EPFL
ENAC TRANSP-OR
Prof. M. Bierlaire

Mathematical Modeling of Behavior
Fall 2020

Analysis of mode choice behavior in the London metropolitan area

Assignment 1

Model 0 [2 points]

Start with a simple model specification. It should include:

1. alternative specific constants, and
2. cost and travel time of the different alternatives associated with generic parameters.

Report both the specification (i.e., the utility functions) and the estimates of the parameters.

The Utility functions are given by

$$U_{walking} = \beta_{time} \cdot dur_{walking}$$
$$U_{cycling} = ASC_{cycling} + \beta_{time} \cdot dur_{cycling}$$
$$U_{driving} = ASC_{driving} + \beta_{time} \cdot dur_{driving} + \beta_{cost} \cdot cost_{driving}$$
$$U_{publictransport} = ASC_{pt} + \beta_{time} \cdot dur_{pt} + \beta_{cost} \cdot cost_{transit}$$

Name	Value
ASC_CYCLING	-3.47
ASC_DRIVING	-1.01
ASC_PT	-0.442
BETA_COST	-0.161
BETA_TIME	-4.7

Figure 1: Biogeme output for Model 0

Comment on the estimation output (e.g., if the signs of the parameters match your expectations).

The signs of the coefficients make sense. Indeed since they are negative, the more travel time or travel cost spent, the lower the utility function. Thus the individual doesn't want to spend too much time or money on the transportation.

Model 1 [2 points]

Use Model 0 as the base model. Include alternative-specific parameters for at least one of the attributes of Model 0, and report both the specification and the estimates of the parameters. Answer to the following questions:

1. *What is the underlying assumption of defining alternative-specific parameters?*

   We assumed that the cost attribute doesn't change according to the transportation mean. However the travel time might change, for example walking 3 hours is more difficult and physically tiring than taking the train for 3 hours. Therefore we estimated different coefficients for each public transport mode. The Utility functions are now given by :

$$U_{walking} = \beta_{timewalking} \cdot dur_{walking}$$

$$U_{cycling} = ASC_{cycling} + \beta_{timecycling} \cdot dur_{cycling}$$

$$U_{driving} = ASC_{driving} + \beta_{timedriving} \cdot dur_{driving} + \beta_{cost} \cdot cost_{driving}$$

$$U_{publictransport} = ASC_{pt} + \beta_{timept} \cdot dur_{pt} + \beta_{cost} \cdot cost_{transit}$$

2. *Comment on the estimation output.* We found that :

Name	Value
ASC_CYCLING	-4.46
ASC_DRIVING	-1.67
ASC_PT	-2.24
BETA_COST	-0.145
BETA_TIME_CYCLING	-4.13
BETA_TIME_DRIVING	-5.31
BETA_TIME_PT	-2.62
BETA_TIME_WALKING	-7.34

Figure 2: Biogeme output for Model 1

   It makes sense, as in the model 0, that all coefficients are negative. As we mentioned in the previous answer, we expected the coefficients of walking to be very low, due to physical effort. For the same reason, the public transport coefficient is the highest, because there is no physical effort in sitting in the bus or train. Actually it can even be prefereable, because you can use your time to read a book, study, listen to music...

3. *Compare Model 0 and Model 1 with a statistical test. Which model is preferred and why? Denote the preferred model as Model* 1_{pref}.

   The new model, with respect to the old one, has exactly 3 degrees of freedom. We make the usual test via the following computation, where $\mathcal{L}_i(w)$ is the final log likelihood corresponding to model i, obtained with Biogeme methods.

$$-2(\mathcal{L}_0(w) - \mathcal{L}_1(w)) = -2(-4784.403 + 4481.136) = 606.534 > 3.84$$

Hence, we reject model 0 with a 0.95 probability (*cf.* χ^2_{df} table).

Model 2 [3 points]

Use Model 1_{pref} as the base model. Include at least an additional attribute and at least one interaction of a socioeconomic variable with either the ASCs or one of the attributes. Report both the specification and the estimates of the parameters.
Answer to the following questions:

1. *What is the underlying assumption of the included attribute(s) and interaction(s)?*

   We assumed that the more someone is old, the less they want to take the bike or walk because it might be tiring. We thus decided to include the attribute *age*. Now, intuitively, the relation between the distance and the age attributes has the same behaviour as the one between the inverse of the income and the cost, as seen in the course. One may expect that distance shall have a greater impact on old people. We established a dependency of those two attributes via a simple multiplication; as a result, the new Utility functions are given by

$$U_{walking} = \beta_{timewalking} \cdot dur_{walking} + \beta_{distanceagewalking} \cdot distance \cdot age$$

$$U_{cycling} = ASC_{cycling} + \beta_{timecycling} \cdot dur_{cycling} + \beta_{distanceagecycling} \cdot distance \cdot age$$

$$U_{driving} = ASC_{driving} + \beta_{timedriving} \cdot dur_{driving} + \beta_{cost} \cdot cost_{driving} + \beta_{distanceagedriving} \cdot distance \cdot age$$

$$U_{publictransport} = ASC_{pt} + \beta_{timept} \cdot dur_{pt} + \beta_{cost} \cdot cost_{transit} + \beta_{distanceagept} \cdot distance \cdot age$$

2. *Comment on the estimation output.* Our results can be found in Figure 3.

