Group 17 – Assignment 2
Marija Nikolic
Yoann Trellu
Ma Yuanfang



# Switzerland Mode Choice - Optima - Multinomial Logit

This case study deals with the estimation of a mode choice behaviour model for inhabitants in Switzerland. The difference with the previous report is the addition of one alternative which is soft mode. In order to estimate the Multinomial Logit (MNL) model we consider complete dataset with three alternatives: public, private transport and soft modes. We started with very simple model (base model), considering that only time and cost are relevant for choice of alternatives private and public transport mode, and that only distance is relevant factor influencing the choice of soft mode. Step by step, we tried many transformations, many improvements of the model specification in order to let «the data talk».

## Interpretation of the estimation results

The estimation results are reported in MNL_model_optima.html.

**Alternative Specific Constants (ASCs)**

One of the alternative specific constants (ASC_PT, ASC_CAR, ASC_MD) has to be normalized to zero for identification purposes. The corresponding alternative is the reference alternative for the ASCs, that is important for the interpretation. As in binary choice, the selection of the referent alternative has no effect on the model other than to shift the values of the estimated constants, preserving their differences. We decided to normalize ASC_MD to zero, and this is actually our «reference alternative» not only for ASCs, but for all socio-economic characteristics later described. ASC_CAR and ASC_PT are both significantly different from zero at 95% level, based on t-test. Sign of ASC_CAR is positive, meaning that everything else being equal alternative private mode is preferred to alternative soft mode.  Sign of ASC_PT is negative, meaning that everything else being equal alternative public mode is less preferred than alternative soft mode.

**Cost parameters**

With base model we tested if there exists taste variation across market segments. The segmentation is made on the income attribute. We first create three market segments as follows: LOW (from 1 to 4000 CHF), MEDIUM (from 4001 to 8000 CHF), and HIGH (more than 8001 CHF).

We estimate a base model on the full data set. Then we estimate the same model for each income group separately. For that purpose we uses [Exclude] section in the model specification to define the observations which should be excluded for the estimation. Obtained values of  Log likelihood functions are presented in Table 1, for each case.

| Model | Final log-likelihood | Number of parameters |
|---|---|---|
| Low income | -189.586 | 6 |
| Medium income | -669.856 | 6 |
| High income | -586.050 | 6 |
| Restricted model | -1245.963 | 6 |

Table 1: Market segmentation test

LR = -2 (Lrestricted – (Llow + Lmedim + Lhihg)) = 399.058 > 21.03 (12 degrees of freedom)

Our null hypothesis was that there is no taste variation, but based on the result of likelihood ratio test we rejected the null hypothesis at a 95% level of confidence, which means that market segmentation on income does exist.

Based on that we performed discrete segmentation of cost attribute. We put those parameters as generic ones, influencing the choice of private and public transport mode. They are all negative and significantly different from zero. This corresponds to our expectations, and this test is relevant. Value of parameter for high income is significantly lower than values of parameters for medium and high income, which makes sense - the more you have money the less you care for your train ticket, gas cost, etc.

**Time parameters**

Our next assumption was that time is perceived in different way in private and public transport mode. That is why we used alternative specific parameters for this variable. In the case of private transportation mode we assumed that time plays role as nonlinear. Therefore we performed Box-Cox test on time with hypothesis that LAMBDA parameter is equals one. Based on t-test value for LAMBDA parameter we concluded that it is significantly different from one, meaning that time in this case influencing the utility in non-linear way. We then tried with piecewise linear segmentation, which gave us better fit, based on adjusted rho-square test. We put a BETA_TIME1_CAR and BETA_TIME2_CAR for the car alternative corresponding to a piecewise linear transformation. We have specified it by dividing the time in [0,50] and [50, max(car_time)]. Estimation results show that those parameters are significantly different from zero, negative which is intuitively correct and the effect of first interval on utility is stronger than the effect of second

one (|-0.0474 | > | -0.0168 |).
For public transportation mode, our assumption was that the trip purpose affects perceived time a lot. Therefore we tried with discrete segmentation of time, based on trip purpose: leisure trip or work trip purpose. This gave us significantly different from zero (based on t-test) and negative parameters, all different and as expected time when purpose is work is more important than time when purpose of the trip is leisure.

