EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2020

Analysis of mode choice behavior in the London metropolitan area

Assignment 2

Group 16 : Palazzo Romain, Zemour Eliott (formerly part of Group 01).

Base Model

In order to achieve the goal of the assignment, we aim to start with a logit model based on the first assignment's best model. We first describe the main updates derived from the corrections provided by the teaching assistants:

- The **distance variable** is added to the model. This was rejected at first because we only tested the relevance of the specification in *linear* form. It eventually appeared that considering a *logarithm* dependency significantly improves the model. The coefficients β_{dist} are defined to be alternative specific.
- The **thresholds of the piecewise linear transformation** of travel time (for car and walk alternatives) are modified to fix endogeneity issue. Indeed, we now use a threshold of 30min for the walk alternative and 15min for the car one. The underlying assumption is the existence of thresholds above which people have a noticeable change in perception of travel time.
- The **age variable** is removed from the model. Our specification included it in a *linear* form but it does not behaviorally make sense. A Box-Cox transformation have been tried in order to take age into account in a more meaningful way but it did not significantly improve the model.

Finally, the base model we shall use for this assignment is the following (with ASC_{walk} normalized to 0):

$$V_{\text{walk}} = \text{ASC}_{\text{walk}} + \text{piecewise}\left(\text{time}_{\text{walk}},\ \text{th} = 0.5\,\text{hrs}\right) \tag{1}$$

$$V_{\text{cycle}} = \text{ASC}_{\text{cycle}} + \beta_{\text{time, cycle}} \cdot \text{time}_{\text{cycle}} + \beta_{\text{dist cycle}} \cdot \log(\text{distance}) \tag{2}$$

$$V_{\text{pt}} = \text{ASC}_{\text{pt}} + \beta_{\text{time, pt}} \cdot \text{time}_{\text{pt}} + \beta_{\text{dist pt}} \cdot \log(\text{distance}) + \beta_{\text{cost}} \cdot \left(\frac{\text{cost}_{\text{pt}}^{\lambda} - 1}{\lambda}\right) \tag{3}$$

$$V_{\text{car}} = \text{ASC}_{\text{car}} + \text{piecewise}\left(\text{time}_{\text{car}},\ \text{th} = 0.25\,\text{hrs}\right) + \beta_{\text{dist car}} \cdot \log(\text{distance}) + \beta_{\text{cost}} \cdot \left(\frac{\text{cost}_{\text{car}}^{\lambda} - 1}{\lambda}\right) + \beta_{\text{dtp}} \cdot \text{dtp} \tag{4}$$

where 'dtp' stands for 'driving traffic percent', 'th' for 'threshold' and the *piecewise* terms are defined as follows:

$$\begin{aligned} \text{piecewise}\,(\text{time}_{\text{walk}},\ \text{threshold} = 0.5\,\text{hrs}) &= \beta_{\text{time, walk>0.5hrs}} \cdot \text{time}_{\text{walk>0.5hrs}} \\ &+ \beta_{\text{time, walk<0.5hrs}} \cdot \text{time}_{\text{walk<0.5hrs}} \quad (5) \\ \text{piecewise}\,(\text{time}_{\text{car}},\ \text{threshold} = 0.25\,\text{hrs}) &= \beta_{\text{time, car>0.25hrs}} \cdot \text{time}_{\text{car>0.25hrs}} \\ &+ \beta_{\text{car, walk<0.25hrs}} \cdot \text{time}_{\text{car<0.25hrs}} \quad (6) \end{aligned}$$

