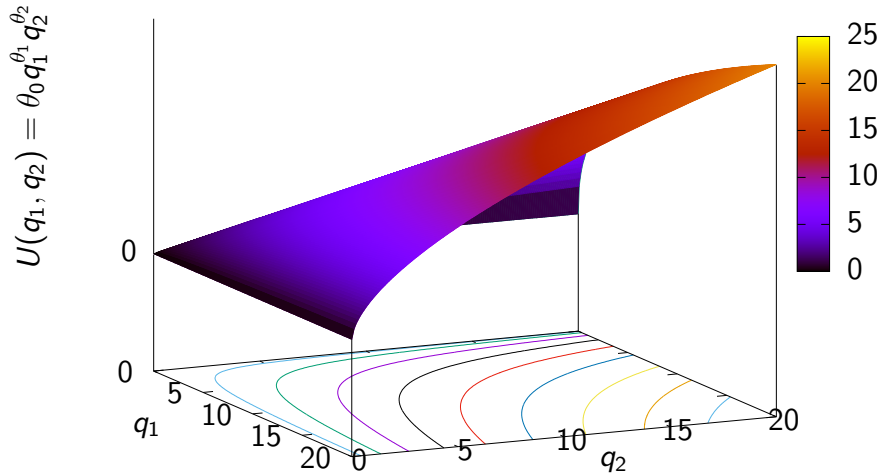
- ▶ Solution of the optimization problem
- ▶ Quantity as a function of prices and budget

$$Q^* = \text{demand}(I, p)$$

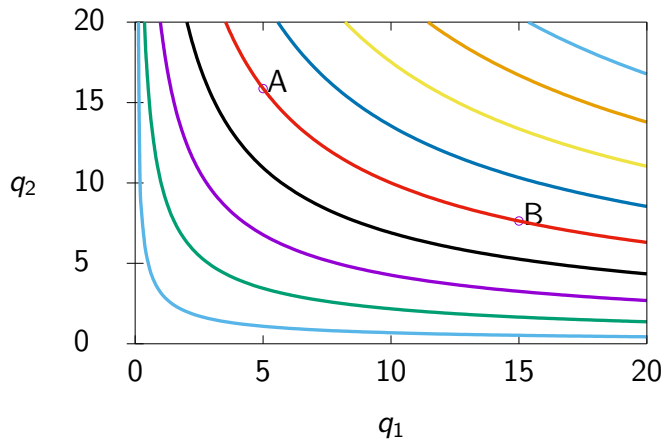
Example: Cobb-Douglas

$$\tilde{U}(Q) = \theta_0 \prod_{\ell=1}^J q_{\ell}^{\theta_{\ell}}$$

Example: Cobb-Douglas



Example



Theoretical foundations – 2.2 Microeconomic consumer theory

Michel Bierlaire

Practice quiz

Derive the demand function for the Cobb-Douglas utility function with two commodities:

$$\tilde{U}(q_1, q_2; \theta) = \theta_0 q_1^{\theta_1} q_2^{\theta_2}, \quad (1)$$

where $\theta = (\theta_0, \theta_1, \theta_2)^T$ is a column vector containing three positive parameters representing the tastes of the consumer.

Theoretical foundations – 2.2 Microeconomic consumer theory

Michel Bierlaire

Solution of the practice quiz

In this exercise, we address the following question: given all the possible values of q_1 and q_2 , which specific quantity of q_1 and quantity of q_2 does the consumer choose? The behavioral assumption is that the consumer wants to maximize her utility. What stops her from consuming an infinite number of goods? These goods have prices and the consumer has a limited budget (I) to spend on the goods.

Consumer behavior can be expressed as an optimization problem where the consumer selects the quantities q_1 and q_2 that maximize her utility \tilde{U} and are compatible with her available budget I :

$$\max_{q_1, q_2} \tilde{U} = \theta_0 q_1^{\theta_1} q_2^{\theta_2} \quad (1)$$

subject to

$$p_1 q_1 + p_2 q_2 = I. \quad (2)$$

The optimal solution of this optimization problem verifies the necessary optimality conditions, based on the Lagrangian function:

$$L(q_1, q_2, \lambda; \theta) = \theta_0 q_1^{\theta_1} q_2^{\theta_2} - \lambda(p_1 q_1 + p_2 q_2 - I), \quad (3)$$

where λ is the Lagrange multiplier.

The Lagrangian somehow turns a constrained optimization problem (1)–(2) into an unconstrained optimization problem where the objective function is (3). In this way, the necessary optimality conditions for unconstrained optimization apply: the first derivatives are equal to zero. Here, the Lagrangian

has three unknowns: q_1 , q_2 and the Lagrange multiplier λ . Therefore,

$$\partial L / \partial q_1 = \theta_0 \theta_1 q_1^{\theta_1 - 1} q_2^{\theta_2} - \lambda p_1 = 0, \quad (4)$$

$$\partial L / \partial q_2 = \theta_0 \theta_2 q_1^{\theta_1} q_2^{\theta_2 - 1} - \lambda p_2 = 0, \quad (5)$$

$$\partial L / \partial \lambda = p_1 q_1 + p_2 q_2 - I = 0. \quad (6)$$

Multiplying (4) by q_1 and (5) by q_2 , we have

$$\theta_0 \theta_1 q_1^{\theta_1} q_2^{\theta_2} - \lambda p_1 q_1 = 0, \quad (7)$$

$$\theta_0 \theta_2 q_1^{\theta_1} q_2^{\theta_2} - \lambda p_2 q_2 = 0. \quad (8)$$

Adding the two and using (6) we obtain

$$\lambda I = \theta_0 q_1^{\theta_1} q_2^{\theta_2} (\theta_1 + \theta_2) \quad (9)$$

or, equivalently,

$$\theta_0 q_1^{\theta_1} q_2^{\theta_2} = \frac{\lambda I}{(\theta_1 + \theta_2)}. \quad (10)$$

Using (10) in (7), we obtain

$$\frac{\lambda p_1 q_1}{\theta_1} = \frac{\lambda I}{(\theta_1 + \theta_2)}. \quad (11)$$

Solving (11) for q_1 , we obtain

$$q_1^* = \frac{\theta_1}{(\theta_1 + \theta_2)} \frac{I}{p_1}. \quad (12)$$

Similarly, we obtain

$$q_2^* = \frac{\theta_2}{(\theta_1 + \theta_2)} \frac{I}{p_2}. \quad (13)$$

Note that the constraints $q_1, q_2 \geq 0$ should also have been included in the optimization problem (1)–(2). As the parameters θ are positive, if the budget is non zero, the optimal quantities are positive, and these constraints are not active at the solution. Therefore, it was appropriate to ignore them.

The Cobb-Douglas function has the property that the demand for a good is only dependent on its own price and independent of the price of any other

good, which is a fairly restrictive assumption. The equations can also be solved for the third unknown, the Lagrange multiplier λ :

$$\lambda = \theta_0(\theta_1 + \theta_2) \left(\frac{\theta_1}{\theta_1 + \theta_2} \right)^{\theta_1} \left(\frac{\theta_2}{\theta_1 + \theta_2} \right)^{\theta_2} \frac{I^{(\theta_1 + \theta_2 - 1)}}{p_1^{\theta_1} p_2^{\theta_2}} \quad (14)$$

The parameter λ is not just a nuisance parameter but has a useful interpretation. Its value is the marginal utility of income, that is the increase in utility that results if income is increased by one unit. Equivalently, λ is equal to the marginal utility of good ℓ ($\partial \tilde{U} / \partial q_\ell$) divided by the marginal cost of good ℓ (equal to p_ℓ in this example) for all goods, or

$$\lambda = \frac{\partial \tilde{U} / \partial q_\ell}{p_\ell} \quad \text{for all goods } \ell. \quad (15)$$

The above equation is directly derived from (4) and (7), that can be written as

$$\partial L / \partial q_\ell = \partial \tilde{U} / \partial q_\ell - \lambda p_\ell = 0. \quad (16)$$

Equation (15) is often described as an optimality condition. Conceptually, at optimal consumption each good should yield the same marginal utility per monetary unit spent. At optimality, if given one extra unit of income to spend, the consumer is indifferent as to which good to purchase more. If the consumer is not indifferent, then she was not at optimality and should adjust her consumption bundle towards the preferred good. The optimality conditions can also be rearranged to state that the marginal rate of substitution of good i for good j is equal to the ratio of the marginal costs of good i relative to good j . For the two commodity case and linear budget constraint, this optimality condition is obtained by calculating the ratio of (15) for $\ell = 1$ and $\ell = 2$ as

$$\frac{\partial \tilde{U} / \partial q_1}{\partial \tilde{U} / \partial q_2} = \frac{p_1}{p_2}. \quad (17)$$

Theoretical foundations

Microeconomic consumer theory

Michel Bierlaire

Introduction to choice models



The case of discrete goods

Microeconomic theory of discrete goods

The consumer

- ▶ selects the quantities of continuous goods: $Q = (q_1, \dots, q_L)$
- ▶ chooses an alternative in a discrete choice set $i = 1, \dots, j, \dots, J$
- ▶ discrete decision vector: (y_1, \dots, y_J) , $y_j \in \{0, 1\}$, $\sum_j y_j = 1$.

Note

- ▶ In theory, one alternative of the discrete choice combines all possible choices made by an individual.
- ▶ In practice, the choice set will be restricted for tractability

Example



Choices

- ▶ House location: discrete choice
- ▶ Car type: discrete choice
- ▶ Number of kilometers driven per year: continuous choice

Discrete choice set

Each combination of a house location and a car is an alternative

Utility maximization

Utility

$$\tilde{U}(Q, y, \tilde{z}^T y)$$

- ▶ Q : quantities of the continuous good
- ▶ y : discrete choice
- ▶ $\tilde{z}^T = (\tilde{z}_1, \dots, \tilde{z}_i, \dots, \tilde{z}_J) \in \mathbb{R}^{K \times J}$: K attributes of the J alternatives
- ▶ $\tilde{z}^T y \in \mathbb{R}^K$: attributes of the chosen alternative
- ▶ θ : vector of parameters

Optimization problem

$$\max_{Q,y} \tilde{U}(Q, y, \tilde{z}^T y)$$

subject to

$$\begin{aligned} p^T Q + c^T y &\leq I \\ \sum_j y_j &= 1 \\ y_j &\in \{0, 1\}, \forall j. \end{aligned}$$

where $c^T = (c_1, \dots, c_i, \dots, c_J)$ is the cost of each alternative

Solving the problem

- ▶ Mixed integer optimization problem
- ▶ No optimality condition
- ▶ Impossible to derive demand functions directly

Solving the problem

Step 1: condition on the choice of the discrete good

- ▶ Fix the discrete good, that is select a feasible y .
- ▶ The problem becomes a continuous problem in Q .
- ▶ Conditional demand functions can be derived:

$$q_{\ell|y} = \text{demand}(I - c^T y, p, \tilde{z}^T y),$$

or, equivalently, for each alternative i ,

$$q_{\ell|i} = \text{demand}(I - c_i, p, \tilde{z}_i).$$

- ▶ $I - c_i$ is the income left for the continuous goods, if alternative i is chosen.
- ▶ If $I - c_i < 0$, alternative i is declared unavailable and removed from the choice set.

Solving the problem

Conditional demand functions

$$\text{demand}(I - c_i, p, \tilde{z}_i), \quad i = 1, \dots, J$$

Conditional indirect utility functions

Substitute the demand functions into the utility:

$$U_i = \tilde{U}(\text{demand}(I - c_i, p, \tilde{z}_i), \tilde{z}_i) = U(I - c_i, p, \tilde{z}_i), \quad i = 1, \dots, J$$

Solving the problem

Step 2: Choice of the discrete good

$$\max_y U(I - c^T y, p, \tilde{z}^T y) \text{ s.t. } \sum_{i=1}^J y_i = 1.$$

- ▶ Enumerate all alternatives.
- ▶ Compute the conditional indirect utility function U_i .
- ▶ Select the alternative with the highest U_i .
- ▶ Note: no income constraint anymore.

Model for individual n

$$\max_y U(I_n - c_n^T y, p_n, \tilde{z}_n^T y)$$

Simplifications

- ▶ S_n : set of characteristics of n , including income I_n .
- ▶ Prices of the continuous goods (p_n) are neglected.
- ▶ c_{in} is considered as another attribute and merged into \tilde{z}_n

$$z_n = \{\tilde{z}_n, c_n\}.$$

$$\max_i U_{in} = U(z_{in}, S_n)$$

Theoretical foundations – 2.3 Example

Michel Bierlaire

Transportation mode choice example.

We illustrate the concept of utility using a simple example of a transportation mode choice, where two alternatives are considered for a commuter trip: car and bus. Each alternative is characterized by two attributes: the travel time and the travel cost, as reported in Table 1

Alternatives	Attributes	
	Travel time (t)	Travel cost (c)
Car (i)	t_i	c_i
Bus (j)	t_j	c_j

Table 1: Attributes of the alternatives

We denote by y_i and y_j the binary variables associated with each alternative:

$$y_i = \begin{cases} 1 & \text{if car is chosen,} \\ 0 & \text{otherwise;} \end{cases}$$

and $y_j = 1 - y_i$, in order to verify the constraint imposing that exactly one alternative is chosen. In terms of the decision problem that the individual is solving, the decision variables and the feasible set are illustrated in Figure 1.

The utility functions associated with each alternative can be written as

$$\begin{aligned} U_i &= -\beta_t t_i - \beta_c c_i, \\ U_j &= -\beta_t t_j - \beta_c c_j, \end{aligned}$$

where $\beta_t > 0$ and $\beta_c > 0$ are parameters.

Note that this specification involves some behavioral assumptions:

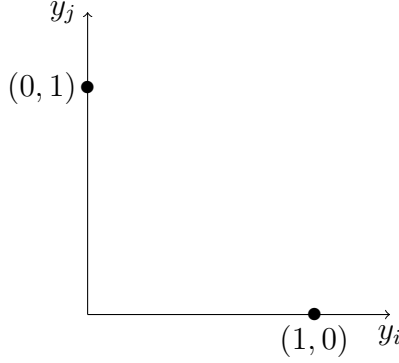


Figure 1: Decision variables and feasible set

- The sign restrictions on the unknown parameters β_t and β_c impose that the value of the utility decreases when one of the variables increases. It is consistent with the behavioral assumption that commuters want to arrive as fast as possible to their destination, at the lowest cost possible.
- The same coefficients are used for both alternatives. This implies that a modification of the travel time has the same impact on the utility of car and on the utility of bus. The same applies for travel cost. This assumption is debatable. It can be argued that an additional minute spent in the bus, with the possibility to sleep, listen to music, or read, may not be perceived the same way as spending one more minute driving the car.

As a representation of the individual's preferences, the utility is defined up to order preserving transformations. For instance, we can divide each utility by a strictly positive number, without modifying their ranking.

$$\begin{aligned} U'_i &= -(\beta_t/\beta_c)t_i - c_i = -\beta t_i - c_i \\ U'_j &= -(\beta_t/\beta_c)t_j - c_j = -\beta t_j - c_j \end{aligned}$$

where $\beta = \beta_t/\beta_c > 0$ is a parameter. Note that this parameter is converting travel time units into travel cost units, so that they can be combined together in the same utility function.

The behavioral assumption is that alternative i is chosen if $U_i \geq U_j$ or,

equivalently, if $U'_i \geq U'_j$. If we ignore ties, we obtain

$$-\beta t_i - c_i < -\beta t_j - c_j,$$

or, equivalently,

$$-\beta(t_i - t_j) < c_i - c_j.$$

Two cases are trivial:

- If $c_j > c_i$ and $t_j > t_i$, the car alternative is both cheaper and faster than the bus alternative. Therefore, $U_i > U_j$ for any $\beta > 0$, and the car is chosen. The car alternative is called a *dominating alternative*.
- Symmetrically, if $c_i > c_j$ and $t_i > t_j$, the car alternative is both more expensive and slower than the bus alternative. Therefore, $U_i < U_j$ for any $\beta > 0$, and the bus is chosen. The car alternative is called a *dominated alternative*.

But what happens when one alternative is cheaper and slower than the other one? In that case, the parameter β captures the trade-off of the decision-maker between the two variables. For instance, assume that the car is cheaper and slower than the bus, that is $c_j > c_i$ and $t_i > t_j$. Alternative j is chosen if

$$-\beta(t_i - t_j) < c_i - c_j,$$

or, as $t_i > t_j$,

$$\beta > \frac{c_j - c_i}{t_i - t_j}.$$

The behavioral question is: Is the traveler willing to pay the extra cost $c_j - c_i$ to save the extra time $t_i - t_j$? The parameter β is capturing this behavioral trade-off. It is called *the value of time*, or the *willingness to pay* to save travel time, and will be discussed in more details later in the course.

This simple example is illustrated in Figure 2. The x -axis corresponds to the difference of travel time $t_i - t_j$. Therefore, negative values correspond to alternative i being faster, and positive values to alternative j being faster. Similarly, the y -axis corresponds to the difference of travel cost $c_i - c_j$. Negative values correspond to alternative i being cheaper, and positive values to alternative j being cheaper.

The north-east quadrant corresponds to situations where alternative j is dominant. Indeed, it is both faster and cheaper. Symmetrically, the south-west quadrant corresponds to situations where alternative i is dominant.

The two other quadrants correspond to situations where there is a trade-off between travel time and travel cost. For a given value of the parameter β , we draw the indifference line, corresponding to situations where the two utilities are equal, that is

$$c_i + \beta t_i = c_j + \beta t_j,$$

or, equivalently,

$$c_i - c_j = -\beta(t_i - t_j).$$

As $\beta > 0$, the slope of this line is negative.

In order to determine the value of β , we collect choice data. We observe a sample of individuals during their commuting trip, and, for each of them, collect:

- the travel time by car t_i ,
- the travel time by bus t_j ,
- the travel cost by car c_i ,
- the travel cost by bus c_j ,
- the alternative actually chosen (i or j).

The data is represented in Figure 3, using the following convention:

- each dot corresponds to an individual,
- the x coordinate of the dot corresponds to the associated value of $t_i - t_j$,
- the y coordinate of the dot corresponds to the associated value of $c_i - c_j$,
- the shape of the dot reveals the choice made by the individual.

Therefore, the objective is to find a value of β , that is, to find a slope of the indifference line, such that all dots corresponding to alternative i lie on one side of the line, and all dots corresponding to alternative j lie on the other side. It is clear that the choice of β in Figure 3 does not achieve that. Moreover, it is relatively easy to figure out that it is impossible to find such a slope. There is at least one dot corresponding to alternative i that is surrounded by dots corresponding to alternative j , and that cannot be separated from them using any line.

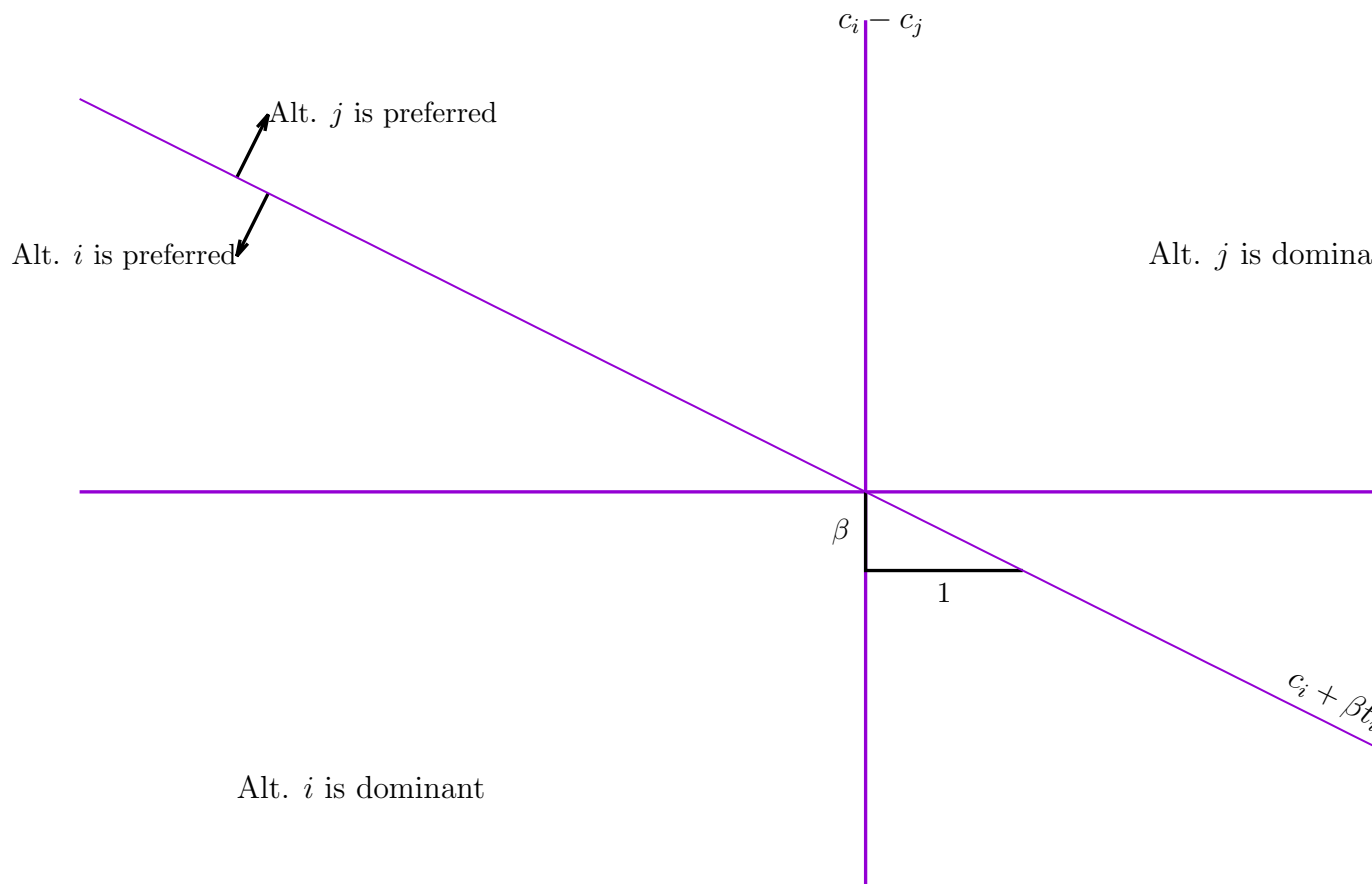


Figure 2: Simple example: two transportation modes

This inconsistency between the behavioral model and the behavioral observations illustrates the limitations of the utility theory, when confronted to data, and motivates to consider the utility as a random variable. The *random utility theory* is discussed next.

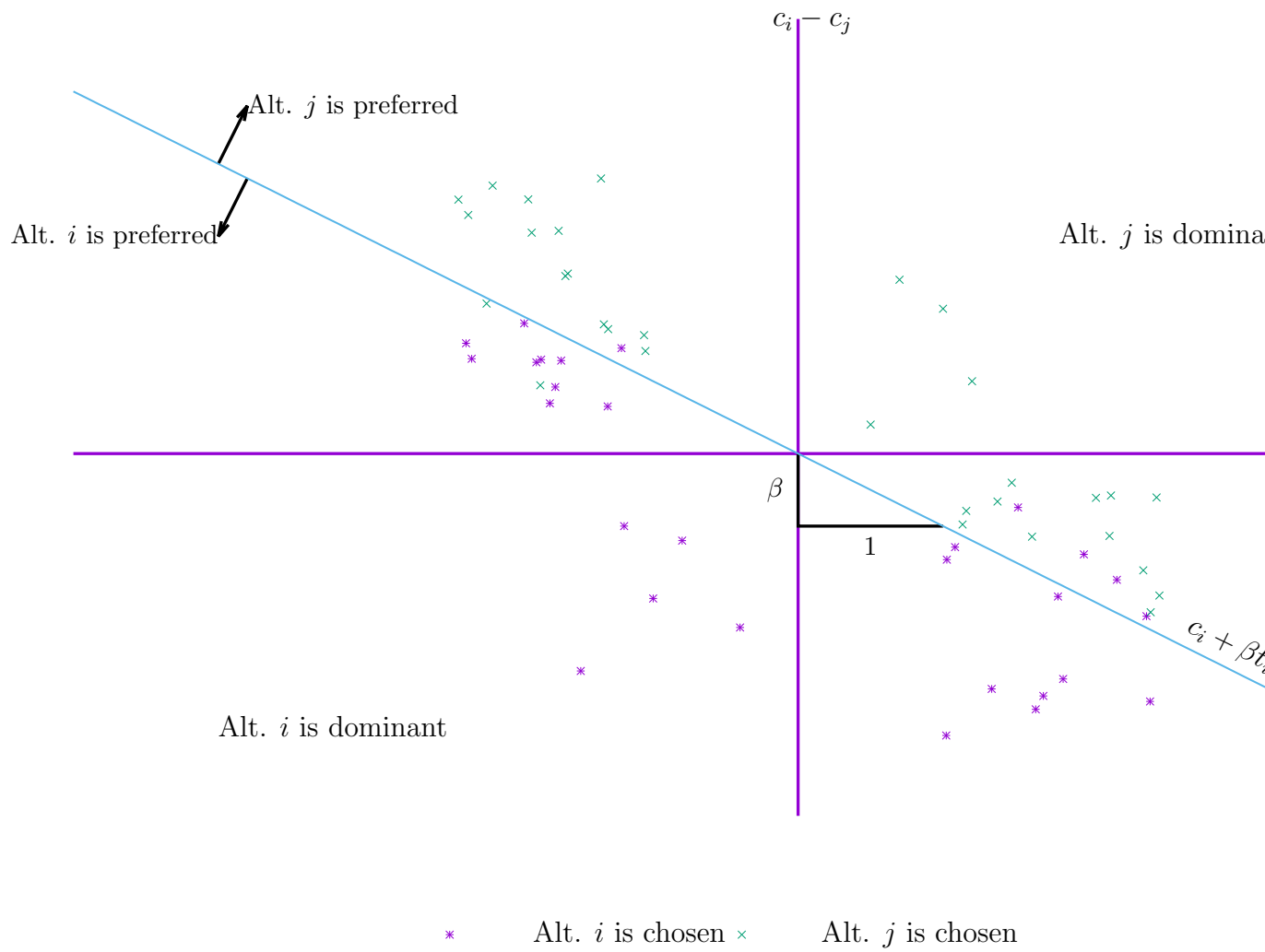


Figure 3: Simple example: two transportation modes with observed choices

Theoretical foundations – 2.3 Example

Michel Bierlaire

Practice quiz.

In order to illustrate the concept of utility, we have introduced a simple example of a transportation mode choice, where two alternatives are considered for a commuter trip: car (alternative i) and bus (alternative j). The utility functions associated with each alternative are written as

$$U_i = -\beta_t t_i - \beta_c c_i, \quad (1)$$

$$U_j = -\beta_t t_j - \beta_c c_j, \quad (2)$$

where t_i and t_j are the travel times of each alternative, c_i and c_j are the travel costs, and $\beta_t > 0$ and $\beta_c > 0$ are parameters.

The same coefficients (β_t and β_c) are used for both alternatives. This implies that a modification of the travel time has the same impact on the utility of car and on the utility of bus. The same applies for travel cost. This assumption is debatable. It can be argued that an additional minute spent in the bus, with the possibility to sleep, listen to music, or read, may not be perceived the same way as spending one more minute driving the car. How would you specify a model where the impact of an additional minute in travel time would be different for the two alternatives?

Theoretical foundations – 2.3 Example

Michel Bierlaire

Solution of the practice quiz.

If we want to capture the idea that an additional minute spent in the bus may not be perceived the same way as spending one more minute driving the car, then the utility functions associated with each alternative have to be written as

$$U_i = -\beta_{ti}t_i - \beta_c c_i, \tag{1}$$

$$U_j = -\beta_{tj}t_j - \beta_c c_j, \tag{2}$$

where $\beta_{ti} > 0$, $\beta_{tj} > 0$, and $\beta_c > 0$ are parameters.

Theoretical foundations – 2.4 Random utility theory

Michel Bierlaire

Probabilistic choice.

Up to now, the theoretical developments have assumed that individuals behave deterministically. Decision-makers are assumed to be all knowing with perfect discriminatory power, able to process information, choose the best choice, and repeat this identical choice under identical circumstances. This is implied by the assumed properties of the preferences, such as completeness, transitivity, and continuity. However, the simple example presented before illustrates that such assumptions may not be fully consistent with real behavior. Actually, there are copious examples both in laboratory experiments and in the field in which it appears that decision-makers do not behave as such. As Tversky (1969) points out, “when faced with repeated choices between x and y , people often choose x in some instances and y in others.” Inspired by the need to explain experimental observations of inconsistent preferences, probabilistic choice theory was developed. In probabilistic choice theory, rather than assuming there is a deterministic process that can be used to establish the choice outcome, it is recognized that the best that can be done is to determine the probability of different choice outcomes given a particular choice situation and decision-maker.

There are several ways of modeling probabilistic choice. In this course, we assume that the source of the stochasticity is due to errors made by the analyst in developing the model. Here the assumption is that while humans *are* deterministic and rational utility maximizers, analysts are unable to understand and model fully all of the relevant factors that affect human behavior. The individual is assumed to be all knowing and rational and select the alternative with the highest utility. However, the utilities are not known to the analyst with certainty and are therefore treated by the analyst

as random variables. This is called the *random utility* approach. The value of the random utility approach is that it provides a link with behavioral theory from microeconomics and therefore a link to the concepts and methods that are useful for both developing model specifications and using the models for analysis.

Formally, the utility that individual n associates with alternative i is a random variable denoted U_{in} . The fact that alternative i is chosen is again associated with the fact that U_{in} is the largest utility. The model is now expressed in a probabilistic way, as follows:

$$P(i|\mathcal{C}_n) = \Pr(U_{in} \geq U_{jn}, \forall j \in \mathcal{C}_n). \quad (1)$$

The most common representation for U_{in} is inspired from linear regression. The utility is separated into two additive parts:

$$U_{in} = V_{in} + \varepsilon_{in}, \quad (2)$$

where V_{in} is called the *deterministic* or *systematic* part of the utility, and ε_{in} is the *error term*. Typically, V_{in} involves the explanatory variables, while distributional assumptions are made on the joint distribution of the random vector of error terms $\varepsilon_n = (\varepsilon_{1n}, \dots, \varepsilon_{J_n n})$

In the rest of the course, we are intuitively deriving concrete models, based on simple assumptions, that are relaxed later on. In the next unit, we provide a general derivation of the model. As it is quite technical, it may be skipped without loss of continuity.

References

Tversky, A. (1969). Intransitivity of preferences, *Psychological Review* **76**(1): 31–48.

Theoretical foundations – 2.4 Random utility theory

Michel Bierlaire

Mathematical derivation of the choice model

We derive here the general random utility model. Although the derivation is quite straightforward, it is also technical. It may be skipped without loss of continuity in the course.

Consider the choice model with J_n alternatives

$$P(i|\mathcal{C}_n) = \Pr(U_{in} \geq U_{jn}, \forall j = 1, \dots, J_n), \quad (1)$$

where

$$U_{in} = V_{in} + \varepsilon_{in}. \quad (2)$$

Denote by

$$\varepsilon_n = (\varepsilon_{1n}, \dots, \varepsilon_{J_n n})$$

the vector of J_n error terms. If ε_n is a multivariate random variable with CDF $F_{\varepsilon_n}(\varepsilon_1, \dots, \varepsilon_{J_n})$ and pdf

$$f_{\varepsilon_n}(\varepsilon_1, \dots, \varepsilon_{J_n}) = \frac{\partial^{J_n} F}{\partial \varepsilon_1 \dots \partial \varepsilon_{J_n}}(\varepsilon_1, \dots, \varepsilon_{J_n}), \quad (3)$$

then

$$\begin{aligned} P_n(i|\mathcal{C}_n) = & \int_{\varepsilon_i=-\infty}^{+\infty} \int_{\varepsilon_1=-\infty}^{V_{in}-V_{1n}+\varepsilon_i} \dots \\ & \int_{\varepsilon_{i-1}=-\infty}^{V_{in}-V_{i-1n}+\varepsilon_i} \int_{\varepsilon_{i+1}=-\infty}^{V_{in}-V_{i+1n}+\varepsilon_i} \dots \\ & \int_{\varepsilon_{J_n}=-\infty}^{V_{1n}-V_{J_n n}+\varepsilon_1} f_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{J_n}}(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{J_n}) d\varepsilon. \end{aligned} \quad (4)$$

and

$$P_n(i|\mathcal{C}_n) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{J_n}}(\dots, V_{in} - V_{(i-1)n} + \varepsilon, \varepsilon, V_{in} - V_{(i+1)n} + \varepsilon, \dots)}{\partial \varepsilon_i} d\varepsilon. \quad (5)$$

Therefore, if the CDF is available in closed form, the choice model is obtained by solving a uni-dimensional integral, which can be done analytically for simple models, and numerically for more complex ones.

Proof. We prove the result for alternative 1 without loss of generality, in order to simplify the notations.

Using (2) into (1), we obtain

$$P(1|\mathcal{C}_n) = \Pr(V_{2n} + \varepsilon_{2n} \leq V_{1n} + \varepsilon_{1n}, \dots, V_{J_n n} + \varepsilon_{J_n n} \leq V_{1n} + \varepsilon_{1n}), \quad (6)$$

or, gathering the random terms on one side, and the deterministic ones on the other side,

$$P_n(1|\mathcal{C}_n) = \Pr(\varepsilon_{2n} - \varepsilon_{1n} \leq V_{1n} - V_{2n}, \dots, \varepsilon_{J_n n} - \varepsilon_{1n} \leq V_{1n} - V_{J_n n}). \quad (7)$$

We consider the following change of variables:

$$\xi_{1n} = \varepsilon_{1n}, \quad \xi_{jn} = \varepsilon_{jn} - \varepsilon_{1n}, \quad j = 2, \dots, J_n, \quad (8)$$

that is, in matrix notations,

$$\xi_n = \begin{pmatrix} \xi_{1n} \\ \xi_{2n} \\ \vdots \\ \xi_{(J_n-1)n} \\ \xi_{J_n n} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ -1 & 1 & \cdots & 0 & 0 \\ & & \vdots & & \\ -1 & 0 & \cdots & 1 & 0 \\ -1 & 0 & \cdots & 0 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{1n} \\ \varepsilon_{2n} \\ \vdots \\ \varepsilon_{(J_n-1)n} \\ \varepsilon_{J_n n} \end{pmatrix} = M \varepsilon_n.$$

Note that the determinant of the change of variables matrix M is 1, so that ε_n and ξ_n have the same pdf. The model in the new variables becomes

$$P_n(1|\mathcal{C}_n) = \Pr(\xi_{2n} \leq V_{1n} - V_{2n}, \dots, \xi_{J_n n} \leq V_{1n} - V_{J_n n}).$$

Therefore,

$$P_n(1|\mathcal{C}_n) = F_{\xi_{1n}, \xi_{2n}, \dots, \xi_{J_n}}(+\infty, V_{1n} - V_{2n}, \dots, V_{1n} - V_{J_n n})$$

from the definition of a cumulative distribution function. As the CDF is obtained by integrating the pdf, we have

$$P_n(1|\mathcal{C}_n) = \int_{\xi_1=-\infty}^{+\infty} \int_{\xi_2=-\infty}^{V_{1n}-V_{2n}} \cdots \int_{\xi_{J_n}=-\infty}^{V_{1n}-V_{J_n n}} f_{\xi_{1n}, \xi_{2n}, \dots, \xi_{J_n}}(\xi_1, \xi_2, \dots, \xi_{J_n}) d\xi.$$

Now we come back to the original variables, exploiting the fact that the pdf of ξ_n and ε_n are identical:

$$P_n(1|\mathcal{C}_n) = \int_{\varepsilon_1=-\infty}^{+\infty} \int_{\varepsilon_2=-\infty}^{V_{1n}-V_{2n}+\varepsilon_1} \cdots \int_{\varepsilon_{J_n}=-\infty}^{V_{1n}-V_{J_n n}+\varepsilon_1} f_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{J_n}}(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_{J_n}) d\varepsilon.$$

By integrating over all dimensions except the first one, we obtain:

$$P_n(1|\mathcal{C}_n) = \int_{\varepsilon_1=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{J_n}}}{\partial \varepsilon_1}(\varepsilon_1, V_{1n} - V_{2n} + \varepsilon_1, \dots, V_{1n} - V_{J_n n} + \varepsilon_1) d\varepsilon_1.$$

□

Theoretical foundations – 2.4 Random utility theory

Michel Bierlaire

Practice quiz.

Consider the general choice model

$$P_n(i|\mathcal{C}_n) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F_{\varepsilon_{1n}, \varepsilon_{2n}, \dots, \varepsilon_{J_n}}(\dots, V_{in} - V_{(i-1)n} + \varepsilon, \varepsilon, V_{in} - V_{(i+1)n} + \varepsilon, \dots)}{\partial \varepsilon_i} d\varepsilon.$$

Derive it for alternative i in the binary case with $\mathcal{C}_n = \{i, j\}$ and the CDF of the error terms is given by

$$F_\varepsilon(\varepsilon_i, \varepsilon_j) = e^{-e^{-\varepsilon_i}} e^{-e^{-\varepsilon_j}}. \quad (1)$$

Hint

The change of variable $t = -e^{-\varepsilon}$ conveniently simplifies the integral.

Theoretical foundations – 2.4 Random utility theory

Michel Bierlaire

Solution of the practice quiz.

The CDF of the error terms is given by

$$F_\varepsilon(\varepsilon_i, \varepsilon_j) = e^{-e^{-\varepsilon_i}} e^{-e^{-\varepsilon_j}}. \quad (1)$$

We have

$$P(i|\{i, j\}) = \int_{\varepsilon=-\infty}^{+\infty} \frac{\partial F_\varepsilon}{\partial \varepsilon_i}(\varepsilon, V_i - V_j + \varepsilon) d\varepsilon. \quad (2)$$

From (1), we have

$$\frac{\partial F_\varepsilon}{\partial \varepsilon_i}(\varepsilon_i, \varepsilon_j) = e^{-e^{-\varepsilon_i}} e^{-e^{-\varepsilon_j}} e^{-\varepsilon_i}. \quad (3)$$

Therefore,

$$\frac{\partial F_\varepsilon}{\partial \varepsilon_i}(\varepsilon, V_i - V_j + \varepsilon) = e^{-e^{-\varepsilon}} e^{-e^{-(V_i - V_j + \varepsilon)}} e^{-\varepsilon} = e^{-e^{-\varepsilon}} e^{-K e^{-\varepsilon}} e^{-\varepsilon} \quad (4)$$

where

$$K = \exp(-(V_i - V_j)). \quad (5)$$

Therefore,

$$P(i|\{i, j\}) = \int_{\varepsilon=-\infty}^{+\infty} e^{-e^{-\varepsilon}} e^{-K e^{-\varepsilon}} e^{-\varepsilon} d\varepsilon. \quad (6)$$

Define

$$t = -e^{-\varepsilon}, \quad dt = e^{-\varepsilon} d\varepsilon,$$

to obtain

$$P(i|\{i, j\}) = \int_{t=-\infty}^0 e^{(1+K)t} dt = \frac{1}{1+K}. \quad (7)$$

Using (5), we obtain the simple expression:

$$P(i|\{i, j\}) = \frac{1}{1 + \exp(-(V_i - V_j))} = \frac{e^{V_i}}{e^{V_i} + e^{V_j}}. \quad (8)$$

This happens to be the binary logit model.

