Multionomial Logit

In this assignment, you will investigate the determinants of high school choice. The data contain information on 422 primary school students who live in Utrecht and were advised to follow the academic track (VWO) in high school. Students can choose between 10 options: 9 high schools in Utrecht or a school outside Utrecht.

The variable id identifies students, school defines the high schools. The dataset includes individual-specific variables, such as gender and test score, as well as individual-school-specific variables, such as distance, sibling, and distance to gymnasium. Table 1 describes the dataset in details.

Table 1: Data description

Variable	Description
id	Student identifier
school	school identifier
gender	gender (0 = male, 1 = female)
test score	test score at the end of the primary school
distance	distance from home to school (km)
sibling	having a sibling in school s
distance to gymnasium	distance from home to the closest gymnasium
choice	Most preferred high school (binary)
gymnasium	The most preferred high school is a gymnasium

You will study if distance to school or having a sibling in a particular school influences school preferences. For that, you need to use the complete dataset. For the estimation, use Stata's ASCLOGIT command.

(*i*) Estimate a multinomial logit model with a choice probability of

$$p_{is} = \frac{\exp\left\{\beta_{0s} + \beta_{1s} \cdot \text{gender}_i + \beta_{2s} \cdot \text{test score}_i\right\}}{\sum_{s=1}^{10} \exp\left\{\beta_{0s} + \beta_{1s} \cdot \text{gender}_i + \beta_{2s} \cdot \text{test score}_i\right\}},$$

where p_{is} is the probability that individual i chooses school s.

(*ii*) Estimate a conditional logit model with the choice probability of

$$p_{is} = \frac{\exp\left\{\beta_0 + \beta_1 \cdot \text{distance}_{is} + \beta_2 \cdot \text{sibling}_{is}\right\}}{\sum_{s=1}^{10} \exp\left\{\beta_0 + \beta_1 \cdot \text{distance}_{is} + \beta_2 \cdot \text{sibling}_{is}\right\}}.$$

How do you interpret the estimated coefficients? Based on your estimates, do you think that students prefer schools that are closer to their home? Do students prefer the high school(s) that their sibling(s) attend?

(*iii*) Now combine the previous specification, and estimate a mixed logit model that includes both individual-specific covariates (such as test score and gender) and alternative-varying regressors (such as distance and sibling).

(*iv*) Explain the independence of irrelevant alternatives (IIA) assumption and discuss if it is relevant in this context.