The parameter estimates are given in Table 1, and the final log-likelihood is:

$$\mathcal{L}_0 = -4178.607 \quad (7)$$

	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value
$\text{ASC}_{\text{cycle}}$	-7.83	2.04	-3.83	0.000128	2.19	-3.58	0.000344
ASC_{car}	-10.3	1.56	-6.62	3.48e-11	1.75	-5.88	4.13e-09
ASC_{pt}	-11.9	1.64	-7.28	3.27e-13	1.79	-6.66	2.73e-11
β_{cost}	-0.198	0.0294	-6.74	1.62e-11	0.0296	-6.69	2.26e-11
β_{dtp}	-2.99	0.246	-12.2	0.0	0.252	-11.9	0.0
$\beta_{\text{dist cycle}}$	0.512	0.326	1.57	0.116	0.348	1.47	0.141
$\beta_{\text{dist car}}$	1.37	0.266	5.15	2.64e-07	0.296	4.62	3.88e-06
$\beta_{\text{dist pt}}$	1.4	0.276	5.06	4.21e-07	0.301	4.64	3.54e-06
$\beta_{\text{time, cycle}}$	-2.89	0.682	-4.24	2.28e-05	0.56	-5.15	2.55e-07
$\beta_{\text{time, pt}}$	-3.72	0.28	-13.3	0.0	0.29	-12.8	0.0
λ	0.341	0.0696	4.9	9.47e-07	0.0698	4.89	1.01e-06
$\beta_{\text{time, car>0.25hrs}}$	-4.4	0.394	-11.2	0.0	0.414	-10.6	0.0
$\beta_{\text{time, car<0.25hrs}}$	-7.59	1.21	-6.28	3.3e-10	1.23	-6.17	6.89e-10
$\beta_{\text{time, walk>0.5hrs}}$	-3.17	0.481	-6.59	4.37e-11	0.766	-4.13	3.57e-05
$\beta_{\text{time, walk<0.5hrs}}$	-6.74	1.1	-6.15	7.65e-10	1.14	-5.91	3.37e-09

Table 1: Parameter estimates for the base model

The distance parameters are positive. Given that this specification is normalized with respect to the walk alternative, this indicates an preference towards other modes than walk as the distance increases. For the following, we shall refer to "public transportation" with the abbreviation "pt".

Modeling

1. **[2.5 points]** Propose and test a nested or cross-nested logit model. Report the nesting structure, the specification and the estimation results. Answer the following questions:
   (a) What is the underlying assumption of the proposed specification?
   (b) Comment the estimation output.
   (c) Compare Model 0 and your new specification using a statistical test. Which model is preferred and why? Denote the preferred model as $\text{Model}_{\text{pref}}$.

Solution to 1.

In this section, we aim to propose a nested logit (NL) model in order to take into account the following nesting structure, separating physically active modes to inactive ones:

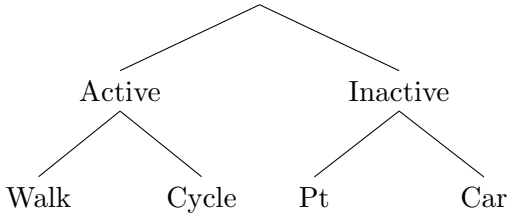

Figure 1: Nesting structure

The underlying assumption of this specification is that the walk alternative is likely to be correlated with the cycle one, since they both involve physical activity. In the same sense, public transportation and car alternatives are seen as both non-physical modes (motorized). Moreover, these nests are also consistent with a nesting structure splitting alternatives according to the (un)sheltered characteristics of transport modes (sensitivity to weather conditions).
Therefore, two scale parameters (μ_{act}, μ_{inact}) associated with the nests are introduced. To be consistent with random utility theory, we infer the inequality $\mu/\mu_m \leq 1$ for all nests m, that is a lower bound $\mu_m \geq 1$ when μ is normalized to 1. The estimation output for this nested model is given in Table 2, and the final log-likelihood is:

