EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2018

EXERCISE SESSION 2 (Solution)

Exercise 1 We want to build a model that predicts the market penetration of electric vehicles (EV) as a function of the income level. We have a sample of 1000 individuals. The data is summarized in Table 1.

	Income			
EV	low	medium	high	
yes	15	50	135	200
no	200	450	150	800
	215	500	285	1000

Table 1: Contingency table of EV ownership conditional on income level

1. Estimate the parameters π_1, π_2 and π_3 using maximum likelihood estimation, where:

$$\begin{aligned}\text{Prob}(\text{EV} = \text{yes} \quad &| \quad \text{income} = \text{low}) = \pi_1 \\ \text{Prob}(\text{EV} = \text{yes} \quad &| \quad \text{income} = \text{medium}) = \pi_2 \\ \text{Prob}(\text{EV} = \text{yes} \quad &| \quad \text{income} = \text{high}) = \pi_3\end{aligned} \tag{1}$$

Hint: write the likelihood function and find its maximum.

Solution: Likelihood: $\mathcal{L}^* = \prod_{n=1}^{N} P(i_n|k_n)$, where $P(\text{EV}_n = \text{yes} \,|\, \text{income}_n = j) = \pi_j$ and $P(\text{EV}_n = \text{no} \,|\, \text{income}_n = j) = 1 - \pi_j$, and $j = 1, 2, 3$ correspond to low, medium and high income, respectively.

In this case, $\mathcal{L}_1^* = \pi_1^{15} \cdot (1-\pi_1)^{200} \cdot \pi_2^{50} \cdot (1-\pi_2)^{450} \cdot \pi_3^{135} \cdot (1-\pi_3)^{150}$
$\mathcal{L}_1 = \ln(\mathcal{L}_1^*) = 15\ln(\pi_1) + 200\ln(1-\pi_1) + 50\ln(\pi_2) + 450\ln(1-\pi_2) + 135\ln(\pi_3) + 150\ln(1-\pi_3)$

$$\max \mathcal{L}_1 \Longleftrightarrow \frac{\partial \mathcal{L}_1}{\partial \pi_j} = 0$$

Thus, model 1 (M1) is characterized as follows:

$$\frac{\partial \mathcal{L}_1}{\partial \pi_1} = 0 \Longleftrightarrow \frac{15}{\pi_1} - \frac{200}{1-\pi_1} = 0 \Longleftrightarrow \pi_1 = \frac{15}{215} = 0.0698$$

$$\pi_2 = \frac{1}{10} = 0.1$$

$$\pi_3 = \frac{135}{285} = 0.474$$

2. Do the values of the parameters make sense?

   Solution: Yes, because they reproduce exactly the observations (maximum is achieved at these values).

3. Calculate the final log likelihood of the model.

   Solution: $\mathcal{L}_1 = 15\ln(0.0698) + 200\ln(1-0.0698) + 50\ln(0.1) + 450\ln(1-0.1) + 135\ln(0.474) + 150\ln(1-0.474) = -414.097$

4. Call the model described above M1. Consider a model with only two income categories: (i) low and medium income and (ii) high income. Call this M2. Now, consider another model with only one income category. Call this M0. Among the three models which one would you choose as the best model? Why?
   Hint: Calculate the final log likelihood associated with each model and perform likelihood ratio tests.

   Solution:

	Income		
EV	low and medium (4)	high (5)	
yes	65	135	200
no	650	150	800
	715	285	1000

We define:

$$\begin{aligned} \text{Prob}(\text{EV} = \text{yes} \mid \text{income} = \text{low, medium}) &= \pi_4 \\ \text{Prob}(\text{EV} = \text{yes} \mid \text{income} = \text{high}) &= \pi_5 \end{aligned} \tag{2}$$

Then, $\mathcal{L}_2^* = \pi_4^{65}(1-\pi_4)^{650}\pi_5^{135}(1-\pi_5)^{150}$
$\mathcal{L}_2 = \ln(\mathcal{L}_2^*) = 65\ln(\pi_4) + 650\ln(1-\pi_4) + 135\ln(\pi_5) + 150\ln(1-\pi_5)$

$$\max \mathcal{L}_2 \iff \frac{\partial \mathcal{L}_2}{\partial \pi_j} = 0$$

Thus, model 2 (M2) is characterized as follows:

$$\pi_4 = 0.\overline{09}$$

$$\pi_5 = 0.474\,(= \pi_3)$$

EV	Income
yes	200
no	800

We define:

$$\text{Prob}(\text{EV} = \text{yes} \mid \text{income} = \text{low, medium, high}) = \text{Prob}(\text{EV} = \text{yes}) = \pi \qquad (3)$$