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 3

Binary choice

Model specification: the error term

Virginie Lurkin

Introduction to choice models



Binary choice model

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

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

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

Choice model

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

Binary choice model

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

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

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

Choice model

$$\begin{aligned} P_n(i|\{i, j\}) &= \Pr(U_{in} \geq U_{jn}) \\ &= \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) \end{aligned}$$

Binary choice model

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

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

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

Choice model

$$\begin{aligned} P_n(i|\{i, j\}) &= \Pr(U_{in} \geq U_{jn}) \\ &= \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) \\ &= \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) \end{aligned}$$

Binary choice model

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

$$\begin{aligned}U_{in} &= V_{in} + \varepsilon_{in} \\U_{jn} &= V_{jn} + \varepsilon_{jn}\end{aligned}$$

Choice model

$$\begin{aligned}P_n(i|\{i, j\}) &= \Pr(U_{in} \geq U_{jn}) \\&= \Pr(V_{in} + \varepsilon_{in} \geq V_{jn} + \varepsilon_{jn}) \\&= \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) \\&= \Pr(\varepsilon_n \leq V_{in} - V_{jn})\end{aligned}$$

where $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$.

Error term

Three assumptions about the random variables ε_{in} and ε_{jn}

1. What's their mean?
2. What's their variance?
3. What's their distribution?

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.

The mean

Change of variables

- ▶ Define $E[\varepsilon_{in}] = \beta_{in}$ and $E[\varepsilon_{jn}] = \beta_{jn}$.
- ▶ Define $\varepsilon'_{in} = \varepsilon_{in} - \beta_{in}$ and $\varepsilon'_{jn} = \varepsilon_{jn} - \beta_{jn}$,
- ▶ so that $E[\varepsilon'_{in}] = E[\varepsilon'_{jn}] = 0$.

Choice model

$$P_n(i|\{i,j\}) = \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in})$$

The mean

Change of variables

- ▶ Define $E[\varepsilon_{in}] = \beta_{in}$ and $E[\varepsilon_{jn}] = \beta_{jn}$.
- ▶ Define $\varepsilon'_{in} = \varepsilon_{in} - \beta_{in}$ and $\varepsilon'_{jn} = \varepsilon_{jn} - \beta_{jn}$,
- ▶ so that $E[\varepsilon'_{in}] = E[\varepsilon'_{jn}] = 0$.

Choice model

$$P_n(i|\{i,j\}) =$$

$$\begin{aligned} \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) &= \\ \Pr(V_{in} - V_{jn} \geq \varepsilon'_{jn} + \beta_{jn} - (\varepsilon'_{in} + \beta_{in})) \end{aligned}$$

The mean

Change of variables

- ▶ Define $E[\varepsilon_{in}] = \beta_{in}$ and $E[\varepsilon_{jn}] = \beta_{jn}$.
- ▶ Define $\varepsilon'_{in} = \varepsilon_{in} - \beta_{in}$ and $\varepsilon'_{jn} = \varepsilon_{jn} - \beta_{jn}$,
- ▶ so that $E[\varepsilon'_{in}] = E[\varepsilon'_{jn}] = 0$.

Choice model

$$P_n(i|\{i,j\}) =$$

$$\begin{aligned} \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) &= \\ \Pr(V_{in} - V_{jn} \geq \varepsilon'_{jn} + \beta_{jn} - (\varepsilon'_{in} + \beta_{in})) &= \\ \Pr(V_{in} + \beta_{in} - (V_{jn} + \beta_{jn}) \geq \varepsilon'_{jn} - \varepsilon'_{in}) \end{aligned}$$

The mean

Change of variables

- ▶ Define $E[\varepsilon_{in}] = \beta_{in}$ and $E[\varepsilon_{jn}] = \beta_{jn}$.
- ▶ Define $\varepsilon'_{in} = \varepsilon_{in} - \beta_{in}$ and $\varepsilon'_{jn} = \varepsilon_{jn} - \beta_{jn}$,
- ▶ so that $E[\varepsilon'_{in}] = E[\varepsilon'_{jn}] = 0$.

Choice model

$$P_n(i|\{i,j\}) =$$

$$\begin{aligned} & \Pr(V_{in} - V_{jn} \geq \varepsilon_{jn} - \varepsilon_{in}) = \\ & \Pr(V_{in} - V_{jn} \geq \varepsilon'_{jn} + \beta_{jn} - (\varepsilon'_{in} + \beta_{in})) = \\ & \Pr(V_{in} + \beta_{in} - (V_{jn} + \beta_{jn}) \geq \varepsilon'_{jn} - \varepsilon'_{in}) = \\ & \Pr(V_{in} + \beta_{in} - (V_{jn} + \beta_{jn}) \geq \varepsilon'_n) \end{aligned}$$

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

Alternative specific constant

- ▶ The mean of each error term can be moved to the deterministic part.
- ▶ It is captured by a parameter, to be estimated from data.
- ▶ It is called the **alternative specific constant** (ASC).
- ▶ Only the mean of the difference of the error terms is identified.

Shift invariance

The choice model is not affected by a uniform shift of all utility functions

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

Alternative Specific Constant

Many equivalent specifications

$$\begin{aligned}U_{in} &= V_{in} + \beta_{in} & +\varepsilon'_{in} \\U_{jn} &= V_{jn} + \beta_{jn} & +\varepsilon'_{jn}\end{aligned}$$

or

$$\begin{aligned}U_{in} &= V_{in} & +\varepsilon'_{in} \\U_{jn} &= V_{jn} + \beta_{jn} - \beta_{in} & +\varepsilon'_{jn}\end{aligned}$$

or

$$\begin{aligned}U_{in} &= V_{in} + \beta_{in} - \beta_{jn} & +\varepsilon'_{in} \\U_{jn} &= V_{jn} & +\varepsilon'_{jn}\end{aligned}$$

In practice

Normalize one constant to zero and estimate $\beta_n = \beta_{jn} - \beta_{in}$

Scale invariance

The choice model is not affected by a uniform scaling of all utility functions

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

The variance

$$\begin{aligned}\text{Var}(\alpha U_{in}) &= \alpha^2 \text{Var}(U_{in}) \\ \text{Var}(\alpha U_{jn}) &= \alpha^2 \text{Var}(U_{jn})\end{aligned}$$

The variance is not identified

- ▶ As any α can be selected arbitrarily, any variance can be assumed.
- ▶ No way to identify the variance of the error terms from data.
- ▶ The scale has to be arbitrarily decided.

In practice

The scale parameter of the assumed distribution is normalized to 1.

The distribution

Assumption

ε_{in} and ε_{jn} are the **maximum** of many r.v. capturing unobservable 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: $EV(\eta, \mu)$, with $\mu > 0$.

The Extreme Value distribution $EV(\eta, \mu)$

Probability density function (pdf)

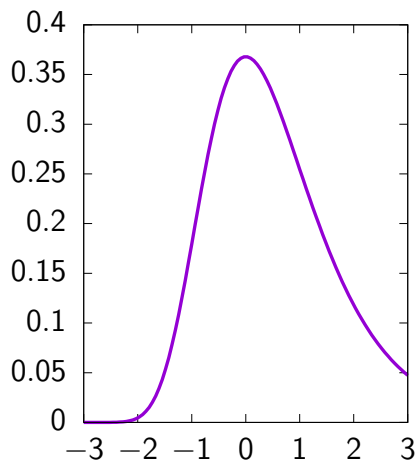
$$f(t) = \mu e^{-\mu(t-\eta)} e^{-e^{-\mu(t-\eta)}}$$

Cumulative distribution function (CDF)

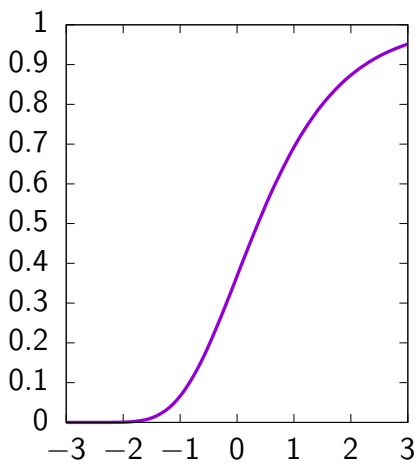
$$\begin{aligned} P(c \geq \varepsilon) = F(c) &= \int_{-\infty}^c f(t) dt \\ &= e^{-e^{-\mu(c-\eta)}} \end{aligned}$$

The Extreme Value distribution

pdf EV(0,1)



CDF EV(0,1)



The Extreme Value distribution

Properties

If

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

then

$$\mathbb{E}[\varepsilon] = \eta + \frac{\gamma}{\mu} \quad \text{and} \quad \text{Var}[\varepsilon] = \frac{\pi^2}{6\mu^2}$$

where γ is Euler's constant.

Euler's constant

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

The distribution

Assumptions

- ▶ ε_{in} and ε_{jn} are i.i.d. Extreme Value.
- ▶ If an alternative specific constant is in the model, their mean can be assumed to be any constant.
- ▶ It is convenient to set the location parameter to 0, so that $E[\varepsilon_{in}] = E[\varepsilon_{jn}] = \gamma/\mu$.

Distributions

$$\varepsilon_{in} \sim EV(0, \mu), \quad \varepsilon_{jn} \sim EV(0, \mu)$$

Problem

We need the distribution of $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$

Logistic distribution

From the properties of the extreme value distribution, we have

$$\begin{aligned}\varepsilon_{in} &\sim \text{EV}(0, \mu) \\ \varepsilon_{jn} &\sim \text{EV}(0, \mu) \\ \varepsilon_n = \varepsilon_{jn} - \varepsilon_{in} &\sim \text{Logistic}(0, \mu)\end{aligned}$$

The Logistic distribution: $\text{Logistic}(\eta, \mu)$

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 $\mu > 0$.

The binary logit model

Choice model

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

The binary logit model

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

Binary choice – 3.1 Model specification: the error term

Michel Bierlaire

Practice quiz.

Consider the utility functions of individual n for two alternatives i and j as follows:

$$U_{in} = V_{in} + \varepsilon_{in}, \quad (1)$$

$$U_{jn} = V_{jn} + \varepsilon_{jn} \quad (2)$$

with the same notations as in the video. The binary probit model is obtained based on the assumption that the error terms are i.i.d. normally distributed across n . Derive the binary probit model $P_n(i)$.

Hints

- Remember that the utility difference matters.
- Remember the definition of a cumulative distribution function (CDF).

Binary choice – 3.1 Model specification: the error term

Michel Bierlaire

Solution of the practice quiz.

The error terms represent everything that is unknown to the analyst. A possible assumption is that all these elements add up to form the error terms. Then, invoking the central limit theorem, they follow a normal distribution.

Suppose that ε_{in} and ε_{jn} are both normal with zero mean, and variance σ_i^2 and σ_j^2 respectively. They are possibly correlated with covariance σ_{ij} . Note that these parameters do not have an index n , to reflect the i.i.d. assumption. They are constant across individuals. Under these assumptions the term $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ is also normally distributed with mean zero and variance $\sigma^2 = \sigma_i^2 + \sigma_j^2 - 2\sigma_{ij}$. We can now solve for the choice probabilities as follows:

$$P_n(i) = \Pr(\varepsilon_{jn} - \varepsilon_{in} \leq V_{in} - V_{jn}) = \Pr(\varepsilon_n \leq V_{in} - V_{jn}), \quad (1)$$

from the random utility model. We now use the fact that ε_n is normally distributed to obtain

$$P_n(i) = \int_{\varepsilon=-\infty}^{V_{in}-V_{jn}} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right] d\varepsilon. \quad (2)$$

By changing the variable $u = \varepsilon/\sigma$ so that $du = d\varepsilon/\sigma$, we obtain a standard normal, and

$$P_n(i) = \frac{1}{\sqrt{2\pi}} \int_{u=-\infty}^{(V_{in}-V_{jn})/\sigma} \exp\left[-\frac{1}{2}u^2\right] du, \quad (3)$$

which is

$$P_n(i) = \Phi\left(\frac{V_{in} - V_{jn}}{\sigma}\right), \quad (4)$$

where $\Phi(\cdot)$ denotes the CDF of a standardized normal distribution.

Binary choice – 3.2 Apply the model on data

Michel Bierlaire

Practice quiz: alternative specific constants.

You have estimated the parameters of the following mode choice model, involving two transportation modes (index n has been dropped for notational convenience):

$$U_{\text{bicycle}} = ASC_{\text{bicycle}} + \beta_{\text{distance}} \cdot \text{distance} + \varepsilon_{\text{bicycle}} \quad (1)$$

$$U_{\text{metro}} = ASC_{\text{metro}} + \beta_{\text{time}} \cdot \text{time}_{\text{metro}} + \beta_{\text{cost}} \cdot \text{cost}_{\text{metro}} + \varepsilon_{\text{metro}} \quad (2)$$

where distance is the distance of the trip in kilometers, $\text{cost}_{\text{metro}}$ is the cost in Swiss francs (CHF) of the trip by metro and $\text{time}_{\text{metro}}$ is the time in minutes of the trip by metro. $\varepsilon_{\text{bicycle}}$ and $\varepsilon_{\text{metro}}$ are random terms.

In order to estimate the model, one of the two alternative specific constants must be normalized to zero. Table 1 reports the estimated parameters for each normalization. However, it is incomplete. First, complete the second column of Table 1 corresponding to the normalization $ASC_{\text{metro}} = 0$.

Parameters	Normalization 1	Normalization 2
ASC_{bicycle}	0	
ASC_{metro}	3	0
β_{distance}	-0.8	
β_{time}	-0.5	
β_{cost}	-1	

Table 1: Estimated parameters

Perform the following tasks for a respondent with a trip of 10 kilometers that takes 20 minutes and costs 2.2 CHF by metro:

1. calculate the choice probabilities in the case of a logit model with the parameter estimates with normalization 1, and the scale parameter set to one,

2. calculate the choice probabilities in the case of a probit model with the parameter estimates with normalization 1, and the scale parameter set to one,
3. calculate the choice probabilities in the case of a logit model with the parameter estimates with normalization 2, and the scale parameter set to one,
4. calculate the choice probabilities in the case of a probit model with the parameter estimates with normalization 2, and the scale parameter set to one.

Binary choice – 3.2 Apply the model on data

Michel Bierlaire

Solution of the practice quiz: alternative specific constants.

In order to complete the table, we have to remember that only the difference between the constants can be identified. This difference should be the same for any normalization. As $ASC_{\text{bicycle}} - ASC_{\text{metro}} = -3$ with the first normalization, it has to be the same for the second one. Therefore, $ASC_{\text{bicycle}} = -3$. The normalization of the constants has no impact on the coefficients of the attributes. Therefore, the β parameters remain unchanged. The result is:

Parameters	Normalization 1	Normalization 2
ASC_{bicycle}	0	-3
ASC_{metro}	3	0
β_{distance}	-0.8	-0.8
β_{time}	-0.5	-0.5
β_{cost}	-1	-1

Now, in order to calculate the choice probabilities with various models, we need first to calculate the utility functions for the scenario that is considered. We have, for the first normalization,

$$\begin{aligned} V_{\text{bicycle}} &= 0 - 0.8 \cdot 10 = -8, \\ V_{\text{metro}} &= 3 - 0.5 \cdot 20 - 1 \cdot 2.2 = -9.2. \end{aligned}$$

And for the second one,

$$\begin{aligned} V_{\text{bicycle}} &= -3 - 0.8 \cdot 10 = -11, \\ V_{\text{metro}} &= 0 - 0.5 \cdot 20 - 1 \cdot 2.2 = -12.2. \end{aligned}$$

1. The logit model with the parameters of normalization 1:

$$P(\text{bicycle}) = \frac{\exp(V_{\text{bicycle}})}{\exp(V_{\text{bicycle}}) + \exp(V_{\text{metro}})} = \frac{\exp(-8)}{\exp(-8) + \exp(-9.2)} = 0.77$$

and

$$P(\text{metro}) = \frac{\exp(V_{\text{metro}})}{\exp(V_{\text{bicycle}}) + \exp(V_{\text{metro}})} = \frac{\exp(-9.2)}{\exp(-8) + \exp(-9.2)} = 0.23.$$

2. The probit model with the parameters of normalization 1:

$$P(\text{bicycle}) = \Phi(V_{\text{bicycle}} - V_{\text{metro}}) = \Phi(-8 + 9.2) = \Phi(1.2) = 0.88$$

and

$$P(\text{metro}) = \Phi(V_{\text{metro}} - V_{\text{bicycle}}) = \Phi(-9.2 + 8) = \Phi(-1.2) = 0.12$$

3. The logit model with the parameters of normalization 2:

$$P(\text{bicycle}) = \frac{\exp(V_{\text{bicycle}})}{\exp(V_{\text{bicycle}}) + \exp(V_{\text{metro}})} = \frac{\exp(-11)}{\exp(-11) + \exp(-12.2)} = 0.77$$

and

$$P(\text{metro}) = \frac{\exp(V_{\text{metro}})}{\exp(V_{\text{bicycle}}) + \exp(V_{\text{metro}})} = \frac{\exp(-12.2)}{\exp(-11) + \exp(-12.2)} = 0.23.$$

4. The probit model with the parameters of normalization 2:

$$P(\text{bicycle}) = \Phi(V_{\text{bicycle}} - V_{\text{metro}}) = \Phi(-11 + 12.2) = \Phi(1.2) = 0.88$$

and

$$P(\text{metro}) = \Phi(V_{\text{metro}} - V_{\text{bicycle}}) = \Phi(-12.2 + 11) = \Phi(-1.2) = 0.12$$

It can be seen that the choice probability does not depend on the normalization of the constants.

Binary choice – 3.2 Apply the model on data

Michel Bierlaire

Practice quiz: scale.

You have estimated the parameters of the following mode choice model, involving two transportation modes (index n has been dropped for notational convenience):

$$U_{\text{bicycle}} = ASC_{\text{bicycle}} + \beta_{\text{distance}} \cdot \text{distance} + \varepsilon_{\text{bicycle}} \quad (1)$$

$$U_{\text{metro}} = ASC_{\text{metro}} + \beta_{\text{time}} \cdot \text{time}_{\text{metro}} + \beta_{\text{cost}} \cdot \text{cost}_{\text{metro}} + \varepsilon_{\text{metro}} \quad (2)$$

where distance is the distance of the trip in kilometers, $\text{cost}_{\text{metro}}$ is the cost in Swiss francs (CHF) of the trip by metro and $\text{time}_{\text{metro}}$ is the time in minutes of the trip by metro. $\varepsilon_{\text{bicycle}}$ and $\varepsilon_{\text{metro}}$ are random terms. The parameter estimates are $ASC_{\text{bicycle}} = 0$, $ASC_{\text{metro}} = 3$, $\beta_{\text{distance}} = -0.8$, $\beta_{\text{time}} = -0.5$ and $\beta_{\text{cost}} = -1$.

Calculate the choice probabilities for a respondent with a trip of 10 kilometers that takes 20 minutes and costs 2.2 CHF by metro in the following cases:

1. using a logit model with scale parameter $\mu = 0.1$,
2. using a logit model with scale parameter $\mu = 10$,
3. using a probit model with scale parameter $\sigma = 0.1$,
4. using a probit model with scale parameter $\sigma = 10$.

Comment on these results.

Binary choice – 3.2 Apply the model on data

Michel Bierlaire

Solution of the practice quiz: scale.

The formulas for the logit model are:

$$P^{\text{logit}}(\text{bicycle}; \mu) = \frac{\exp(\mu V_{\text{bicycle}})}{\exp(\mu V_{\text{bicycle}}) + \exp(\mu V_{\text{metro}})},$$

and

$$P^{\text{logit}}(\text{metro}; \mu) = 1 - P^{\text{logit}}(\text{bicycle}; \mu).$$

The formulas for the probit model are:

$$P^{\text{probit}}(\text{bicycle}; \sigma) = \Phi\left(\frac{V_{\text{bicycle}} - V_{\text{metro}}}{\sigma}\right),$$

and

$$P^{\text{probit}}(\text{metro}; \sigma) = 1 - P^{\text{probit}}(\text{bicycle}).$$

In order to calculate the choice probabilities with various models, we need first to calculate the utility functions for the scenario that is considered. We have

$$\begin{aligned} V_{\text{bicycle}} &= 0 - 0.8 \cdot 10 = -8, \\ V_{\text{metro}} &= 3 - 0.5 \cdot 20 - 1 \cdot 2.2 = -9.2. \end{aligned}$$

1. The logit model with $\mu = 0.1$:

$$\begin{aligned} P^{\text{logit}}(\text{bicycle}; 0.1) &= \frac{\exp(\mu V_{\text{bicycle}})}{\exp(\mu V_{\text{bicycle}}) + \exp(\mu V_{\text{metro}})} \\ &= \frac{\exp(-0.8)}{\exp(-0.8) + \exp(-0.92)} \\ &= 0.53, \end{aligned}$$

and

$$\begin{aligned}
P^{\text{logit}}(\text{metro}; 0.1) &= \frac{\exp(V_{\text{metro}})}{\exp(V_{\text{bicycle}}) + \exp(V_{\text{metro}})} \\
&= \frac{\exp(-0.92)}{\exp(-0.8) + \exp(-0.92)} \\
&= 0.47.
\end{aligned}$$

2. The logit model with $\mu = 10$:

$$\begin{aligned}
P^{\text{logit}}(\text{bicycle}; 10) &= \frac{\exp(\mu V_{\text{bicycle}})}{\exp(\mu V_{\text{bicycle}}) + \exp(\mu V_{\text{metro}})} \\
&= \frac{\exp(-80)}{\exp(-80) + \exp(-92)} \\
&= 0.999994,
\end{aligned}$$

and

$$\begin{aligned}
P^{\text{logit}}(\text{metro}; 10) &= \frac{\exp(V_{\text{metro}})}{\exp(V_{\text{bicycle}}) + \exp(V_{\text{metro}})} \\
&= \frac{\exp(-92)}{\exp(-80) + \exp(-92)} \\
&= 0.000006.
\end{aligned}$$

3. The probit model with $\sigma = 0.1$:

$$P^{\text{probit}}(\text{bicycle}; 0.1) = \Phi((V_{\text{bicycle}} - V_{\text{metro}})/\sigma) = \Phi(12) \approx 1,$$

and

$$P^{\text{probit}}(\text{metro}; 0.1) = \Phi(V_{\text{metro}} - V_{\text{bicycle}}) = \Phi(-12) \approx 0.$$

4. The probit model with $\sigma = 10$:

$$P^{\text{probit}}(\text{bicycle}; 10) = \Phi((V_{\text{bicycle}} - V_{\text{metro}})/\sigma) = \Phi(0.12) = 0.55,$$

and

$$P^{\text{probit}}(\text{metro}; 10) = \Phi(V_{\text{metro}} - V_{\text{bicycle}}) = \Phi(-0.12) = 0.45.$$

The results are summarized in the following table:

	Logit		Probit	
	$\mu = 0.1$	$\mu = 10$	$\sigma = 0.1$	$\sigma = 10$
$P(\text{bicycle})$	0.53	0.999994	1	0.55
$P(\text{metro})$	0.47	0.000006	0	0.45

This exercise illustrates the importance of the scale parameter in the calculation of the choice probability. If the β 's are set, varying the scale parameter modifies the choice probability. As the scale is normalized to one at the estimation stage, it is tempting to forget about it. But it is embedded in the values of the coefficients.

The results also illustrate the limiting cases of the models:

Probit There are two limiting cases of a probit model of special interest, both involving extreme values of the scale parameter. The first case is for $\sigma \rightarrow 0$:

$$\lim_{\sigma \rightarrow 0} P_n(i) = \begin{cases} 1 & \text{if } V_{in} - V_{jn} > 0, \\ 0 & \text{if } V_{in} - V_{jn} < 0; \end{cases}$$

this is, as $\sigma \rightarrow 0$, the variance vanishes and the choice model is deterministic. On the other hand, when $\sigma \rightarrow \infty$, the choice probabilities become 1/2. Intuitively the model predicts equal probability of choice for each alternative irrespectively of V_{metro} and V_{bicycle} .

Logit As with probit, if the V 's are set, there are two limiting cases of a binary logit model that are of special interest. The first case is for $\mu \rightarrow \infty$:

$$\lim_{\mu \rightarrow \infty} P_n(i) = \begin{cases} 1 & \text{if } V_{in} - V_{jn} > 0, \\ 0 & \text{if } V_{in} - V_{jn} < 0; \end{cases}$$

that is, as $\mu \rightarrow \infty$, the variance vanishes and the choice model is deterministic. On the other hand, when $\mu \rightarrow 0$, the choice probabilities become 1/2.

Note that, for probit, the variance is proportional to the (squared) scale parameter, while for logit, it is inversely proportional.

Binary choice – 3.3 Maximum likelihood estimation

Michel Bierlaire

Maximum likelihood estimation.

We now estimate the values of the unknown parameters β_1, \dots, β_K from a sample of observations drawn at random from the population. Each observation of this sample consists of the following:

1. An indicator variable defined as

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

2. Two vectors of explanatory variables $x_{in} = h(z_{in}, S_n)$ and $x_{jn} = h(z_{jn}, S_n)$, each containing K values.

For notational convenience, we also define $y_{jn} = 1 - y_{in}$.

As an example, consider a transportation mode choice problem (train or car), where the utility functions are specified as reported in Table 1. Consider also the sample of 3 individuals presented in Table 2.

Using the above notations, we have

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

The values of the variables x are:

$$\begin{aligned} x_{i1} &= (1 & 5 & 0 & 1.17 & 0 & 0 & 1 & 0 & 0), \\ x_{j1} &= (0 & 40 & 0 & 0 & 2.5 & 0 & 0 & 0 & 0), \\ x_{i2} &= (1 & 8.33 & 2 & 0 & 0 & 0 & 0 & 1 & 1), \\ x_{j2} &= (0 & 7.8 & 0 & 0 & 1.75 & 1 & 0 & 0 & 0), \\ x_{i3} &= (1 & 3.2 & 0 & 2.55 & 0 & 0 & 0 & 1 & 0), \\ x_{j3} &= (0 & 40 & 0 & 0 & 2.67 & 0 & 0 & 0 & 0). \end{aligned}$$

	Car	Train
β_1	1	0
β_2	cost of trip by car	cost of trip by train
β_3	travel time by car (hours) if trip purpose is work, 0 otherwise	0
β_4	travel time by car (hours) if trip purpose is not work, 0 otherwise	0
β_5	0	travel time by train (hours)
β_6	0	1 if first class is preferred, 0 otherwise
β_7	1 if commuter is male, 0 otherwise	0
β_8	1 if commuter is the main earner in the family, 0 otherwise	0
β_9	1 if commuter had a fixed arrival time, 0 otherwise	0

Table 1: Specification table of the binary mode choice model

The choice model is

$$P_n(i) = \frac{e^{V_{in}}}{e^{V_{in}} + e^{V_{jn}}}, \quad (1)$$

where

$$V_{in} = \sum_{k=1}^K \beta_k x_{ink} \quad (2)$$

$$V_{jn} = \sum_{k=1}^K \beta_k x_{jnk}. \quad (3)$$

Given a sample of N observations, we want to find estimates $\hat{\beta}_1, \dots, \hat{\beta}_K$ that have some or all of the desirable properties of statistical estimators. We consider in detail the most widely used estimation procedure — maximum likelihood. The maximum likelihood estimators have the following desired properties:

	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
Gender	M	F	F
Trip purpose	Not work	Work	Not work
Class	Second	First	Second
Main earner	No	Yes	Yes
Arrival time	Variable	Fixed	Variable
Choice	Train	Car	Car

Table 2: A sample of three individuals

1. They are consistent in the sense of convergence to true values as sample size gets larger.
2. They are asymptotically normally distributed in the sense of the Central Limit Theorem.
3. They are asymptotically efficient, and hence their variance attains the Cramer-Rao lower bound.

The maximum likelihood estimation procedure is conceptually quite straightforward. It consists in identifying the value of the unknown parameters such that the joint probability of the observed choices as predicted by the model is the highest possible. This joint probability is called the *likelihood* of the sample. And it is a function of the parameters of the model.

In the above example, the likelihood of the sample of 3 individuals is calculated as follows:

- individual 1 has chosen the car, and this choice is predicted by the model with probability $P_1(i)$,
- individual 2 has chosen the train, and this choice is predicted by the model with probability $P_2(j)$,
- individual 3 has chosen the train, and this choice is predicted by the model with probability $P_3(j)$.

Consequently, the probability that the model predicts all three observations is

$$\mathcal{L}^*(\beta_1, \dots, \beta_9) = P_1(i)P_2(j)P_3(j). \quad (4)$$

If this value is calculated for $\beta_k = 0$, $k = 1, \dots, K$, we obtain

$$\mathcal{L}^* = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = 0.125. \quad (5)$$

If this value is calculated for the values of

$$\beta = (3.04, -0.0527, -2.66, -2.22, -0.576, 0.961, -0.850, 0.383, -0.624),$$

we have

$$\mathcal{L}^* = 0.947 \cdot 0.924 \cdot 0.225 = 0.197. \quad (6)$$

This value of the likelihood is higher. But we do not know if it is the highest possible.

This can be generalized to a sample of N observations assumed to be independently drawn from the population. As discussed above, the likelihood of the sample is the product of the likelihoods (or probabilities) of the individual observations. It is defined as follows:

$$\mathcal{L}^*(\beta_1, \beta_2, \dots, \beta_K) = \prod_{n=1}^N P_n(i)^{y_{in}} P_n(j)^{y_{jn}}, \quad (7)$$

where $P_n(i)$ and $P_n(j)$ are functions of β_1, \dots, β_K . Note that each factor represents the choice probability of the chosen alternative. Indeed,

$$P_n(i)^{y_{in}} P_n(j)^{y_{jn}} = \begin{cases} P_n(i) & \text{if } y_{in} = 1, y_{jn} = 0 \\ P_n(j) & \text{if } y_{in} = 0, y_{jn} = 1. \end{cases}$$

It is more convenient to analyze the logarithm of \mathcal{L}^* , denoted as \mathcal{L} and called the *log likelihood*, because the logarithm of a product of elements is easier to manipulate, being equal to the sum of the logarithms of the elements. Moreover, the value of the likelihood is always between 0 and 1, and usually very small, especially when N is large. The range of values of the log likelihood is much larger, as it can take any negative value (from $-\infty$ to 0) and can be represented better in computers. The log likelihood is written as follows:

$$\mathcal{L}(\beta_1, \dots, \beta_K) = \sum_{n=1}^N (y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)). \quad (8)$$

where β is the vector with entries β_1, \dots, β_K . We are looking for estimates $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$ that solve

$$\max \mathcal{L}(\hat{\beta}) = \mathcal{L}(\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K), \quad (9)$$

where $\hat{\beta}$ is the vector with entries $\hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_K$. The optimization problem is solved using dedicated algorithms.

If a solution exists, it must satisfy the necessary first order conditions:

$$\frac{\partial \mathcal{L}}{\partial \beta_k}(\hat{\beta}) = \sum_{n=1}^N \left(y_{in} \frac{\partial P_n(i)/\partial \beta_k}{P_n(i)} + y_{jn} \frac{\partial P_n(j)/\partial \beta_k}{P_n(j)} \right) = 0, \quad k = 1, \dots, K, \quad (10)$$

or in vector form

$$\frac{\partial \mathcal{L}}{\partial \beta}(\hat{\beta}) = 0. \quad (11)$$

The term $\partial \mathcal{L}(\hat{\beta})/\partial \beta$ is the vector of first derivatives of the log likelihood function with respect to the unknown parameters, evaluated at the estimated value of the parameters. Each entry k of the vector $\partial \mathcal{L}(\hat{\beta})/\partial \beta$ represents the slope of the multi-dimensional log likelihood function along the corresponding k th axis. If $\hat{\beta}$ corresponds to a maximum of the function, all these slopes must be zero, justifying (10).

Solving the optimization problem requires an iterative procedure. It starts with arbitrary values for the parameters (provided by the analyst, or all set to zero if no value can be guessed). If the first derivatives of the log likelihood function are zero, a solution has been found. If not, they provide information about the slope of the function, and a direction of “hill-climbing” can be identified. This direction is followed for a while, until a new set of values is found, corresponding to a higher log likelihood. The process is restarted from this new set of values, until convergence to the maximum is reached.

A family of algorithms commonly used in practice is called *Newton's method*. At each iteration ℓ , a quadratic model of the log likelihood function is built around the current iterate $\beta^{(\ell)}$. This quadratic model is such that the value of the model and of its first and second derivatives are the same at $\beta^{(\ell)}$ as the log likelihood function:

$$m(\beta; \beta^{(\ell)}) = \mathcal{L}(\beta^{(\ell)}) + (\beta - \beta^{(\ell)})^T \nabla \mathcal{L}(\beta^{(\ell)}) + \frac{1}{2} (\beta - \beta^{(\ell)})^T \nabla^2 \mathcal{L}(\beta^{(\ell)}) (\beta - \beta^{(\ell)}), \quad (12)$$

where $\nabla\mathcal{L}(\beta^{(\ell)})$ is the gradient, that is the vector of the first derivatives of the log likelihood function evaluated at $\beta^{(\ell)}$, and $\nabla^2\mathcal{L}(\beta^{(\ell)})$ is the matrix of the second derivatives of the log likelihood function evaluated at $\beta^{(\ell)}$. The k th entry of $\mathcal{L}(\beta^{(\ell)})$ is $\partial\mathcal{L}(\beta^{(\ell)})/\partial\beta_k$, and the entry in the k th row and the m th column of $\nabla^2\mathcal{L}(\beta^{(\ell)})$ is

$$\frac{\partial^2\mathcal{L}(\beta^{(\ell)})}{\partial\beta_k\partial\beta_m}. \quad (13)$$

The approximation of the log likelihood function by the quadratic model is illustrated in Figure 1 for a log likelihood function with only one parameter, where both the log likelihood function and the quadratic model at $\beta^{(\ell)}$ are displayed. Note that both functions coincide at $\beta^{(\ell)}$, and have the same slope (first derivative) and curvature (second derivative) at that point. The next iterate is selected as the value of the parameters maximizing the quadratic model, that is

$$\beta^{(\ell+1)} = \beta^{(k)} - \nabla^2\mathcal{L}(\beta^{(\ell)})^{-1}\nabla(\beta^{(\ell)}), \quad (14)$$

as illustrated in Figures 1 and 2 for two successive iterations.

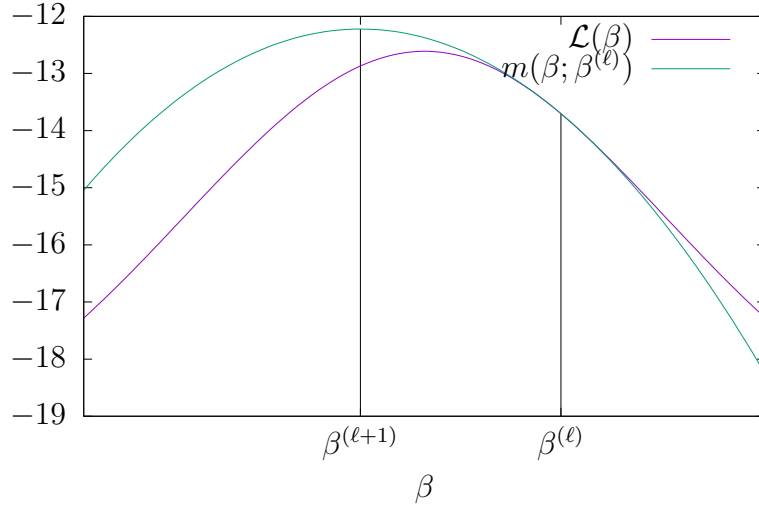


Figure 1: Illustration of Newton's method for optimization

It is numerically obtained by solving the system of linear equations

$$\nabla^2\mathcal{L}(\beta^{(\ell)})d = -\nabla(\beta^{(\ell)}), \quad (15)$$

to obtain the direction d , and then calculating

$$\beta^{(\ell+1)} = \beta^{(\ell)} + d. \quad (16)$$

The procedure continues until the gradient is sufficiently close to zero, depending on the level of precision that is required. In practice, it happens when the norm of the gradient is below a user-specified threshold Γ , that is

$$\left\| \frac{\partial \mathcal{L}(\beta)}{\partial \beta} \right\| = \sqrt{\sum_k \left(\frac{\partial \mathcal{L}(\beta)}{\partial \beta_k} \right)^2} \leq \Gamma.$$

A typical value for Γ is 10^{-6} .

Actually, the method described above is not guaranteed to converge, and variants involving a scaled version of d have to be used, that is

$$\beta^{(\ell+1)} = \beta^{(\ell)} + \alpha d, \quad \alpha > 0. \quad (17)$$

We refer the reader to Bierlaire (2015) for more details on optimization algorithms.

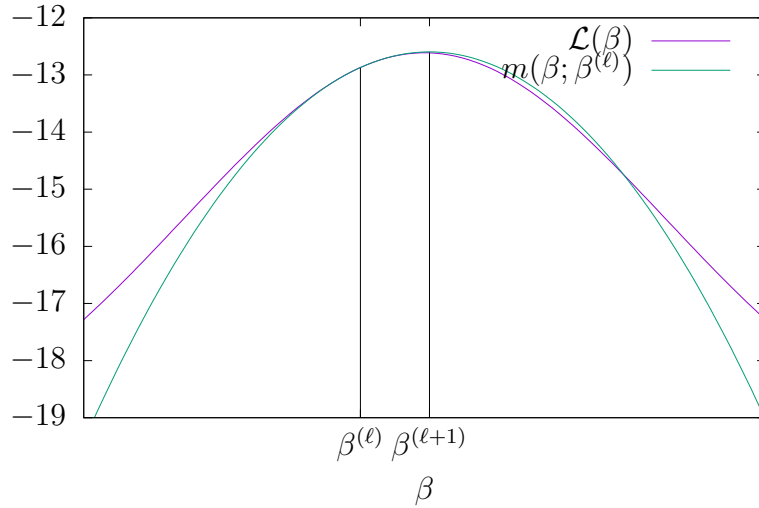


Figure 2: Illustration of Newton's method for optimization: second iteration

References

Bierlaire, M. (2015). *Optimization: Principles and Algorithms*, EPFL Press.

Binary choice – 3.3 Maximum likelihood estimation

Michel Bierlaire

Practice quiz.

Calculate the first order optimality conditions of the maximum likelihood optimization problem for the binary logit model with a linear-in-parameters specification of the utility functions.

Binary choice – 3.3 Maximum likelihood estimation

Michel Bierlaire

Solution of the practice quiz.