$$\mathcal{L}_{\text{nested}} = -4177.749 \tag{8}$$

	Value	Std err	t-test	p-value	Rob. Std err	Rob. t-test	Rob. p-value
$\text{ASC}_{\text{cycle}}$	-8.14	5.9	-1.38	0.168	6.73	-1.21	0.227
ASC_{car}	-10.1	1.65	-6.11	1.02e-09	1.93	-5.22	1.76e-07
ASC_{pt}	-11.5	1.75	-6.57	5.04e-11	1.98	-5.8	6.49e-09
β_{cost}	-0.172	0.0332	-5.17	2.32e-07	0.0354	-4.85	1.24e-06
β_{dtp}	-2.71	0.313	-8.65	0.0	0.307	-8.84	0.0
$\beta_{\text{dist cycle}}$	0.561	0.706	0.794	0.427	0.787	0.713	0.476
$\beta_{\text{dist car}}$	1.34	0.287	4.67	2.98e-06	0.337	3.98	6.91e-05
$\beta_{\text{dist pt}}$	1.36	0.299	4.53	5.76e-06	0.345	3.93	8.38e-05
$\beta_{\text{time, cycle}}$	-2.59	1.15	-2.25	0.0242	1.2	-2.17	0.0299
$\beta_{\text{time, pt}}$	-3.28	0.42	-7.8	6.22e-15	0.434	-7.56	4.17e-14
λ	0.332	0.0699	4.75	2.03e-06	0.0702	4.73	2.26e-06
μ_{act}	1.0	0.416	2.4	0.0162	0.491	2.04	0.0418
μ_{inact}	1.15	0.128	8.98	0.0	0.132	8.68	0.0
$\beta_{\text{time, car>0.25hrs}}$	-3.84	0.551	-6.98	3.05e-12	0.595	-6.46	1.08e-10
$\beta_{\text{time, car<0.25hrs}}$	-6.83	1.26	-5.41	6.46e-08	1.32	-5.19	2.1e-07
$\beta_{\text{time, walk>0.5hrs}}$	-2.99	0.958	-3.12	0.00181	1.13	-2.64	0.00838
$\beta_{\text{time, walk<0.5hrs}}$	-6.52	1.37	-4.76	1.9e-06	1.65	-3.95	7.93e-05

Table 2: Parameter estimates for the nested logit model

The parameter estimates are close to the ones found for the base model (Table 1). One can check this result by computing the maximum relative change in the parameters:

$$\text{max relative change} = \max \frac{|\text{estimate}_{\text{logit}} - \text{estimate}_{\text{nested}}|}{|\text{estimate}_{\text{logit}}|} \times 100 = 13.4\% \tag{9}$$

The maximal (relative) change occurs for β_{cost}, varying from -0.198 to -0.172.
Concerning the alternative specific parameters (ASC, β_{dist}, β_{time}), the order relations are preserved (e.g $\text{ASC}_{\text{cycle}} > \text{ASC}_{\text{car}} > \text{ASC}_{\text{pt}}$). We also notice that the scale parameter of the active nest μ_{act} is found to be pushed to its lower bound (1.0) meanwhile μ_{inact} is also close to 1. This indicates a lack of correlation between walk and cycle alternatives (within the active nest) and a small one between public transportation and car.
As the model 0 is a restricted version of the NL model, we perform a likelihood ratio test. The null hypothesis is that there is no correlation within the nests, and therefore the definition of nests is meaningless. More precisely, the scale parameters associated with both nests are equal:

$$H_0 : \mu_{\text{act}} = \mu_{\text{inact}} := \mu = 1$$

Labelling the base model as the restricted (R) model and the nested model as the unrestricted (U) one, it can be shown that under H_0, $-2(\mathcal{L}_R - \mathcal{L}_U) \sim \chi^2_{(K_U - K_R)}$ where K denotes the number of parameters in a model. Therefore, according to the likelihood ratio test, we will reject H_0 with level of confidence $(1-\alpha)$ if:

$$-2(\mathcal{L}_R - \mathcal{L}_U) > \chi^2_{(1-\alpha,\, df)} \tag{10}$$

where $df = K_U - K_R$.
In our case, $K_R = 15$ parameters are estimated in the base model while $K_U = 17$ for the NL one. From Eqs.7,8, we obtain the following statistic:

$$-2(-4177.749 + 4178.607) = 1.716 < \chi^2_{(0.95,2)} = 5.991$$

Therefore, we cannot reject H_0 with the level of significance 95%. The definition of nests does not improve significantly the model and we shall keep Model 0 as the $\text{Model}_{\text{pref}}$.

Forecasting

Using $\text{Model}_{\text{pref}}$, perform the following tasks. Make sure to justify all your calculations.