$\mathcal{L}_0^* = \pi^{200}(1-\pi)^{800}$
$\mathcal{L}_0 = \ln(\mathcal{L}_0^*) = 200\ln(\pi) + 800\ln(1-\pi)$

Thus, model 0 (M0) is characterized as follows:

$$\max \mathcal{L}_0 \iff \frac{\partial \mathcal{L}_3}{\partial \pi} = 0 \iff \frac{200}{\pi} - \frac{800}{1-\pi} = 0 \iff \pi = 0.2$$

We note that M0 $\subset$ M2 $\subset$ M1, where $\subset$ means a restricted version of. Now we perform the pairwise comparisons to determine which model we would choose as the best model.

- **M0 against M2:** H_0 : "$\pi_4 = \pi_5 = \pi$", where M0 is the restricted model and M2 the unrestricted model.
  M0: $\mathcal{L}_R = \mathcal{L}_0 = -500.402$
  M2: $\mathcal{L}_U = \mathcal{L}_2 = -414.967$
  $-2(\mathcal{L}_R - \mathcal{L}_U) = 170.87 \overset{?}{>} \mathcal{X}^2_{0.99,1} = 6.635$ (yes)
  $\Rightarrow$ we reject H_0 (at 99% level of confidence) $\Rightarrow$ we keep M2
- **M2 against M1:** H_0 : "$\pi_1 = \pi_2 = \pi_4$", where M2 is the restricted model and M1 the unrestricted model.
  M2: $\mathcal{L}_R = \mathcal{L}_2 = -414.967$
  M1: $\mathcal{L}_U = \mathcal{L}_1 = -414.097$
  $-2(\mathcal{L}_R - \mathcal{L}_U) = 1.74 \overset{?}{>} \mathcal{X}^2_{0.99,1} = 6.635$ (no)
  $\Rightarrow$ we cannot reject H_0 (at 99% level of confidence) $\Rightarrow$ we keep M2 (simpler model)

5. Suppose now that after some economical growth, the income distribution is as follows: 75 individuals with low income, 400 individuals with medium income and 525 with high income. Use the best model to predict the market penetration of EV for this scenario.

   Solution:

   - **Low and medium income:** $75 + 400 = 475$ individuals, $\pi_4 = 0.\overline{09}$
   - **High income:** 525 individuals, $\pi_5 = 0.474$

   Penetration rate: $0.\overline{09} \cdot \frac{475}{1000} + 0.474 \cdot \frac{525}{1000} = 29.20\%$

6. Could we have used linear regression instead of discrete choice models in the context above? If so, what would have been the dependent variable?

   Solution: The main motivation behind using discrete choice models is the fact that the dependent variable we are modeling is discrete (to have or not to have an electric vehicle), as opposed to linear regression, where the dependent variable is continuous. An example of linear regression in this context could be the number of electric vehicles per household.

Exercise 2 In a mode choice experiment the follow utility functions are defined for private motorized modes (pmm) and public transportation (pt):

$$\begin{aligned} U_{pmm,n} &= -\beta_c \cdot \text{cost}_{pmm,n} - \beta_t \cdot \text{time}_{pmm,n} \\ U_{pt,n} &= -\beta_c \cdot \text{cost}_{pt,n} - \beta_t \cdot \text{time}_{pt,n} \end{aligned} \tag{4}$$

where $\text{cost}_{pmm,n}$ and $\text{cost}_{pt,n}$ are the cost of the trip by private motorized modes and public transportation respectively for individual n in CHF, and $\text{time}_{pmm,n}$ and $\text{time}_{pt,n}$ are their travel times in minutes.

Our sample contains the following 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. What is the value of time according to the model in [CHF/h]?