The first order necessary optimality conditions are

$$\frac{\partial \mathcal{L}}{\partial \beta_k}(\hat{\beta}) = 0, \quad k = 1, \dots, K, \quad (1)$$

and the log likelihood is

$$\mathcal{L}(\beta_1, \dots, \beta_K) = \sum_{n=1}^N (y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)). \quad (2)$$

In the case of the logit model

$$P_n(i) = \frac{e^{V_{in}}}{e^{V_{in}} + e^{V_{jn}}}, \quad (3)$$

where

$$V_{in} = \sum_k \beta_k x_{ink}, \quad (4)$$

and x_{in} is the vector of explanatory variables associated with alternative i . Note that

$$\ln P_n(i) = V_{in} - \ln(e^{V_{in}} + e^{V_{jn}}). \quad (5)$$

Therefore,

$$\frac{\partial \ln P_n(i)}{\partial V_{in}} = 1 - P_n(i), \quad (6)$$

$$\frac{\partial \ln P_n(i)}{\partial V_{jn}} = -P_n(j). \quad (7)$$

Consequently, the partial derivative of (2) with respect to V_{in} is

$$\frac{\partial \mathcal{L}}{\partial V_{in}} = \sum_{n=1}^N (y_{in}(1 - P_n(i)) - y_{jn}P_n(i)) = \sum_{n=1}^N y_{in} - P_n(i), \quad (8)$$

as $y_{in} + y_{jn} = 1$. The partial derivative of (2) with respect to β_k is

$$\frac{\partial \mathcal{L}}{\beta_k} = \sum_{n=1}^N \sum_i \frac{\partial \mathcal{L}}{\partial V_{in}} \frac{\partial V_{in}}{\partial \beta_k} = \sum_{n=1}^N (y_{in} - P_n(i))x_{ink} + (y_{jn} - P_n(j))x_{jnk}, \quad (9)$$

as the utility function (4) is linear. As $y_{in} + y_{jn} = 1$ and $P_n(i) + P_n(j) = 1$, this simplifies to

$$\frac{\partial \mathcal{L}}{\beta_k} = \sum_{n=1}^N (y_{in} - P_n(i))(x_{ink} - x_{jnk}). \quad (10)$$

Therefore, the first order necessary optimality conditions are

$$\sum_{n=1}^N (y_{in} - P_n(i))(x_{ink} - x_{jnk}) = 0, \quad \forall k = 1, \dots, K. \quad (11)$$

Binary choice – 3.3 Maximum likelihood estimation

Michel Bierlaire

Output of the estimation.

We explain here the various outputs from the maximum likelihood estimation procedure.

Solution of the maximum likelihood estimation

The main outputs of the maximum likelihood estimation procedure are

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

Most estimation software packages provide additional information after the estimation, in order to help appreciating the quality of the results. We summarize the most common ones here.

Variance-covariance matrix of the estimates

In addition to play a role in the optimization algorithm, the second derivatives matrix of the log likelihood function $\nabla^2 \mathcal{L}(\beta)$ is also used to compute an estimate of the variance-covariance matrix of the parameter estimates, from which standard errors, t statistics and p values are generated.

Under relatively general conditions, the asymptotic variance-covariance matrix of the maximum likelihood estimates is given by the Cramer-Rao bound

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

From the second order optimality conditions, this matrix is negative definite if the local maximum is unique, which is the algebraic equivalent of the local strict concavity of the log likelihood function.

Since we do not know the actual values of the parameters at which to evaluate the second derivatives, or the distribution of x_{in} and x_{jn} over which to take their expected value, we estimate the variance-covariance matrix by evaluating the second derivatives at the estimated parameters $\hat{\beta}$ and the sample distribution of x_{in} and x_{jn} instead of their true distribution. Thus we use

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

as a consistent estimator of the matrix of second derivatives. Denote this matrix as \hat{A} . Therefore, an estimate of the Cramer-Rao bound (1) is given by

$$\hat{\Sigma}_{\beta}^{\text{CR}} = -\hat{A}^{-1}. \quad (3)$$

If the matrix \hat{A} is negative definite then $-\hat{A}$ is invertible and the Cramer-Rao bound is positive definite. Note that this may not always be the case, as it depends on the model and the sample.

Another consistent estimator of the (negative of the) second derivatives matrix can be obtained by the matrix of the cross-products of first derivatives as follows:

$$-E \left[\frac{\partial^2 \mathcal{L}(\beta)}{\partial \beta \partial \beta^T} \right] \approx \sum_{n=1}^n \nabla L_n(\hat{\beta}) \nabla L_n(\hat{\beta})^T = \hat{B}, \quad (4)$$

where

$$\nabla L_n(\hat{\beta}) = \nabla (y_{in} \ln P_n(i) + y_{jn} \ln P_n(j)) \quad (5)$$

is the gradient vector of the log likelihood of observation n . As the gradient $\nabla L_n(\hat{\beta})$ is a column vector of dimension $K \times 1$, and its transpose $\nabla L_n(\hat{\beta})^T$ is a row vector of size $1 \times K$, the product $\nabla L_n(\hat{\beta}) \nabla L_n(\hat{\beta})^T$ appearing for each observation n in (4) is a rank one matrix of size $K \times K$. The approximation \hat{B} is employed by the BHHH algorithm (Berndt et al., 1974). It can also provide an estimate of the variance-covariance matrix:

$$\hat{\Sigma}_{\beta}^{\text{BHHH}} = \hat{B}^{-1}, \quad (6)$$

although this estimate is rarely used. Instead, \hat{B} is used to derive a third consistent estimator of the variance-covariance matrix of the parameters, defined as

$$\hat{\Sigma}_\beta^R = (-\hat{A})^{-1} \hat{B} (-\hat{A})^{-1} = \hat{\Sigma}_\beta^{CR} (\hat{\Sigma}_\beta^{BHHH})^{-1} \hat{\Sigma}_\beta^{CR}. \quad (7)$$

It is called the *robust* estimator, or sometimes the *sandwich* estimator, due to the form of equation (7).

When the true likelihood function is maximized, these estimators are asymptotically equivalent, and the Cramer-Rao bound (1) should be preferred (Kauermann and Carroll, 2001). When other consistent estimators are used, different from the maximum likelihood, the robust estimator (7) must be used (White, 1982).

Standard errors

Consider an estimate $\hat{\beta}_k$ of the parameter β_k , and consider $\hat{\Sigma}_\beta$ an estimate of the variance-covariance matrix of the estimates (typically, the Rao-Cramer bound or the robust estimator, as described above). The standard error of the parameter is defined as

$$\sigma_k = \sqrt{\hat{\Sigma}_\beta(k, k)}, \quad (8)$$

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

t statistics

Consider an estimate $\hat{\beta}_k$ of the parameter β_k , and σ_k its standard error. Its t statistic is defined as

$$t_k = \frac{\hat{\beta}_k}{\sigma_k}. \quad (9)$$

It is 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

$$|t_k| \geq 1.96. \quad (10)$$

p value

Consider an estimate $\hat{\beta}_k$ of the parameter β_k , and t_k its t statistic. The p value is calculated as

$$p_k = 2(1 - \Phi(t_k)), \quad (11)$$

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

It conveys the exact same information as the t statistic, presented in a different way. 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 $\beta_k = 0$. The null hypothesis can be rejected with level of confidence $1 - p_k$.

Goodness of fit

Unlike linear regression, 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.

Clearly, an obvious measure is the log likelihood itself. It is common to compare it with a benchmark model. For instance, consider a trivial model with no parameter, associating a probability of 50% with each of the two alternatives:

$$P_n(i) = P_n(j) = \frac{1}{2}.$$

The log likelihood of the sample is therefore

$$\mathcal{L}(0) = \log\left(\frac{1}{2^N}\right) = -N \log(2),$$

where N is the number of observations. It can be used to calculate the likelihood ratio statistic:

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

It is called as such because it is the logarithm of the ratio of the respective likelihood values.

The statistic is used to test the null hypothesis H_0 that the estimated model is equivalent to the equal probability model. Under H_0 , $-2(\mathcal{L}(0) - \mathcal{L}(\hat{\beta}))$ is asymptotically distributed as χ^2 with K degrees of freedom.

It can also be used to compute a normalized measure of goodness of fit:

$$\rho^2 = 1 - \frac{\mathcal{L}(\hat{\beta})}{\mathcal{L}(0)}. \quad (12)$$

Such a measure has been derived to somehow mimic the R^2 in linear regression. However, in this case, it is not the square of anything. If the estimated model has the same log likelihood as the equal probability model, $\rho^2 = 0$. If the estimated model perfectly fits the data, that is if $\mathcal{L}(\hat{\beta}) = 0$, then $\rho^2 = 1$. As mentioned above, the value itself cannot be interpreted, and it must be used only to compare two models. In particular, unlike linear regression, it is possible to have a good model with a low value of ρ^2 , and a bad model with a high value.

An important limitation of this goodness of fit measure is that it is monotonic in the number of parameters of the model. It means that ρ^2 mechanically increases each time an additional variable is added to the model, even if this variable does not explain anything. Therefore, the following corrected measure is often preferred:

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

References

- Berndt, E. K., Hall, B. H., Hall, R. E. and Hausman, J. A. (1974). Estimation and inference in nonlinear structural models, *Annals of Economic and Social Measurement* **3/4**: 653–665.
- Kauermann, G. and Carroll, R. (2001). A note on the efficiency of sandwich covariance matrix estimation, *Journal of the American Statistical Association* **96**(456).
- White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**: 1–25.

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 4

My first model with Pythonbiogeme

Specification and estimation of the parameters

Michel Bierlaire

Introduction to choice models



The case study

Swissmetro

- ▶ a revolutionary mag-lev underground system in Switzerland,
- ▶ 500 km/h.



swissmetro.ch

Transportation mode choice

1. Train
2. Swissmetro
3. Car

The model

Variables

- ▶ Travel time: TRAIN_TT, SM_TT, CAR_TT
- ▶ Travel cost: TRAIN_CO, SM_CO, CAR_CO
- ▶ Yearly subscription: GA

Utility functions

- ▶ $ASC_TRAIN + B_TIME * TRAIN_TT + B_COST * TRAIN_CO * (GA = 0)$
- ▶ $B_TIME * SM_TT + B_COST * SM_CO * (GA = 0)$
- ▶ $ASC_CAR + B_TIME * CAR_TT + B_COST * CAR_CO$

Biogeme: an open-source software for estimating choice models – 4.2 Using Biogeme

Michel Bierlaire

Practice quiz: back to the simple example

Consider the simple example about the choice between purchasing an electric car or not, presented in Section 1.2. The data is summarized in the contingency table below

	Age		
	20–39	40–64	65+
Electric	65	55	5
Not electric	835	1045	495

The model has three parameters:

- $P(\text{electric} \mid \text{age } 20\text{--}39) = \pi_1$,
- $P(\text{electric} \mid \text{age } 40\text{--}64) = \pi_2$,
- $P(\text{electric} \mid \text{age } 65+) = \pi_3$.

We want to calculate the maximum likelihood estimates of these parameters using Biogeme. To do so, perform the following steps:

1. Prepare a data file, called for instance **small.dat**. The first row should contain the name of each variable. Make sure that these names do not contain blank spaces, and start with a letter. Remember that these names are case sensitive, so that “Electric” is not the same as “electric”. Each subsequent row should be associated with a cell of the contingency table. So the first question is: how many columns should this file contain, and what are the names of the corresponding variables?

2. Download the file `maxlike_question.py` (from the edX webpage) and use it as a template for the model specification file. Edit the file and include the formula for the contribution of each row of the data file to the log likelihood function.
3. Estimate the parameters π_1 , π_2 , π_3 using Biogeme.
4. Open the HTML output file and verify that the values obtained are the same as reported in the course.

Biogeme: an open-source software for estimating choice models – 4.2 Using Biogeme

Michel Bierlaire

Solution to the practice quiz: back to the simple example

1. Each cell of the contingency table contains three pieces of information: the age, the electric car ownership, and the number of corresponding individuals in the sample. Therefore, we define three headers in the data file: **Age**, **Electric** and **Number**. Download the file **small.dat** (from the edX webpage).
2. Download the model specification file **maxlike.py** (from the edX webpage) and read the included comments.
3. $\pi_1 = 0.0722$, $\pi_2 = 0.05$, $\pi_3 = 0.01$.
4. Download the output file **maxlike.html** (from the edX webpage).

Biogeme: an open-source software for estimating choice models – 4.2 Using Biogeme

Michel Bierlaire

Practice quiz: specification coding

Description

The goal of this exercise is to become familiar with the Python syntax in Biogeme and learn how to code various hypotheses regarding the factors influencing the choice for a given application of interest. In order to achieve this goal you will specify and estimate mode choice models for the *Swissmetro* case study. In each step of the exercise you will be asked to provide basic interpretations of the specifications you tested.

The data is provided in the file `swissmetro.dat` (from the edX webpage). The case study includes 3 alternatives, namely train, swissmetro and car. The choice variable is coded as 1 if the individual chose train, 2 if the individual chose swissmetro and 3 if the individual chose car. A choice variable equal to 0 indicates that we don't know what the individual chose (see `swissmetroDescription.pdf` (from the edX webpage) for a complete description of the data set.) In this exercise we consider a binary choice model between train and car. To do so, we exclude observations for which the **CHOICE** variable is equal to 0 or 2. Furthermore, we only consider work related trips. To do so, we exclude observations for which **PURPOSE** is different than 1 (commuter) or 3 (business). The following code is included in the model specification file:

```
exclude = (TRAIN_AV_SP == 0) + (CAR_AV_SP == 0) +  
( CHOICE == 0 ) + ( CHOICE == 2 ) +  
(( PURPOSE != 1 ) * ( PURPOSE != 3 )) > 0
```

```
BIOGEME_OBJECT.EXCLUDE = exclude
```

Note that the output of a logical operator is 1 if true and 0 if false. Therefore, the “+” acts as a “or” and the “*” acts as a “and” in the above formula.

1 Base model

Download and estimate the `v422_binary_SM_base.py` (from the edX webpage) file with the example model specification. It is a binary logit model between car and train. This is your base model. Use it as template to perform the following exercises.

2 Alternative specific attributes — I

Create a specification file `v422_binary_SM_specific.py` for a binary logit model with *alternative specific coefficients* for the cost variables of the two alternatives, i.e. car cost and train cost, in the utility functions of the car and train alternatives, respectively.

What behavioral assumption of the base model is relaxed by including the alternative specific parameters for the cost variables?

3 Alternative specific attributes — II

Copy the `v422_binary_SM_specific.py` into a new file called `v422_binary_SM_specificHeadway.py`. Edit this file to include the *train headway* `TRAIN_HE` in the utility function of the train alternative. The “headway” is the time separating the departure of two consecutive trains. It is actually the inverse of the frequency of the line. Estimate the model and answer the following questions:

1. What is the underlying behavioral assumption associated with the inclusion of that attribute?
2. Is the sign of the coefficient estimate consistent with your expectations? Why?

4 Socioeconomic characteristics — I

Copy the `v422_binary_SM_specificHeadway.py` into a new file called `v422_binary_SM_specificHeadwaySocioec.py`. Edit this file to include the following *interactions of socioeconomic variables* with the alternative specific constant (ASC), in the utility function of the train alternative:

1. define a variable **SENIOR** for people above the age of 65 and interact it with the ASC, and
2. interact the variable for people owning a Swiss annual season ticket **GA** with the ASC.

Estimate the model and answer the following questions:

1. What does the specification with the variables for (i) senior people and (ii) season ticket owners in the utility of the train alternative capture?
2. What does the sign of the parameter estimate for senior people reflect with respect to the preferences of older people in the sample?
3. Is the sign of the parameter estimate associated with the season ticket owners according to your expectations? Why?

5 Socioeconomic characteristics — II

Copy the `v422_binary_SM_specificHeadwaySocioec.py` into a new file called `v422_binary_SM_specificHeadwaySocioec2.py`. Edit this file to include the two variables of the previous exercise (**SENIOR** and **GA**) in both utilities and estimate the model once again. Answer the following questions:

1. What happens when you include the socioeconomic variables in both utilities?
2. Why does it happen?

Biogeme: an open-source software for estimating choice models – 4.2 Using Biogeme

Michel Bierlaire

Solution to practice quiz: specification coding

2 Alternative specific attributes — I

Download the model specification file

`v422_binary_SM_specific.py` (from the edX webpage)

and the corresponding output file

`v422_binary_SM_specific.html` (from the edX webpage)

The specification relaxes the assumption that the influence of the cost on the utility is the same for each alternative.

3 Alternative specific attributes — II

Download the model specification file

`v422_binary_SM_specificHeadway.py` (from the edX webpage)

and the corresponding output file

`v422.binary_SM_specificHeadway.html` (from the edX webpage)

1. The behavioral assumption is that the train headway actually influences the choice between car and train. We include the `TRAIN_HE` variable in the utility function of the train alternative and estimate the corresponding coefficient `B_HE`.
2. Yes. The negative estimate of the headway coefficient `B_HE` indicates that the higher the headway, i.e. the lower the frequency of service, the lower the utility of train.

4 Socioeconomic characteristics — I

Download the model specification file

`v422_binary_SM_specificHeadwaySocioec.py` (from the edX webpage)

and the corresponding output file

`v422_binary_SM_specificHeadwaySocioec.html` (from the edX webpage)

1. This specification associates a different alternative specific constant with different individuals in the sample, based on their socio-economic characteristics. We assume that (i) the age and (ii) the ownership of an annual season ticket for train have an influence on the choice. That is, (i) older people may have a preference towards a specific mode, and (ii) people with an annual season ticket for train have a preference towards train. We test this assumption by including the variables **SENIOR** and **GA** in the utility of the train alternative and estimating the corresponding coefficient **B_SENIOR** and **B_GA**. Note that, the observations, for which the variable **AGE** is unknown (coded as 6), are removed from the estimation.
2. The positive sign of the age coefficient **B_SENIOR** reflects a preference of older individuals for the *train* alternative with respect to car.
3. The coefficient related to the ownership of a **GA** is positive, as expected. It reflects a preference for the train alternative with respect to car for travelers possessing a season ticket.

5 Socioeconomic characteristics — II

Download the model specification file

`v422_binary_SM_specificHeadwaySocioec2.py` (from the edX webpage)

and the corresponding output file

`v422_binary_SM_specificHeadwaySocioec2.html` (from the edX webpage)

1. The model is unidentified. Note that Biogeme is able to estimate the model. The fact that it is unidentified is detected in the results by the huge standard errors associated with the coefficients **B_GA** and **B_SENIOR**. Also, at the very end of the output file, the model is reported to be unidentified, and Biogeme names the parameters causing problem.
2. This happens because only differences in utility matters. That is, if a constant is added to the utility of all alternatives, the alternative with the highest utility remains the same. As the two variables **SENIOR** and **GA** are individual specific, they do not change over the two alternatives in the choice set. Their effect cancels out when included in both utilities and therefore cannot be identified.

Biogeme: an open-source software for estimating choice models – 4.2 Using Biogeme

Michel Bierlaire

Practice quiz: logit and probit

The objective of this exercise is to build and estimate a binary *probit* model using Biogeme. You will continue working with the *Swissmetro* dataset provided in the file `swissmetro.dat` (from the edX webpage). You shall perform the following tasks:

1. Download the specification of the logit model: `v423_binaryLogitSM.py` (from the edX webpage)
2. Prepare the file `v423_binaryProbitSM.py` file for a binary probit model with the exact same specification of utility functions as the binary logit model. To do so, copy the file `v423_binaryLogitSM.py` into `v423_binaryProbitSM.py` and edit it. **Hint:** the Biogeme syntax for calculating the normal CDF of x is

`bioNormalCdf(x),`

where x is any valid Biogeme formula.

3. Estimate the parameters of the binary logit model.
4. Estimate the parameters of the binary probit model.
5. Can you directly compare the estimates of the parameters that you obtain from the binary probit model with the ones that you obtained from the binary logit? If not, what do you need to take under consideration, and *how*, when interpreting the model results?

Biogeme: an open-source software for estimating choice models – 4.2 Using Biogeme

Michel Bierlaire

Solution to the practice quiz: logit and probit

- The model specification file for the binary probit model is

`v423_binaryProbitSM.py` (from the edX webpage).

- The estimation results for logit are available in the file

`v423_binaryLogitSM.html` (from the edX webpage).

- The estimation results for probit are available in the file

`v423_binaryProbitSM.html` (from the edX webpage).

The normalization of the variance of the error terms must be taken under consideration when comparing the estimates obtained from the two models. The coefficients of the logit model are $\pi/\sqrt{3}$ larger than those of the probit model due to the difference in normalization. Therefore, the scale difference must be taken into account to correctly interpret the sensitivity to the cost and time attributes implied by the two models.

The parameter estimates of the two models are reported in Table 1, where the last column contains the estimates of the logit model divided by $\pi/\sqrt{3}$. These values can be compared with the estimates of the probit model. It is important to note that the scaled logit estimates and the probit estimates are not equal. They are simply comparable. The logit and the probit are two different models, based on different assumptions.

Another possibility to compare the parameters from one model to the next is to divide all the parameters by one of the parameters (typically, the cost parameter, so that the results can be interpreted in monetary units.) Note that, because of this ratio, the scales cancel out and the values can be compared.

		Logit	Probit	Logit / $(\pi/\sqrt{3})$
1	Car cte.	-1.24	-0.55	-0.684
2	Travel cost (car)	-2.40	-0.985	-1.32
3	Travel time (car)	-1.13	-0.651	-0.623
4	Headway	-0.00581	-0.00332	-0.00320
5	Travel cost (train)	-1.11	-0.543	-0.612
6	Travel time (train)	-0.394	-0.195	-0.217

Table 1: Probit and logit: comparison of the estimates

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 5

Choice with multiple alternatives

Derivation of the logit model

Michel Bierlaire

Introduction to choice models



The choice set

For all $i \in \mathcal{C}_n$

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

- ▶ What is \mathcal{C}_n ?
- ▶ What is ε_{in} ?
- ▶ What is V_{in} ?

Choice set

Universal choice set

- ▶ All potential alternatives for the population
- ▶ Restricted to relevant alternatives

Mode choice

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

Choice set

Individual's choice set

- ▶ 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

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

Choice model

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

Choice with multiple alternatives

Derivation of the logit model

Michel Bierlaire

Introduction to choice models



The error term

For all $i \in \mathcal{C}_n$

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

- ▶ What is \mathcal{C}_n ?
- ▶ What is ε_{in} ?
- ▶ What is V_{in} ?

Error terms

Logit: same assumptions as for binary logit

ε_{in} are

- ▶ independent and
- ▶ identically distributed,
- ▶ extreme value $EV(0, \mu)$.

Comments

- ▶ Independence: across i and n .
- ▶ Identical distribution: same scale parameter μ across i and n .
- ▶ For estimation, scale is normalized: $\mu = 1$.

Choice with multiple alternatives – 5.1

Derivation of the logit model

Michel Bierlaire

Mathematical derivation.

To derive the logit model, we consider the following ingredients:

- a choice set for each individual n : $\mathcal{C}_n = \{1, \dots, J_n\}$, and
- a utility function for each individual and each alternative: $U_{in} = V_{in} + \varepsilon_{in}$.

We assume that the error terms ε_{in} are

- independent, both across alternatives and individuals, and
- Extreme Value distributed, with the same parameters for each individual and each alternative:

$$\varepsilon_{in} \sim \text{EV}(0, \mu). \quad (1)$$

These assumptions are summarized by the statement “i.i.d. Extreme Value”, where “i.i.d.” stands for *independent and identically distributed*.

The choice model is

$$P(i|\mathcal{C}_n) = \Pr(V_{in} + \varepsilon_{in} \geq \max_{j=1, \dots, J_n} V_{jn} + \varepsilon_{jn}). \quad (2)$$

The idea of the derivation is to consider this model as a binary logit model, as we have already derived its specification. In order to be chosen, alternative i must have a utility larger than all other alternatives. Now, consider within the set $\mathcal{C}_n \setminus \{i\}$ composed of the other alternatives, the one associated with

the highest utility. We do not know which specific alternative achieves this, but we know that its utility is

$$U_n^* = \max_{j \in \mathcal{C}_n \setminus \{i\}} U_{in} = \max_{j \in \mathcal{C}_n \setminus \{i\}} (V_{jn} + \varepsilon_{jn}). \quad (3)$$

Therefore, the choice model can be written:

$$P(i|\mathcal{C}_n) = \Pr(U_{in} \geq U_n^*), \quad (4)$$

that involves only two alternatives. In order to derive the choice model, we need to know the distribution of U_n^* . From a property of the extreme value distribution (see property 6 in the appendix below), and the fact that all error terms are i.i.d., we have that

$$U_n^* \sim \text{EV} \left(\frac{1}{\mu} \ln \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}, \mu \right). \quad (5)$$

Equivalently (see property 4 in the appendix), we can write

$$U_n^* = V_n^* + \varepsilon_n^* \quad (6)$$

where

$$V_n^* = \frac{1}{\mu} \ln \sum_{j=2}^{J_n} e^{\mu V_{jn}} \quad (7)$$

and

$$\varepsilon_n^* \sim \text{EV}(0, \mu). \quad (8)$$

Consequently, (4) is a binary logit model and

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + e^{\mu V_n^*}} \quad (9)$$

where

$$V_n^* = \frac{1}{\mu} \ln \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}. \quad (10)$$

We have

$$e^{\mu V_n^*} = e^{\ln \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}} = \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}, \quad (11)$$

and (9) can be written

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{e^{\mu V_{in}} + \sum_{j \in \mathcal{C}_n \setminus \{i\}} e^{\mu V_{jn}}}, \quad (12)$$

to finally obtain

$$\frac{e^{\mu V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}}}. \quad (13)$$

This is the logit model. Interestingly, it is a straightforward extension of the binary logit model, where the sum at the denominator involves now all alternatives in the choice set.

Properties of the extreme value distribution

The extreme value distribution with location parameter η and scale parameter μ has the following properties:

1. The mode is η .
2. The mean is $\eta + \frac{\gamma}{\mu}$, where

$$\gamma = - \int_0^{+\infty} e^{-x} \ln x dx \approx 0.5772 \quad (14)$$

is Euler's constant.

3. The variance is $\frac{\pi^2}{6\mu^2}$.
4. If $\varepsilon \sim \text{EV}(\eta, \mu)$, then

$$a\varepsilon + b \sim \text{EV}(a\eta + b, \frac{\mu}{a}),$$

where $a, b \in \mathbb{R}$, $a > 0$.

5. If $\varepsilon_a \sim \text{EV}(\eta_a, \mu)$ and $\varepsilon_b \sim \text{EV}(\eta_b, \mu)$ are independent with the same scale parameter μ , then

$$\varepsilon = \varepsilon_a - \varepsilon_b \sim \text{Logistic}(\eta_a - \eta_b, \mu),$$

namely

$$f_{\varepsilon}(\xi) = \frac{\mu e^{-\mu(\xi-\eta_a+\eta_b)}}{(1 + e^{-\mu(\xi-\eta_a+\eta_b)})^2}, \quad (15)$$

$$F_{\varepsilon}(\xi) = \frac{1}{1 + e^{-\mu(\xi-\eta_a+\eta_b)}}, \quad \mu > 0, -\infty < \xi < \infty. \quad (16)$$

$$(17)$$

6. If $\varepsilon_i \sim \text{EV}(\eta_i, \mu)$, for $i = 1, \dots, J$, and ε_i are independent with the same scale parameter μ , then

$$\varepsilon = \max_{i=1, \dots, J} \varepsilon_i \sim \text{EV}(\eta, \mu) \quad (18)$$

where

$$\eta = \frac{1}{\mu} \ln \sum_{i=1}^J e^{\mu \eta_i}. \quad (19)$$

It is important to note that this property holds only if all ε_i have the same scale parameter μ . As ε follows an extreme value distribution, its expected value is

$$E[\varepsilon] = \eta + \frac{\gamma}{\mu}.$$

Equivalently,

$$\eta = E[\varepsilon] - \frac{\gamma}{\mu}.$$

Therefore, (19) provides the expected value of the maximum, up to a constant.

Choice with multiple alternatives – 5.1

Derivation of the logit model

Michel Bierlaire

Note on the scale parameter.

The logit model is

$$P(i|\mathcal{C}_n) = \frac{e^{\mu V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{\mu V_{jn}}}. \quad (1)$$

The scale parameter μ is not identified from data. If

$$V_{in} = \sum_k \beta_k x_{ink}, \quad (2)$$

the quantity involved in the logit model is

$$\mu V_{in} = \sum_k \mu \beta_k x_{ink}. \quad (3)$$

Only the products $\mu \beta_k$ are identified. It is therefore common to normalize the parameter μ to 1 and to estimate the coefficients β_k .

The fact that the scale parameter is not identified does not mean that it does not exist. This is particularly important to remember at the stage where the model is applied in a specific context.

It is interesting to investigate the extreme cases for the scale parameters.

When the value of μ goes to zero, that is when the variance of the error terms goes to infinity, the systematic parts of the utility V_{in} do not play a role anymore, and the model assigns the same probability to each alternative:

$$\lim_{\mu \rightarrow 0} P(i|\mathcal{C}_n; \mu) = \frac{1}{J_n}, \quad \forall i \in \mathcal{C}_n. \quad (4)$$

When the value of μ goes to infinity, that is when the variance of the error terms goes to zero, the model becomes fully deterministic:

$$\begin{aligned}\lim_{\mu \rightarrow \infty} P(i|C_n; \mu) &= \lim_{\mu \rightarrow \infty} \frac{1}{1 + \sum_{j \neq i} e^{\mu(V_{jn} - V_{in})}} \\ &= \begin{cases} 1 & \text{if } V_{in} > \max_{j \neq i} V_{jn}, \\ 0 & \text{if } V_{in} < \max_{j \neq i} V_{jn}. \end{cases} \end{aligned} \quad (5)$$

The above formula does not treat the case of ties. Ties do not matter in a probabilistic context, as the probability that they occur is zero. As this specific case is deterministic, ties matter. Suppose that the maximum utility is achieved by J_n^* alternatives, that is

$$V_{in} = \max_{j \in C_n} V_{jn}, \quad i = 1, \dots, J_n^*. \quad (6)$$

In that case, each of them has the same probability to be chosen, that is

$$\lim_{\mu \rightarrow \infty} P(i|C_n; \mu) = \frac{1}{J_n^*}, \quad i = 1, \dots, J_n^*, \quad (7)$$

and

$$\lim_{\mu \rightarrow \infty} P(i|C_n; \mu) = 0, \quad i = J_n^* + 1, \dots, J_n. \quad (8)$$

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Systematic part of the utility function

For all $i \in \mathcal{C}_n$

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

- ▶ What is \mathcal{C}_n ?
- ▶ What is ε_{in} ?
- ▶ What is V_{in} ?

Systematic part of the utility function

$$V_{in} = V(z_{in}, S_n)$$

- ▶ z_{in} is a vector of attributes of alternative i for individual n
- ▶ S_n is a vector of socio-economic characteristics of n

Functional form: linear utility

Notation

$$x_{in} = (z_{in}, S_n)$$

Linear-in-parameters utility functions

$$V_{in} = V(z_{in}, S_n) = V(x_{in}) = \sum_k \beta_k (x_{in})_k$$

Not as restrictive as it may seem

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Practice quiz: model

Consider the following model specification.

$$U_{\text{walk},n} = \text{ASC}_{\text{walk}} + \beta_{\text{distance}} \cdot \text{distance}_n + \varepsilon_{\text{walk},n} \quad (1)$$

$$U_{\text{bicycle},n} = \text{ASC}_{\text{bicycle}} + \beta_{\text{distance}} \cdot \text{distance}_n + \varepsilon_{\text{bicycle},n} \quad (2)$$

$$U_{\text{car},n} = \text{ASC}_{\text{car}} + \beta_{\text{time}} \cdot \text{time}_{\text{car},n} + \beta_{\text{cost}} \cdot \text{cost}_{\text{car},n} + \varepsilon_{\text{car},n} \quad (3)$$

$$U_{\text{bus},n} = \beta_{\text{time}} \cdot \text{time}_{\text{bus},n} + \beta_{\text{cost}} \cdot \text{cost}_{\text{bus},n} + \varepsilon_{\text{bus},n} \quad (4)$$

The values of the parameters are shown in Table 1.

Parameter	value
ASC_{walk}	-2.42
$\text{ASC}_{\text{bicycle}}$	-3.62
ASC_{car}	-4.55
β_{distance}	-4.53
β_{time}	-2.76
β_{cost}	0.25

Table 1: Estimation results

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Solution to the practice quiz: model

1 Model specification

1. $\mathcal{C} = \{\text{walk, bicycle, car, bus}\}$
2. $|\mathcal{C}_n| = 4 - 1 = 3$
3. Deterministic part of the utility functions:
 - $V_{\text{walk},n} = \text{ASC}_{\text{walk}} + \beta_{\text{distance}} \cdot \text{distance}_n$
 - $V_{\text{bicycle},n} = \text{ASC}_{\text{bicycle}} + \beta_{\text{distance}} \cdot \text{distance}_n$
 - $V_{\text{car},n} = \text{ASC}_{\text{car}} + \beta_{\text{time}} \cdot \text{time}_{\text{car},n} + \beta_{\text{cost}} \cdot \text{cost}_{\text{car},n}$
 - $V_{\text{bus},n} = \beta_{\text{time}} \cdot \text{time}_{\text{bus},n} + \beta_{\text{cost}} \cdot \text{cost}_{\text{bus},n}$
4. As no assumption about the distribution of the error terms has been specified, there is not enough information to know the type of model.
 - ☐ `logit`,
 - ☐ `probit`,
 - ☒ I don't know.

2 Model parameters

1. As the coefficient of travel is negative, the higher the travel time of an alternative, the **lower** its utility.

☐ ~~true~~,

☒ false.

2. As the coefficient of travel cost is positive, the higher the travel cost of an alternative, the higher its utility.

☒ true,

☐ ~~false~~.

3. As the coefficient of distance is negative, the higher the travel distance of an alternative, the **lower** its utility.

☐ ~~true~~,

☒ false.

4. We would expect the three variables (travel time, travel cost and distance) to be associated with negative coefficients. The positive cost coefficient implies that an increase of the cost of an alternative would increase its attractiveness. It is not consistent with our expectations.

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Quantitative explanatory variables

Quantitative attributes

Numerical and continuous

- ▶ $(z_{in})_k \in \mathbb{R}, \forall i, n, k$
- ▶ Associated with a specific unit
- ▶ Vary across both i and n .

Examples

- ▶ Auto in-vehicle time (in min.)
- ▶ Transit in-vehicle time (in min.)
- ▶ Auto out-of-pocket cost (in cents)
- ▶ Transit fare (in cents)
- ▶ Walking time to the bus stop (in min.)

Straightforward modeling

Quantitative attributes

- ▶ V_{in} is unitless
- ▶ Therefore, β depends on the unit of the associated attribute
- ▶ Example: consider two specifications

$$\begin{aligned}V_{in} &= \beta_1 TT_{in} + \dots \\V_{in} &= \beta'_1 TT'_{in} + \dots\end{aligned}$$

- ▶ If TT_{in} is a number of minutes, the unit of β_1 is 1/min
- ▶ If TT'_{in} is a number of hours, the unit of β'_1 is 1/hour
- ▶ Both models are equivalent, but the estimated value of the coefficient will be different

$$\beta_1 TT_{in} = \beta'_1 TT'_{in} \implies \frac{TT_{in}}{TT'_{in}} = \frac{\beta'_1}{\beta_1} = 60$$

Quantitative attributes

Generic vs alternative specific

$$\begin{aligned}V_{in} &= \beta_1 TT_{in} + \dots \\V_{jn} &= \beta_1 TT_{jn} + \dots\end{aligned}$$

or

$$\begin{aligned}V_{in} &= \beta_1 TT_{in} + \dots \\V_{jn} &= \beta_2 TT_{jn} + \dots\end{aligned}$$

Modeling assumption: a minute has/has not the same marginal utility whether it is incurred on the auto or bus mode

Quantitative socio-eco. characteristics

Numerical and continuous

- ▶ $(S_n)_k \in \mathbb{R}, \forall n, k$
- ▶ Associated with a specific unit
- ▶ Vary only across n , not across i .

Examples

- ▶ Annual income (in KCHF)
- ▶ Age (in years)

Modeling heterogeneity

Behavioral assumption

- ▶ Individuals have different taste parameters.
- ▶ The difference is explained by one socio-economic characteristic.

$$V_{in} = \beta_{1n}z_{in} + \dots$$

where

$$\beta_{1n} = \beta_{1n}(\text{income}_n).$$

Modeling heterogeneity

Interaction

Typical definition of β_{1n} :

$$\beta_{1n} = \beta_1 \text{income}_n$$

$$V_{in} = \beta_{1n} z_{in} + \dots = \beta_1 \text{income}_n z_{in} + \dots = \beta_1 x_{in} + \dots$$

where

$$x_{in} = \text{income}_n z_{in}$$

Modeling heterogeneity

Behavioral assumption

- ▶ Individuals have different taste parameters.
- ▶ The difference is explained by **several** socio-economic characteristics.

$$V_{in} = \beta_{1n}z_{in} + \dots$$

where

$$\beta_{1n} = \beta_{1n}(\text{income}_n, \text{age}_n).$$

Modeling heterogeneity

Interaction

Typical definition of β_{1n} :

$$\beta_{1n} = \beta_1 \text{income}_n \text{age}_n$$

$$V_{in} = \beta_{1n} z_{in} + \dots = \beta_1 \text{income}_n \text{age}_n z_{in} + \dots = \beta_1 x_{in} + \dots$$

where

$$x_{in} = \text{income}_n \text{age}_n z_{in}$$

Modeling heterogeneity

Creativity and relevance

- ▶ Several functional forms can be investigated.
- ▶ For instance, if z_{in} is the cost variables, we write

$$\beta_{cn} = \beta_c / \text{income}_n$$

- ▶ Indeed, in this case, the new variable can be interpreted as the share of the income dedicated to this purchase:

$$x_{in} = z_{in} / \text{income}_n$$

Modeling heterogeneity: alternative specific constants

ASCs can also vary across individuals

Base model

$$\begin{aligned}V_{1n} &= \beta_x x_{1n1} + \beta_1 + \dots \\V_{2n} &= \beta_x x_{2n1} + \beta_2 + \dots \\V_{3n} &= \beta_x x_{3n1} + \dots\end{aligned}$$

Heterogeneous specification

$$\begin{aligned}V_{1n} &= \beta_x x_{1n1} + \beta_{1n} + \dots \\V_{2n} &= \beta_x x_{2n1} + \beta_{2n} + \dots \\V_{3n} &= \beta_x x_{3n1} + \dots\end{aligned}$$

where

$$\beta_{in} = \beta_i \text{income}_n$$

Modeling heterogeneity: alternative specific constants

Heterogeneous specification

$$\begin{aligned}V_{1n} &= \beta_x x_{1n1} + \beta_{1n} + \dots \\V_{2n} &= \beta_x x_{2n1} + \beta_{2n} + \dots \\V_{3n} &= \beta_x x_{3n1} + \dots\end{aligned}$$

where

$$\beta_{in} = \beta_i \text{income}_n$$

$$\begin{aligned}V_{1n} &= \beta_x x_{1n1} + \beta_1 \text{income}_n + \dots \\V_{2n} &= \beta_x x_{2n1} + \beta_2 \text{income}_n + \dots \\V_{3n} &= \beta_x x_{3n1} + \dots\end{aligned}$$

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Practice quiz: parameters

Consider a mode choice model with three alternatives: (i) bicycle, (ii) walk and (iii) metro. The specification includes alternative specific constants (ASCs) and a coefficient associated with the effect of travel time (β_{time}):

$$\begin{aligned} V_{\text{bicycle},n} &= \text{ASC}_{\text{bicycle}} + \beta_{\text{time}} \cdot \text{travel time}_{\text{bicycle},n}, \\ V_{\text{walk},n} &= \text{ASC}_{\text{walk}} + \beta_{\text{time}} \cdot \text{travel time}_{\text{walk},n}, \\ V_{\text{metro},n} &= \text{ASC}_{\text{metro}} + \beta_{\text{time}} \cdot \text{travel time}_{\text{metro},n}. \end{aligned}$$

The travel time attribute in the dataset is expressed in minutes. The alternative specific constant of the bicycle alternative has been normalized to zero, and the value of the other parameters estimated from data. The estimates are:

- ASC_{walk} : -0.01,
- $\text{ASC}_{\text{metro}}$: 0.2,
- β_{time} : -0.05.