2. **[0.75 points]** Assume that stratified random sampling was used to produce the sample you are working with. We consider the following strata:

   - **S1:** Females aged 40 years or younger
   - **S2:** Females aged 41 years or older
   - **S3:** Males aged 40 years or younger
   - **S4:** Males aged 41 years or older

   Table 3 gives the size of each category in the full population. Report the size and weight of each stratum in your sample.

	Age ≤ 40	Age > 40
Male	2'676'249	1'633'263
Female	2'599'058	1'765'143

Table 3: London population estimates in 2015 (Source: ONS)

Solution to 2.

Once $\text{Model}_{\text{pref}}$ has been estimated, it must be used to derive useful indicators. We assume that stratified random sampling was used to produce the sample, and consider the strata defined above. Since some groups are proportionally more represented in the sample than they are in the population, it has to be taken into account when inferring quantities related to the population from the same quantities calculated with the sample. This is done by computing the weights (w_i) associated with each stratum, and related with their sizes S_i in the sample:

$$w_i = \frac{N_i}{N}\frac{S}{S_i}, \qquad i = 1, ..., 4 \tag{11}$$

where $S = \sum_{i=1}^{4} S_i$ is the total size of the sample and $N = \sum_{i=1}^{4} N_i$, N_i taken from Table 3. Moreover, Eq.11 is normalized such that:

$$\sum_{k=1}^{4} w_k S_k = S \tag{12}$$

Finally, the size and weight of each stratum is reported in Table 4. The results indicate that the sample 3 is less represented in our data than it actually is in the population. In order to take into account this sampling bias, the associated weight w_3 is above 1. There is a higher proportion of other strata (S1, S2, S4) in our data than in the population: their weights are less than 1.

	Sample 1	Sample 2	Sample 3	Sample 4
Size	1509	1168	1262	1061
Weight	0.993	0.871	1.222	0.887

Table 4: Sample sizes and weights of the strata

3. **[1.25 points]** Compute the aggregate market shares for each mode.
   (a) Comment the obtained results.
   (b) Compare the predicted market shares with the actual choices.

Solution to 3.

The market share corresponds to the portion of a market owned by a an alternative. In LPMC, the available alternatives are walk, bike, public transport and car and the choice set $\mathcal{C}_n$ contains all of these for each individual n. An estimator of the market share of

alternative i, W_i in the population is given by:

$$W_i = \frac{1}{S} \sum_{n=1}^{S} w_n \cdot P_n\big(i|x_n\big) \tag{13}$$

where S is the sample size and $P_n(i|x_n)$ the choice model that has been estimated from data (the probability that individual n chooses alternative i given x_n). Within the sample, the estimation of the market share does not need to take into account the sampling biases: it consists in unweighted aggregated probabilities. We denote these two quantities as "weighted market share" and "sample market share" for the population and the sample, respectively.
Using simulations, a confidence interval at level 90% is computed for the estimator on each alternative. The number of draws for this interval is taken to be 100 (as suggested by default by Biogeme). The results obtained for market shares are presented in Table 5, where we can observe a clear market dominance by pt and car alternatives. Meanwhile, the cycling alternative market share is well below the others, indicating that our model predicts a very low use of this alternative in the population. This result can be interpreted as an indicator of the poor quality of cycling infrastructures in London, or that this transport mode is perceived as less agreeable than any other modes by the main part of the population.

Moreover, a market share based on the actual choice (LPMC is a revealed-preference dataset) is computed and can be compared to the predicted sample market share. The two measures are very close to each other (actually equal up to 2 digits), indicating the accuracy of $\text{Model}_{\text{pref}}$ within the sample. The importance of aggregation can be appreciated when it comes to observe the share of users choosing the bike alternative while there was a higher probability for other modes: 3.06 %. Indeed, predicting the market share by always selecting the alternative with the highest probability would lead to a null share for the bike alternative, which is incorrect regarding the stated preferences (actual choice).

Finally, the estimated market shares for the population (weighted) are also close to the actual choices: the largest difference appears to be small (0.34%) and happens for the pt alternative. The differences arise from the sampling biases and the first column of Table 5 aims to predict the actual market share for the whole population.