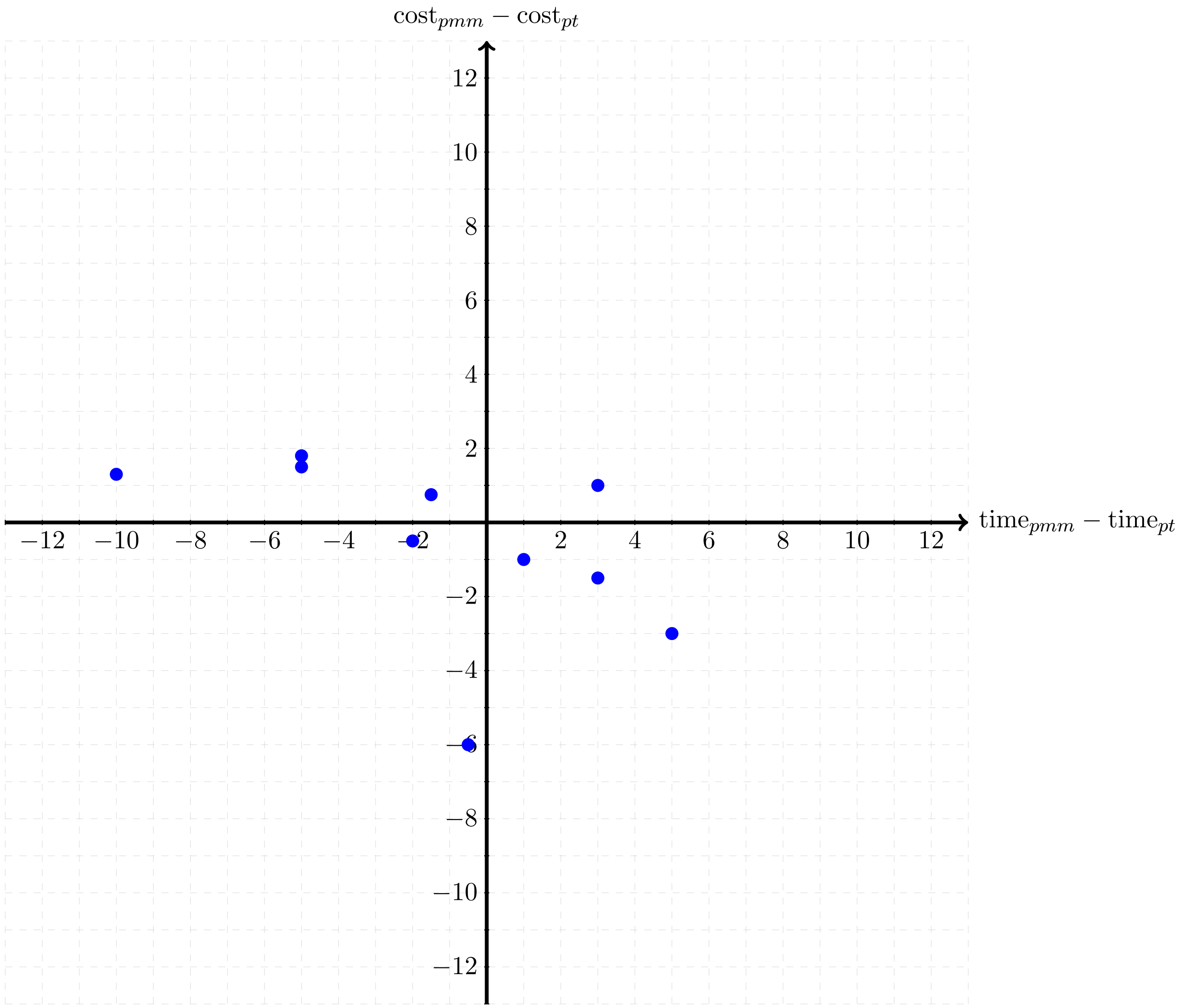
   Solution: The value of time is calculated as follows (notice the required conversion of units as the time is expressed in minutes in this exercise):

$$\text{VOT} = \frac{\beta_t}{\beta_c} \cdot 60 = \frac{0.363}{1.38} \cdot 60 = 15.78 \text{ [CHF/h]}$$

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 the plot provided in the following page.

   Solution: The differences in time are the following: Our sample contains the following 10 observations:

Individual	Choice	$\text{time}_{pmm} - \text{time}_{pt}$	$\text{cost}_{pmm} - \text{cost}_{pt}$
1	pmm	-10	1.3
2	pt	-5	1.8
3	pmm	5	-3
4	pmm	-2	-0.5
5	pt	-1.5	0.75
6	pt	-5	1.5
7	pt	3	1
8	pt	1	-1
9	pt	3	-1.5
10	pmm	-0.5	-6