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Solution to the practice quiz: parameters

1. In order to obtain an equivalent model where ASC_{walk} is zero, each constant must be increased by 0.01. The coefficient β_{time} is not affected.
 - ASC_{bicycle} : 0.01,
 - ASC_{walk} : 0,
 - ASC_{metro} : 0.21
 - β_{time} : -0.05.
2. The coefficient of travel time must be multiplied by 60.
 - ASC_{bicycle} : 0,
 - ASC_{walk} : -0.01,
 - ASC_{metro} : 0.2,
 - β_{time} : -3.

For instance, for a trip of one hour, the contribution of travel time to the utility is -3 for both models.

3. The socio-economic characteristic is interacted with the alternative specific constant, but is not normalized in this model. Increasing all $\beta_{\text{man},i}$ parameters by the same quantity does not change the choice probability and, therefore, the log likelihood. Consequently, there is an infinite number of maxima of the log likelihood function, and the second derivative matrix is singular at the solution. A proper specification would be

$$\begin{aligned}
V_{\text{bicycle},n} &= \text{ASC}_{\text{bicycle}} + \beta_{\text{time}} \cdot \text{travel time}_{\text{bicycle},n} \\
V_{\text{walk},n} &= \text{ASC}_{\text{walk}} + \beta_{\text{time}} \cdot \text{travel time}_{\text{walk},n} + \beta_{\text{man,walk}} \cdot \text{man}_n, \\
V_{\text{metro},n} &= \text{ASC}_{\text{metro}} + \beta_{\text{time}} \cdot \text{travel time}_{\text{metro},n} + \beta_{\text{man,metro}} \cdot \text{man}_n.
\end{aligned}$$

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Qualitative explanatory variables

Qualitative attributes

Examples

- ▶ Level of comfort for the train
- ▶ Reliability of the bus
- ▶ Color
- ▶ Shape
- ▶ etc...

Modeling

Identify all possible levels of the variable

- ▶ Very comfortable,
- ▶ Comfortable,
- ▶ Rather comfortable,
- ▶ Not comfortable.

Select a base level

- ▶ Very comfortable,
- ▶ Comfortable,
- ▶ Rather comfortable,
- ▶ Not comfortable.

Modeling

Introduce a 0/1 attribute for all levels except the base case

- ▶ z_c for comfortable
- ▶ z_{rc} for rather comfortable
- ▶ z_{nc} for not comfortable

	z_c	z_{rc}	z_{nc}
very comfortable	0	0	0
comfortable	1	0	0
rather comfortable	0	1	0
not comfortable	0	0	1

If a qualitative attribute has K levels, we introduce $K - 1$ binary variables (0/1) in the model

Modeling

Utility function

$$V_{in} = \beta_c z_c + \beta_{rc} z_{rc} + \beta_{nc} z_{nc} + \dots$$

Note

The choice of the base level is arbitrary.

Qualitative characteristics

Examples

- ▶ Sex
- ▶ Education
- ▶ Professional status
- ▶ etc.

Modeling heterogeneity

Behavioral assumption

- ▶ Individuals have different taste parameters.
- ▶ The difference is explained by a qualitative socio-economic characteristic.

$$V_{in} = \beta_{1n}z_{in} + \dots$$

where

$$\beta_{1n} = \beta_{1n}(\text{education}_n).$$

Modeling heterogeneity

Segmentation

- ▶ Assume that there are K levels for the qualitative variable (e.g. education).
- ▶ They characterize K segments in the population.
- ▶ Define

$$\delta_{kn} = \begin{cases} 1 & \text{if individual } n \text{ is associated with level } k \\ 0 & \text{otherwise} \end{cases}$$

- ▶ Introduce a parameter β_1^k for each level and define

$$\beta_{1n} = \sum_{k=1}^K \beta_1^k \delta_{kn}$$

Modeling heterogeneity

Segmentation

$$V_{in} = \beta_{1n}z_{in} + \cdots = \sum_{k=1}^K \beta_1^k \delta_{kn} z_{in} + \cdots = \sum_{k=1}^K \beta_1^k x_{ink} + \cdots$$

where

$$x_{ink} = \delta_{kn} z_{in}$$

Segmentation with several variables

Example

- ▶ Gender (M,F)
- ▶ House location (metro, suburb, perimeter areas)
- ▶ 6 segments: (M, m) , (M, s) , (M, p) , (F, m) , (F, s) , (F, p) .

Segmentation

Specification

$$\beta_{M,m} TT_{M,m} + \beta_{M,s} TT_{M,s} + \beta_{M,p} TT_{M,p} + \\ \beta_{F,m} TT_{F,m} + \beta_{F,s} TT_{F,s} + \beta_{F,p} TT_{F,p} +$$

$TT_i = TT$ if indiv. belongs to segment i , and 0 otherwise

Remarks

- ▶ For a given individual, exactly one of these terms is non zero.
- ▶ The number of segments grows exponentially with the number of variables.

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Practice quiz: qualitative variables

Consider a mode choice model between car and metro. The specification includes an alternative specific constant for metro, the travel time for each alternative, and the level of comfort for metro. The level of comfort is a discrete variable that can take four values: *very comfortable*, *comfortable*, *rather comfortable* and *not comfortable*. Using the level *very comfortable* as the base case, it is included in the utility function through three dummy variables $z_{c,n}$, $z_{rc,n}$ and $z_{nc,n}$ defined as follows:

Level of comfort for n	$z_{c,n}$	$z_{rc,n}$	$z_{nc,n}$
very comfortable	0	0	0
comfortable	1	0	0
rather comfortable	0	1	0
not comfortable	0	0	1

The model specification is

$$\begin{aligned}
 V_{car,n} &= \beta_{time} \cdot \text{Travel time}_{car,n} \\
 V_{metro,n} &= ASC_{metro} + \beta_{time} \cdot \text{Travel time}_{car,n} + \beta_c \cdot z_{c,n} + \beta_{rc} \cdot z_{rc,n} + \beta_{nc} \cdot z_{nc,n}.
 \end{aligned}$$

The estimates of the parameters are

- ASC_{metro} : 0.55,
- β_{time} : -0.231,
- β_c : -0.90,
- β_{rc} : -1.00,

- β_{nc} : -2.00 .

Consider now the exact same model, where the level *comfortable* is considered as the base case for the comfort variable.

1. Define the dummy variables and their coding.
2. Write the model specification.
3. Provide the estimates of the parameters.

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Solution to the practice quiz: qualitative variables

1. We use three dummy variables $z_{vc,n}$, $z_{rc,n}$ and $z_{nc,n}$ defined as follows:

Level of comfort for n	$z_{vc,n}$	$z_{rc,n}$	$z_{nc,n}$
very comfortable	1	0	0
comfortable	0	0	0
rather comfortable	0	1	0
not comfortable	0	0	1

2. The model specification is

$$\begin{aligned}
 V_{car,n} &= \beta_{time} \cdot \text{Travel time}_{car,n} \\
 V_{metro,n} &= ASC_{metro} + \beta_{time} \cdot \text{Travel time}_{car,n} + \beta_{vc} \cdot z_{vc,n} + \beta_{rc} \cdot z_{rc,n} + \beta_{nc} \cdot z_{nc,n}.
 \end{aligned}$$

3. The estimates of the parameters are

- ASC_{metro} : 0.55,
- β_{time} : -0.231,
- β_{vc} : 0.90,
- β_{rc} : -0.10,
- β_{nc} : -1.10.

Only the coefficients of the dummy variables have changed. Considering that β_{vc} has been normalized to zero in the first model, and β_c in the second, their value has increased by 0.90, so that the coefficient for the level “comfortable” becomes 0.

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Nonlinear specifications: data preprocessing

Behavioral motivation

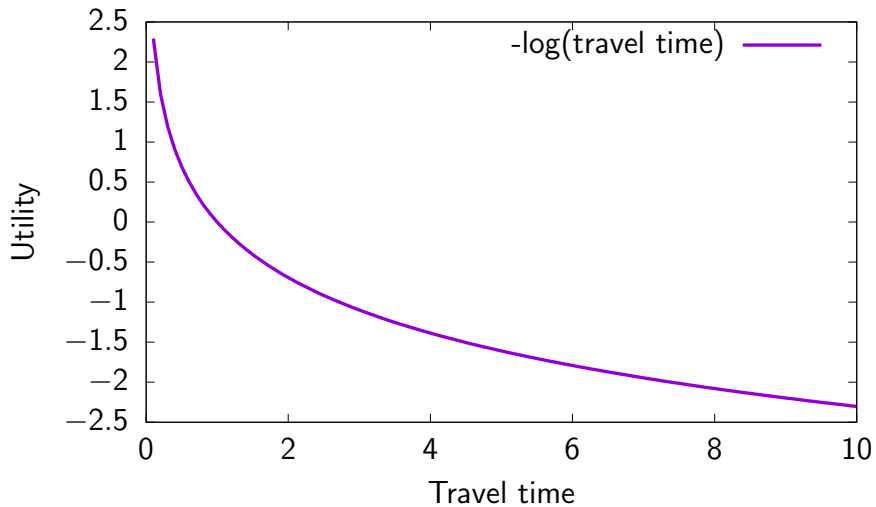
Example with travel time

- ▶ Compare a trip of 5 min with a trip of 10 min
- ▶ Compare a trip of 120 min with a trip of 125 min
- ▶ Utility difference: $\beta_T \times 5$ min, in both cases.

Behavioral assumption

One more minute of travel is not perceived the same way for short trips as for long trips

Behavioral motivation



Nonlinear transformations of the variables

Assumption 1: the marginal impact of travel time is constant

$$V_{in} = \beta_T \text{time}_{in} + \dots$$

Assumption 2: the marginal impact of travel time decreases with travel time

$$V_{in} = \beta_T \ln(\text{time}_{in}) + \dots$$

Remarks

- ▶ It is still a linear-in-parameters form
- ▶ The unit, the value, and the interpretation of β_T are different

Nonlinear transformations of the variables

Data can be preprocessed to account for nonlinearities

$$V_{in} = V(h(z_{in}, S_n)) = \sum_k \beta_k (h(z_{in}, S_n))_k$$

It is linear-in-parameter, even with h nonlinear.

Note

Interactions between attributes and socio-economic characteristics are a special case of h

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



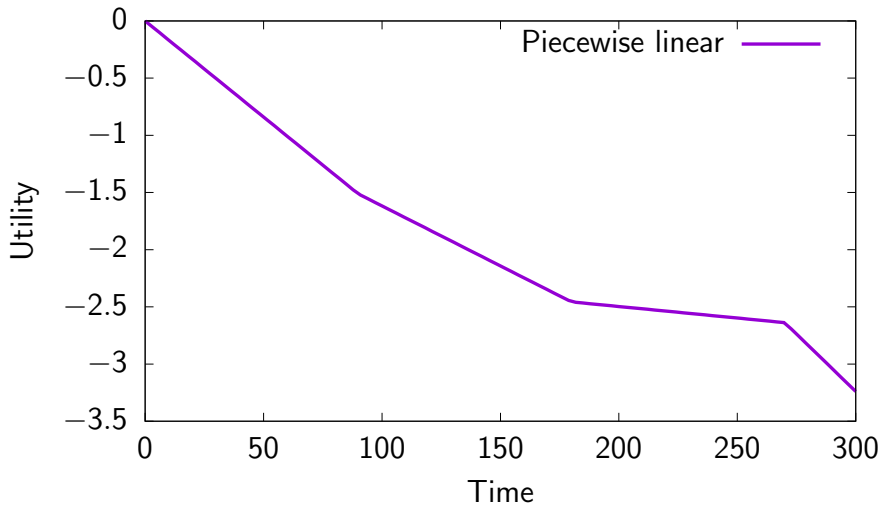
Nonlinear specifications: piecewise linear specification

Piecewise linear specification

Again: sensitivity to travel time varies with travel time

- ▶ Log transform is not the only specification
- ▶ Another possibility: split the range of values of the variable
 - ▶ Short trips: 0–90 min.
 - ▶ Medium strips: 90–180 min.
 - ▶ Long trips: 180–270 min.
 - ▶ Very long trips: 270 min. and more
- ▶ Each category is associated with a different coefficient.

Piecewise linear specification

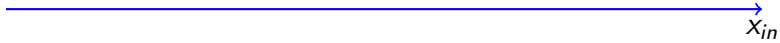


Piecewise linear specification

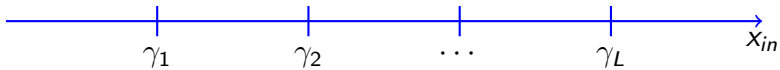
Procedure

- ▶ Select breakpoints $\gamma_1 < \gamma_2 < \dots < \gamma_L$
- ▶ Define new variables

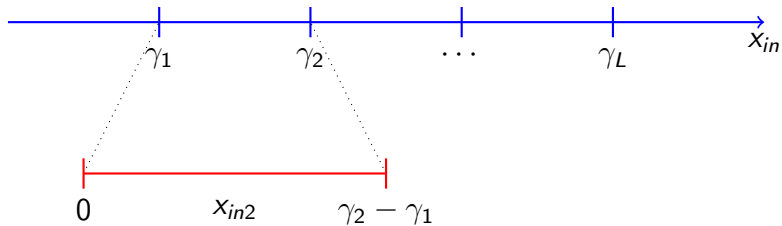
Piecewise linear specification



Piecewise linear specification



Piecewise linear specification



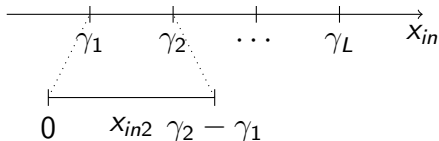
Piecewise linear specification

Formulation

$$x_{in1} = \begin{cases} x_{in} & \text{if } x_{in} < \gamma_1 \\ \gamma_1 & \text{otherwise} \end{cases}$$

$$x_{in\ell} = \begin{cases} 0 & \text{if } x_{in} < \gamma_{\ell-1} \\ x_{in} - \gamma_{\ell-1} & \text{if } \gamma_{\ell-1} \leq x_{in} < \gamma_{\ell} \\ \gamma_{\ell} - \gamma_{\ell-1} & \text{otherwise} \end{cases}$$

$$x_{inL} = \begin{cases} 0 & \text{if } x_{in} < \gamma_L \\ x_{in} - \gamma_L & \text{otherwise} \end{cases}$$



Piecewise linear specification

Equivalent formulations

$$x_{in1} = \begin{cases} x_{in} & \text{if } x_{in} < \gamma_1 \\ \gamma_1 & \text{otherwise} \end{cases}$$

$$x_{in1} = \min(x_{in}, \gamma_1)$$

$$x_{in\ell} = \begin{cases} 0 & \text{if } x_{in} < \gamma_{\ell-1} \\ x_{in} - \gamma_{\ell-1} & \text{if } \gamma_{\ell-1} \leq x_{in} < \gamma_{\ell} \\ \gamma_{\ell} - \gamma_{\ell-1} & \text{otherwise} \end{cases}$$

$$x_{in\ell} = \max(0, \min(x_{in} - \gamma_{\ell-1}, \gamma_{\ell} - \gamma_{\ell-1}))$$

$$x_{inL} = \begin{cases} 0 & \text{if } x_{in} < \gamma_L \\ x_{in} - \gamma_L & \text{otherwise} \end{cases}$$

$$x_{inL} = \max(0, x_{in} - \gamma_L)$$

Piecewise linear specification

Examples

$\gamma_1 = 90, \gamma_2 = 180, \gamma_3 = 270$.

x_{in}	50	100	200	300
x_{in1}	50	90	90	90
x_{in2}	0	10	90	90
x_{in3}	0	0	20	90
x_{in4}	0	0	0	30

$\gamma_1 = 1, \gamma_2 = 5, \gamma_3 = 10$.

x_{in}	0.5	4	8	12
x_{in1}	0.5	1	1	1
x_{in2}	0	3	4	4
x_{in3}	0	0	3	5
x_{in4}	0	0	0	2

Utility function

$$V_{in} = \sum_{\ell=1}^L \beta_{\ell} x_{in\ell}$$

Box-Cox transforms

Box and Cox (1964)

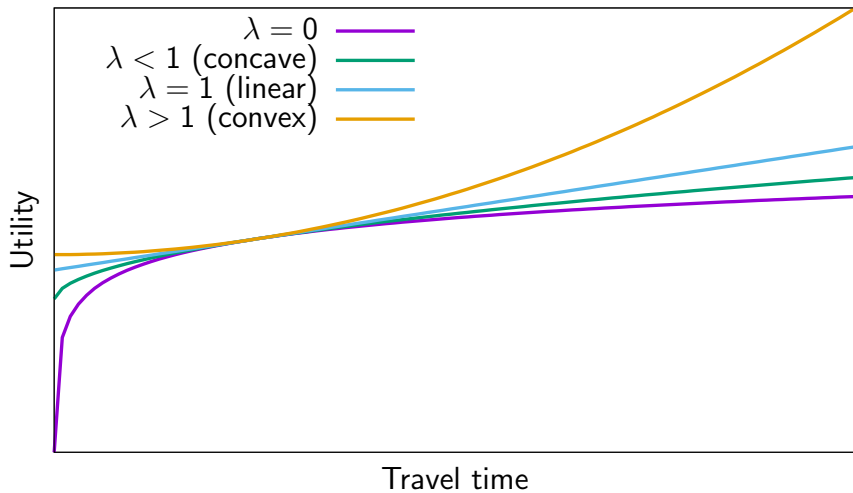
$$V_{in} = \beta x_{in}(\lambda) + \dots$$

where

$$x_{in}(\lambda) = \begin{cases} \frac{x_{in}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x_{in} & \text{if } \lambda = 0. \end{cases}$$

and $x_{in} > 0$.

Box-Cox transforms



Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Nonlinear specifications: heterogeneity

Heterogeneity

Interaction

$$V_{in} = \beta_{in} z_{in} + \dots$$

Linear interaction

$$\beta_{in} = \beta \text{ income}_n$$

Nonlinear interaction

$$\beta_{in} = \beta \text{ income}_n^\lambda, \text{ where } \lambda = \frac{\partial \beta_{in}}{\partial \text{income}_n} \frac{\text{income}_n}{\beta_{in}}$$

Nonlinear interactions

Remarks

- ▶ λ must be estimated
- ▶ Utility is not linear-in-parameters anymore
- ▶ Use a reference value for the socio-economic characteristic:

$$\beta_{in} = \beta \left(\frac{\text{income}_n}{\text{refIncome}} \right)^\lambda$$

- ▶ Reference value is arbitrary
- ▶ Several (continuous) characteristics can be combined:

$$\beta_{in} = \beta \left(\frac{\text{income}_n}{\text{refIncome}} \right)^{\lambda_1} \left(\frac{\text{age}_n}{\text{refAge}} \right)^{\lambda_2}$$

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Nonlinear specifications: heteroscedasticity

Heteroscedasticity

Logit is homoscedastic

- ▶ ε_{in} i.i.d. across both i and n .
- ▶ In particular, they all have the same variance.

Motivation

- ▶ People may have different level of knowledge (e.g. taxi drivers)
- ▶ Different sources of data

Heteroscedasticity

Data

- ▶ G groups in the population.
- ▶ Each individual n belongs to exactly one group g .
- ▶ Characterized by indicators:

$$\delta_{ng} = \begin{cases} 1 & \text{if } n \text{ belongs to } g, \\ 0 & \text{otherwise} \end{cases}$$

and $\sum_g \delta_{ng} = 1$, for all n .

Heteroscedasticity

Assumption: variance of error terms is different across groups

Consider individual n_1 belonging to group 1, and individual n_2 belonging to group 2.

$$\begin{aligned}U_{in_1} &= V_{in_1} + \varepsilon_{in_1} \\U_{in_2} &= V_{in_2} + \varepsilon_{in_2}\end{aligned}$$

and $\text{Var}(\varepsilon_{in_1}) \neq \text{Var}(\varepsilon_{in_2})$

Modeling

Without loss of generality:

$$\text{Var}(\varepsilon_{in_1}) = \alpha_2^2 \text{Var}(\varepsilon_{in_2})$$

Heteroscedasticity

Modeling: scale parameters

$$\begin{aligned} U_{in_1} &= V_{in_1} + \varepsilon_{in_1} = V_{in_1} + \varepsilon'_{in_1} \\ \alpha_2 U_{in_2} &= \alpha_2 V_{in_2} + \alpha_2 \varepsilon_{in_2} = \alpha_2 V_{in_2} + \varepsilon'_{in_2} \end{aligned}$$

Variance

$$\text{Var}(\varepsilon'_{in_2}) = \text{Var}(\alpha_2 \varepsilon_{in_2})$$

Heteroscedasticity

Modeling: scale parameters

$$\begin{aligned}U_{in_1} &= V_{in_1} + \varepsilon_{in_1} = V_{in_1} + \varepsilon'_{in_1} \\ \alpha_2 U_{in_2} &= \alpha_2 V_{in_2} + \alpha_2 \varepsilon_{in_2} = \alpha_2 V_{in_2} + \varepsilon'_{in_2}\end{aligned}$$

Variance

$$\begin{aligned}\text{Var}(\varepsilon'_{in_2}) &= \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \alpha_2^2 \text{Var}(\varepsilon_{in_2})\end{aligned}$$

Heteroscedasticity

Modeling: scale parameters

$$\begin{aligned}U_{in_1} &= V_{in_1} + \varepsilon_{in_1} = V_{in_1} + \varepsilon'_{in_1} \\ \alpha_2 U_{in_2} &= \alpha_2 V_{in_2} + \alpha_2 \varepsilon_{in_2} = \alpha_2 V_{in_2} + \varepsilon'_{in_2}\end{aligned}$$

Variance

$$\begin{aligned}\text{Var}(\varepsilon'_{in_2}) &= \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \alpha_2^2 \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \text{Var}(\varepsilon_{in_1})\end{aligned}$$

Heteroscedasticity

Modeling: scale parameters

$$\begin{aligned}U_{in_1} &= V_{in_1} + \varepsilon_{in_1} = V_{in_1} + \varepsilon'_{in_1} \\ \alpha_2 U_{in_2} &= \alpha_2 V_{in_2} + \alpha_2 \varepsilon_{in_2} = \alpha_2 V_{in_2} + \varepsilon'_{in_2}\end{aligned}$$

Variance

$$\begin{aligned}\text{Var}(\varepsilon'_{in_2}) &= \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \alpha_2^2 \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \text{Var}(\varepsilon_{in_1}) \\ &= \text{Var}(\varepsilon'_{in_1})\end{aligned}$$

Heteroscedasticity

Modeling: scale parameters

$$\begin{aligned}U_{in_1} &= V_{in_1} + \varepsilon_{in_1} = V_{in_1} + \varepsilon'_{in_1} \\ \alpha_2 U_{in_2} &= \alpha_2 V_{in_2} + \alpha_2 \varepsilon_{in_2} = \alpha_2 V_{in_2} + \varepsilon'_{in_2}\end{aligned}$$

Variance

$$\begin{aligned}\text{Var}(\varepsilon'_{in_2}) &= \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \alpha_2^2 \text{Var}(\alpha_2 \varepsilon_{in_2}) \\ &= \text{Var}(\varepsilon_{in_1}) \\ &= \text{Var}(\varepsilon'_{in_1})\end{aligned}$$

ε'_{in_1} and ε'_{in_2} can be assumed i.i.d.

Heteroscedasticity

Modeling: utility function

$$\mu_n V_{in} + \varepsilon_{in}$$

where

$$\mu_n = \sum_{g=1}^G \delta_{ng} \alpha_g$$

and α_g , $g = 1, \dots, G$ are unknown parameters to be estimated from data.

Remarks

- ▶ Even if $V_{in} = \sum_j \beta_j x_{jin}$ is linear-in-parameters, $\mu_n V_{in} = \sum_j \mu_n \beta_j x_{jin}$ is not.
- ▶ Normalization: one α_g must be normalized.

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Practice quiz

Consider two variables involved in a choice model:

- c_{in} , the price of alternative i for individual n , and,
- I_n , the monthly income of individual n .

You are asked to propose various specifications for the deterministic part of the utility function of alternative i and decision-maker n including the price variable. It is important that the utility function is always continuous in c_{in} and I_n .

1. The impact of the price on the utility function is proportional to its value.
2. The marginal effect of price on the utility function varies with price.
3. The impact of the price on the utility function is proportional, but the factor of proportionality is different for prices below and above 25 CHF.
4. The impact of price on the utility function varies nonlinearly with income.

Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Solution to the practice quiz

1. It is the classical linear specification:

$$V_{in} = \dots + \beta_c c_{in} + \dots .$$

2. The derivative of the utility function with respect to price should depend on price. Any nonlinear specification would do. A typical specification involves the log:

$$V_{in} = \dots + \beta_c \log(c_{in}) + \dots .$$

3. It is a piecewise linear specification:

$$V_{in} = \dots + \beta_{<25} c_{in,1} + \beta_{\geq 25} c_{in,2} + \dots ,$$

where

$$c_{in,1} = \begin{cases} c_{in} & \text{if } c_{in} < 25 \\ 25 & \text{otherwise} \end{cases}$$

and

$$c_{in,2} = \begin{cases} 0 & \text{if } c_{in} < 25 \\ c_{in} - 25 & \text{otherwise} \end{cases}$$

4. The beta coefficient must depend nonlinearly on income. Classical specifications include:

$$V_{in} = \dots \beta_c \log(I_n) c_{in} \dots ,$$

and

$$V_{in} = \cdots \beta_c \left(\frac{I_n}{I_n^{ref}} \right)^\lambda c_{in} \cdots .$$

Choice with multiple alternatives

Specification of the deterministic part

Michel Bierlaire

Introduction to choice models



Nonlinear specifications: Box-Cox transforms

Box-Cox transforms

Box and Cox (1964)

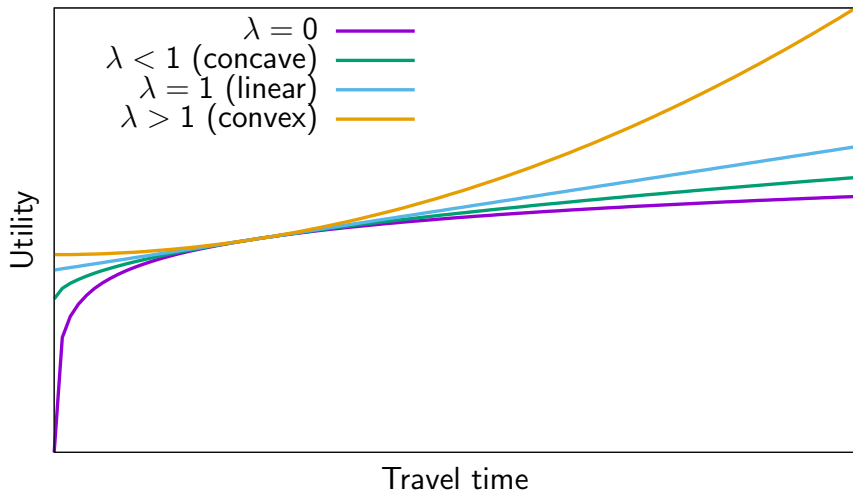
$$V_{in} = \beta x_{in}(\lambda) + \dots$$

where

$$x_{in}(\lambda) = \begin{cases} \frac{x_{in}^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln x_{in} & \text{if } \lambda = 0. \end{cases}$$

and $x_{in} > 0$.

Box-Cox transforms



Choice with multiple alternatives – 5.2

Specification of the deterministic part

Michel Bierlaire

Box-Cox transforms

The Box-Cox transform of a positive variable x , introduced by Box and Cox (1964), is defined as

$$x(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0. \end{cases} \quad (1)$$

Note that

$$\lim_{\lambda \rightarrow 0} \frac{x^\lambda - 1}{\lambda} = \log x, \quad (2)$$

so that $x(\lambda)$ is continuous [Verify]. It can be embedded in the specification of a utility function:

$$V_{in} = \beta_k x_{ink}(\lambda) + \dots, \quad (3)$$

where both β_k and λ are estimated from data. Such a specification is not linear-in-parameters. Its flexibility allows to let the data tell if the variable is involved in a linear way ($\lambda = 1$), a logarithmic way ($\lambda = 0$) or as a power law.

If the variable x may take negative values, Box and Cox (1964) propose to shift it before the transform is applied:

$$x(\lambda, \alpha) = \begin{cases} \frac{(x + \alpha)^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x + \alpha) & \text{if } \lambda = 0, \end{cases} \quad (4)$$

where $\alpha > -x$.

There are other ways to impose the positivity of the argument of the transform. For instance, Manly (1976) suggests to use an exponential:

$$x(\lambda) = \begin{cases} \frac{e^{x\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ x & \text{if } \lambda = 0, \end{cases} \quad (5)$$

while John and Draper (1980) propose to use the absolute value:

$$x(\lambda) = \begin{cases} \text{sign}(x) \frac{(|x|+1)^\lambda-1}{\lambda} & \text{if } \lambda \neq 0 \\ \text{sign}(x) \log(|x|+1) & \text{if } \lambda = 0. \end{cases} \quad (6)$$

A more complex transform has been proposed by Yeo and Johnson (2000):

$$x(\lambda) = \begin{cases} \frac{(x+1)^\lambda-1}{\lambda} & \text{if } \lambda \neq 0, x \geq 0; \\ \log(x+1) & \text{if } \lambda = 0, x \geq 0; \\ \frac{(1-x)^{2-\lambda}-1}{\lambda-2} & \text{if } \lambda \neq 2, x < 0; \\ -\log(1-x) & \text{if } \lambda = 2, x < 0. \end{cases} \quad (7)$$

Plenty of references are available in the literature. We refer the reader to Sakia (1992) for a review, and to Zarembka (1990) for a discussion in terms of model specification.

References

- Box, G. E. and Cox, D. R. (1964). An analysis of transformations, *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 211–252.
- John, J. and Draper, N. (1980). An alternative family of transformations, *Applied Statistics* pp. 190–197.
- Manly, B. (1976). Exponential data transformations, *The Statistician* pp. 37–42.
- Sakia, R. M. (1992). The box-cox transformation technique: A review, *Journal of the Royal Statistical Society. Series D (The Statistician)* **41**(2): 169–178.
URL: <http://www.jstor.org/stable/2348250>

- Yeo, I.-K. and Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry, *Biometrika* **87**(4): 954–959.
- Zarembka, P. (1990). Transformation of variables in econometrics, *Econometrics*, Springer, pp. 261–264.

Choice with multiple alternatives – 5.3 An example: mode choice in Switzerland

Michel Bierlaire

Description of the data.

This example deals with the estimation of a mode choice behavior model for inhabitants in Switzerland using revealed preference data. The survey was conducted between 2009 and 2010 for CarPostal, the public transport branch of the Swiss Postal Service. The main purpose of this survey is to collect data for analyzing the travel behavior of people in low-density areas, where CarPostal typically serves. A following study proposes new public transport alternatives according to the respondents' willingness to pay for these potential services in order to increase the market share of public transport.

Data collection

The survey covers French and German speaking areas of Switzerland. Questionnaires were sent to people living in rural area by mail. The respondents were asked to register all the trips performed during a specified day. The collected information consists of origin, destination, cost, travel time, chosen mode and activity at the destination. Moreover, we collected socio-economic information about the respondents and their households.

1124 completed surveys were collected. For each respondent, cyclic sequences of trips (starting and ending at the same location) are detected and their main transport mode is identified. The resulting data base includes 1906 sequences of trips linked with psychometric indicators and socio-economic attributes of the respondents. It should be noticed that each observation is a sequence of trips that starts and ends at home. A respondent may have several sequences of trips in a day.

Variables and descriptive statistics

The variables are described in Tables 1, 2, 3, 4 and 5. The attitudinal statements are written in Tables 6 and 7. A summary of descriptive statistics for the main variables is given in Table 8.

Given the presence of missing data (coded as -1) an additional table summarizing the three main affected variables (TripPurpose, ReportedDuration, age) after removing the missing cases is presented (see Table 9).

We refer the reader to Atasoy et al. (2013) for an analysis based on this data set.

References

Atasoy, B., Glerum, A. and Bierlaire, M. (2013). Attitudes towards mode choice in switzerland, *disP - The Planning Review* **49**(2): 101–117.

Table 1: Description of variables

Name	Description
ID	Identifier of the respondent who described the trips in the loop.
NbTransf	The total number of transfers performed for all trips of the loop, using public transport (ranging from 1-9).
TimePT	The duration of the loop performed in public transport (in minutes).
WalkingTimePT	The total walking time in a loop performed in public transports (in minutes).
WaitingTimePT	The total waiting time in a loop performed in public transports (in minutes).
TimeCar	The total duration of a loop made using the car (in minutes).
CostPT	Cost for public transports (full cost to perform the loop).
MarginalCostPT	The total cost of a loop performed in public transports, taking into account the ownership of a seasonal ticket by the respondent. If the respondent has a “GA” (full Swiss season ticket), a seasonal ticket for the line or the area, this variable takes value zero. If the respondent has a half-fare travel card, this variable corresponds to half the cost of the trip by public transport..
CostCarCHF	The total gas cost of a loop performed with the car in CHF.
CostCar	The total gas cost of a loop performed with the car in euros.
TripPurpose	The main purpose of the loop: 1 =Work-related trips; 2 =Work- and leisure-related trips; 3 =Leisure related trips. -1 represents missing values
TypeCommune	The commune type, based on the Swiss Federal Statistical Office 1 =Centers; 2 =Suburban communes; 3 =High-income communes; 4 =Peri-urban communes; 5 =Touristic communes; 6 =Industrial and tertiary communes; 7 =Rural and commuting communes; 8 =Agricultural and mixed communes; 9 =Agricultural communes
UrbRur	Binary variable, where: 1 =Rural; 2 =Urban.
ClassifCodeLine	Classification of the type of bus lines of the commune: 1 =Center; 2 =Centripetal; 3 =Peripheral; 4 =Rabatement.

Table 2: Description of variables

Name	Description
frequency	Categorical variable for the frequency: 1 =Low frequency, < 12 pairs of trips per day; 2 =Low-middle frequency, 13 - 20 pairs of trips per day; 3 =Middle-high frequency, 21-30 pairs of trips per day; 4 =High frequency, > 30 pairs of trips per day.
NbTrajects	Number of trips in the loop
Region OR Coderegion-CAR	Region where the commune of the respondent is situated. These regions are denoted by CarPostal as follows: 1 =Vaud; 2 =Valais; 3 =Delemont; 4 =Bern; 5 =Basel, Aargau, Olten; 6 =Zurich; 7 =Eastern Switzerland; 8 =Graubunden.
distance_km	Total distance performed for the loop.
Choice	Choice variable: 0 = public transports (train, bus, tram, etc.); 1 = private modes (car, motorbike, etc.); 2 = soft modes (bike, walk, etc.).
InVehicleTime	Time spent in (on-board) the transport modes only (discarding walking time and waiting time), -1 if missing value.
ReportedDuration	Time spent for the whole loop, as reported by the respondent. -1 represents missing values
LangCode	Language of the commune where the survey was conducted: 1 =French; 2 =German.
age	Age of the respondent (in years) -1 represents missing values.
DestAct	The main activity at destination: 1 is work, 2 is professional trip, 3 is studying, 4 is shopping, 5 is activity at home, 6 is eating/drinking, 7 is personal business, 8 is driving someone, 9 is cultural activity or sport, 10 is going out (with friends, restaurant, cinema, theater), 11 is other and -1 is missing value.
FreqTripHouseh	Frequency of trips related to the household (drive someone, like kids, or shopping), 1 is never, 2 is several times a day, 3 is several times a week, 4 is occasionally, -1 is for missing data and -2 if respondent didn't answer to any opinion questions.

Table 3: Description of variables

Name	Description
ModeToSchool	Most often mode used by the respondent to go to school as a kid (> 10), 1 is car (passenger), 2 is train, 3 is public transport, 4 is walking, 5 is biking, 6 is motorbike, 7 is other, 8 is multiple modes, -1 is for missing data and -2 if respondent didn't answer to any opinion questions.
ResidChild	Main place of residence as a kid (< 18), 1 is city center (large town), 2 is city center (small town), 3 is suburbs, 4 is suburban town, 5 is country side (village), 6 is countryside (isolated), -1 is for missing data and -2 if respondent didn't answer to any opinion questions.
FreqCarPar	Frequency of the usage of car by the respondent's parents (or adults in charge) during childhood (< 18), 1 is never, 2 is occasionally, 3 is regularly, 4 is exclusively, -1 is for missing data and -2 if respondent didn't answer to any opinion questions.
FreqTrainPar	Frequency of the usage of train by the respondent's parents (or adults in charge) during childhood (< 18), 1 is never, 2 is occasionally, 3 is regularly, 4 is exclusively, -1 is for missing data and -2 if respondent didn't answer to any opinion questions.
FreqOthPar	Frequency of the usage of tram, bus and other public transport (not train) by the respondent's parents (or adults in charge) during childhood (< 18), 1 is never, 2 is occasionally, 3 is regularly, 4 is exclusively, -1 is for missing data and -2 if respondent didn't answer to any opinion questions.
NbHousehold	Number of persons in the household. -1 for missing value.
NbChild	Number of kids (< 15) in the household. -1 for missing value.
NbCar	Number of cars in the household. -1 for missing value.
NbMoto	Number of motorbikes in the household. -1 for missing value.
NbBicy	Number of bikes in the household. -1 for missing value.
NbBicyChild	Number of bikes for kids in the household. -1 for missing value.