	Weighted market share [%]	Sample market share [%]	Actual choice [%]
Walk	18.13 [16.74, 19.59]	18.32 [16.79, 19.85]	18.32
Cycle	3.07 [2.50, 3.91]	3.06 [2.48, 3.80]	3.06
Pt	35.37 [33.04, 37.81]	35.02 [32.74, 37.83]	35.02
Car	43.43 [40.62, 46.04]	43.60 [40.58, 46.17]	43.60

Table 5: Market share and actual choice

4. **[2.5 points]** Consider the following scenarios: (1) an increase in car cost by 15% and (2) a decrease in public transport cost by 15%.

   (a) Compare the mode shares obtained with each policy with the original mode shares.

   (b) Which scenario is the most effective policy if the goal is to decrease the share of car?

(c) Which scenario reports the highest public transportation total revenue? Is it higher than the public transportation total revenue obtained when not applying any of the policies?

Solution to 4.

In this part, we aim to study the market share evolution as a variation of the public transportation or driving cost. Two scenarios are considered: (1) an increase in car cost by 15% and (2) a decrease in public transport cost by 15%. This is done by multiplying the associated cost column by a factor (1 ± 0.15) in the dataset, and then by simulating the market share using the choice model $\text{Model}_{\text{pref}}$. In each case the market share and the public transportation total revenue is computed.

Case 1 - Increase in car cost :

The estimated market shares are shown in Table 6. With an increase in car cost by 15%, the car share decreased by 0.65% while other modes have increased their share. The largest increase is found for the public transportation (+0.35%). This indicates that pt services are the main winners resulting from this policy and might be interpreted as the fact that pt is the main rival of the car alternative in London. Nevertheless, the loss in market share by the car alternative seems quite small compared an increase in cost by 15%, suggesting that the car cost is not a dominating variable for the decision makers

In this scenario, the total public transport revenue in the sample is:

$$R_{\text{pt}}^{(1)} = £\,3331.98 \tag{14}$$

	Original share [%]		Car cost increase [%]	
Walk	18.13	[16.74; 19.59]	18.23	[16.72; 19.72]
Cycle	3.07	[2.50; 3.91]	3.10	[2.52; 3.82]
Pt	35.37	[33.04; 37.81]	35.72	[33.57; 38.22]
Car	43.43	[40.62; 46.04]	42.95	[40.09; 45.56]

Table 6: Comparison of the market share with and without an increase in car cost by 15%

Case 2 - Decrease in pt cost :

In this scenario, the policy considered is a decrease in pt cost by 15%. This leads to the results of Table 7 where one can notice a decrease in all mode shares except for the pt one. Its market share increased by 0.43%, while the change in car share is found to be -0.34%. This is consistent with the behavioral expectations and $\text{Model}_{\text{pref}}$ ($\beta_{\text{cost}} < 0$): a decrease in pt cost will only increase the public transportation utility.

In this scenario, the total public transport revenue in the sample is:

$$R_{\text{pt}}^{(2)} = £\,2850.07 \tag{15}$$

	Original share [%]	Pt cost decrease [%]
Walk	18.13 [16.74; 19.59]	18.07 [16.67; 19.48]
Cycle	3.07 [2.50; 3.91]	3.03 [2.46; 3.72]
Pt	35.37 [33.04; 37.81]	35.80 [33.41; 38.34]
Car	43.43 [40.62; 46.04]	43.09 [40.45; 45.85]

Table 7: Comparison market share with and without a pt cost decrease of 15%

Therefore, even though the pt market share is higher in the second scenario, the most effective policy to decrease the car share is the first one: an increase in car cost by 15%. This result can be used to target the optimal environmental policy in order to reduce the use of car in London.
Finally, the pt revenue obtained without applying any policy is $R_{pt}^{(0)} = £\,3299.53$. Thus, the highest total revenue for public transportation is reached for the increase in car cost by 15%. From the point of view of the pt services, the 2$^{\text{nd}}$ scenario reveals a trade-off between increasing the market share and maximizing the total revenue.