   Now, it seems like describing this dependency with specific parameters for each transport type wasn't entirely justified. Indeed, the estimated paramaters are the same and thus a generic paramatrisation of the *age-distance* dependency seems more appropriate. In simple words, we conclude that distance has a greater impact on the choice of old people, but the quantification of this impact does not vary along the transport mode. At last, we note that those estimators have a positive value, which we think, given that we strongly expect a negative contribution from the distance attribute, has to do with a complex non-linear dependency with the age parameter.

3. *Compare Model* 1_{pref} *and Model 2 with a statistical test. Which model is preferred and why? Denote the preferred model as Model* 2_{pref}.

   The new model, with respect to the old one, has exactly 4 degrees of freedom. We make the usual test via the following computation, where $\mathcal{L}_i(w)$ is the final log likelihood corresponding to model i, obtained with Biogeme methods.

Name	Value
ASC_CYCLING	-4.46
ASC_DRIVING	-1.63
ASC_PT	-2.32
BETA_COST	-0.149
BETA_DISTANCE_AGE_CYCLING	0.603
BETA_DISTANCE_AGE_DRIVING	0.603
BETA_DISTANCE_AGE_PT	0.603
BETA_DISTANCE_AGE_WALKING	0.603
BETA_TIME_CYCLING	-4.13
BETA_TIME_DRIVING	-6.6
BETA_TIME_PT	-2.14
BETA_TIME_WALKING	-6.2

Figure 3: Biogeme output for Model 2

$$-2(\mathcal{L}_1(w) - \mathcal{L}_2(w)) = -2(-4481.136 + 4428.231) = 105.81 > 9.49$$

Hence, we reject model 1 with a 0.95 probability (*cf.* χ^2_{df} table).

Model 3 [3 points]

Use Model 2_{pref} as the base model. Include at least one appropriate non-linear specification for one of the variables, and report both the specification and the estimates of the parameters. Answer to the following questions:

1. *What is the underlying assumption of the included non-linear specification(s)?*

   In this last model, we assumed that the relation between the public transport utility and its corresponding duration was not linear. This is justified by the natural feeling that a 5 minutes difference doesn't have the same impact on a choice when added to a total of 5 minutes and one of 5 hours. It the second case, the difference is almost insignificant. The method chosen to describe this non-linearity is the piecewise affine dependency implementation. We were careful however, not to include too many thresholds, as a high number of degrees of freedom of the new model could have led us to stick with model 2.

$$U_{walking} = \beta_{timewalking} \cdot dur_{walking} + \beta_{distanceagewalking} \cdot distance \cdot age$$

$$U_{cycling} = ASC_{cycling} + \beta_{timecycling} \cdot dur_{cycling} + \beta_{distanceagecycling} \cdot distance \cdot age$$

$$U_{driving} = ASC_{driving} + \beta_{timedriving} \cdot dur_{driving} + \beta_{cost} \cdot cost_{driving} + \beta_{distanceagedriving} \cdot distance \cdot age$$

$$U_{publictransport} = ASC_{pt} + \beta_{timept} \cdot dur_{pt} + \beta_{cost} \cdot cost_{transit} + \beta_{distanceagept} \cdot distance \cdot age + f(dur_pt, [\beta_{1,init}, \beta_{2,init}, \beta_{3,init}]),$$

where $f(dur_pt, [\beta_{1,init}, \beta_{2,init}, \beta_{3,init}])$ represents the piecewise formula, and is a sum of the form $\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3$, because we have chosen to include exactly 4 thresholds.

2. *Comment on the estimation output.*

Name	Value
ASC_CYCLING	-4.35
ASC_DRIVING	-1.52
ASC_PT	-2.89
BETA_COST	-0.144
BETA_DISTANCE_AGE_CYCLING	0.113
BETA_DISTANCE_AGE_DRIVING	0.113
BETA_DISTANCE_AGE_PT	0.113
BETA_DISTANCE_AGE_WALKING	0.113
BETA_TIME_CYCLING	-4.13
BETA_TIME_DRIVING	-6.59
BETA_TIME_WALKING	-5.99
beta_dur_pt_0.23164244444329998_0.6949273333299	-1.79
beta_dur_pt_0.6949273333299_more	-3.3
beta_dur_pt_lessthan_0.23164244444329998	0.692

Figure 4: Biogeme output for Model 3

The estimators β_1 and β_2 are negative, but the third one is not : it seems like people care a lot about the duration for the first 3 thresholds, but after a long duration, time doesn't have that much of a strong impact, as expected (even though an almost zero value for β_3 would have been more natural). We obtain a final log likelihood of -4417.152, which, given the low number of added parameters with respect to the previous model, makes this new model more adapted (according to the test of question 3). We thus obtained a sequence of nested models with increasing efficiency.

3. *Compare Model 2_{pref} and Model 3 with a statistical test. Which model is preferred and why?*

The new model, with respect to the old one, has exactly 3 degrees of freedom (indeed, we implemented 4 thresholds for the piecewise formula). We make the usual test via the following computation, where $\mathcal{L}_i(w)$ is the final log likelihood corresponding to model i, obtained with Biogeme methods.

$$-2(\mathcal{L}_2(w) - \mathcal{L}_3(w)) = -2(-4428.231 + 4417.152) = 22.158 > 7.81$$

Hence, we reject model 2 with a 0.95 probability (*cf.* χ^2_{df} table).