**Distance parameter**
Our assumption about soft mode is that the distance is only relevant for soft transportation mode because cost, for instance should be of greater importance in the other choices. That is why we only put it in the soft mode. Within the base model we included distance as linear, and then we assumed interaction between distance and age. Reason for that assumption is that we believe that distance is perceived in different way by young and old people. Based on estimation results we concluded that proposed interaction is not appropriate (parameter associated to that interaction term was not significantly different from zero).
The next assumption was that this variable plays role as nonlinear in our model. Logarithm transformation was used. To test our assumption we performed Cox test, described later in text. Based on results of this test in combination with our judgment we concluded that our assumption was not appropriate (details are described later). After all mentioned attempts, we decided to keep distance variable as linear. Parameter associated to it is significantly different from zero and negative, meaning that the utility perceived by the decision maker for soft mode alternative decreases with increase in distance.

**Number of trajectories parameters**
One could suspect that a high number of trajectories would decrease the utility of public transportation mode, which is the case in our model. The coefficient is negative, meaning that the higher the number of trajectories performed in the trip loop using public transport mode, the lower utility for choosing this alternative. Value of t-test: -7.01 for this coefficient tells us that it is significantly different from zero, as we assumed.

We have also estimated a model where the number of trajectories is interacted with the number of transfers**.** Regarding this interaction, the associated parameter is significantly different from zero and positive, meaning that greater the ratio (less number of transfers) more attractive is utility. This term interacting with number of trajectories and as a result we have that loops consisting of the same number of trajectories have a different effect on utility function according to the number of transfers performed.

**Gender, type of region and language region parameters**
Since the socio-economic characteristics of decision-maker do not vary across alternatives and only differences in utilities matter, we need to choose one alternative as referent (MD in our case). For gender {man, woman}, type of region {urban, rural} and language region {French, German} characteristics we chose woman, rural and German language region as referent levels. Dummy variables for man, urban and French language region are included it in the deterministic part of utilities for private and public transport mode. Interpretation of the estimated coefficients for the remaining alternatives has to be done with respect to the reference alternative, which arbitrarily is alternative MD. We assumed that the type of region, language and gender affect differently each alternative so they are included as alternative specific.
The coefficient of the French language region for private mode are statistically significant different from zero and positive, indicating that a priori people from French speaking region tend to prefer the private transport mode more than people from German language region. Based on estimation results, men have tendency to favor alternative private mode, compared to women. For people from urban regions private mode alternative is less preferred than from people from rural ones. Note that, for men from French speaking urban regions the constant becomes 0.406 + 0.253 + 1.16 -0.399 = 0.761 for private mode. Nothing could be said about the public transportation mode regarding those characteristics, since they were not significant and therefore we decided to fix them to zero.

**Comparison of different model specifications**
In order to test if significant improvement is achieved by using nonlinear transformation of distance (log) we performed Cox test. Therefore we constructed a composite specification (Mc) with 16 coefficients from which both, the model (M1) with distance as linear and the model (M2) with the distance as nonlinear variable can be derived, as a special cases. Then we performed two likelihood ratio tests for each of the two restricted models against the composite model. -2(L(Mr) - L(Mu)) is Chi-squared distributed with K = 1 degrees of freedom in this case.

**M1** H0 : BETA_DIST = 0

-2(L(M1) - L(Mc)) = -2(-1197.222 + 1162.914) = 68.616 > 3.84 (95%)

The result tells us that we cannot reject the null hypothesis H0: the linear model cannot be rejected.

**M2** H0 : BETA_DIST_LOG= 0

-2(L(M2) - L(Mc)) = -2(-1175.571 + 1162.914) = 25.314 > 3.84 (95%)

The result tells us that we cannot reject the null hypothesis H0: the nonlinear model cannot be rejected. The test is then indecisive and the model with the highest adjusted likelihood ratio index $\bar{\rho}^2$ should be preferred, which in this case means that model with nonlinear transformation of distance variable is prefered (M2 = 0.428, M1 = 0.421).
However, we decided to keep model specification in which distance plays role as linear. Reason for that is the fact that with nonlinear specification of distance cost parameters increases as the income rises, which does not make sense (people with higher income care about money spent for transport more than people with low incomes). In our opinion model M1 explains data better than M2.