5. **[1 point]** Calculate the average value of time for car and public transportation. Comment the obtained results.

Solution to 5.

In this section, we aim to calculate the average value of time for car and public transportation. The value of time (VOT) corresponds to the price a traveler is willing to pay to decrease the travel time. This is mathematically defined by Eq.16 and will be expressed in GBP/hour:

$$\text{VOT}_{i,n} = \frac{(\partial V_{i,n}/\partial t_{i,n})(c_{i,n}, t_{i,n})}{(\partial V_{i,n}/\partial c_{i,n})(c_{i,n}, t_{i,n})} \tag{16}$$

where $c_{i,n}$, $t_{i,n}$ are the cost and travel time of alternative i and individual n, respectively. The average value of time for the sample is obtained from the weighted average of the sample, using the weights defined in Table 4.

For LPMC, only pt and car alternatives have a non-zero travel cost. Moreover, $\text{Model}_{\text{pref}}$ implies a non-linearity in the cost variable and therefore the value of time is not in the form $\beta_{\text{time}}/\beta_{\text{cost}}$ and one must be careful with the data. Indeed, the derivative of a Box-Cox transformation (w.r.t cost) at zero is ambiguous, and one must ignore the individuals for which pt cost is null. Considering this 'filter', the resulting sample is large of 3395 individuals and this implies a re-normalization of the weights in order to satisfy Eq.12.
The results obtained are shown in Table 8 and the confidence intervals are computed using the same number of draws as before (100).

	Average VOT [GBP/hour]	Confidence interval (90%)
Car	36.73	[23.39; 53.69]
Pt	31.37	[21.41; 44.77]

Table 8: Average value of time and confidence interval for pt and car alternatives.

The average value of time is above 30 GBP/hour for both alternatives, indicating that saving two minutes of travel time is worth more than £1 in London. Moreover, we found a larger value of time for the car alternative than the public transportation. On average, individuals are willing to pay more in order to decrease the travel time of the car alternative.

6. **[2 points]** Compute the aggregate cost elasticities. Report the normalization factors alongside the obtained results, and comment.

Solution to 6.

The aim of the aggregate elasticities is to anticipate the variation of the choice of a person due to the change in value of a variable. In this case, the variable of interest is the cost c. In choice model, this can be understood as how the market share is impacted by a variation of cost. The aggregate cost elasticity for alternative i, E_i, is defined in Eq.17 with W_i being the market share of the alternative i, c_n the cost for the person n and the other terms as before:

$$E_i = \sum_{n=1}^{S} \frac{\partial W_i}{\partial c}\Big|_{c=c_n} \times \frac{c_n}{W_i} \times \frac{w_n P(i|x_n)}{\sum_{j=1}^{S} w_j P(i|x_j)} S \tag{17}$$

Here again, $S = 3995$ since the individuals for which pt cost is null are ignored (because they are not influenced by a proportional variation of cost). Using this, one can also define the normalisation factor of alternative i, N_i:

$$N_i = \sum_{n=1}^{S} w_n P(i|x_n) \tag{18}$$

After having computed these formulas, one obtain the result presented in Table 9:

	Elasticity	Normalization factor
Car	-0.078	1497.14
Pt	-0.11	1193.39

Table 9: Elasticity and normalization factor for pt and car alternatives.

As a result, both elasticites are negative, meaning that an increase of cost will decrease the market share of the the car or the public transport. The value of the elasticity has to be compared with the evolution of the market share with an increase or decrease of cost. This study was realised in Question 4. Thus according to the results shown above, when the car cost increases by 15%, the variation of the market share is $\frac{42.95-43.42}{43.42} \times 100 = -1.08\%$ so the empirical elasticity is $E_{car} = \frac{-1.08}{15} = -0.072$. In the same way, for the public transport, one obtain $E_{pt} = -0.083$. Therefore, these empirical measures are close to the computed elasticity. Then, by an increase of travel cost, the market share of public transportation will be more affected than the one of car. The normalization factor for a given alternative is the expected of users in an unbiased sample of size $S = 3395$. Thus, one expect more people to take car than public transport in London.