Table 4: Description of variables

Name	Description
NbComp	Number of computers in the household. -1 for missing value.
NbTV	Number of TVs in the household. -1 for missing value.
Internet	Internet connection, 1 is yes, 2 is no. -1 for missing value.
NewsPaperSubs	Newspaper subscription, 1 is yes, 2 is no. -1 for missing value.
NbCellPhones	Number of cell phones in the household (total). -1 for missing value.
NbSmartPhone	Number of smartphones in the household (total). -1 for missing value.
HouseType	House type, 1 is individual house (or terraced house), 2 is apartment (and other types of multi-family residential), 3 is independent room (subletting). -1 for missing value.
OwnHouse	Do you own the place where you are living? 1 is yes, 2 is no. -1 for missing value.
NbRoomsHouse	Number of rooms in your house. -1 for missing value.
YearsInHouse	Number of years spent in the current house. -1 for missing value.
Income	Net monthly income of the household in CHF. 1 is less than 2500, 2 is from 2501 to 4000, 3 is from 4001 to 6000, 4 is from 6001 to 8000, 5 is from 8001 to 10'000 and 6 is more than 10'001. -1 for missing value.
Gender	Gender of the respondent, 1 is man, 2 is woman. -1 for missing value.
BirthYear	Year of birth of the respondent. -1 for missing value.
Mothertongue	Mothertongue. 1 for German or Swiss German, 2 for French, 3 for other, -1 for missing value.
FamilSitu	Familiar situation: 1 is single, 2 is in a couple without children, 3 is in a couple with children, 4 is single with your own children, 5 is in a co-location, 6 is with your parents and 7 is for other situations. -1 for missing values.
OccupStat	What is your occupational status? 1 is for full-time paid professional activity, 2 for partial-time paid professional activity, 3 for searching a job, 4 for occasional employment, 5 for no paid job, 6 for homemaker, 7 for invalidity leave, 8 for student and 9 for retired. -1 for missing values.
SocioProfCat	To which of the following socio-professional categories do you belong? 1 is for top managers, 2 for intellectual professions, 3 for freelancers, 4 for intermediate professions, 5 for artisans and salespersons, 6 for employees, 7 for workers and 8 for others. -1 for missing values.

Table 5: Description of variables

Name	Description
Education	Highest education achieved. As mentioned by Wikipedia in English: "The education system in Switzerland is very diverse, because the constitution of Switzerland delegates the authority for the school system mainly to the cantons. The Swiss constitution sets the foundations, namely that primary school is obligatory for every child and is free in public schools and that the confederation can run or support universities." (source: Wikipedia, accessed April 16, 2013). It is thus difficult to translate the survey that was originally in French and German. The possible answers in the survey are: 1. Unfinished compulsory education: education is compulsory in Switzerland but pupils may finish it at the legal age without succeeding the final exam. 2. Compulsory education with diploma 3. Vocational education: a three or four-year period of training both in a company and following theoretical courses. Ends with a diploma called "Certificat fédéral de capacité" (i.e., "professional baccalaureate"). 4. A 3-year generalist school giving access to teaching school, nursing schools, social work school, universities of applied sciences or vocational education (sometime in less than the normal number of years). It does not give access to universities in Switzerland 5. High school: ends with the general baccalaureate exam. The general baccalaureate gives access automatically to universities. 6. Universities of applied sciences, teaching schools, nursing schools, social work schools: ends with a Bachelor and sometimes a Master, mostly focus on vocational training 7. Universities and institutes of technology: ends with an academic Bachelor and in most cases an academic Master 8. PhD thesis
HalfFareST	Is equal to 1 if the respondent has a half-fare travel card and to 2 if not.
LineRelST	Is equal to 1 if the respondent has a line-related season ticket and 2 if not.
GenAbST	Is equal to 1 if the respondent has a GA (full Swiss season ticket) and 2 if not.
AreaRelST	Is equal to 1 if the respondent has an area-related season ticket and 2 if not.
OtherST	Is equal to 1 if the respondent has a season ticket that was is not in the list and 2 if not.
CarAvail	Represents the availability of a car for the respondent: 1 is always, 2 is sometime, 3 is never. -1 for missing value.

Table 6: Attitude questions. Coding: 1= strongly disagree, 2=disagree, 3=neutral, 4= agree, 5= strongly agree, 6=not applicable, -1= missing value, -2= all answers to attitude questions missing

Name	Description
Envir01	Fuel price should be increased to reduce congestion and air pollution.
Envir02	More public transportation is needed, even if taxes are set to pay the additional costs.
Envir03	Ecology disadvantages minorities and small businesses.
Envir04	People and employment are more important than the environment.
Envir05	I am concerned about global warming.
Envir06	Actions and decision making are needed to limit greenhouse gas emissions.
Mobil01	My trip is a useful transition between home and work.
Mobil02	The trip I must do interferes with other things I would like to do.
Mobil03	I use the time of my trip in a productive way.
Mobil04	Being stuck in traffic bores me.
Mobil05	I reconsider frequently my mode choice.
Mobil06	I use my current mean of transport mode because I have no alternative.
Mobil07	In general, for my activities, I always have a usual mean of transport.
Mobil08	I do not feel comfortable when I travel close to people I do not know.
Mobil09	Taking the bus helps making the city more comfortable and welcoming.
Mobil10	It is difficult to take the public transport when I travel with my children.
Mobil11	It is difficult to take the public transport when I carry bags or luggage.
Mobil12	It is very important to have a beautiful car.
Mobil13	With my car I can go wherever and whenever.
Mobil14	When I take the car I know I will be on time.
Mobil15	I do not like looking for a parking place.
Mobil16	I do not like changing the mean of transport when I am traveling.
Mobil17	If I use public transportation I have to cancel certain activities I would have done if I had taken the car.
Mobil18	CarPostal bus schedules are sometimes difficult to understand.
Mobil19	I know very well which bus/train I have to take to go where I want to.
Mobil20	I know by heart the schedules of the public transports I regularly use ⁸

Table 7: Attitude questions. Coding: 1= strongly disagree, 2=disagree, 3=neutral, 4= agree, 5= strongly agree, 6=not applicable, -1= missing value, -2= all answers to attitude questions missing.

Name	Description
Mobil21	I can rely on my family to drive me if needed
Mobil22	When I am in a town I don't know I feel strongly disoriented
Mobil23	I use the internet to check the schedules and the departure times of buses and trains.
Mobil24	I have always used public transports all my life
Mobil25	When I was young my parents took me to all my activities
Mobil26	I know some drivers of the public transports that I use
Mobil27	I think it is important to have the option to talk to the drivers of public transports.
ResidCh01	I like living in a neighborhood where a lot of things happen.
ResidCh02	The accessibility and mobility conditions are important for the choice of housing.
ResidCh03	Most of my friends live in the same region I live in.
ResidCh04	I would like to have access to more services or activities.
ResidCh05	I would like to live in the city center of a big city.
ResidCh06	I would like to live in a town situated in the outskirts of a city.
ResidCh07	I would like to live in the countryside.
LifSty01	I always choose the best products regardless of price.
LifSty02	I always try to find the cheapest alternative.
LifSty03	I can ask for services in my neighborhood without problems.
LifSty04	I would like to spend more time with my family and friends.
LifSty05	Sometimes I would like to take a day off .
LifSty06	I can recognize the social status of other travelers by looking at their cars.
LifSty07	The pleasure of having something beautiful consists in showing it.
LifSty08	For me the car is only a practical way to move.
LifSty09	I would like to spend more time working.
LifSty10	I do not like to be in the same place for too long.
LifSty11	I always plan my activities well in advance
LifSty12	I like to experiment new or different situations
LifSty13	I am not afraid of unknown people
LifSty14	My schedule is rather regular.

Table 8: Descriptive statistics of the main variables (no data excluded)

	nbr. cases	nbr. null	min	max	median	mean	std.dev
age	1906	0	-1	88	47	46.48	18.57
Choice	1906	536	0	2	1	0.78	0.54
TypeCommune	1906	0	1	9	6	5.39	1.99
UrbRur	1906	0	1	2	2	1.51	0.5
ClassifCodeLine	1906	0	1	4	4	3.17	0.97
LangCode	1906	0	1	2	2	1.74	0.44
CoderegionCAR	1906	0	1	8	5	4.58	2.08
CostCarCHF	1906	5	0	67.65	2.98	5.76	8.34
distance_km	1906	1	0	519	18.75	40.38	62.6
TimeCar	1906	28	0	494	26	40.68	47.61
TimePT	1906	7	0	745	85	107.88	86.52
frequency	1906	0	1	4	3	2.84	1.09
ID	1906	0	10350017	96040538	44690042	45878800	23846908
InVehicleTime	1906	66	-128	631	40.5	55.13	57.78
MarginalCostPT	1906	270	0	230	5.6	11.11	16.13
NbTrajects	1906	0	1	9	2	2.04	1.05
NbTransf	1906	644	0	14	2	2.01	2.17
Region	1906	0	1	8	5	4.58	2.08
ReportedDuration	1906	3	-1	855	35	57.73	72.47
TripPurpose	1906	0	-1	3	2	1.94	1.18
WaitingTimePT	1906	693	0	392	5	13.13	22.07
WalkingTimePT	1906	17	0	213	33	39.63	28

Table 9: Descriptive statistics of the main variables affected by missing data
(observations with -1 excluded)

	nbr. cases	nbr.null	min	max	median	mean	std.dev
age	1791	0	16	88	48	49.53	14.59
ReportedDuration	1835	3	0	855	37	60	72.92
TripPurpose	1783	0	1	3	3	2.14	0.92

Choice with multiple alternatives – 5.3 Mode choice in Switzerland

Michel Bierlaire

Practice quiz: reproduce the base model

The objective of this practice quiz is to reproduce the results of the base logit model published in the paper

Attitudes towards mode choice in Switzerland, Atasoy et al. (2013)
[Click here].

A preliminary version of the paper is available as a technical report:

Atasoy et al. (2011) [Click here].

The model has three alternatives:

1. private motorized modes (PMM), including car, motorbike and taxi,
2. public transport (PT), including bus, train and car postal, and
3. slow modes (SM), including walking and bike.

The specification of the utilities is presented in Table 1.

We ask you to reproduce the results of the base model. To do so, perform the following steps:

1. Download the Optima dataset provided in the file `optima.dat` (from the edX webpage).
2. Check the definition of the variables in the file `optimaDescription.pdf` (from the edX webpage).

Table 1: Specification of the utility functions

Parameter	V_{PMM}	V_{PT}	V_{SM}
ASC_{CAR}	1	-	-
ASC_{PT}	-	-	-
ASC_{SM}	-	-	1
β_{cost}	CostCarCHF	MarginalCostPT	-
$\beta_{timeCar}$	TimeCar	-	-
β_{timePT}	-	TimePT	-
$\beta_{distance}$	-	-	distance_km
β_{nbCars}	NbCar * (NbCar > 0)	-	-
$\beta_{nbChild}$	NbChild * (NbChild > 0)	-	-
β_{french}	LangCode = 1	-	-
β_{work}	TripPurpose = 1 or TripPurpose = 2	-	-
β_{urban}	-	UrbRur = 2	-
$\beta_{student}$	-	OccupStat = 8	-
$\beta_{nbBikes}$	-	-	NbBicy * (NbBicy > 0)

3. Download the `v532_optima_template.py` (from the edX webpage) file and use it as a template to create the file `v532_optima_base.py` with the model specification defined in Table 1.
4. Estimate the parameters of the base model.
5. Compare your results with the ones reported in Atasoy et al. (2013) [[Click here](#)]. You must obtain the same parameter values and final log likelihood value.

References

- Atasoy, B., Glerum, A. and Bierlaire, M. (2011). Attitudes towards mode choice in switzerland, *Technical Report TRANSP-OR 110502*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.
- Atasoy, B., Glerum, A. and Bierlaire, M. (2013). Attitudes towards mode choice in switzerland, *disP - The Planning Review* **49**(2): 101–117.

Choice with multiple alternatives – 5.3 Mode choice in Switzerland

Michel Bierlaire

Solution to the practice quiz: reproduce the base model

- The model specification file that reproduces the results of the base model presented in Atasoy et al. (2013) [Click here] and Atasoy et al. (2011) [Click here] is

`v532_optima_base.py` (from the edX webpage).

- The estimation results are available in the file

`v532_optima_base.html` (from the edX webpage).

- The parameter estimates of the model are reported in Tables 1 and 2.

References

- Atasoy, B., Glerum, A. and Bierlaire, M. (2011). Attitudes towards mode choice in switzerland, *Technical Report TRANSP-OR 110502*, Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne.
- Atasoy, B., Glerum, A. and Bierlaire, M. (2013). Attitudes towards mode choice in switzerland, *disP - The Planning Review* **49**(2): 101–117.

Table 1: Estimation results: parameters estimates

Parameter number	Description	Coeff. estimate	Robust std. error	t -stat	p -value
1	ASC_{CAR}	-0.413	0.173	-2.39	0.02
2	ASC_{SM}	-0.470	0.369	-1.27	0.20
3	β_{cost}	-0.0592	0.0105	-5.61	0.00
4	$\beta_{distance}$	-0.227	0.0531	-4.28	0.00
5	β_{french}	1.09	0.159	6.89	0.00
6	β_{nbCars}	1.00	0.0972	10.30	0.00
7	$\beta_{nbChild}$	0.154	0.0647	2.37	0.02
8	$\beta_{nbBikes}$	0.347	0.0548	6.34	0.00
9	$\beta_{student}$	3.21	0.344	9.34	0.00
10	$\beta_{timeCar}$	-0.0299	0.00604	-4.96	0.00
11	β_{timePT}	-0.0121	0.00265	-4.55	0.00
12	β_{urban}	0.286	0.123	2.33	0.02
13	β_{work}	-0.582	0.116	-5.01	0.00

Table 2: Estimation results: summary statistics

Number of observations = 1906
 Number of excluded observations = 359
 Number of estimated parameters = 13
 $\mathcal{L}(\beta_0) = -2093.955$
 $\mathcal{L}(\hat{\beta}) = -1067.356$

Choice with multiple alternatives – 5.4

Maximum likelihood estimation

Michel Bierlaire

The likelihood function for logit.

The maximum likelihood estimation method is exactly the same for logit with multiple alternatives, as for the binary logit model. The logit model is

$$P_n(i|\mathcal{C}_n) = \frac{e^{V_{in}}}{\sum_{j \in \mathcal{C}_n} e^{V_{jn}}}. \quad (1)$$

The log likelihood of a sample is

$$\mathcal{L}(\beta_1, \dots, \beta_K) = \sum_{n=1}^N \left(\sum_{i \in \mathcal{C}_n} y_{in} \ln P_n(i|\mathcal{C}_n) \right), \quad (2)$$

where $y_{in} = 1$ if individual n has chosen alternative i , 0 otherwise. Using (1) into (2), we obtain

$$\mathcal{L}(\beta_1, \dots, \beta_K) = \sum_{n=1}^N \sum_{i \in \mathcal{C}_n} y_{in} \left(V_{in} - \ln \sum_{j \in \mathcal{C}_n} e^{V_{jn}} \right). \quad (3)$$

The maximum likelihood estimation amounts to find the vector β solving the optimization problem

$$\max_{\beta \in \mathbb{R}^K} \mathcal{L}(\beta). \quad (4)$$

In the case of logit, it can be shown (McFadden, 1974) that, if the utility function is linear in the parameters, the log likelihood function is globally concave and does not exhibit local maxima (under some relatively weak conditions).

The necessary first-order optimality conditions impose that the partial derivatives with respect to each parameter is equal to zero. The k th partial derivative is

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{n=1}^N \sum_{i \in \mathcal{C}_n} y_{in} \left(\frac{\partial V_{in}}{\partial \beta_k} - \sum_{j \in \mathcal{C}_n} P_n(j) \frac{\partial V_{jn}}{\partial \beta_k} \right) \text{ for } k = 1, \dots, K. \quad (5)$$

Distributing the y_{in} , we obtain

$$\frac{\partial \mathcal{L}}{\partial \beta_k} = \sum_{n=1}^N \sum_{i \in \mathcal{C}_n} (y_{in} - P_n(i)) \frac{\partial V_{in}}{\partial \beta_k} \text{ for } k = 1, \dots, K. \quad (6)$$

For a linear-in-parameters logit, it is

$$\sum_{n=1}^N \sum_{i \in \mathcal{C}_n} (y_{in} - P_n(i)) x_{ink}, \text{ for } k = 1, \dots, K. \quad (7)$$

Setting these equations to zero leads to the necessary first-order optimality conditions:

$$\sum_{n=1}^N \sum_{i \in \mathcal{C}_n} y_{in} x_{ink} = \sum_{n=1}^N \sum_{i \in \mathcal{C}_n} P_n(i) x_{ink}. \quad (8)$$

It means that the expected value of each attribute of the chosen alternative must be the same when computed in the sample or with the choice model.

The reader can also verify that the second derivatives of \mathcal{L} are given by

$$\frac{\partial^2 \mathcal{L}}{\partial \beta_k \partial \beta_\ell} = - \sum_{n=1}^N \sum_{i \in \mathcal{C}_n} P_n(i) \left(x_{ink} - \sum_{j \in \mathcal{C}_n} x_{jnk} P_n(j) \right) \left(x_{in\ell} - \sum_{j \in \mathcal{C}_n} x_{jn\ell} P_n(j) \right). \quad (9)$$

References

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior, in P. Zarembka (ed.), *Frontiers in Econometrics*, Academic Press, New York, pp. 105–142.

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 6

Testing – 6.1 Specification testing

Michel Bierlaire

A short reminder on hypothesis testing

Hypothesis testing is a method to contradict a theoretical assumption using data. In his seminal book on design of experiments, Fisher (1937) uses the example of a lady who pretends to be able to tell if the milk has been poured before or after the tea in a cup just by tasting it. This is the theoretical assumption. An experiment is organized, where the lady is tasting several cups of tea, and reports each time if the milk has been poured first or not. The hypothesis to be tested, called the *null hypothesis* and often denoted H_0 , is that the provided responses are purely random.

Hypothesis testing has some analogy with a court trial. In that context, the theoretical assumption is that an individual has committed a felony. The null hypothesis to be tested is that she is innocent. The main principle is that the defendant is presumed innocent until proved guilty. Similarly, the null hypothesis is considered correct, until the data provide sufficient evidences that it is not.

Mathematically, the test of the hypothesis consists in identifying a statistic calculated from the data that has a known distribution under the null hypothesis. If the value of the statistic lies in the tail of the distribution, that is, if the probability that such a value occurs is low under the null hypothesis, it is rejected, acknowledging that there is a non zero probability that an error is made. In the tea tasting example, the probability to give a correct answer k times among n trials is given by a binomial distribution:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad (1)$$

where X is a random variable representing the number of successes, and p is the probability of a success. The null hypothesis corresponds to $p = 0.5$. Consider an experiment with 8 trials. It is easy to calculate that, under

the null hypothesis, the number of correct answers would be between zero and six 96.5% of the time. Therefore, if it happens that seven or eight correct answers are provided, the null hypothesis can be rejected with 3.5% of confidence. If this is not considered sufficient for an evidence, it is possible to be more strict. As there is 99.6% chance to obtain 7 correct answers or less, the analyst could decide to reject the null hypothesis only when 8 correct answers are provided, with confidence 0.4%.

The level of confidence is important. Indeed, because of the randomness associated with the data generation process, the outcome of an hypothesis test (that is, rejecting the null hypothesis or not), may be incorrect.

There are two types of potential errors, as illustrated in Table 1:

- Type I errors occur when the null hypothesis is true, and rejected by the test. Using the analogy with the court trial, it corresponds to sending an innocent to jail. It is sometimes called a “false positive”. We denote α the conditional probability:

$$P(H_0 \text{ is rejected} | H_0 \text{ is true}) = \alpha. \quad (2)$$

It is called the *significance level* of the test.

- Type II errors occur when the null hypothesis is false, but not rejected by the test. Using the analogy with the court trial, it corresponds to releasing a culprit. It is sometimes called a “false negative”. We denote β the conditional probability:

$$P(H_0 \text{ is not rejected} | H_0 \text{ is false}) = \beta. \quad (3)$$

In practice, the analyst decides on α . The value $1 - \beta$, that is

$$P(H_0 \text{ is rejected} | H_0 \text{ is false}) = 1 - \beta, \quad (4)$$

is called the *power* of the test. Clearly, for a given data set, decreasing α has the consequence to increase β . The extreme case is to never reject the hypothesis, so that $\alpha = 0$.

Back to our tea tasting example, suppose that the lady has actually the ability to identify if the milk has been poured first or last, with 80% of success rate, and consider again the two tests presented above.

1. The first test rejects the hypothesis when there are 7 or 8 correct answers. It fails to reject the (incorrect) null hypothesis $\beta = 49.7\%$ of the time (verify using the binomial distribution with $p = 0.8$.) The power of the test is 50.3% .
2. The second test rejects the hypothesis when there are 8 correct answers. It fails to reject the (incorrect) null hypothesis $\beta = 83.2\%$ of the time. The power of the test is only 16.8% .

We refer the reader to textbooks in statistics (such as Larsen and Marx (2001, Chapter 6) for a comprehensive introduction to hypothesis testing.

	Accept H_0	Reject H_0
H_0 is true		Type I error (prob. α)
H_0 is false	Type II error (prob. β)	

Table 1: Type of errors in hypothesis testing

References

- Fisher, R. A. (1937). *The design of experiments*, Oliver And Boyd, Edinburgh London.
- Larsen, R. J. and Marx, M. L. (2001). *An introduction to mathematical statistics and its applications*, Vol. 2, 3rd edn, Prentice-Hall, Upper Saddle River, NJ.

Testing

Specification testing

Michel Bierlaire

Introduction to choice models



Differences from classical hypothesis testing

Classical hypothesis testing: example

Null hypothesis (H_0)

A simple hypothesis contradicting a theoretical assumption.

Lady testing tea

- ▶ Theory: a lady is able to tell if the milk has been poured before or after the tea in a cup.
- ▶ H_0 : the outcome of the taste is purely random.



Specification testing: example

Null hypothesis (H_0)

A simple hypothesis contradicting a theoretical assumption.

Explanatory variable



- ▶ Theory: a variable explains the choice behavior.
- ▶ H_0 : the coefficient of the variable is zero.

Errors in hypothesis testing

Errors in hypothesis testing

Type I error

Errors in hypothesis testing

Type I error

Type II error

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.

Type II error

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.

Type II error

- ▶ H_0 accepted and H_0 false.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.

Type II error

- ▶ H_0 accepted and H_0 false.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.

Type II error

- ▶ H_0 accepted and H_0 false.
- ▶ Omit a relevant variable.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.
- ▶ Loss of efficiency.

Type II error

- ▶ H_0 accepted and H_0 false.
- ▶ Omit a relevant variable.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.
- ▶ Loss of efficiency.

Type II error

- ▶ H_0 accepted and H_0 false.
- ▶ Omit a relevant variable.
- ▶ Specification error.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.
- ▶ Loss of efficiency.
- ▶ Cost: C_I .

Type II error

- ▶ H_0 accepted and H_0 false.
- ▶ Omit a relevant variable.
- ▶ Specification error.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.
- ▶ Loss of efficiency.
- ▶ Cost: C_I .

Type II error

- ▶ H_0 accepted and H_0 false.
- ▶ Omit a relevant variable.
- ▶ Specification error.
- ▶ Cost: $C_{II} \gg C_I$.

Errors in hypothesis testing

Type I error

- ▶ H_0 rejected and H_0 true.
- ▶ Include an irrelevant variable.
- ▶ Loss of efficiency.
- ▶ Cost: C_I .

Type II error

- ▶ H_0 accepted and H_0 false.
- ▶ Omit a relevant variable.
- ▶ Specification error.
- ▶ Cost: $C_{II} \gg C_I$.

Note

In classical hypothesis testing, $C_I \approx C_{II}$

Impact of an error

Impact of an error

Probability of an error

$$P(\text{Type I}) =$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true})$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = \frac{P(H_0 \text{ rejected} | H_0 \text{ true})}{P(H_0 \text{ true})}$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = \underbrace{P(H_0 \text{ rejected} | H_0 \text{ true})}_{\alpha} P(H_0 \text{ true})$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α λ

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α λ

$$P(\text{Type II}) =$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false})$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Expected cost

Expected cost =

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Expected cost

$$\text{Expected cost} = P(\text{Type I}) \quad C_I \quad + \quad P(\text{Type II}) \quad C_{II}$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Expected cost

$$\begin{aligned} \text{Expected cost} &= P(\text{Type I}) C_I + P(\text{Type II}) C_{II} \\ &= \alpha \lambda C_I + \beta (1 - \lambda) C_{II} \end{aligned}$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Expected cost

$$\begin{aligned} \text{Expected cost} &= P(\text{Type I}) C_I + P(\text{Type II}) C_{II} \\ &= \alpha \lambda C_I + \beta (1 - \lambda) C_{II} \end{aligned}$$

Classical hypothesis testing

$$\lambda \approx 1, C_I \approx C_{II}$$

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Expected cost

$$\begin{aligned} \text{Expected cost} &= P(\text{Type I}) C_I + P(\text{Type II}) C_{II} \\ &= \alpha \lambda C_I + \beta (1 - \lambda) C_{II} \end{aligned}$$

Classical hypothesis testing

$\lambda \approx 1$, $C_I \approx C_{II}$: prefer small α .

Impact of an error

Probability of an error

$$P(\text{Type I}) = P(H_0 \text{ rejected} | H_0 \text{ true}) \quad P(H_0 \text{ true})$$

α

λ

$$P(\text{Type II}) = P(H_0 \text{ accepted} | H_0 \text{ false}) \quad P(H_0 \text{ false})$$

β

$(1 - \lambda)$

Expected cost

$$\begin{aligned} \text{Expected cost} &= P(\text{Type I}) C_I + P(\text{Type II}) C_{II} \\ &= \alpha \lambda C_I + \beta (1 - \lambda) C_{II} \end{aligned}$$

Specification testing

$\lambda \approx 0.5$, $C_{II} \gg C_I$: larger α can be used.

Testing – 6.1 Specification testing

Michel Bierlaire

Practice quiz

1. If the hypothesis test does not reject the null hypothesis, we can conclude that the null hypothesis is true.
 - (a) True
 - (b) False
2. When we reject a true null hypothesis, we commit a Type I error.
 - (a) True
 - (b) False
3. When the null hypothesis is false and is not rejected, you make a type II error.
 - (a) True
 - (b) False
4. The power of a test is the probability of rejecting the null hypothesis when it is true.
 - (a) True
 - (b) False
5. If the level of significance α of a test is increased, the power of the test decreases.
 - (a) True
 - (b) False

6. If a null hypothesis is rejected at the level of significance 0.01, it is also rejected at the level of significance 0.05.
- (a) True
 - (b) False
7. For a given level of significance, if the sample size is increased, the power of the test decreases.
- (a) True
 - (b) False

Testing – 6.1 Specification testing

Michel Bierlaire

Solution to practice quiz

1. If the hypothesis test does not reject the null hypothesis, we can conclude that the null hypothesis is true.
 - (a) True
 - (b) False

Correct answer: False. Hypothesis testing is never fully conclusive. There is always a possibility to make the wrong decision. When the null hypothesis is not rejected, it is because there is not enough evidence to reject it. It does not mean that it is true. Similarly, a trial may fail to convict a guilty criminal.

2. When we reject a true null hypothesis, we commit a Type I error.
 - (a) True
 - (b) False

Correct answer: True. A Type I error occurs when the null hypothesis is rejected when it is true.

3. When the null hypothesis is false and is not rejected, you make a type II error.
 - (a) True
 - (b) False

Correct answer: True. A Type II error occurs if we fail to reject the null hypothesis when it is false.

4. The power of a test is the probability of rejecting the null hypothesis when it is true.

- (a) True
- (b) False

Correct answer: False. The probability of rejecting the null hypothesis when it is true is the Type I error. The power of a test is the probability of rejecting the null hypothesis when it is false.

5. If the level of significance of a test is increased, the power of the test decreases.

- (a) True
- (b) False

Correct answer: False. As the level of significance α increases, the test rejects the null hypothesis more often. Therefore, if the null hypothesis happens to be false, the risk β to make a mistake decreases. Consequently, the power $(1 - \beta)$ increases.

6. If a null hypothesis is rejected at the level of significance 0.01, it is also rejected at the level of significance 0.05.

- (a) True
- (b) False

Correct answer: True. The larger the level of significance, the most likely it is to reject the null hypothesis.

7. For a given level of significance, if the sample size is increased, the power of the test decreases.

- (a) True
- (b) False

Correct answer: False. As we have more information with a larger sample, we can only do better. The power of the test increases.

Informal tests are designed to identify early inconsistencies between the estimated model and a priori expectations. We describe here the two most common tests performed in practice.

For many variables in the utility function, we have a clear idea about the sign of their coefficient. For instance, the coefficient of the cost variable is always expected to be negative. Indeed, everything else being equal, a cheaper alternative is preferred to a more expensive one. If the estimated value of the coefficient does not have the expected sign, the issue must be investigated.

The second test consists in changing the units of the utility function, typically into monetary units. Suppose that the cost variable c_{in} , expressed in CHF, appears in the utility function:

$$U_{in} = \beta_c c_{in} + \beta_1 x_{in1} + \beta_2 x_{in2} + \dots \quad (1)$$

The values of the coefficients are impossible to interpret as such. Not only their units are not intuitive, but they are also confounded with the scale parameter.

As the utility function has no unit, the units of the coefficient β_c are 1/CHF. Therefore, if the utility function is divided by β_c , it is expressed in CHF:

$$U'_{in} = c_{in} + \frac{\beta_1}{\beta_c} x_{in1} + \frac{\beta_2}{\beta_c} x_{in2} + \dots \quad (2)$$

The ratio of the parameters can now easily be interpreted, as it is expressed in monetary units. Moreover, as the coefficients at the numerator and the denominator are scaled in the same way, the scale cancels out.

A typical example in transportation is the coefficient of the travel time variable. If travel time is expressed in minutes, the coefficient β_t of travel time is in 1/minute. In the utility function expressed in monetary units, the ratio β_t/β_c is expressed in CHF/minute. It is called the *value of time*, or the willingness to pay for travel time savings. The value of time is often reported in the literature. Therefore, the value that is obtained after estimation can be compared to published value.

But the interpretation of the willingness to pay is not restricted to the value of time. The willingness to pay to improve the value of other variables can also be calculated and interpreted.

Testing – 6.2 Informal tests

Michel Bierlaire

Practice quiz

In a mode choice experiment with two alternatives, the following utility functions are specified for private motorized mode (pmm) and public transportation (pt):

$$\begin{aligned} U_{pmm,n} &= -\beta_c \cdot \text{cost}_{pmm,n} - \beta_t \cdot \text{time}_{pmm,n} + \varepsilon_{pmm,n} \\ U_{pt,n} &= -\beta_c \cdot \text{cost}_{pt,n} - \beta_t \cdot \text{time}_{pt,n} + \varepsilon_{pt,n} \end{aligned} \quad (1)$$

where $\text{cost}_{pmm,n}$ and $\text{cost}_{pt,n}$ are the cost of the trip by private motorized mode and public transportation respectively for individual n in CHF, and $\text{time}_{pmm,n}$ and $\text{time}_{pt,n}$ are the corresponding travel times in minutes. The error terms $\varepsilon_{pmm,n}$ and $\varepsilon_{pt,n}$ are i.i.d. Extreme Value: $\text{EV}(0, 1)$.

We have a sample containing 10 observations:

Individual	Choice	time_{pmm}	time_{pt}	cost_{pmm}	cost_{pt}
1	pmm	10	20	2.3	1
2	pt	5	10	2.3	0.5
3	pmm	35	30	9	12
4	pmm	20	22	1.5	2
5	pt	6	7.5	2	1.25
6	pt	10	15	5	3.5
7	pt	8	5	3	2
8	pt	19	18	4	5
9	pt	22	19	7	8.5
10	pmm	8	8.5	3	9

The parameter estimates are $\beta_c = 1.38$ and $\beta_t = 0.363$

1. Can you check if the value of time makes sense, given that Axhausen et al. (2008) report values ranging from 17.73 CHF/h to 50.23 CHF/h for the value of time?

2. Plot these observations where the x -axis is $\text{time}_{pmm} - \text{time}_{pt}$ and the y -axis is $\text{cost}_{pmm} - \text{cost}_{pt}$. Use a different shape for the marker depending on the observed choice.
3. Add to the previous plot the line $-\beta_c \cdot \text{cost}_{pmm} - \beta_t \cdot \text{time}_{pmm} = -\beta_c \cdot \text{cost}_{pt} - \beta_t \cdot \text{time}_{pt}$. What does its slope represent?

References

Axhausen, K., Hess, S., Koenig, A., Abay, G., Bates, J. and Bierlaire, M. (2008). Income and distance elasticities of values of travel time savings: new swiss results, *Transport Policy* **15**(3): 173–185.

Testing – 6.2 Informal tests

Michel Bierlaire

Solution to the practice quiz

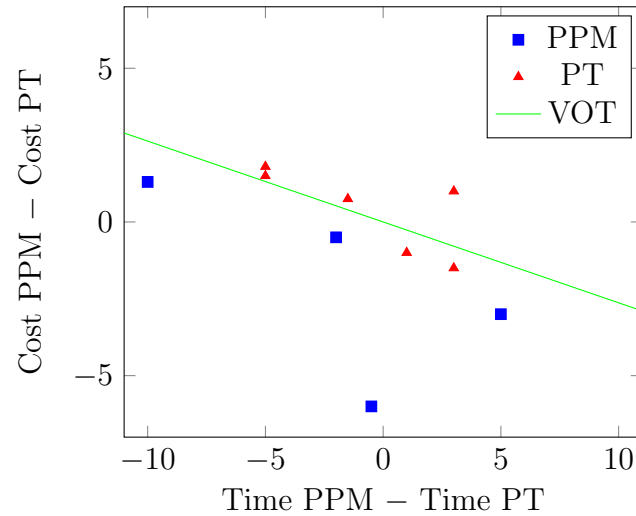
1. The value of time in [CHF/minute] is

$$\text{VOT} = \frac{-\beta_t}{-\beta_c} = \frac{0.363}{1.38} = 0.263 \text{ [CHF/min]}$$

and to obtain it in [CHF/h] we need to multiply it by 60, so

$$\text{VOT} = 15.78 \text{ [CHF/h]}$$

This value is lower than the value reported in the literature, but the level of magnitude is similar, which is acceptable.



2. The plot is:

Note that, due to the presence of the error terms, the indifference line does not separate exactly the PPM observations from the PT.

3. The slope is the value of time.

Testing – 6.3 t -tests

Michel Bierlaire

The Boeing data set.

Before we start investigating the methodology for the t -test, we consider another data set that will be used for illustration.

These data come from an Internet choice survey conducted by the Boeing Company in the Fall of 2004. Boeing was interested in understanding the sensitivity that air passengers have toward the attributes of an airline itinerary, such as fare, travel time, transfers, legroom, and aircraft. It was executed on a sample of the customers of an Internet airline booking service. The Internet service takes a specific user request for travel in a city pair and interrogates the web sites of airlines that provide service in that market, returning to the user a compiled list of available itineraries. While that interrogation is taking place, randomly selected customers were recruited to be surveyed.

A typical page of the survey instrument is shown in Figure 1. The respondent was offered three choices based on the origin-destination market request that the respondent entered into the itinerary search engine. The first alternative is always a non-stop flight, the second always a flight with 1 stop on the same airline, and the third is always a flight with 1 stop and a change of airline. The respondent was asked to rank the available choices as well as given the option to decline all of the stated options. Demographic data collected included age, gender, income, occupation, and education. Situational variables that were identified included: a) the desired departure time; b) trip purpose; c) who is paying for the trip; and d) the number in the travel party. All trips were for origin-destination city pairs in the United States.