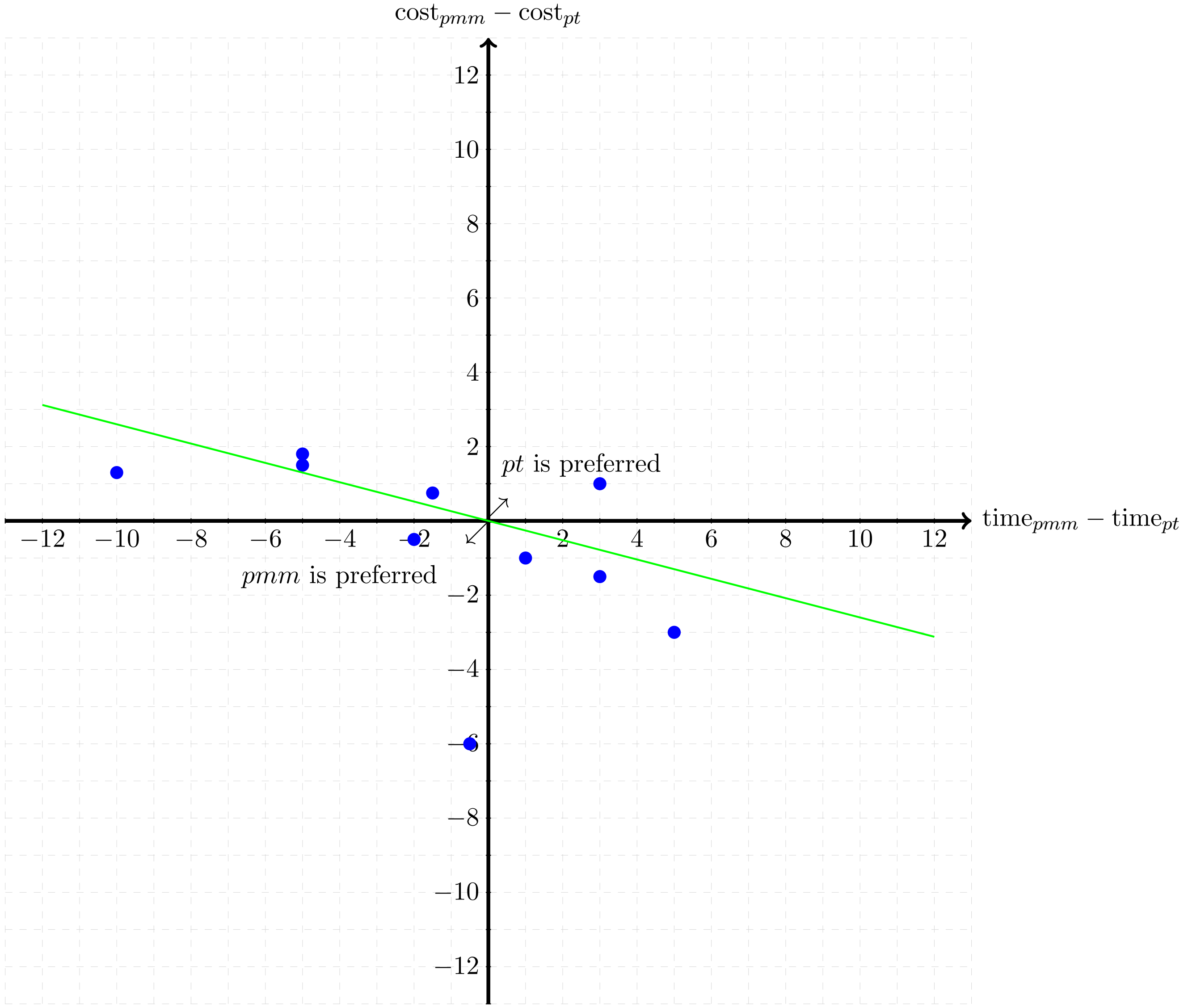


3. Add to the previous plot the line $-\beta_c \cdot \text{cost}_{pmm} - \beta_t \cdot \text{time}_{pmm} = -\beta_c \cdot \text{cost}_{pt} - \beta_t \cdot \text{time}_{pt}$. What does its slope represent?

   Solution: The line can be written as follows:

$$\text{cost}_{pmm} - \text{cost}_{pt} = -\frac{\beta_t}{\beta_c}(\text{time}_{pmm} - \text{time}_{pt})$$

The slope represents the change in the y-value ($\text{cost}_{pmm} - \text{cost}_{pt}$) per unit change in the x-value ($\text{time}_{pmm} - \text{time}_{pt}$). Thus, it represents the value of time, as each minute less in time corresponds to an additional cost of $\frac{\beta_t}{\beta_c}$.



mpp/ yo/ tr