Pick Your Preferred Flight			
<p>Three flight options are described for your trip from Chicago to San Diego. These are options that might be available on this route or might be new options actively being considered for this route as well as replacing some options that are offered now. The options differ from each other in one or more of the features described on the left.</p> <p>Please evaluate these options, assuming that everything about the options is the same except these particular features. Indicate your choices at the bottom of the appropriate column and press the Continue button.</p>			
FEATURES	Non-Stop (Option 1)	1 Stop (Option 2)	1 Stop (Option 3)
Departure time (local)	6:00 PM	4:30 PM	6:00 PM
Arrival time (local)	8:14 PM	8:44 PM	9:44 PM
Total time in air	4 hr 14 min	4 hr 44 min	4 hr 44 min
Total trip time	4 hr 14 min	6 hr 14 min	5 hr 44 min
Legroom <input type="checkbox"/>	typical legroom	2-in more of legroom	4-in more of legroom
Airline [Airplane]	Depart Chicago Continental Airlines [8737] to San Diego	Depart Chicago Southwest Airlines [A320], connecting with Southwest Airlines [MD80] to San Diego	Depart Chicago Northwest Airlines [MD80], connecting with American Airlines [DC9] to San Diego
Fare	\$565	\$485	\$620
<p>1. Which is MOST attractive? <input type="radio"/> Option 1 <input type="radio"/> Option 2 <input type="radio"/> Option 3</p>			
<p>2. Which is LEAST attractive? <input type="radio"/> Option 1 <input type="radio"/> Option 2 <input type="radio"/> Option 3</p>			
<p>3. If these were the ONLY three options available, I would NOT make this trip by air. <input type="radio"/> Yes <input type="radio"/> No</p>			

Figure 1: The choice of airline itinerary: example of survey instrument

Variable	Description
Subj_ID	Unique identifier for each respondent.
Subj_Male	1 if male, 2 otherwise
Subj_Age	Age, (1 = Less than 18 years, 2 = 18-24 years, 3 = 25-34 years, 4 = 35-44 years, 5 = 45-54 years, 6 = 55-64 years, 7 = 65-74 years, 8 = 75 years or older)
Subj_Occupation	Occupation (01 = Executive and Managerial, 02 = Professional, 03 = Technicians and related support, 04 = Sales, 05 = Administrative support, 06 = Services, 07 = Precision production, craft, repair, 08 = Machine operators, assemblers, inspectors, 09 = Transportation and material moving, 10 = Handlers, cleaners, helpers, 11 = Farming, forestry, and fishing, 12 = Armed forces)
Subj_Income	Annual income in 100\$
Subj_IncomeMissing	Income is missing
Subj_Education	Education (01 = Less than High School Diploma, 02 = High School Graduate, 03 = Some college, No Degree, 04 = Associate Degree - Occupational, 05 = Associate Degree - Academic, 06 = Bachelors Degree, 07 = Masters Degree, 08 = Professional Degree, 09 = Doctorate Degree)

Table 1: The choice of airline itinerary: description of respondent specific variables

Variable	Description
SP1_MostAttractive	SP survey response to “Which is MOST attractive”
SP2_LeastAttractive	SP survey response to “Which is LEAST attractive”
SP3_NotByAir	SP survey response to “If these were the ONLY three options available, I would NOT make this trip by air” (1=yes, 2=no)

Table 2: The choice of airline itinerary: description of survey responses

Variable	Description
Trip_Purpose	Trip purpose (1=business, 2=leisure, 3=attending conference/seminar/training, 4=both business and leisure)
Trip_TravelerPays	1 if the traveler is paying for the trip, 2 if it is his employer, 3 if it is a third party
Trip_IdealDepartureTime	Respondents ideal departure time (hours after midnight)
Trip_PartySize	Number of persons traveling
Trip_OrigMinGMT	Origin city time zone (minutes from GMT (Greenwich Mean Time))
Trip_DestMinGMT	Destination city time zone (minutes from GMT)
Trip_BaseFlightTime	Flight time for shortest non-stop itinerary in minutes
Trip_Miles	Length of itinerary in miles
Trip_Direction	Direction of itinerary (1=East to West, 2=West to East, 3=North-South)

Table 3: The choice of airline itinerary: description of trip specific attributes

Variable	Description
OptX_DepTimeHrs	Option X: Departure time, local (hours after midnight)
OptX_ArrTimeHrs	Option X: Arrival time, local (hours after midnight)
OptX_TotalTimeInAir	Option X: Total time in air (hours)
OptX_TotalTriptime	Option X: Total trip time (hours)
OptX_Legroom	Option X: Legroom in Inches, -2 = 2 inches less than typical, 0 = typical, 2 = 2 inches more than typical, 4 = 4 inches more than typical
OptX_AirlineA	Option X: Airline for first leg (only known to arbitrary airline number for proprietary reasons)
OptX_AirlineB	Option X: Airline for second leg (if there exists a second leg) (only known to arbitrary airline number for proprietary reasons)
OptX_AirplaneA	Option X: Airplane for first leg (only known to arbitrary airplane number for proprietary reasons)
OptX_AirplaneB	Option X: Airplane for second leg (if there exists a second leg) (only known to arbitrary airplane number for proprietary reasons)
OptX_Fare	Option X: Fare (\$)
OptX_SchedDelayEarly	Option X: Schedule delay (hours) - early departure (calculated from OptX_DepTimeHrs and Trip_IdealDepartureTime)
OptX_SchedDelayLate	Option X: Schedule delay (hours) - late departure (calculated from OptX_DepTimeHrs and Trip_IdealDepartureTime)

Table 4: The choice of airline itinerary: description of alternative specific attributes where X corresponds to the choice option (1),(2) and (3)

Variable	Average	St. Dev.	Min	Max
Subj_ID	1807.97	1043.03	1.00	3613.00
Subj_Male	0.50	0.50	0	1.00
Subj_Age	3.95	1.14	1.00	8.00
Subj_Occupation	2.54	1.89	1.00	12.00
Subj_Income	107.08	81.40	10.00	350.00
Subj_IncomeMissing	0.1	0.30	0.00	1.00
Subj_Education	5.88	1.71	1.00	9.00
Trip_Purpose	2.04	0.77	1.00	4.00
Trip_TravelerPays	1.20	0.46	1.00	3.00
Trip_IdealDepartureTime	12.75	4.99	0	23.75
Trip_PartySize	1.70	0.99	1.00	5.00
Trip_OrigMinGMT	382.18	82.07	300.00	480.00
Trip_DestMinGMT	397.34	82.86	300.00	480.00
Trip_BaseFlightTime	224.14	95.15	40.00	381.00
Trip_Miles	1568.43	783.79	119.00	2719.00
Trip_Direction	1.91	0.87	1.00	3.00
SP1_MostAttractive	1.45	0.73	1.00	3.00
SP2_LeastAttractive	2.36	0.68	1.00	3.00
SP3_NotByAir	1.60	0.54	1.00	2.00
Opt1_DepTimeHrs	11.72	3.34	6.00	18.00
Opt1_ArrTimeHrs	15.21	3.35	7.67	21.63
Opt1_TotalTimeInAir	3.73	1.59	0.67	6.35
Opt1_TotalTiptime	3.73	1.59	0.67	6.35
Opt1_Legroom	0.92	2.24	-2.00	4.00
Opt1_AirlineA	4.52	2.60	1.00	11.00
Opt1_AirlineB	0.00	0.00	0.00	0.00
Opt1_AirplaneA	4.57	2.30	1.00	8.00
Opt1_AirplaneB	0.00	0.00	0.00	0.00
Opt1_Fare	405.65	199.84	80	1330
Opt1_SchedDelayEarly	2.04	3.98	0.00	17.00
Opt1_SchedDelayLate	2.28	2.91	0.00	21.38

Table 5: The choice of airline itinerary: descriptive statistics of variables

Variable	Average	St. Dev.	Min	Max
Opt2_DepTimeHrs	11.67	3.35	6.00	18.00
Opt2_ArrTimeHrs	16.92	3.36	9.17	24.10
Opt2_TotalTimeInAir	4.23	1.59	1.16	6.85
Opt2_TotalTriptime	5.50	1.67	1.83	8.85
Opt2_Legroom	0.96	2.25	-2.00	4.00
Opt2_AirlineA	4.68	2.67	1.00	11.00
Opt2_AirlineB	0.00	0.00	0.00	0.00
Opt2_AirplaneA	4.46	2.32	1.00	8.00
Opt2_AirplaneB	4.40	2.34	1.00	8.00
Opt2_Fare	407.07	200.93	80.00	1390.00
Opt2_SchedDelayEarly	1.92	4.05	0.00	17.75
Opt2_SchedDelayLate	2.75	2.81	0.00	23.38
Opt3_DepTimeHrs	11.66	3.34	6.00	18.00
Opt3_ArrTimeHrs	16.89	3.41	9.25	24.03
Opt3_TotalTimeInAir	4.24	1.59	1.16	6.85
Opt3_TotalTriptime	5.48	1.67	1.92	8.85
Opt3_Legroom	1.06	2.25	-2.00	4.00
Opt3_AirlineA	4.63	2.61	1.00	11.00
Opt3_AirlineB	4.73	2.67	1.00	11.00
Opt3_AirplaneA	4.49	2.33	1.00	8.00
Opt3_AirplaneB	4.52	2.27	1.00	8.00
Opt3_Fare	405.20	197.65	80.00	1275.00
Opt3_SchedDelayEarly	1.92	3.98	0.00	17.00
Opt3_SchedDelayLate	2.73	2.78	0.00	22.97

Table 6: The choice of airline itinerary: descriptive statistics of variables

Testing – 6.3 t -tests

Michel Bierlaire

Bootstrap

The calculation of the t statistics is described in Week 3. It relies on approximations of the variance-covariance matrix of the estimates: the Cramer-Rao bound, and the robust/sandwich estimator. These approximations are derived from theoretical developments.

Alternatively, the variance covariance matrix can be approximated empirically using simulation. The technique, called *bootstrapping*, is described by Algorithm 1.

Algorithm 1 Approximate the variance-covariance matrix by bootstrapping

- 1: Consider a sample of N observations.
 - 2: **for** $r = 1, \dots, R$ **do**
 - 3: Draw N observations from the sample with replacement.
 - 4: Calculate the maximum likelihood estimates $\hat{\beta}_r$ using the drawn sample.
 - 5: **end for**
 - 6: Calculate the empirical variance-covariance matrix of the vectors $\hat{\beta}_r$, $r = 1, \dots, R$.
-

Testing

t -tests

Michel Bierlaire

Introduction to choice models



Usage of the t -tests

t -test

Question

Is the parameter θ significantly different from a given value θ^* ?

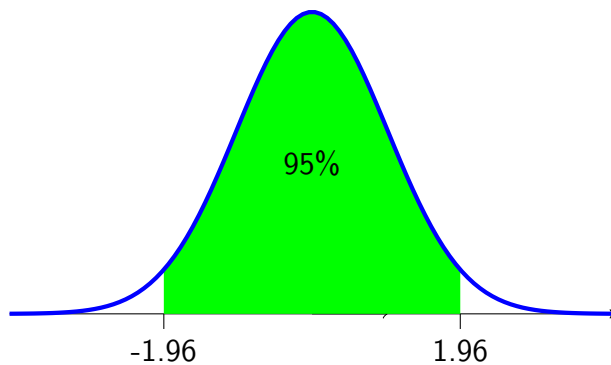
- ▶ $H_0 : \theta = \theta^*$
- ▶ $H_1 : \theta \neq \theta^*$

Statistic (assuming maximum likelihood estimator)

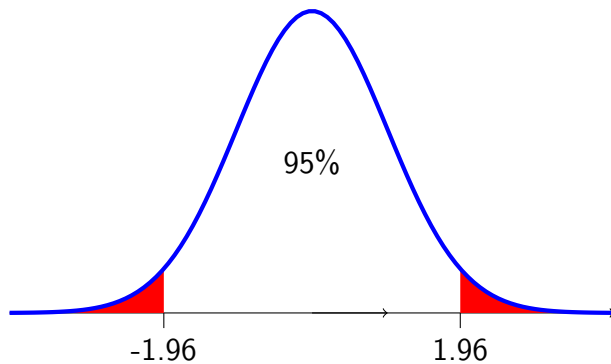
Under H_0 , if $\hat{\theta}$ is normally distributed with known variance σ^2

$$\frac{\hat{\theta} - \theta^*}{\sigma} \sim N(0, 1).$$

t -test: under H_0



t -test: if the statistic lies outside



H_0 is rejected at the 5% level.

Applying the test

Statistic

$$P(-1.96 \leq \frac{\hat{\theta} - \theta^*}{\sigma} \leq 1.96) = 0.95 = 1 - 0.05$$

Decision

H_0 can be rejected at the 5% level ($\alpha = 0.05$) if

$$\left| \frac{\hat{\theta} - \theta^*}{\sigma} \right| \geq 1.96.$$

Comments

- ▶ If $\hat{\theta}$ asymptotically normal
- ▶ If variance unknown
- ▶ A t test should be used with N degrees of freedom.
- ▶ When $N \geq 30$, the Student t distribution is well approximated by a $N(0, 1)$

p value

- ▶ probability to get a t statistic at least as large (in absolute value) as the one reported, under the null hypothesis
- ▶ it is calculated as

$$p = 2(1 - \Phi(t))$$

where $\Phi(\cdot)$ is the CDF of the standard normal.

- ▶ the null hypothesis is rejected when the p -value is lower than the significance level (typically 0.05)

Comparing two coefficients

Hypothesis

$$H_0 : \beta_1 = \beta_2.$$

Statistic

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_1 - \hat{\beta}_2)}}$$

where

$$\text{Var}(\hat{\beta}_1 - \hat{\beta}_2) = \text{Var}(\hat{\beta}_1) + \text{Var}(\hat{\beta}_2) - 2 \text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$$

Distribution

Under H_0 , distributed as $N(0, 1)$.

Testing – 6.3 t -tests

Michel Bierlaire

Practice quiz

The parameters of a model have been estimated using the Boeing dataset: `boeing.dat` (from the edX webpage), described in `BoeingDescription.pdf` (from the edX webpage). The output of the estimation is available in the file

`v634.Boeing.M0.html` (from the edX webpage).

Answer the following questions by performing t -tests:

1. Test the null hypothesis that the true value of the coefficient of the variable “being early” is zero?
2. Test the null hypothesis that the coefficients of the variables “elapsed time” for alternatives “non stop” and “one stop–same airline” are equal.
3. Test the null hypothesis that the coefficients of the variables “elapsed time” for alternatives “non stop” and “one stop–multiple airlines” are equal.
4. Test the null hypothesis that the coefficients of the variables “elapsed time” for alternatives “one-stop–same airline” and “one stop–multiple airlines” are equal.

Testing – 6.3 t -tests

Michel Bierlaire

Solution to the practice quiz

The estimation results are available in the file

`v634_Boeing_M0.html` (from the edX webpage).

1. Testing the null hypothesis that the true value of the coefficient of the variable “being early” is zero requires a t -test. The t statistic of parameter `SchedDelayEarly` is -8.02 which is larger in absolute value than 2.56, so the null hypothesis can be rejected at the 1% level. Actually, the fact that the p value is so small that the two first digits after the decimal point are zero, is a sign that the hypothesis can be safely rejected at any reasonable level. The variable plays a role in the model.
2. The next three questions require a t -test to compare two coefficients β_i and β_j . The null hypothesis is that both parameters are equal ($H_0 : \beta_i = \beta_j$) and the t -statistic is given by

$$\frac{\hat{\beta}_i - \hat{\beta}_j}{\sqrt{\text{Var}(\hat{\beta}_i - \hat{\beta}_j)}}$$

where

$$\text{Var}(\hat{\beta}_i - \hat{\beta}_j) = \text{Var}(\hat{\beta}_i) + \text{Var}(\hat{\beta}_j) - 2\text{Cov}(\hat{\beta}_i, \hat{\beta}_j).$$

The variance of a parameter is the square of its standard error. The complete variance-covariance matrix can be found in `v634_Boeing_M0.html` (from the edX webpage). It is reported in Table 1 for the involved coefficients.

B_ElapsedTime_1	B_ElapsedTime_2	0.00627
B_ElapsedTime_1	B_ElapsedTime_3	0.00600
B_ElapsedTime_2	B_ElapsedTime_3	0.00553

Table 1: Covariances for the involved coefficients

The three t -tests are applied below.

H_0 : B_ElapsedTime_1=B_ElapsedTime_2

$$\frac{\hat{\beta}_1 - \hat{\beta}_2}{\sqrt{\text{Var}(\hat{\beta}_1 - \hat{\beta}_2)}} = \frac{-0.341 - (-0.291)}{\sqrt{0.00729 + 0.00676 - 2 \times 0.00627}} = -1.28,$$

and the p -value is 0.2. Note that these two values are readily available in the HTML file. The null hypothesis can be rejected only at the 20% level. It is therefore reasonable not to reject it.

3. H_0 : B_ElapsedTime_2=B_ElapsedTime_3

$$\frac{\hat{\beta}_2 - \hat{\beta}_3}{\sqrt{\text{Var}(\hat{\beta}_2 - \hat{\beta}_3)}} = \frac{-0.291 - (-0.310)}{\sqrt{0.00676 + 0.00643 - 2 \times 0.00553}} = 0.41,$$

and the p -value is 0.68. Note that these two values are readily available in the HTML file. The null hypothesis can be rejected only at the 68% level. It is therefore reasonable not to reject it.

4. H_0 : B_ElapsedTime_1=B_ElapsedTime_3

$$\frac{\hat{\beta}_1 - \hat{\beta}_3}{\sqrt{\text{Var}(\hat{\beta}_1 - \hat{\beta}_3)}} = \frac{-0.341 - (-0.310)}{\sqrt{0.00729 + 0.00643 - 2 \times 0.006}} = -0.74,$$

and the p -value is 0.46. Note that these two values are readily available in the HTML file. The null hypothesis can be rejected only at the 46% level. It is therefore reasonable not to reject it.

In conclusion, we have no evidence from the data that suggests that the coefficient of the variable “elapsed time” is alternative specific. Consequently, in such circumstances, it may be worth investigating a model with a generic elapsed time, that will be more parsimonious.

The general motivation of the likelihood ratio test is to investigate parsimonious versions of a given specification, by introducing linear restrictions on the parameters. The null hypothesis of the test is that the parsimonious, or restricted, model is the true model. If it is rejected, the unrestricted model is preferred.

It can be shown (Wilks, 1938) that under the null hypothesis H_0 , the statistic

$$-2(\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_U)) \sim \chi^2_{(K_U - K_R)}, \quad (1)$$

where

- $\mathcal{L}(\hat{\beta}_R)$ is the log likelihood of the restricted model,
- $\mathcal{L}(\hat{\beta}_U)$ is the log likelihood of the unrestricted model,
- K_R is the number of parameters in the restricted model, and,
- K_U is the number of parameters in the unrestricted model.

The simple hypothesis (that is, the restricted model is correct) and the composite hypothesis (that is, the unrestricted model is correct) are said to be *nested*¹, because the former can be obtained from the latter using linear restrictions. Note that the test can only be applied for nested hypotheses.

The test can be written in terms of likelihood. As

$$\mathcal{L}(\hat{\beta}_R) = \log \mathcal{L}^*(\hat{\beta}_R) \text{ and } \mathcal{L}(\hat{\beta}_U) = \log \mathcal{L}^*(\hat{\beta}_U), \quad (2)$$

we can write (1) as

$$-2 \log \frac{\mathcal{L}^*(\hat{\beta}_R)}{\mathcal{L}^*(\hat{\beta}_U)} \sim \chi^2_{(K_U - K_R)}, \quad (3)$$

that explains the name of the test.

References

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses, *Ann. Math. Statist.* **9**(1): 60–62.

URL: <http://dx.doi.org/10.1214/aoms/1177732360>

¹This has nothing to do with the nested logit model.

Testing

Likelihood ratio test

Michel Bierlaire

Introduction to choice models



Applications of the likelihood ratio test

Benchmarking

Unrestricted model

$$V_{in} = \beta_1 x_{ink} + \dots$$

$$V_{jn} = \beta_2 x_{jnk} + \dots$$

$$\vdots$$

Restricted model

Equal probability model

$$V_{in} = 0$$

$$V_{jn} = 0$$

$$\vdots$$

Restrictions

$$\beta_k = 0, \forall k$$

Benchmarking

Log likelihood of the unrestricted model

$$\mathcal{L}(\hat{\beta})$$

Log likelihood of the restricted model

$$P_{in} = 1/J_n, \forall i \in \mathcal{C}_n, \forall n$$

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

Statistic

$$-2(\mathcal{L}(0) - \mathcal{L}(\hat{\beta})) \sim \chi_K^2$$

Benchmarking revisited

Unrestricted model

$$V_{in} = \beta_1 x_{ink} + \dots$$

$$V_{jn} = \beta_2 x_{jnk} + \dots$$

$$\vdots$$

Restricted model

Only alternative specific constants

$$V_{in} = \beta_i$$

$$V_{jn} = \beta_j$$

$$\vdots$$

Restrictions

All coefficients but the constants are constrained to zero.

Benchmarking revisited

Log likelihood of the unrestricted model

$$\mathcal{L}(\hat{\beta})$$

Log likelihood of the restricted model

$$P_{in} = N_i/N \quad \forall i \in \mathcal{C}, \forall n$$

$$\mathcal{L}(c) = \sum_{i=1}^J N_i \log(N_i/N)$$

Statistic

$$-2(\mathcal{L}(c) - \mathcal{L}(\hat{\beta})) \sim \chi_d^2 \text{ with } d = K - J + 1$$

Benchmarking

Classical output of estimation software

Summary statistics

Number of observations = 2544

$$\mathcal{L}(0) = -2794.870$$

$$\mathcal{L}(c) = -2203.160$$

$$\mathcal{L}(\hat{\beta}) = -1640.525$$

$$-2[\mathcal{L}(0) - \mathcal{L}(\hat{\beta})] = 2308.689$$

Test of generic attributes

Unrestricted model

Alternative specific

$$V_{in} = \beta_{1i}x_{ink} + \dots$$

$$V_{jn} = \beta_{1j}x_{jnk} + \dots$$

\vdots

Restricted model

Generic

$$V_{in} = \beta_1x_{ink} + \dots$$

$$V_{jn} = \beta_1x_{jnk} + \dots$$

\vdots

Restriction

$$\beta_{1i} = \beta_{1j} = \dots$$

Test of generic attributes

Log likelihood of the unrestricted model

$$\mathcal{L}(\hat{\beta}_{AS})$$

Log likelihood of the restricted model

$$\mathcal{L}(\hat{\beta}_G)$$

Statistic

$$-2(\mathcal{L}(\hat{\beta}_G) - \mathcal{L}(\hat{\beta}_{AS})) \sim \chi_d^2 \text{ with } d = K_{AS} - K_G$$

Test of taste variations

Segmentation

- ▶ Classify the data into G groups. Size of group g : N_g .
- ▶ The same specification is considered for each group.
- ▶ A different set of parameters is estimated for each group.

Test of taste variations

N_1	N_2	N_3	N_4	N
-------	-------	-------	-------	-----

$$\mathcal{L}_{N_1}(\hat{\beta}^1) \mathcal{L}_{N_2}(\hat{\beta}^2)$$

$$\mathcal{L}_{N_3}(\hat{\beta}^3)$$

$$\mathcal{L}_{N_4}(\hat{\beta}^4)$$

$$\sum_{g=1}^G \mathcal{L}_{N_g}(\hat{\beta}^g)$$

Test of taste variations

Unrestricted model

Group specific coefficients

$$V_{in} = \sum_{g=1}^G (\delta_{ng} \beta_{1g}) x_{ink} + \dots$$

$$V_{jn} = \sum_{g=1}^G (\delta_{ng} \beta_{2g}) x_{jnk} + \dots$$

\vdots

Restricted model

Generic coefficients

$$V_{in} = \beta_1 x_{ink} + \dots$$

$$V_{jn} = \beta_2 x_{jnk} + \dots$$

\vdots

Restrictions

$$\beta_{k1} = \beta_{k2} = \dots = \beta_{kG}, \forall k.$$

Test of taste variations

Log likelihood of the unrestricted model

$$\sum_{g=1}^G \mathcal{L}_{N_g}(\hat{\beta}^g)$$

Log likelihood of the restricted model

$$\mathcal{L}_N(\hat{\beta})$$

Statistic

$$-2 \left[\mathcal{L}_N(\hat{\beta}) - \sum_{g=1}^G \mathcal{L}_{N_g}(\hat{\beta}^g) \right] \sim \chi_d^2 \text{ with } d = \sum_{g=1}^G K - K = (G - 1)K.$$

Tests of nonlinear specifications

Unrestricted model

Power series

$$V_{in} = \sum_{\ell=1}^L \beta_{1\ell} \frac{x_{ink}^\ell}{x_{\text{ref}}} + \dots$$

$$V_{jn} = \beta_2 x_{jnk} + \dots$$

\vdots

Restricted model

Linear specification

$$V_{in} = \beta_1 x_{ink} + \dots$$

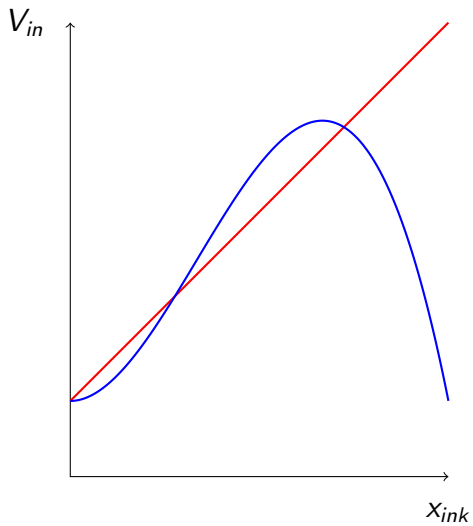
$$V_{jn} = \beta_2 x_{jnk} + \dots$$

\vdots

Restrictions

$$\beta_{12} = \beta_{13} = \dots = \beta_{1L} = 0$$

Power series



Test of nonlinear specifications

Log likelihood of the unrestricted model

$$\mathcal{L}(\hat{\beta}_U)$$

Log likelihood of the restricted model

$$\mathcal{L}(\hat{\beta}_R)$$

Statistic

$$-2 \left[\mathcal{L}(\hat{\beta}_R) - \mathcal{L}(\hat{\beta}_U) \right] \sim \chi_d^2 \text{ with } d = L - 1$$

Notes

- ▶ Usually not behaviorally meaningful
- ▶ Danger of overfitting
- ▶ Polynomials are most of the time inappropriate for extrapolation due to oscillation
- ▶ Other nonlinear specifications can be used for testing
 - ▶ Piecewise linear
 - ▶ Box-Cox

Testing – 6.4 Likelihood ratio test

Michel Bierlaire

Practice quiz

In a mode choice case study, consider the models with the following utility specifications (where the index n related to the individual has been dropped to simplify the notations):

1. Linear with generic coefficients

$$\begin{aligned}U_{\text{car}} &= ASC_{\text{car}} + \beta_{\text{tt}} \cdot \text{tt}_{\text{car}} + \beta_{\text{tc}} \cdot \text{tc}_{\text{car}} + \varepsilon_{\text{car}}, \\U_{\text{pt}} &= \beta_{\text{tt}} \cdot \text{tt}_{\text{pt}} + \beta_{\text{tc}} \cdot \text{tc}_{\text{pt}} + \varepsilon_{\text{pt}}.\end{aligned}$$

2. Linear with alternative specific coefficients

$$\begin{aligned}U_{\text{car}} &= ASC_{\text{car}} + \beta_{\text{tt,car}} \cdot \text{tt}_{\text{car}} + \beta_{\text{tc,car}} \cdot \text{tc}_{\text{car}} + \varepsilon_{\text{car}}, \\U_{\text{pt}} &= \beta_{\text{tt,pt}} \cdot \text{tt}_{\text{pt}} + \beta_{\text{tc,pt}} \cdot \text{tc}_{\text{pt}} + \varepsilon_{\text{pt}}.\end{aligned}$$

3. Power series

$$\begin{aligned}U_{\text{car}} &= ASC_{\text{car}} + \beta_{\text{tt}} \cdot \text{tt}_{\text{car}} + \beta_{\text{tc}} \cdot \text{tc}_{\text{car}} + \beta_{\text{tc.squared}} \cdot \text{tc}_{\text{car}}^2 + \varepsilon_{\text{car}}, \\U_{\text{pt}} &= \beta_{\text{tt}} \cdot \text{tt}_{\text{pt}} + \beta_{\text{tc}} \cdot \text{tc}_{\text{pt}} + \beta_{\text{tc.squared}} \cdot \text{tc}_{\text{pt}}^2 + \varepsilon_{\text{pt}}.\end{aligned}$$

4. Box-cox

$$\begin{aligned}U_{\text{car}} &= ASC_{\text{car}} + \beta_{\text{tt.boxcox}} \cdot \frac{(\text{tt}_{\text{car}} - 1)^\lambda}{\lambda} + \beta_{\text{tc}} \cdot \text{tc}_{\text{car}} + \varepsilon_{\text{car}}, \\U_{\text{pt}} &= \beta_{\text{tt.boxcox}} \cdot \frac{(\text{tt}_{\text{pt}} - 1)^\lambda}{\lambda} + \beta_{\text{tc}} \cdot \text{tc}_{\text{pt}} + \varepsilon_{\text{pt}}.\end{aligned}$$

5. Logarithm

$$\begin{aligned}U_{\text{car}} &= ASC_{\text{car}} + \beta_{\text{tt}} \cdot \text{tt}_{\text{car}} + \beta_{\text{tc.log}} \cdot \log(\text{tc}_{\text{car}}) + \varepsilon_{\text{car}}, \\U_{\text{pt}} &= \beta_{\text{tt}} \cdot \text{tt}_{\text{pt}} + \beta_{\text{tc.log}} \cdot \log(\text{tc}_{\text{pt}}) + \varepsilon_{\text{pt}}.\end{aligned}$$

6. Piecewise linear

$$\begin{aligned}
U_{\text{car}} &= ASC_{\text{car}} + \beta_{\text{tt}, < 15} \cdot \text{tt}_{\text{car}, < 15} + \beta_{\text{tt}, [15, 60)} \cdot \text{tt}_{\text{car}, [15, 60)} + \beta_{\text{tt}, \geq 60} \cdot \text{tt}_{\text{car}, \geq 60} \\
&\quad + \beta_{\text{tc}} \cdot \text{tc}_{\text{car}} + \varepsilon_{\text{car}}, \\
U_{\text{pt}} &= \beta_{\text{tt}, < 15} \cdot \text{tt}_{\text{pt}, < 15} + \beta_{\text{tt}, [15, 60)} \cdot \text{tt}_{\text{pt}, [15, 60)} + \beta_{\text{tt}, \geq 60} \cdot \text{tt}_{\text{pt}, \geq 60} + \beta_{\text{tc}} \cdot \text{tc}_{\text{pt}} + \varepsilon_{\text{pt}}.
\end{aligned}$$

where for $i \in \{\text{car}, \text{pt}\}$

$$\begin{aligned}
\text{tt}_{i, < 15} &= \begin{cases} \text{tt}_i, & \text{if } \text{tt}_i < 15 \\ 15, & \text{otherwise,} \end{cases} \\
\text{tt}_{\text{car}, [15, 60)} &= \begin{cases} 0, & \text{if } \text{tt}_i < 15 \\ \text{tt}_i - 15, & \text{if } \text{tt}_i \in [15, 60) \\ 60, & \text{otherwise,} \end{cases} \\
\text{tt}_{i, \geq 60} &= \begin{cases} 0, & \text{if } \text{tt}_i < 60 \\ \text{tt}_i - 60, & \text{otherwise.} \end{cases}
\end{aligned}$$

where tt_{car} and tt_{pt} are the travel times in minutes by car and public transportation respectively, tc_{car} and tc_{pt} are the travel costs in CHF of car and public transportation respectively; ASC_{car} , β 's and λ are parameters to be estimated; and $\varepsilon_{\text{car}}, \varepsilon_{\text{pt}} \stackrel{iid}{\sim} EV(0, 1)$.

When we want to test two of these models, when can we apply the likelihood ratio test?

Testing – 6.4 Likelihood ratio test

Michel Bierlaire

Solution to the practice quiz

- Models 1 and 2 can be compared using the likelihood ratio test: the linear restrictions are $\beta_{tt,car} = \beta_{tt,pt}$ and $\beta_{tc,car} = \beta_{tc,pt}$.
- Models 1 and 3 can be compared using the likelihood ratio test: the linear restriction is $\beta_{tc_squared} = 0$.
- Models 1 and 4 can be compared using the likelihood ratio test: the linear restriction is $\lambda = 1$.
- Models 1 and 6 can be compared using the likelihood ratio test: the linear restrictions are $\beta_{tt,<15} = \beta_{tt,[15,60)} = \beta_{tt,\geq 60}$.

No other pair of models can be compared using a likelihood ratio test, as none of them can be obtained from the other one using linear restrictions.

Testing – 6.4 Likelihood ratio test

Michel Bierlaire

Practice quiz: Boeing

Question

1. Consider the two models (M1_leisure and M2) for which the python-biogeme files have been provided:
 - `v643_Boeing_M1_leisure.py` (from the edX webpage),
 - `v643_Boeing_M2.py` (from the edX webpage).

Both models can be compared using a likelihood ratio test. Justify why, explicitly writing the linear restrictions. Which model is better according to a likelihood ratio test?

2. Consider models M0 and M3, for which the pythonbiogeme files have been provided:
 - `v643_Boeing_M0.py` (from the edX webpage),
 - `v643_Boeing_M3.py` (from the edX webpage).

Is the elapsed time perceived differently depending on the alternative considered? Perform a likelihood ratio test to answer, and justify why you can use it by explicitly writing the linear restrictions.

3. Consider model M1, for which the pythonbiogeme file has been provided. Is there taste variation across all parameters depending on the trip purpose? *Hint: use models*
 - `v643_Boeing_M1_leisure.py` (from the edX webpage),
 - `v643_Boeing_M1_nonleisure.py` (from the edX webpage) and

- *v643_Boeing_M1.py* (from the edX webpage)

Testing – 6.4 Likelihood ratio test

Michel Bierlaire

Solution to practice quiz: Boeing

1. The estimation results of models M1_leisure and M2 are available in the files
 - v643_Boeing_M1_leisure.html (from the edX webpage), and
 - v643_Boeing_M2.html (from the edX webpage).

We can compare both by using a likelihood ratio test (LRT) since we can define H_0 using linear restrictions:

- LegroomOpt1_Male = LegroomOpt23_Male,
- LegroomOpt1_Female = LegroomOpt23_Female,
- FareOverIncome = 0.

Following the notations from the previous lecture,

- $\mathcal{L}(\hat{\beta}_U) = -1640.525$,
- $\mathcal{L}(\hat{\beta}_R) = -1652.573$,
- $K_U = 15$, and
- $K_R = 12$.

So the likelihood ratio test is

$$-2(-1652.573 + 1640.525) = 24.096.$$

The number of degrees of freedom for the test is $K_U - K_R = 3$, which happens also to be the number of linear restrictions. The threshold for the test at 5% level is therefore $\chi^2_{3,0.05} = 7.81$. Since $24.096 > 7.81$, we can reject H_0 at the 5% level. According to the LRT, the hypothesis that the two models are equivalent can be rejected. Therefore, the unrestricted model (M1_leisure) is preferred.

2. The estimation results of models M0 and M3 are available in

- `v643_Boeing_M0.html` (from the edX webpage), and
- `v643_Boeing_M3.html` (from the edX webpage).

The LRT can be used since we can define H_0 using linear restrictions: `B_ElapsedTime_1=B_ElapsedTime_2=B_ElapsedTime_3`. Following the notations from the previous lecture,

- $\mathcal{L}(\hat{\beta}_R) = -1642.796$,
- $\mathcal{L}(\hat{\beta}_U) = -1641.932$,
- $K_U = 15$, and,
- $K_R = 13$.

So the likelihood ratio test is

$$-2(-1642.796 + 1641.932) = 1.728.$$

The number of degrees of freedom for the test is $K_U - K_R = 2$. The threshold of interest is here $\chi^2_{2,0.05} = 5.99$. Since $1.728 < 5.99$, we cannot reject the hypothesis that the two models are equivalent at the 5% level. Therefore, the restricted model (M3) is preferred, as it is more parsimonious.

Note that it yields the same conclusion as the practice quiz in the previous section when we used a t -test. This is not always the case. It may happen that testing the same hypothesis with two different tests yields to different conclusions.

3. The likelihood ratio test can also be used to test variations between market segments in the following way. We have two groups of travelers: leisure (2544) and non leisure (1065). We estimate the exact same model specification on

- the full population (`v643_Boeing_M1.html` (from the edX webpage)),
- leisure travelers only (`v643_Boeing_M1_leisure.html` (from the edX webpage)),

- non leisure travelers only (`v643.Boeing M1_nonleisure.html` (from the edX webpage)).

The LRT can be used since the null hypothesis is defined using linear restrictions, that is each parameter estimated for the first segment is equal to the corresponding parameter estimated for the other segment. The number of restrictions is therefore equal to the number of parameters in the original model.

Table 1 summarizes the data extracted from the estimation results, needed to perform a likelihood ratio test. Note that the *unrestricted* values are obtained from the sum of the rows *leisure* and *non leisure*: Therefore, the likelihood ratio test is:

Model	$\mathcal{L}(\hat{\beta})$	Sample size	K
Restricted	-2300.453	3609	15
Leisure	-1640.525	2544	15
Non leisure	-629.080	1065	15
Unrestricted	-2269.605	3609	30

Table 1: Summary of the estimation results

$$-2(-2300.453 + 2269.605) = 61.696.$$

The threshold of interest is here $\chi^2_{15,0.05} = 25.00$. We can therefore reject the null hypothesis that there is no taste variation across trip purposes.

Testing

Non nested hypotheses

Michel Bierlaire

Introduction to choice models



The Cox test

Non nested hypotheses

Nested hypotheses

- ▶ Restricted and unrestricted models
- ▶ Linear restrictions
- ▶ H_0 : restricted model is correct
- ▶ Test: likelihood ratio test

Non nested hypotheses

- ▶ Need to compare two models
- ▶ None of them is a restriction of the other
- ▶ Likelihood ratio test cannot be used

Example

Model 1

$$V_{in} = \beta_1 x_{ink} + \cdots$$

$$V_{jn} = \beta_2 x_{jnk} + \cdots$$

$$\vdots$$

Model 2

$$V_{in} = \beta_1 \log(x_{ink}) + \cdots$$

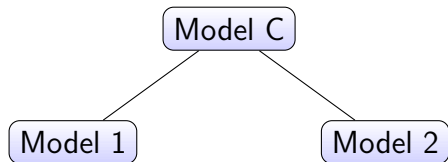
$$V_{jn} = \beta_2 \log(x_{jnk}) + \cdots$$

$$\vdots$$

Cox test

Back to nested hypotheses

- ▶ We want to test model 1 against model 2
- ▶ We generate a composite model C such that both models 1 and 2 are restricted cases of model C.



Example

Model 1

$$V_{in} = \beta_1 x_{ink} + \dots$$

$$V_{jn} = \beta_2 x_{jnk} + \dots$$

$$\vdots$$

Model 2

$$V_{in} = \beta_1 \log(x_{ink}) + \dots$$

$$V_{jn} = \beta_2 \log(x_{jnk}) + \dots$$

$$\vdots$$

Model C

$$V_{in} = \beta_{11} x_{ink} + \beta_{12} \log(x_{ink}) + \dots$$

$$V_{jn} = \beta_{21} x_{jnk} + \beta_{22} \log(x_{jnk}) + \dots$$

$$\vdots$$

Cox test

Testing

- ▶ We test 1 against C using the likelihood ratio test
- ▶ We test 2 against C using the likelihood ratio test

Conclusions

C against 1	C against 2	Conclusion
1 is not rejected	2 is rejected	Prefer 1
1 is rejected	2 is not rejected	Prefer 2
1 is rejected	2 is rejected	Develop better models
1 is not rejected	2 is not rejected	Use another test

Testing – 6.5 Non nested hypotheses

Michel Bierlaire

Davidson and McKinnon J test.

A disadvantage of the Cox test is the need to estimate a model with a potentially very large number of parameters. We now describe the J test developed by Davidson and MacKinnon (1981) which is a general solution to the selection between two non-nested models. The J test is in general preferred to the Cox test. As we will see, it is also subject to the same four outcomes.

This is a general treatment based on generating artificial regressions that embed two competing non nested model formulations to explain a given dependent variable. Consider two specifications:

$$M_1 : U_{in} = V_{in}^{(1)}(x_{in}; \beta) + \varepsilon_{in}^{(1)}, \quad (1)$$

$$M_2 : U_{in} = V_{in}^{(2)}(x_{in}; \gamma) + \varepsilon_{in}^{(2)}. \quad (2)$$

To choose between model 1 in Equation (1) and model 2 in Equation (2), we consider the following composite specification:

$$M_C : U_{in} = (1 - \alpha)V_{in}^{(1)}(x_{in}; \beta) + \alpha V_{in}^{(2)}(x_{in}; \gamma) + \varepsilon_{in}. \quad (3)$$

Intuitively, the idea is to test the competing models against the composite model in equation (3). Note that if $\alpha = 0$, the model collapses to the model M_1 while with $\alpha = 1$, the composite model collapses to the model M_2 . The major problem is that very often, the composite model cannot be estimated. Namely, there may be exact multicollinearity among the explanatory variables. Moreover, the α coefficient may not be identified.

The J test solution to this problem is to replace the unknown parameters not being tested by consistent estimates. In order to test M_1 , one could consider the following composite model:

$$M_C : U_{in} = (1 - \alpha)V_{in}^{(1)}(x_{in}; \beta) + \alpha V_{in}^{(2)}(x_{in}; \hat{\gamma}) + \varepsilon_{in}, \quad (4)$$

where model 2 in Equation (2) has been previously estimated, and $\hat{\gamma}$ is the vector of estimates. Thus, $V_{in}^{(2)}(x_{in}; \hat{\gamma})$ corresponds to the fitted systematic utility of model 2 and represents in this artificial model a single variable associated with the parameter α . Under the null hypothesis that model 1 is correct, the true value of α in the composite model is 0. The objective is then to test if $\alpha = 0$ using a t test. This would involve estimating model 1 with the additional variable computed as $V_{in}^{(2)}(x_{in}; \hat{\gamma})$.

In order to test M_2 , one could instead consider the following composite model:

$$M_C : U_{in} = (1 - \alpha)V_{in}^{(1)}(x_{in}; \hat{\beta}) + \alpha V_{in}^{(2)}(x_{in}; \gamma) + \varepsilon_{in}. \quad (5)$$

where model 1 in Equation (1) has been previously estimated. Thus, $V_{in}^{(1)}(x_{in}; \hat{\beta})$ corresponds to the fitted systematic utility of model 1 and represents in this artificial model a single variable associated with the parameter $(1 - \alpha)$. Under the null hypothesis that model 2 is correct, the true value of α is 1. The objective is then to test if $\alpha = 1$ using a t test.

The J test has the same four outcomes presented for the Cox test, that is

- M_1 is rejected and M_2 is not rejected. Then, it is reasonable to prefer model 2.
- M_1 is not rejected and M_2 is rejected. Then it is reasonable to prefer model 1.
- M_1 and M_2 are rejected. This indicates that better models should be developed.
- Neither M_1 nor M_2 can be rejected. Then, in this case, the data does not seem to be informative enough to distinguish between the two competing models, and the $\bar{\rho}^2$ should be used.

It should now be clear that one of the two models does not have to represent the truth. Both models could be unsatisfactory.

References

Davidson, R. and MacKinnon, J. (1981). Several tests for model specification in the presence of alternative hypotheses, *Econometrica: Journal of the Econometric Society* pp. 781–793.

Testing – 6.5 Non nested hypotheses

Michel Bierlaire

Adjusted likelihood ratio index.

The likelihood ratio index is defined as

$$\rho^2 = 1 - \frac{\mathcal{L}(\hat{\beta})}{\mathcal{L}(0)}, \quad (1)$$

and called “rho-squared” by analogy to the R^2 in regression analysis. Note that it is not the square of anything. And even if it generally lies between zero (when $\mathcal{L}(\hat{\beta}) = \mathcal{L}(0)$) and one (when the model perfectly fits the data and $\mathcal{L}(\hat{\beta}) = 0$), there is no absolute interpretation of its value.

For the same estimation data set, the ρ^2 of a model always increases or at least stays the same whenever new variables are added to the model, even if they are actually irrelevant. This is the motivation to use the adjusted likelihood ratio index (rho bar squared):

$$\bar{\rho}^2 = 1 - \frac{\mathcal{L}(\hat{\beta}) - K}{\mathcal{L}(0)}. \quad (2)$$

where K is the number of unknown parameters in the model.

The adjusted likelihood ratio index $\bar{\rho}^2$ can be used for testing non nested hypotheses of discrete choice models. Under the null hypothesis that model 1 is the true specification, compared to model 2, the following holds asymptotically:

$$\Pr(\bar{\rho}_2^2 \geq z + \bar{\rho}_1^2) \leq \Phi\{-\sqrt{-2z\mathcal{L}(0) + (K_1 - K_2)}\}, \quad z > 0, \quad (3)$$

where

- $\bar{\rho}_\ell^2$ is the the adjusted likelihood ratio index for model ℓ ,

- K_ℓ is the number of parameters in model ℓ ,
- $\Phi(\cdot)$ is the standard normal cumulative distribution function.

This can be used to answer the following question: if we select the model with the largest $\bar{\rho}^2$, what is the probability that we make a mistake? In other words, if the true model is model 1, what is the probability that $\bar{\rho}_2^2$ is larger than $z + \bar{\rho}_1^2$, for some threshold z ?

When all N observations in the sample have all J alternatives, we have

$$\mathcal{L}(0) = -N \log J \quad (4)$$

the bound becomes

$$\Pr(\bar{\rho}_2^2 \geq z + \bar{\rho}_1^2) \leq \Phi\{-\sqrt{2Nz \log J + (K_1 - K_2)}\}, \quad z > 0. \quad (5)$$

Consider an example with $J = 2$ alternatives, $N = 300$ observations and two models with the same number of parameters ($K_1 = K_2$). Then we have

$$\Pr(\bar{\rho}_2^2 \geq 0.0001 + \bar{\rho}_1^2) \leq 41.92\% \quad (6)$$

$$\Pr(\bar{\rho}_2^2 \geq 0.001 + \bar{\rho}_1^2) \leq 25.95\% \quad (7)$$

$$\Pr(\bar{\rho}_2^2 \geq 0.01 + \bar{\rho}_1^2) \leq 2.07\% \quad (8)$$

For $z = 0.1$ and above, the probability is practically zero.

Testing – 6.5 Non nested hypotheses

Michel Bierlaire

Practice quiz

Consider the specification files

- `v655_Boeing_M1_leisure.py` (from the edX webpage) and
- `v655_Boeing_M1_leisure_sq.py` (from the edX webpage),

and the associate dataset: `boeing.dat` (from the edX webpage). You shall use them to estimate the parameters of the models and to perform the following tasks:

1. Compare both models using a Cox test. Which of the two models should be preferred?
2. Compare both models using a Davidson and McKinnon J -test. Which of the two models should be preferred?

Testing – 6.5 Non-nested hypotheses

Michel Bierlaire

Solution to the practice quiz

1. None of the two models can be obtained using linear restrictions from the other one. Consequently, we cannot use a likelihood ratio test. Instead we use a Cox test. To do so, we define a composite model that contains a combination of the specifications of each model:

`v655_Boeing_M1_leisure_composite.py` (from the edX webpage).

By design, each of the two models to be tested can be obtained using linear restrictions of the composite model. Therefore, we can apply two likelihood ratio tests. The estimation results are found in the files

- `v655_Boeing_M1_leisure.html` (from the edX webpage),
- `v655_Boeing_M1_leisure_sq.html` (from the edX webpage), and
- `v655_Boeing_M1_leisure_composite.html` (from the edX webpage).

Table 1 shows a summary of the final log likelihood of the three models. Finally, Table 2 shows a summary of the likelihood ratio tests between the composite model and the other two. As *M1_leisure_sq* is rejected, and not *M1_leisure*, the latter is preferred.

2. We can use two Davidson and McKinnon *J*-tests to compare these models, as described below. The first test considers the following null hypothesis:

H_0 : the linear specification is correct.

We generate the Pythonbiogeme file for this test as follows:

Model	$\mathcal{L}(\hat{\beta})$	K
M1_leisure	-1640.525	15
M1_leisure_sq	-1649.407	15
M1_leisure_composite	-1640.487	17

Table 1: Final log likelihood and number of parameters for the three models

	Statistic	Threshold	Outcome
leisure <i>vs.</i> composite	0.076	5.99	Cannot reject M1_leisure
leisure_sq <i>vs.</i> composite	17.84	5.99	Reject M1_leisure_sq

Table 2: Likelihood ratio test between models

- (a) Copy the file

`v655_Boeing_M1_leisure.py` (from the edX webpage)

and rename it to *v655_Boeing_M1_leisure_Jtest.py*.

- (b) Add to this file a parameter to be estimated, α .

- (c) Copy the values of the estimated parameters from

`v655_Boeing_M1_leisure_sq_param.py` (from the edX webpage),

that has been created by Pythonbiogeme. Do not give them the same name, so that there is no conflict with the parameters that are going to be estimated. For example, add *_quadratic* to their names.

- (d) Make sure that these parameters are not re-estimated, either by assigning their value to a Python variable:

`B_SM_TIME_LOG_M3 = -1.47858`

or, if you use the Beta statement, by setting the 5th arguments to 1 instead of 0:

`B_SM_TIME_LOG_M3 =
Beta('B_SM_TIME_LOG_M3', -1.47858, -10, 10, 1)`

- (e) Add to this file the utility functions from the quadratic specification, and rename them, so that there is no conflict with the utility

functions that are already in the file. For example, add *_quadratic* to their names. Remember to use the renamed parameters.

- (f) Write the expression of the utility functions by combining the linear one, multiplied by $(1 - \alpha)$ and the quadratic one, multiplied by α .

You can find the Pythonbiogeme file created as explained above in

`v655_Boeing_M1_leisure_Jtest.py` (from the edX webpage).

The estimation results of the model used to test this are available in the file

`v655_Boeing_M1_leisure_Jtest.html` (from the edX webpage).

Under H_0 , the true value of α is 0, so we can use a t -test. The t -statistic for α is -0.23, which is smaller in absolute value than 1.96. Therefore we cannot reject the null hypothesis that the linear specification is correct.

- 3. The second test considers the following null hypothesis:

H_0 : the quadratic specification is correct.

The Pythonbiogeme file to test this can be generated analogously. It is available here:

`v655_Boeing_M1_leisure_sq_Jtest.py` (from the edX webpage)

The estimation results of the model used to test this hypothesis are available in the file

`v655_Boeing_M1_leisure_sq_Jtest.html` (from the edX webpage).

Under H_0 , the true value of α is 0, so we can again use a t -test. The t -statistic for α is 3.89, which is larger than 1.96. Therefore the null hypothesis that the quadratic specification is correct can be rejected.

Those two tests suggest to prefer the linear model *M1_leisure*. Note that we reach here the same conclusion as by using a Cox test. It is not always necessarily the case.

Testing

Prediction tests

Michel Bierlaire

Introduction to choice models



Outlier analysis

Outlier analysis

Procedure

- ▶ Apply the model on the sample
- ▶ Examine observations where the predicted probability is the smallest for the observed choice
- ▶ Test model sensitivity to outliers, as a small probability has a significant impact on the log likelihood
- ▶ Potential causes of low probability:
 - ▶ Coding or measurement error in the data
 - ▶ Model misspecification
 - ▶ Inexplicable variation in choice behavior

Coding or measurement error in the data

Look for signs of data errors

- ▶ Travel time is negative
- ▶ Number is coded as a string
- ▶ etc.

Correct or remove the observation

- ▶ Go back to the original survey
- ▶ Correct only if you are certain

Model misspecification

Improve the specification

- ▶ Seek clues of missing variables from the observation.
- ▶ Why is the model associating such a low probability for this choice?
- ▶ Did we forget to account for age, income, or any other variable ?
- ▶ Should a nonlinear specification be investigated?
- ▶ Use a behavioral intuition.

Inexplicable variation in choice behavior

Keep the observation

- ▶ If no acceptable explanation is found, keep the observation.
- ▶ Avoid overfitting of the model to the data.
- ▶ The model should reflect how people behave, not how they should behave.

Testing

Prediction tests

Michel Bierlaire

Introduction to choice models



Out of sample validation

Cross-validation

Motivation

- ▶ Purpose of the model: prediction.
- ▶ Is the model able to predict?

Cross-validation

Motivation

- ▶ Purpose of the model: prediction.
- ▶ Is the model able to predict?



Cross-validation

Motivation

- ▶ Purpose of the model: prediction.
- ▶ Is the model able to predict?

Estimation 80%	Validation 20%
-------------------	-------------------

Methodology

Split the sample

- ▶ Decide the size of the validation set (e.g. 20%)
- ▶ Draw randomly an estimation set and a validation set.
- ▶ Repeat R times.

Evaluate

- ▶ For each pair of estimation/validation set...
- ▶ Estimate the parameters of the model with the estimation set.
- ▶ Calculate a measure of fit of the estimated model on the validation set.
- ▶ Typically, the likelihood.
- ▶ Calculate the average measure of fit.
- ▶ Select the model with the highest average fit on the validation sets.

Testing – 6.6 Prediction tests

Michel Bierlaire

Goodness of fit measures

Various goodness of fit measures have been proposed in the literature to perform the “out of sample” prediction test. However, we follow Diersen and Manfredo (1998) and Norwood et al. (2001), and suggest to use the likelihood of the validation set as the measure for comparing the quality of the models. If i_n is the alternative chosen by individual n in the validation sample, the log likelihood of the sample is

$$\sum_{n=1}^N \log P(i_n|x_n), \quad (1)$$

where x_n is the vector of explanatory variables for individual n .

Another useful measure is the expected number of correctly predicted observations:

$$\sum_{n=1}^N P(i_n|x_n). \quad (2)$$

Note that this number is often incorrectly calculated by counting a success when the chosen alternative is the one with the highest probability given by the model. This should be avoided by all means.

References

- Diersen, M. A. and Manfredo, M. R. (1998). Forecast evaluation: A likelihood scoring method, *Proceedings of the NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management, Chicago, IL*.

Norwood, B., Ferrier, P. and Lusk, J. (2001). Model selection criteria using likelihood functions and out-of-sample performance, *NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management, St. Louis, Missouri, April*, pp. 23–24.

Introduction to choice models

Michel Bierlaire – Virginie Lurkin

Week 7

Forecasting

Aggregation

Michel Bierlaire

Introduction to choice models



Aggregation

Motivation



- ▶ Prediction about a single individual is of little use in practice.
- ▶ Need for indicators about aggregate demand.
- ▶ Typical application: aggregate market shares.

Aggregation

- ▶ Disaggregate model:

$$P_n(i|x_n; \theta)$$

- ▶ Obtain x_n for each individual n in the population.
- ▶ Question: why is \mathcal{C}_n omitted?

Aggregate market shares

Number of individuals choosing alternative i

$$N_T(i) = \sum_{n=1}^{N_T} P_n(i|x_n; \theta).$$

Share of the population choosing alternative i

$$W(i) = \frac{1}{N_T} \sum_{n=1}^{N_T} P(i|x_n; \theta) = \mathbb{E}[P(i|x_n; \theta)].$$

Aggregation

Population	Alternatives				Total
	1	2	...	J	
1	$P(1 x_1; \theta)$	$P(2 x_1; \theta)$...	$P(J x_1; \theta)$	1
2	$P(1 x_2; \theta)$	$P(2 x_2; \theta)$...	$P(J x_2; \theta)$	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	$P(1 x_N; \theta)$	$P(2 x_N; \theta)$...	$P(J x_N; \theta)$	1
Total	$N(1)$	$N(2)$...	$N(J)$	N

Large table

When the table has too many rows...

apply sample enumeration.

When the table has too many columns...

apply micro simulation.

Forecasting

Aggregation

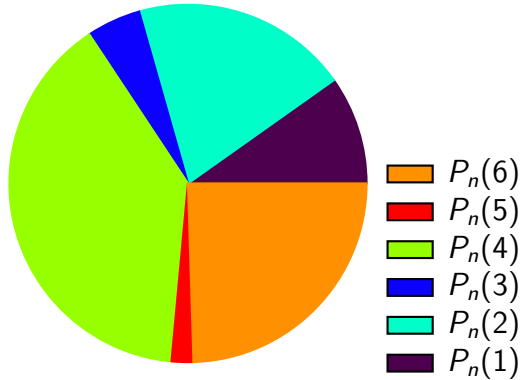
Michel Bierlaire

Introduction to choice models



Microsimulation

Microsimulation



Microsimulation

Simulated choice

- ▶ For each observation, draw R times from the choice model.
- ▶ Define $\hat{y}_{inr} = 1$ if alternative i has been generated by draw r , 0 otherwise.
- ▶ Approximation:

$$P_n(i|x_n; \theta) \approx \frac{1}{R} \sum_{r=1}^R \hat{y}_{inr}.$$

Warning

It is **invalid** to select the alternative with the highest probability.

Aggregate market shares

Number of individuals choosing alternative i

$$N(i) = \frac{1}{R} \sum_{n=1}^N \sum_{i=1}^R \hat{y}_{inr}.$$

Share of the population choosing alternative i

$$N(i) = \frac{1}{N} \frac{1}{R} \sum_{n=1}^N \sum_{i=1}^R \hat{y}_{inr}.$$

Microsimulation

For each r

Population	Alternatives				Total
	1	2	\dots	J	
1	\hat{y}_{11r}	\hat{y}_{21r}	\dots	\hat{y}_{J1r}	1
2	\hat{y}_{12r}	\hat{y}_{22r}	\dots	\hat{y}_{J2r}	1
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
N	\hat{y}_{1Nr}	\hat{y}_{2Nr}	\dots	\hat{y}_{JNr}	1
Total	$N(1)$	$N(2)$	\dots	$N(J)$	N

Microsimulation

In practice

Population	Draw			
	1	2	\dots	R
1	i_{11}	i_{12}	\dots	i_{1R}
2	i_{21}	i_{22}	\dots	i_{2R}
\vdots	\vdots	\vdots	\vdots	\vdots
N	i_{N1}	i_{N2}	\dots	i_{NR}

Given a choice model $P(i|x_n)$ that has been estimated from data, the predicted share of the population of N individuals choosing alternative i is given by

$$W(i) = \frac{1}{N} \sum_{n=1}^N P(i|x_n; \theta) = \mathbb{E} [P(i|x_n; \theta)]. \quad (1)$$

In practice, the population is often too large for the analyst to have access to each x_n vector, or even to their distribution. We introduce here a practical method to estimate $W(i)$ called *sample enumeration*.

The idea is to draw a sample from the population. It is actually possible to use the same sample used for the estimation of the parameters, but only if it consists of revealed preference data, that is data where the actual choice has been observed. Stated preferences data, where respondents are exposed to hypothetical scenarios, cannot be used for aggregation and prediction.

It is usually infeasible in practice to collect a purely random sample, where each individual in the population has exactly the same probability to be considered. A method called *stratified random sampling* is more realistic to implement.

It consists in partitioning the population into G mutually exclusive and collectively exhaustive groups, each called a stratum. This segmentation is motivated by the logistic of the data collection, and by the objectives of the survey. For instance, each stratum can be a geographical territory (a city, a county, etc.), where a local coordinator can be assigned. Or the partition can be organized by age, because we are interested in the impact of age on the choice behavior, like in the simple example presented in the beginning of the course.

Once the partitioning is defined, we sample S_g observations in each stratum g , using simple random sampling. The total size of the sample is $S = \sum_{g=1}^G S_g$.

Contrarily to simple random sampling, stratified sampling generates samples where some groups are proportionally more represented in the sample than they are in the population. This has to be taken into account when inferring quantities related to the population from the same quantities calculated with the sample.

To do that, each group is associated with a weight:

$$\omega_g = \frac{N_g}{N} \frac{S}{S_g} = \frac{\text{share of persons in segment } g \text{ in the population}}{\text{share persons in segment } g \text{ in the sample}}. \quad (2)$$

As each individual n belongs to exactly one stratum g , we define

$$\omega_n = \sum_{g=1}^G \delta_{ng} \omega_g, \quad (3)$$

where $\delta_{ng} = 1$ if individual n belongs to stratum g , and 0 otherwise.

Therefore, an estimate of the predicted share (1) of the population choosing alternative i is given by

$$\widehat{W}(i) = \frac{1}{S} \sum_{n=1}^S \omega_n P(i|x_n; \theta). \quad (4)$$

Forecasting – 7.1 Aggregation

Michel Bierlaire

Description of the simulation file

Once a model has been estimated, it can be used to derive useful indicators. PythonBiogeme provides a simulation feature for this purpose. In this document, we summarize the procedure. Use it as a reference when you work on the practice quiz.

We refer to the specification file `.py` containing all the instructions as the *simulation file*. This file is generated as follows:

1. Consider a model estimation file (called `model.py`, say) that has been already treated by Biogeme. In particular, the file `model_param.py` has been generated.
2. Generate a copy of `model.py` and rename it (called `simul.py`, say).
3. Replace all `Beta` statements by the equivalent statements including the estimated values. These statements can be found in the file

`model_param.py`

that is generated automatically when the parameters of the model are estimated.

4. Copy and paste the code for the sensitivity analysis, which can also be found in the file `model_param.py`:
 - the names of the parameters: the line starting with `names=...`
 - the values of the variance-covariance matrix: the line starting with `values=...`

- the definition of the matrix itself, for instance:

```
vc = bioMatrix(9,names,values)
BIOGEME OBJECT.VARCOVAR = vc
```

5. Make sure that the weight is properly defined by the statement
BIOGEME.OBJECT.WEIGHT =
6. Replace the statement related to the estimation
BIOGEME.OBJECT.ESTIMATE = Sum(1,'obsIter')
by the statement for simulation:
BIOGEME.OBJECT.SIMULATE = Enumerate(simulate,'obsIter').
7. The `simulate` variable must be a dictionary describing what has to be calculated during the sample enumeration. For example, if we want to calculate the choice probability of each alternative of a binary choice model for each individual in the sample, we define the `simulate` variable as follows:

```
simulate = {'Prob. alt 1': prob_1, 'Prob. alt 2': prob_2},
```

where `prob_1` and `prob_2` have been defined with the appropriate formulas. Each entry of this dictionary corresponds to an indicator that must be calculated, that is composed of two element: a key and a formula. The key of the entry is a string, that is used for the reporting. The value must be a valid formula describing the calculation. In this example, we define `prob_1` and `prob_2` as

```
prob_1 = bioLogit(V,av,1)
prob_2 = bioLogit(V,av,2)
```

which calculates the choice probability of each alternative as provided by the logit model.

8. We note that the simulated indicators are reported in alphabetical order. If a specific order is desired, the keys can be modified to do so. For instance:

```
simulate = {'01 Prob.alt 2': prob_2, '02 Prob.alt 1': prob_1}
```

The simulation is performed using the statement:

`pythonbiogeme simul dataset.dat`

that generates the file `simul.html` including the following sections:

- The preamble reports information about the version of PythonBiogeme, useful URLs and the names of the files involved in the run.
- Statistics: this section is the same as for the estimation, and reports the requested statistics.
- The simulation report, which contains two parts: the *detailed records* and the *aggregate values*.
 - The *detailed records* is a table where each row corresponds to an entry in the sample file, and each (group of) column(s) to an entry in the dictionary defined by the statement

`BIOGEME_OBJECT.SIMULATE.`

The group of columns contains the calculated indicator, as well as the 90% confidence interval for this indicator, if requested. It is calculated using simulation. As the estimates have been obtained from maximum likelihood, they are (asymptotically) normally distributed. Therefore, Biogeme draws instance of the parameters from a multivariate normal distribution $N(\hat{\beta}, \hat{\Sigma})$, where $\hat{\beta}$ is the vector of estimated parameters, and $\hat{\Sigma}$ is the variance-covariance matrix defined by the `BIOGEME_OBJECT.VARCOVAR` statement. The number of draws is controlled by the parameter:

`NbrOfDrawsForSensitivityAnalysis.`

The requested indicator is calculated for each realization, and the 5% and the 95% quantiles of the obtained simulated values are reported to generate the 90% confidence interval. Note that the confidence interval is reported only if the statement

`BIOGEME_OBJECT.VARCOVAR = vc`

is present. If you do not need the confidence intervals, simply remove this statement from the `.py` file.

- The *aggregate values* section reports the aggregate indicators. Denote by z_n the value of the indicator calculated for individual n (such as the probability that individual n chooses alternative 1, for instance), and by w_n the weight associated with individual n to correct for sampling biases. Then, the following aggregate values are reported, together with the associated confidence interval (if requested):

* Total:

$$\sum_{n=1}^{N_s} z_n. \quad (1)$$

* Weighted total:

$$\sum_{n=1}^{N_s} w_n z_n. \quad (2)$$

* Average:

$$\frac{1}{N_s} \sum_{n=1}^{N_s} z_n. \quad (3)$$

* Weighted average:

$$\frac{1}{N_s} \sum_{n=1}^{N_s} w_n z_n. \quad (4)$$

* Non zeros:

$$\sum_{n=1}^{N_s} \delta(z_n \neq 0), \quad (5)$$

where

$$\delta(z_n \neq 0) = \begin{cases} 1 & \text{if } z_n \neq 0, \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

* Non zeros average:

$$\frac{\sum_{n=1}^{N_s} z_n}{\sum_{n=1}^{N_s} \delta(z_n \neq 0)}. \quad (7)$$

* Weighted non zeros average:

$$\frac{\sum_{n=1}^{N_s} w_n z_n}{\sum_{n=1}^{N_s} \delta(z_n \neq 0)}. \quad (8)$$

* Minimum:

$$\min_n z_n. \quad (9)$$

* Maximum:

$$\max_n z_n. \quad (10)$$

Forecasting – 7.1 Aggregation

Michel Bierlaire

Practice quiz

Consider the logit model included in `v715_optima_model.py` (from the edX webpage) for the *Optima* case study (`optima.dat` (from the edX webpage)). The deterministic terms of the utility functions are defined as:

$$\begin{aligned}V_{PT} &= \beta_{\text{TIME_FULLTIME}} \cdot \text{TimePT_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimePT_scaled} \cdot \text{notfulltime} \\ &\quad \beta_{\text{COST}} \cdot \text{MarginalCostPT_scaled} \\ V_{CAR} &= ASC_{CAR} + \beta_{\text{TIME_FULLTIME}} \cdot \text{TimeCar_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimeCar_scaled} \cdot \text{notfulltime} + \\ &\quad \beta_{\text{COST}} \cdot \text{CostCarCHF_scaled} \\ V_{SM} &= ASC_{SM} + \beta_{\text{DIST_MALE}} \cdot \text{distance_km_scaled} \cdot \text{male} + \\ &\quad \beta_{\text{DIST_FEMALE}} \cdot \text{distance_km_scaled} \cdot \text{female} + \\ &\quad \beta_{\text{DIST_UNREPORTED}} \cdot \text{distance_km_scaled} \cdot \text{unreportedGender}\end{aligned}$$

where ASC_{CAR} , ASC_{SM} , $\beta_{\text{TIME_FULLTIME}}$, $\beta_{\text{TIME_OTHER}}$, $\beta_{\text{DIST_MALE}}$, $\beta_{\text{DIST_FEMALE}}$ and $\beta_{\text{DIST_UNREPORTED}}$ are parameters, `TimePT_scaled`, `MarginalCostPT_scaled`, `TimeCar_scaled`, `CostCarCHF_scaled` and `distance_km_scaled` are the scaled variables of the corresponding variables in the dataset, and `fulltime`, `notfulltime`, `male`, `female` and `unreportedGender` are socio-economic characteristics.

Calculate the predicted market shares for this model with stratified random sampling for the three alternatives (public transportation, car and soft modes) in `PythonBiogeme`.

Hints

- The sum of the weights has to be computed for normalization purposes. To do so, the following instruction has been included in `v715_optima_model.py`:

```
BIOGEME_OBJECT.STATISTICS['Sum of weights'] =
    Sum(Weight,'obsIter').
```

The result is reported in the `.html` file generated after running Python-Biogeme on `v715_optima_model.py` (from the edX webpage).

- Check the sample size on the generated `.html` file. Note that it is not equivalent to the number of entries in the dataset because we are excluding the observations that satisfy certain conditions.
- A new variable (`theWeight`) needs to be defined in order to normalize the existing weights (`Weight`). The normalized weights are calculated by multiplying the original weights by the sample size (`xxxx`) and dividing them by the total sum of weights (`yyyy`):

```
theWeight = Weight * xxxx/yyyy
BIOGEME_OBJECT.WEIGHT = theWeight
```

- The probability statements have to be defined in the simulation file for each alternative, and the log likelihood does not need to be calculated, that is, you can replace `l = bioLogLogit(V,av,Choice)` in the simulation file by statements of the form `prob_PT = bioLogit(V,av,0)` (this is for the public transportation alternative).

Forecasting – 7.1 Aggregation

Michel Bierlaire

Solution of the practice quiz

Note that the weights had to be normalized because their sum has to be equal to the sample size. The output of the model estimation is available in the file `v715_optima_model.html` (from the edX webpage), where it is reported that the sum of weights is 0.804451 and the sample size is 1899. Therefore, we include the following statements in the simulation file:

```
theWeight = Weight * 1899/0.804451
BIOGEME_OBJECT.WEIGHT = theWeight
```

The complete simulation file is available here: `v715_optima_simul.py` (from the edX webpage). It contains all the instructions to perform the simulation. This file is run as usual in PythonBiogeme and generates all the quantities that have been specified in the `simulate` variable. The output file is `v715_optima_simul.html` (from the edX webpage).

The market shares we are interested in can be found in the row “Weighted average”:

- Public transportation: 28.70%
- Car: 65.22%
- Soft modes: 6.082%

Forecasting – 7.2 Forecasting and confidence intervals

Michel Bierlaire

Practice quiz: forecasting

Consider the logit model specified in the file `v715_optima_model.py` (from the edX webpage) for the *Optima* case study (`optima.dat` (from the edX webpage)). The deterministic parts of the utility functions are defined as:

$$\begin{aligned} V_{PT} &= \beta_{\text{TIME_FULLTIME}} \cdot \text{TimePT_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimePT_scaled} \cdot \text{notfulltime} \\ &\quad \beta_{\text{COST}} \cdot \text{MarginalCostPT_scaled} \\ V_{CAR} &= ASC_{CAR} + \beta_{\text{TIME_FULLTIME}} \cdot \text{TimeCar_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimeCar_scaled} \cdot \text{notfulltime} + \\ &\quad \beta_{\text{COST}} \cdot \text{CostCarCHF_scaled} \\ V_{SM} &= ASC_{SM} + \beta_{\text{DIST_MALE}} \cdot \text{distance_km_scaled} \cdot \text{male} + \\ &\quad \beta_{\text{DIST_FEMALE}} \cdot \text{distance_km_scaled} \cdot \text{female} + \\ &\quad \beta_{\text{DIST_UNREPORTED}} \cdot \text{distance_km_scaled} \cdot \text{unreportedGender} \end{aligned}$$

where ASC_{CAR} , ASC_{SM} , $\beta_{\text{TIME_FULLTIME}}$, $\beta_{\text{TIME_OTHER}}$, $\beta_{\text{DIST_MALE}}$, $\beta_{\text{DIST_FEMALE}}$ and $\beta_{\text{DIST_UNREPORTED}}$ are parameters, TimePT_scaled , $\text{MarginalCostPT_scaled}$, TimeCar_scaled , CostCarCHF_scaled and $\text{distance_km_scaled}$ are the scaled variables of the corresponding variables in the dataset, and fulltime , notfulltime , male , female and unreportedGender are socio-economic characteristics.

Test the effect of the increase in the gas cost by forecasting the market shares of the different alternatives with stratified random sampling. To do so, complete the following table:

	Increase in the gas cost					
	5%	10%	15%	20%	25%	30%
Public transportation						
Car						
Soft modes						

Hints

- Consider the simulation file from the previous practice quiz `v715_optima_simul.py` (from the edX webpage).
- Define a new variable capturing the increase of the gas cost (before the utility statements). For instance, for a 5% increase, it is defined in Pythonbiogeme as

```
CostCarCHF_scaled_increased = DefineVariable
('CostCarCHF_scaled_increased', CostCarCHF_scaled*1.05),
```

where `CostCarCHF_scaled` is the scaled variable for the original gas cost variable.

- Replace the scaled gas cost variable by the new variable wherever it appears in the utility statements.
- Run the resulting file as usual. You should name the file associated with each scenario differently. For instance, for a 5% increase it can be named `v721_optima_increase_05.py`.

Forecasting – 7.2 Forecasting and confidence intervals

Michel Bierlaire

Solution of the practice quiz: forecasting

The specification files and the output files for each of the different scenarios are :

5% v721_optima_increase_05.py (from the edX webpage)
v721_optima_increase_05.html (from the edX webpage)
10% v721_optima_increase_10.py (from the edX webpage)
v721_optima_increase_10.html (from the edX webpage)
15% v721_optima_increase_15.py (from the edX webpage)
v721_optima_increase_15.html (from the edX webpage)
20% v721_optima_increase_20.py (from the edX webpage)
v721_optima_increase_20.html (from the edX webpage)
25% v721_optima_increase_25.py (from the edX webpage)
v721_optima_increase_25.html (from the edX webpage)
30% v721_optima_increase_30.py (from the edX webpage)
v721_optima_increase_30.html (from the edX webpage)

The market shares in each output file can be found in the table “Aggregate values”, in row “Weighted average”. They are reported in the following table.

	Increase in the gas cost					
	5%	10%	15%	20%	25%	30%
Public transportation	29.00%	29.30%	29.60%	29.90%	30.20%	30.50%
Car	64.91%	64.59%	64.28%	63.97%	63.65%	63.32%
Soft modes	6.096%	6.110%	6.125%	6.139%	6.153%	6.167%

Forecasting – 7.2 Forecasting and confidence intervals

Michel Bierlaire

Practice quiz: confidence intervals

Consider the logit model specified in `v715_optima_model.py` (from the edX webpage) for the *Optima* case study (data file: `optima.dat` (from the edX webpage)). The deterministic terms of the utility functions are defined as:

$$\begin{aligned} V_{PT} &= \beta_{\text{TIME_FULLTIME}} \cdot \text{TimePT_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimePT_scaled} \cdot \text{notfulltime} \\ &\quad \beta_{\text{COST}} \cdot \text{MarginalCostPT_scaled} \\ V_{CAR} &= ASC_{CAR} + \beta_{\text{TIME_FULLTIME}} \cdot \text{TimeCar_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimeCar_scaled} \cdot \text{notfulltime} + \\ &\quad \beta_{\text{COST}} \cdot \text{CostCarCHF_scaled} \\ V_{SM} &= ASC_{SM} + \beta_{\text{DIST_MALE}} \cdot \text{distance_km_scaled} \cdot \text{male} + \\ &\quad \beta_{\text{DIST_FEMALE}} \cdot \text{distance_km_scaled} \cdot \text{female} + \\ &\quad \beta_{\text{DIST_UNREPORTED}} \cdot \text{distance_km_scaled} \cdot \text{unreportedGender} \end{aligned}$$

where ASC_{CAR} , ASC_{SM} , $\beta_{\text{TIME_FULLTIME}}$, $\beta_{\text{TIME_OTHER}}$, $\beta_{\text{DIST_MALE}}$, $\beta_{\text{DIST_FEMALE}}$ and $\beta_{\text{DIST_UNREPORTED}}$ are parameters, TimePT_scaled , $\text{MarginalCostPT_scaled}$, TimeCar_scaled , CostCarCHF_scaled and $\text{distance_km_scaled}$ are the scaled variables of the corresponding variables in the dataset, and fulltime , notfulltime , male , female and unreportedGender are socio-economic characteristics.

Obtain the 90% confidence intervals of the predicted market shares (with stratified random sampling) for the scenarios of increase of the gas cost considered in the previous practice quiz, i.e., 5%, 10%, 15%, 20%, 25% and 30%. For each alternative and each scenario, provide the 90% confidence interval.

Hints

- Consider the files `v721_optima_increase_XX.py` (where `XX` is the increase in the gas cost) from the previous quiz.
- The following statements of the code for the sensitivity analysis are necessary here:

```
vc = bioMatrix(8,names,values)
BIOGEME_OBJECT.VARCOVAR = vc
```

Forecasting – 7.2 Forecasting and confidence intervals

Michel Bierlaire

Solution of the practice quiz: confidence intervals

We need to uncomment the statements that were commented out in the previous quizzes:

```
vc = bioMatrix(27,names,values)
BIOGEME_OBJECT.VARCOVAR = vc
```

For each file from the previous quiz we have included these statements in order to report the 5% and the 95% quantile of the obtained simulated values, which generate the 90% confidence interval. The .py files have been called as follows:

- 5%: v721_optima_increase_05_ci.py (from the edX webpage)
- 10%: v721_optima_increase_10_ci.py (from the edX webpage)
- 15%: v721_optima_increase_15_ci.py (from the edX webpage)
- 20%: v721_optima_increase_20_ci.py (from the edX webpage)
- 25%: v721_optima_increase_25_ci.py (from the edX webpage)
- 30%: v721_optima_increase_30_ci.py (from the edX webpage)

The confidence intervals for each alternative are the following:

Public transportation:

- 5%: [25.32%, 32.64%]

- 10%: [25.59%, 32.95%]
- 15%: [25.86%, 33.26%]
- 20%: [26.12%, 33.57%]
- 25%: [26.39%, 33.89%]
- 30%: [26.65%, 34.21%]

Car:

- 5%: [60.24%, 68.69%]
- 10%: [59.91%, 68.41%]
- 15%: [59.59%, 68.12%]
- 20%: [59.26%, 67.84%]
- 25%: [58.93%, 67.56%]
- 30%: [58.60%, 67.27%]

Soft modes:

- 5%: [4.336%, 9.442%]
- 10%: [4.344%, 9.469%]
- 15%: [4.353%, 9.496%]
- 20%: [4.361%, 9.523%]
- 25%: [4.367%, 9.551%]
- 30%: [4.378%, 9.578%]

Forecasting – 7.3 Indicators

Michel Bierlaire

Willingness to pay and value of time

If the model contains a cost or price variable, it is possible to analyze the trade-off between any variable and money. It reflects the *willingness* of the decision maker *to pay* for a modification of another variable of the model.

A typical example in transportation is the *value of time*, that is the price that a traveler is willing to pay to decrease the travel time.

- Let c_{in} be the cost of alternative i for individual n .
- Let x_{in} be the value of another variable of the model.
- Let $V_{in}(c_{in}, x_{in})$ be the value of the utility function associated by individual n with alternative i .
- Consider a scenario where the variable under interest takes the value $x_{in} + \delta_{in}^x$.
- We denote by δ_{in}^c the additional cost that would achieve the same utility, that is

$$V_{in}(c_{in} + \delta_{in}^c, x_{in} + \delta_{in}^x) = V_{in}(c_{in}, x_{in}). \quad (1)$$

The willingness to pay is the additional cost per unit of x , that is

$$\delta_{in}^c / \delta_{in}^x. \quad (2)$$

Its calculation involves solving equation (1).

If the variable x_{in} is continuous, and if V_{in} is differentiable in x_{in} and c_{in} , we can invoke Taylor's theorem

$$\begin{aligned} V_{in}(c_{in}, x_{in}) &= V_{in}(c_{in} + \delta_{in}^c, x_{in} + \delta_{in}^x) \\ &\approx V_{in}(c_{in}, x_{in}) + \delta_{in}^c \frac{\partial V_{in}}{\partial c_{in}}(c_{in}, x_{in}) + \delta_{in}^x \frac{\partial V_{in}}{\partial x_{in}}(c_{in}, x_{in}) \end{aligned}$$

to obtain

$$\frac{\delta_{in}^c}{\delta_{in}^x} = -\frac{(\partial V_{in}/\partial x_{in})(c_{in}, x_{in})}{(\partial V_{in}/\partial c_{in})(c_{in}, x_{in})}. \quad (3)$$

If x_{in} and c_{in} appear linearly in the utility function, that is if

$$V_{in}(c_{in}, x_{in}) = \beta_c c_{in} + \beta_x x_{in} + \dots \quad (4)$$

then the willingness to pay involves the ratio of the coefficients:

$$\frac{\delta_{in}^c}{\delta_{in}^x} = -\frac{(\partial V_{in}/\partial x_{in})(c_{in}, x_{in})}{(\partial V_{in}/\partial c_{in})(c_{in}, x_{in})} = -\frac{\beta_x}{\beta_c}. \quad (5)$$

The above equation is the willingness to pay for an *increase* of the value of the variable x_{in} . If this increase improves the utility of the alternative, then $\beta_x > 0$. As $\beta_c < 0$, the willingness to pay is positive. In the context of value of time, we want to calculate the willingness to pay to *decrease* the travel time. Therefore, we have

$$\text{VOT}_{in} = \delta_{in}^c / (-\delta_{in}^t) = \frac{(\partial V_{in}/\partial t_{in})(c_{in}, t_{in})}{(\partial V_{in}/\partial c_{in})(c_{in}, t_{in})}. \quad (6)$$

If V is linear in these variables, we have

$$\text{VOT}_{in} = \delta_{in}^c / (-\delta_{in}^t) = \frac{\beta_t}{\beta_c}. \quad (7)$$

The willingness to pay is negative when a decrease of the cost compensates a modification of another variable that decreases the utility. In this case, it is sometimes called *willingness to accept*.

The above derivation, based on converting trade-offs into monetary units, is the most common one. Similar quantities can be derived by using other variables than cost as the reference.

Forecasting – 7.3 Indicators

Michel Bierlaire

Willingness to pay and value of time: practice quiz

Consider the following utility function corresponding to the public transportation alternative i in a mode choice context:

$$V_{in} = -0.0704 \text{ TC}_{in} - 0.0117 \text{ TT}_{in} - 0.0594 \text{ Transfers}_{in} \quad (1)$$

where

- TC_{in} is the travel cost for individual n ,
- TT_{in} is the travel time for individual n ,
- Transfers_{in} is the number of transfers for individual n .

Consider an individual n confronted with the following values:

- $\text{TC}_{in} = 12.4$ CHF,
- $\text{TT}_{in} = 85$ minutes,
- $\text{Transfers}_{in} = 2$.

What is the value of time for this individual?

1. Correct: $-0.0117 / -0.0704 = 0.166$ CHF/min
2. 6.02 CHF/min
3. 14.6 CHF/min
4. 6.85 CHF/min

Forecasting – 7.3 Indicators

Michel Bierlaire

Practice quiz

Consider the logit model included in `v715_optima_model.py` (from the edX webpage) for the *Optima* case study (dataset: `optima.dat` (from the edX webpage)). The deterministic terms of the utility functions are defined as:

$$\begin{aligned}V_{PT} &= \beta_{\text{TIME_FULLTIME}} \cdot \text{TimePT_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimePT_scaled} \cdot \text{notfulltime} \\ &\quad \beta_{\text{COST}} \cdot \text{MarginalCostPT_scaled} \\ V_{CAR} &= ASC_{CAR} + \beta_{\text{TIME_FULLTIME}} \cdot \text{TimeCar_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimeCar_scaled} \cdot \text{notfulltime} + \\ &\quad \beta_{\text{COST}} \cdot \text{CostCarCHF_scaled} \\ V_{SM} &= ASC_{SM} + \beta_{\text{DIST_MALE}} \cdot \text{distance_km_scaled} \cdot \text{male} + \\ &\quad \beta_{\text{DIST_FEMALE}} \cdot \text{distance_km_scaled} \cdot \text{female} + \\ &\quad \beta_{\text{DIST_UNREPORTED}} \cdot \text{distance_km_scaled} \cdot \text{unreportedGender}\end{aligned}$$

where ASC_{CAR} , ASC_{SM} , $\beta_{\text{TIME_FULLTIME}}$, $\beta_{\text{TIME_OTHER}}$, $\beta_{\text{DIST_MALE}}$, $\beta_{\text{DIST_FEMALE}}$ and $\beta_{\text{DIST_UNREPORTED}}$ are parameters, TimePT_scaled , $\text{MarginalCostPT_scaled}$, TimeCar_scaled , CostCarCHF_scaled and $\text{distance_km_scaled}$ are the scaled variables of the corresponding variables in the dataset, and fulltime , notfulltime , male , female and unreportedGender are socio-economic characteristics.

Perform the following tasks:

1. Calculate for each individual in the sample the value of time for public transportation and for car in CHF/hour;
2. Provide an estimate of the average value of time in the population;
3. Analyze the distribution of the value of time in the sample.

Hints

- The value of time is calculated as

$$\text{VOT}_{in} = \frac{(\partial V_{in} / \partial t_{in})(c_{in}, t_{in})}{(\partial V_{in} / \partial c_{in})(c_{in}, t_{in})}, \quad (1)$$

where c_{in} and t_{in} are the cost and travel time of alternative i and individual n , respectively. In PythonBiogeme, the calculation of derivatives is written as follows (example for public transportation):

```
VOT_PT = Derive(V_PT, 'TimePT')/Derive(V_PT, 'MarginalCostPT')
```

We can add these statements to the simulation file `v715_optima_simul.py` (from the edX webpage).

- The `DefineVariable` operator is designed to preprocess the data file, and can be seen as a way to add another column in the data file, defining a new variable. However, when used, the functional relationship between the new variable and the original one is lost. Therefore, PythonBiogeme is not able to properly calculate the derivatives. For instance, one of the variable of interest is `TimePT`, not `TimePT_scaled`, and their relationship must be explicitly known to correctly calculate the derivatives. Consequently, all statements such as
`TimePT_scaled = DefineVariable('TimePT_scaled', TimePT/200)`
should be replaced by statements such as
`TimePT_scaled = TimePT/200`
in order to maintain the analytical structure of the formula to be derived.
- In order to analyze the distribution of the value of time in the sample, identify the socioeconomic characteristic(s) that play(s) a role in the calculation of the value of time and report its value together with the value of time.

Forecasting – 7.3 Indicators

Michel Bierlaire

Solution of the practice quiz

1. The file `v735_optima_VOT01.py` (from the edX webpage) has been defined from the simulation file `v715_optima_simul.py` (from the edX webpage) by adding the following statements:
`VOT_PT = Derive(V_PT, 'TimePT')/Derive(V_PT, 'MarginalCostPT')`
`VOT_CAR = Derive(V_CAR, 'TimeCar')/Derive(V_CAR, 'CostCarCHF')`

In order to get the value of time in CHF/hour, we can easily convert the original values by multiplying them by 60. We can include this formula directly in the `simulate` variable:

```
simulate = {'01 PT: Value of time (CHF/h)': 60*VOT_PT, '02  
Car: Value of time (CHF/h)': 60*VOT_CAR}
```

The results can be found in `v735_optima_VOT01.html` (from the edX webpage).

2. The estimate of the average value of time in the population is obtained from the weighted average of the sample. The aggregate values are found in the “Weighted average” row of the report file. The estimate of the value of time is equal to

3.59CHF/hour,

with a confidence interval of

[1.66,6.66].

Note that this value is abnormally low, which is a sign of a potential poor specification of the model. Note also that, with this specification, the value of time is the same for car and public transportation, as the coefficients of the time and cost variables are generic.

3. It is important to look at the distribution of the willingness to pay in the population/sample. The detailed records of the report file allow to do so. It is easy to drag and drop the `.html` file into your favorite spreadsheet software in order to perform additional statistics. In this example, the value of time takes two values, depending on the employment status of the individual.

The file `v735_optima_VOT01.py` (from the edX webpage) has been adapted to account for the employment status. The resulting file, `v735_optima_VOT02.py` (from the edX webpage), contains an additional instruction in the `simulate` variable to report if the employment status is full time or not:

```
simulate = {'01 PT: Value of time (CHF/h)': 60*VOT_PT, '02  
Car: Value of time (CHF/h)': 60*VOT_CAR, '03 Full time':  
fulltime}
```

The output is available in `v735_optima_VOT02.html` (from the edX webpage). One can easily identify the value of time for full time employees (`fulltime = 1`) and for not full time employees (`fulltime = 0`) by looking at the corresponding rows on the resulting `.html` file:

- Full time: 6.37 CHF/hour (confidence interval: [4.02, 10.08])
- Not full time: 2.00 CHF/hour (confidence interval: [0.32, 4.70]).

Even though the values are still low, the obtained results are as expected in the sense that the value of time for full time employees is higher, that is, full-time employees are willing to pay more to save 1 hour of transportation than part-time employees.

Forecasting – 7.3 Indicators

Michel Bierlaire

Practice quiz: elasticities

Consider the logit model included in `v715_optima_model.py` (from the edX webpage) for the *Optima* case study (dataset: `optima.dat` (from the edX webpage)). The deterministic terms of the utility functions are defined as:

$$\begin{aligned} V_{PT} &= \beta_{\text{TIME_FULLTIME}} \cdot \text{TimePT_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimePT_scaled} \cdot \text{notfulltime} \\ &\quad \beta_{\text{COST}} \cdot \text{MarginalCostPT_scaled} \\ V_{CAR} &= ASC_{CAR} + \beta_{\text{TIME_FULLTIME}} \cdot \text{TimeCar_scaled} \cdot \text{fulltime} + \\ &\quad \beta_{\text{TIME_OTHER}} \cdot \text{TimeCar_scaled} \cdot \text{notfulltime} + \\ &\quad \beta_{\text{COST}} \cdot \text{CostCarCHF_scaled} \\ V_{SM} &= ASC_{SM} + \beta_{\text{DIST_MALE}} \cdot \text{distance_km_scaled} \cdot \text{male} + \\ &\quad \beta_{\text{DIST_FEMALE}} \cdot \text{distance_km_scaled} \cdot \text{female} + \\ &\quad \beta_{\text{DIST_UNREPORTED}} \cdot \text{distance_km_scaled} \cdot \text{unreportedGender} \end{aligned}$$

where ASC_{CAR} , ASC_{SM} , $\beta_{\text{TIME_FULLTIME}}$, $\beta_{\text{TIME_OTHER}}$, $\beta_{\text{DIST_MALE}}$, $\beta_{\text{DIST_FEMALE}}$ and $\beta_{\text{DIST_UNREPORTED}}$ are parameters, TimePT_scaled , $\text{MarginalCostPT_scaled}$, TimeCar_scaled , CostCarCHF_scaled and $\text{distance_km_scaled}$ are the scaled variables of the corresponding variables in the dataset, and fulltime , notfulltime , male , female and unreportedGender are socio-economic characteristics.

Calculate the following indicators:

1. Estimate of the aggregate direct point elasticities for the population:
 - (a) Elasticity of the share of public transportation with respect to travel time by public transportation.

- (b) Elasticity of the share of public transportation with respect to marginal cost of public transportation.
 - (c) Elasticity of the share of car with respect to travel time of car.
 - (d) Elasticity of the share of car with respect to cost of car.
 - (e) Elasticity of the share of slow modes with respect to the trip length.
2. Estimate of the aggregate cross point elasticities for the population:
- (a) Elasticity of the share of public transportation with respect to the travel time by car.
 - (b) Elasticity of the share of public transportation with respect to the cost of car.
 - (c) Elasticity of the share of car with respect to the travel time by public transportation.
 - (d) Elasticity of the share of car with respect to the marginal cost of public transportation.

Hints

- Since the aggregate point elasticities are obtained by aggregating the disaggregate elasticities, the normalization factors need to be calculated. To do so, we include the following statements to the simulation file `v715_optima_simul.py` and run the resulting file:


```
BIOGEME.OBJECT.STATISTICS['Normalization for elasticities PT']
= Sum(theWeight * prob_PPT , 'obsIter')
BIOGEME.OBJECT.STATISTICS['Normalization for elasticities CAR']
= Sum(theWeight * prob_CAR , 'obsIter')
BIOGEME.OBJECT.STATISTICS['Normalization for elasticities SM']
= Sum(theWeight * prob_SM , 'obsIter')
```
- A simulation file with the statements for the aggregate elasticities of interest has to be created. For instance, if we want to calculate the aggregate elasticity of the choice of public transport with respect to its travel time, the following instructions need to be included:


```
normalization_pt = ...
elas_pt_time = Derive(prob_PPT, 'TimePT') * TimePT / prob_PPT
```

`'Agg. Elast. PT - Time': elas_pt_time * prob_PT / normalization_pt`

where the first statement corresponds to the normalization factor, the second statement calculates the disaggregate elasticity of the choice of public transportation with respect to its marginal cost and the third statement is the entry to the simulation dictionary that is designed to calculate the aggregate elasticities.

- Note that the weights have not been included in the formula for the aggregate elasticity, so that the values of the aggregate elasticities can be found in the row “Weighted total” of the generated `.html` file.

Forecasting – 7.3 Indicators

Michel Bierlaire

Solution of the practice quiz: elasticities

The file `v736_optima_elasticities01.py` (from the edX webpage) has been defined from the `v715_optima_simul.py` (from the edX webpage) by adding the following statements in order to obtain the normalization factors:

```
BIOGEME_OBJECT.STATISTICS['Normalization for elasticities PT'] =  
Sum(theWeight * prob_PT , 'obsIter')  
BIOGEME_OBJECT.STATISTICS['Normalization for elasticities CAR'] =  
Sum(theWeight * prob_CAR , 'obsIter')  
BIOGEME_OBJECT.STATISTICS['Normalization for elasticities SM'] =  
Sum(theWeight * prob_SM , 'obsIter')
```

The values are found in the output file `v736_optima_elasticities01.html` (from the edX webpage):

- Public transportation: 545.061,
 - Car: 1238.44,
 - Slow modes: 115.504.
1. We then create a new file `v736_optima_elasticities02.py` (from the edX webpage) that contains the normalization factors, the disaggregate elasticities that need to be aggregated and the resulting aggregate elasticities. The values are found in the output file `v736_optima_elasticities02.html` (from the edX webpage):
 - (a) Agg. Elast. PT - Time PT = -0.2412,
 - (b) Agg. Elast. PT - Cost PT = -0.3131,
 - (c) Agg. Elast. Car - Time Car = -0.4147,
 - (d) Agg. Elast. Car - Time Car = -0.0948,

(e) Agg. Elast. SM - Distance SM = -1.017.

2. We replace the previous direct elasticities' instructions by the corresponding cross ones in the file `v736_optima_elasticities03.py` (from the edX webpage) and proceed in the same way as before. The values are found in the output file `v736_optima_elasticities03.html` (from the edX webpage):

- (a) Agg. Elast. PT - Time car: 0.0885,
- (b) Agg. Elast. PT - Cost car: 0.2056,
- (c) Agg. Elast. Car - Time PT: 0.1018,
- (d) Agg. Elast. Car - Cost PT: 0.1314.

Forecasting – 7.3 Indicators

Michel Bierlaire

Consumer surplus

The consumer surplus is the difference between what a consumer is willing to pay for a good and what she actually pays for the good. If the reader is not familiar with the concept of consumer surplus, we suggest to read a primer on the topic in a text book on microeconomics, such as Nicholson and Snyder (2007). The change in consumer surplus is often used to evaluate public policy decisions. For example, the impact on consumers of changing emissions regulations or increasing investments in the public transit system.

It is equal to the area under the demand curve and above the market price. In classical microeconomics, the demand curve gives the price of a good as a function of the quantity consumed. In discrete choice, the demand for individual n is characterized by the choice probability. Also, the role of price is taken by the utility of the good. In Figure 1 the choice probability is represented on the x -axis, and the y -axis represents the negative utility of alternative i , $-V_i$. The minus sign helps in obtaining the same interpretation as in classical microeconomic: going up the axis corresponds to a deterioration.

Simulating a future increase of the utility of item i (that is, a decrease of the quantity $-V_i$), while the utility of other alternatives is constant, the change in consumer surplus is represented by the filled area. For binary logit, this area can be calculated by the following integral, where the index n has been dropped to simplify the notations:

$$\int_{V_i^1}^{V_i^2} P(i|V_i, V_j) dV_i = \int_{V_i^1}^{V_i^2} \frac{e^{\mu V_i}}{e^{\mu V_i} + e^{\mu V_j}} dV_i \quad (1)$$

which is

$$\frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j}) - \frac{1}{\mu} \ln(e^{\mu V_i^1} + e^{\mu V_j}). \quad (2)$$

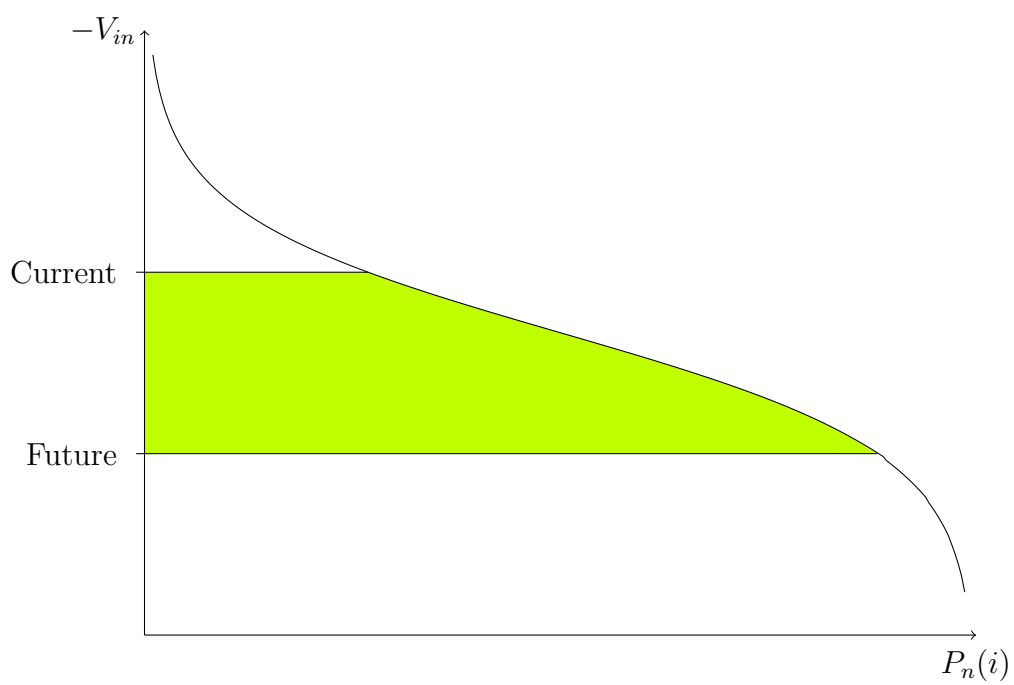


Figure 1: Illustration of the consumer surplus for binary logit

To generalize this result, we calculate the difference in an individual's consumer surplus between two situations corresponding to vectors of systematic utilities V^1 and V^2 as follows:

$$\sum_{i \in \mathcal{C}} \int_{V^1}^{V^2} P(i|V) dV_i, \quad (3)$$

where the choice probability is denoted as conditional on the vector V of systematic utilities in order to make the dependency explicit. Note that the term $P(i|V)dV_i$ corresponds to the filled area in Figure 1 where the change in the utility of i is infinitesimal. The difficulty here is that the utility of **all** alternatives are modified. Therefore, the integral in (3) is a line integral. And there are infinitely many ways to move from utility vector V^1 to utility vector V^2 in the J -dimensional space. This is illustrated for an example with two alternatives in Figures 2 and 3. In order to simplify the calculation of the integral, we consider paths that are updating each coordinate at a time. In Figure 2, the path moves first from $V^1 = (V_i^1, V_j^1)$ to (V_i^2, V_j^1) , and then from (V_i^2, V_j^1) to $(V_i^2, V_j^2) = V^2$. The path in Figure 3 moves first from $V^1 = (V_i^1, V_j^1)$ to (V_i^1, V_j^2) , and then from (V_i^1, V_j^2) to $(V_i^2, V_j^2) = V^2$.

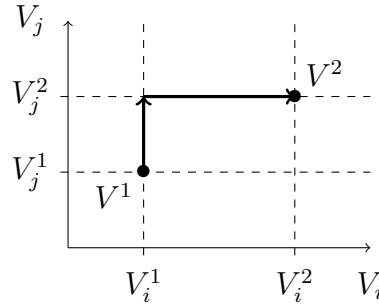


Figure 2: Moving from utility vector V_1 to utility vector V_2 : first path

For the example with two alternatives, where the path in Figure 2 is followed, the integral (3) is

$$\int_{V_i^1}^{V_i^2} P(i|V_i, V_j^1) dV_i + \int_{V_j^1}^{V_j^2} P(j|V_i^2, V_j) dV_j. \quad (4)$$

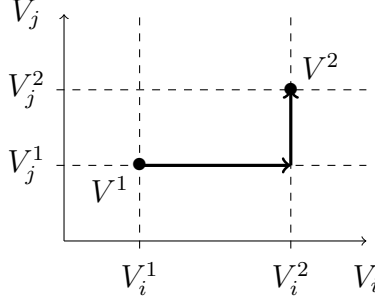


Figure 3: Moving from utility vector V_1 to utility vector V_2 : second path

Assuming now a binary logit model, the first integral is

$$\int_{V_i^1}^{V_i^2} \frac{e^{\mu V_i}}{e^{\mu V_i} + e^{\mu V_j^1}} dV_i. \quad (5)$$

Let $t = e^{\mu V_i} + e^{\mu V_j^1}$ so that $dt = \mu e^{\mu V_i} dV_i$, we obtain

$$\frac{1}{\mu} \int_{e^{\mu V_i^1} + e^{\mu V_j^1}}^{e^{\mu V_i^2} + e^{\mu V_j^1}} \frac{dt}{t} = \frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j^1}) - \frac{1}{\mu} \ln(e^{\mu V_i^1} + e^{\mu V_j^1}). \quad (6)$$

The second integral is

$$\int_{V_j^1}^{V_j^2} \frac{e^{\mu V_j}}{e^{\mu V_i^2} + e^{\mu V_j}} dV_j. \quad (7)$$

Let $t = e^{\mu V_i^2} + e^{\mu V_j}$ so that $dt = \mu e^{\mu V_j} dV_j$, we obtain

$$\frac{1}{\mu} \int_{e^{\mu V_i^2} + e^{\mu V_j^1}}^{e^{\mu V_i^2} + e^{\mu V_j^2}} \frac{dt}{t} = \frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j^2}) - \ln(e^{\mu V_i^2} + e^{\mu V_j^1}). \quad (8)$$

Adding (6) and (8), we obtain the difference of logsum, that is

$$\frac{1}{\mu} \ln(e^{\mu V_i^2} + e^{\mu V_j^2}) - \frac{1}{\mu} \ln(e^{\mu V_i^1} + e^{\mu V_j^1}). \quad (9)$$

Calculating the integral following the path described in Figure 3 leads to the exact same result. We say that the calculation of the integral is *path*

independent. Not all line integrals are path independent. If the choice model happens to have equal cross-derivatives, that is

$$\frac{\partial P(i|V, \mathcal{C})}{\partial V_j} = \frac{\partial P(j|V, \mathcal{C})}{\partial V_i}, \quad \forall i, j \in \mathcal{C}, \quad (10)$$

the calculation of the integral in (3) is path independent.

The logit model, as well as several choice models used in practice, have this property. Therefore, for logit, we can select the path of integration. We calculate (3) using a path that first updates the utility of alternative 1, then of alternative 2, and so on until J . The k th term integrates over V_k , with all utilities of alternatives 1 to $k-1$ set at the level V^2 , while all utilities of alternatives $k+1$ to J are set at the level V^1 . This term writes

$$\int_{V_k^1}^{V_k^2} \frac{e^{\mu V_k}}{\sum_{j=1}^{k-1} e^{\mu V_j^2} + e^{\mu V_k} + \sum_{j=k+1}^J e^{\mu V_j^1}} dV_k = \frac{1}{\mu} \ln \left(\sum_{j=1}^{k-1} e^{\mu V_j^2} + e^{\mu V_k^2} + \sum_{j=k+1}^J e^{\mu V_j^1} \right) - \frac{1}{\mu} \ln \left(\sum_{j=1}^{k-1} e^{\mu V_j^2} + e^{\mu V_k^1} + \sum_{j=k+1}^J e^{\mu V_j^1} \right). \quad (11)$$

When summing up over k , most terms of the sum over alternatives cancel out, and the difference of consumer surplus for the logit model is

$$\frac{1}{\mu} \ln \sum_{j \in \mathcal{C}} e^{\mu V_j^2} - \frac{1}{\mu} \ln \sum_{j \in \mathcal{C}} e^{\mu V_j^1}. \quad (12)$$

which is the difference among expected maximum utilities in the two situations. When the choice set changes from \mathcal{C}^1 to \mathcal{C}^2 , the result of (3) is

$$\frac{1}{\mu} \ln \sum_{j \in \mathcal{C}^2} e^{\mu V_j^2} - \frac{1}{\mu} \ln \sum_{j \in \mathcal{C}^1} e^{\mu V_j^1}. \quad (13)$$

We refer the reader to Neuburger (1971), Small and Rosen (1981), Haneemann (1984), McConnell (1995), Dagsvik and Karlström (2005) for more detailed discussions about consumer surplus.

References

- Dagsvik, J. K. and Karlström, A. (2005). Compensating variation and hick-sian choice probabilities in random utility models that are nonlinear in income, *Review of Economic Studies* **72**(1): 57–76.
- Hanemann, W. (1984). Welfare evaluations in contingent valuation experiments with discrete responses, *American journal of agricultural economics* **66**(3): 332–341.
- McConnell, K. E. (1995). Consumer surplus from discrete choice models, *Journal of Environmental Economics and Management* **29**(3): 263 – 270.
URL: <http://www.sciencedirect.com/science/article/B6WJ6-45S92WM-N/2/a252034061e51b518591926a6914d672>
- Neuburger, H. (1971). User benefit in the evaluation of transport and land use plans, *Journal of Transport Economics and Policy* **5**(1): 52–75.
- Nicholson, W. and Snyder, C. M. (2007). *Microeconomic Theory: Basic Principles and Extensions*, South Western/Thomson.
- Small, K. A. and Rosen, H. S. (1981). Applied welfare economics with discrete choice models, *Econometrica* **49**(1): 105–30.
URL: <http://ideas.repec.org/a/ecm/emetrp/v49y1981i1p105-30.html>

Forecasting – 7.3 Indicators

Michel Bierlaire

Practice quiz: consumer surplus

Consider the following logit model for the *Optima* case study. The deterministic terms of the utility functions are defined as:

$$\begin{aligned}V_{PT} &= \beta_{\text{TIME}} \cdot \text{TimePT} + \beta_{\text{COST}} \cdot \text{MarginalCostPT} \\V_{CAR} &= ASC_{CAR} + \beta_{\text{TIME}} \cdot \text{TimeCar} + \beta_{\text{COST}} \cdot \text{CostCarCHF} \\V_{SM} &= ASC_{SM} + \beta_{\text{DIST}} \cdot \text{distance_km}\end{aligned}$$

where ASC_{CAR} , ASC_{SM} , β_{TIME} , β_{COST} and β_{DIST} are parameters; and TimePT, MarginalCostPT, TimeCar, CostCarCHF and distance.km are variables in the dataset. The parameter estimates are included in Table 1.

Parameter number	Description	Coeff. estimate	Robust Asympt. std. error	<i>t</i> -stat	<i>p</i> -value
1	ASC_{CAR}	0.301	0.102	2.96	0.00
2	ASC_{SM}	-0.0337	0.296	-0.11	0.91
3	β_{COST}	-0.0753	0.0139	-5.43	0.00
4	β_{DIST}	-0.198	0.0491	-4.03	0.00
5	β_{TIME}	-0.00478	0.00144	-3.31	0.00

Table 1: Estimates of the parameters

The public authorities are planning to do some investments in a train line in order to make public transportation more attractive. The population concerned by the train line is facing the following alternatives:

- PT: MarginalCostPT = 3.5 CHF, TimePT = 25 min.,

- Car: $\text{CostCarCHF} = 7.5 \text{ CHF}$, $\text{TimeCar} = 10 \text{ min.}$, and
- SM: $\text{distance_km} = 15 \text{ km}$.

Three different scenarios are considered by the authorities to improve the attractiveness of the train line:

1. decreasing its cost by 5%,
2. decreasing its travel time by 5 minutes,
3. decreasing its travel time by 10 minutes, while increasing its cost by 10%.

In order to decide the best scenario for the sake of the travelers, the increase in consumer surplus must be calculated for each scenario.

1. Calculate the consumer surplus of each traveler for the three scenarios in utility units.
2. What scenario increases the most the welfare of the travelers?
3. Calculate the consumer surplus of each traveler for the three scenarios in CHF.

Hint: Use a spreadsheet or statistical software to perform the calculations.

Forecasting – 7.3 Indicators

Michel Bierlaire

Solution of the practice quiz: consumer surplus

In order to calculate the consumer surplus we need to compute the deterministic terms of the utility functions for the present and the future scenarios before and after the decreased attribute(s) for each scenario. The results are included in the following table:

Alt.	Scenario 1		Scenario 2		Scenario 3	
	V^{before}	V^{after}	V^{before}	V^{after}	V^{before}	V^{after}
PT	-0.3831	-0.3699	-0.3831	-0.3592	-0.3831	-0.3616
Car	-0.3116	-0.3116	-0.3116	-0.3116	-0.3116	-0.3116
SM	-3.004	-3.004	-3.004	-3.004	-3.004	-3.004

Note that the deterministic terms of the utility for car and SM remain unchanged.

The consumer surplus of each scenario for the population of interest is then computed as follows (note that μ has been normalized to 1):

$$\ln(e^{V_{\text{PT}}^{\text{after}}} + e^{V_{\text{Car}}^{\text{after}}} + e^{V_{\text{SM}}^{\text{after}}}) - \ln(e^{V_{\text{PT}}^{\text{before}}} + e^{V_{\text{Car}}^{\text{before}}} + e^{V_{\text{SM}}^{\text{before}}}).$$

1. The difference of consumer surplus in terms of utility units:
 - (a) Cost decrease scenario: $0.3871 - 0.3810 = 0.006160$ utility units.
 - (b) Time decrease scenario: $0.3922 - 0.3810 = 0.01120$ utility units.
 - (c) Time decrease and cost increase scenario: $0.3910 - 0.3810 = 0.01005$ utility units.
2. The scenario increasing the most the consumer surplus is the time decrease scenario.

3. The difference of consumer surplus in terms of monetary units is obtained by dividing the above quantities by the cost coefficient:

- (a) Cost decrease scenario:

$$0.006160 / -0.0753 = -0.082.$$

Therefore, the consumer surplus of each traveler increases by 8.2 cents. It is interesting to compare this value with the amount of the cost decrease, that is 17.5 cents.

- (b) Time decrease scenario:

$$0.01120 / -0.0753 = -0.149.$$

Therefore, the consumer surplus of each traveler increases by 14.9 cents.

- (c) Time decrease and cost increase scenario:

$$0.01005 / -0.0753 = -0.133.$$

Therefore, the consumer surplus of each traveler increases by 13.3 cents.

Forecasting

Revenue maximization

Michel Bierlaire

Introduction to choice models



Revenue maximization

Revenue

Supplier i

- ▶ Consider the supplier of alternative i in the market.
- ▶ The price offered to individual n is p_{in} .
- ▶ The expected revenue generated by individual n is

$$p_{in}P(i|x_n, p_{in}; \theta)$$

- ▶ The total expected revenue is therefore

$$\sum_{n=1}^N p_{in}P(i|x_n, p_{in}; \theta)$$

Revenue maximization

Solve the problem

$$\max_{p_{i1}, \dots, p_{iN}} \sum_{n=1}^N p_{in} P(i | x_n, p_{in}; \theta)$$

Notes

- ▶ In practice, prices are often the same for the population, or for large groups.
- ▶ It assumes that the rest of the market is not affected.
- ▶ In practice, it is likely that the competition will also adjust the prices

Illustrative example

Binary logit model

$$\begin{aligned}V_{in} &= \beta_{pn}p_{in} - 0.5 \\V_{jn} &= \beta_{pn}p_{jn}\end{aligned}$$

so that

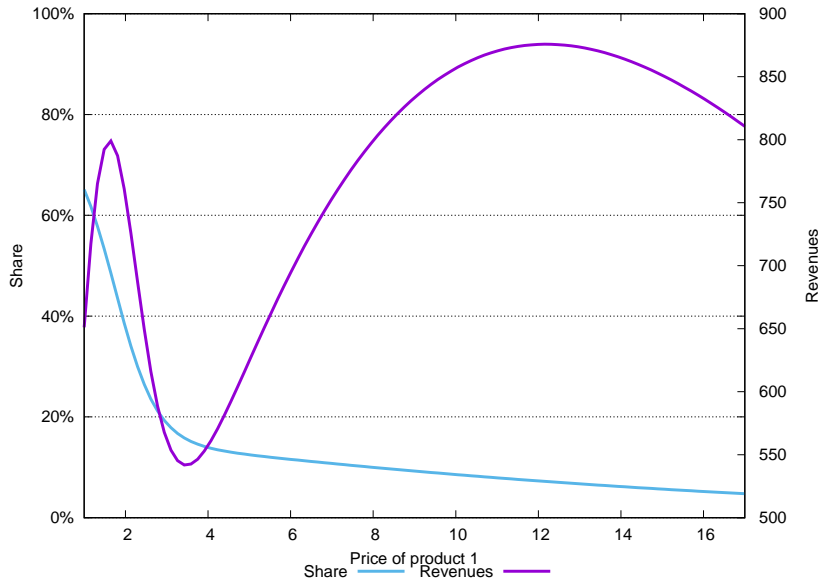
$$P_n(i|p_{in}, p_{jn}) = \frac{e^{\beta_{pn}p_{in}-0.5}}{e^{\beta_{pn}p_{in}-0.5} + e^{\beta_{pn}p_{jn}}}$$

Two groups in the population

- ▶ Group 1: $\beta_{pn} = -2$, $N_1 = 600$
- ▶ Group 2: $\beta_{pn} = -0.1$, $N_2 = 400$

Assume that $p_{jn} = 2$, $\forall n$.

Illustrative example



Forecasting – 7.4 Revenue maximization

Michel Bierlaire

Practice quiz: revenue maximization

Consider the binary logit model described in the previous video, where the utility of the two competing alternatives i and j are defined as

$$\begin{aligned}V_{in} &= \beta_{pn}p_i - 0.5, \\V_{jn} &= \beta_{pn}p_j,\end{aligned}$$

where p_i and p_j are the prices of each alternative, and β_{pn} is the price coefficient for individual n . Assume that you are in control of alternative i , and have to decide its price. We also assume that the price for the competing alternative j is fixed: $p_j = 2$. Note that the absence of an index n on the price variable indicates that the prices are the same for each individual.

The population, composed of 1000 individuals, is heterogeneous in its preferences, and composed of three groups. For each group, the value of the price coefficient β_{pn} and the number of individuals have been determined to be:

1. $\beta_{pn} = -1$, $N_1 = 300$,
2. $\beta_{pn} = -0.5$, $N_2 = 300$, and
3. $\beta_{pn} = -0.1$, $N_3 = 400$.

The price levels that can be considered for p_i are $\{1, 3, 5, 7, 9, 11, 13\}$. Answer the following questions:

1. What is the price that maximizes your expected revenue?
2. What is the price that maximizes the expected market share of alternative i ?

3. What is the price that maximizes your expected revenue obtained from individuals in group 1?
4. What is the price that maximizes your expected revenue obtained from individuals in group 2?
5. What is the price that maximizes your expected revenue obtained from individuals in group 3?

Hints

We recommend to use a spreadsheet and complete a table where each row corresponds to a price level, and each column corresponds to a relevant indicator:

- the market shares,
- the expected revenues of group 1,
- the expected revenues of group 2,
- the expected revenues of group 3,
- the total expected revenues.

Forecasting – 7.4 Revenue maximization

Michel Bierlaire

Solution of the practice quiz: revenue maximization

We start by calculating the deterministic part of the utility of both alternatives. They are defined as

$$\begin{aligned}V_{in} &= \beta_{pn}p_i - 0.5, \\V_{jn} &= \beta_{pn}p_j,\end{aligned}$$

where β_{pn} is the coefficient of the cost associated with individual n and $p_j = 2$. The values for V_{in} can be found in the following table:

p_i	V_{in}		
	$n \in \text{Group 1}$	$n \in \text{Group 2}$	$n \in \text{Group 3}$
1	-1.5	-1	-0.6
3	-3.5	-2	-0.8
5	-5.5	-3	-1
7	-7.5	-4	-1.2
9	-9.5	-5	-1.4
11	-11.5	-6	-1.6
13	-13.5	-7	-1.8

As the price is fixed for the competing alternative, the values of V_{jn} do not vary among price levels:

1. $n \in \text{Group 1}$: $V_{jn} = -2$,
2. $n \in \text{Group 2}$: $V_{jn} = -1$, and
3. $n \in \text{Group 3}$: $V_{jn} = -0.2$.

We can now calculate the probabilities for both alternatives. The probabilities of individual n to choose each alternative are calculated as:

$$P_n(i|p_i, p_j) = \frac{e^{V_{in}}}{e^{V_{in}} + e^{V_{jn}}},$$

$$P_n(j|p_i, p_j) = 1 - P_n(i|p_i, p_j).$$

Their values for the different groups and price levels are shown in the following table.

p_i	$P_n(i p_i, p_j)$			$P_n(j p_i, p_j)$		
	$n \in \text{G. 1}$	$n \in \text{G. 2}$	$n \in \text{G. 3}$	$n \in \text{G. 1}$	$n \in \text{G. 2}$	$n \in \text{G. 3}$
1	0.622	0.500	0.401	0.378	0.500	0.599
3	0.182	0.269	0.354	0.818	0.731	0.646
5	0.0293	0.119	0.310	0.971	0.881	0.690
7	0.00407	0.0474	0.269	0.996	0.953	0.731
9	0.000553	0.018	0.231	0.999	0.982	0.769
11	0.0000748	0.00669	0.198	1.00	0.993	0.802
13	0.0000101	0.00247	0.168	1.00	0.998	0.832

Denote G_k the set of individuals belonging to group k . The market share of alternative i within group k is

$$\frac{1}{N_k} \sum_{n \in G_k} P_n(i|p_i, p_j) = \frac{1}{N_k} N_k P_n(i|p_i, p_j) = P_n(i|p_i, p_j),$$

as the choice model is the same for all individuals in a group. Therefore, we can denote it as

$$P_k(i|p_i, p_j).$$

Thus, the market shares of each alternative within group k are directly the probabilities from the above table. The total market share of alternative i in the population is then calculated as

$$\frac{1}{\sum_{k=1}^3 N_k} \sum_{k=1}^3 N_k P_k(i|p_i, p_j).$$

Once the market shares have been calculated, we can obtain the revenue generated by group k as follows:

$$p_i N_k P_k(i|p_i, p_j).$$

The total revenue is therefore obtained as:

$$p_i \sum_{k=1}^3 N_k P_k(i|p_i, p_j).$$

The following table contains the revenue of alternative i per group, the market share of alternative i and the total revenue.

p_i	Market share (%)	Rev. G. 1	Rev. G. 2	Rev. G. 3	Total Rev.
1	49.7	186.7	150.0	160.5	497.3
3	27.7	164.2	242.0	425.2	831.4
5	16.9	43.97	178.8	620.1	842.8
7	12.3	8.547	99.59	753.0	861.2
9	9.82	1.493	48.56	833.3	883.4
11	8.12	0.247	22.09	870.4	892.7
13	6.79	0.03951	9.643	873.5	883.2

1. The price that maximizes your expected revenue is 11.
2. The price that maximizes the expected market share of alternative i is 1.
3. The price that maximizes your expected revenue obtained from individuals in group 1 is 1.
4. The price that maximizes your expected revenue obtained from individuals in group 2 is 3.
5. The price that maximizes your expected revenue obtained from individuals in group 3 is 13.