Assignment DEEQA EEE: discrete choice models

You received a dataset with 5158 students deciding to go to school in two periods. In period 1 they can attend high school. If successful, they can attend college in period 2.

We observe their decisions ("school") in every "period" and some characteristics. Note that ability measures are normalized to be mean 0 and standard deviation 1 in the original database (but then discretized in your database). "parentcollege" is a dummy=1 if at least one of the child's parents went to college. "dist" is distance to the closest higher education institution in kilometers. The dummy "degree" indicates if the student graduated from high school.

Assume a 2-period model: in period 1, students can only attend high school or drop out, in period 2 they can only attend college or drop out. They need a high school degree to be able to go to college. Also assume they do not get any utility from obtaining a high school degree (except through the fact that it allows for college enrollment). You can assume that students differ because of observable characteristics, as well as unobservables that are uncorrelated over time. You don't need to worry about standard errors but focus on the estimates.

I recommend using Stata but you can use the software you prefer. Please hand in your code, a log file and a document with your answers to the questions below. Make sure your document also explains how you do things, not just the final result. For some questions, there is more than 1 way to answer it (e.g. you can choose some functional form for the utility function) so results can differ depending on these choices.

The deadline is **October 10, 2021**. You can work individually or in groups of 2.

This assignment counts for 3 points out of 20 for this course (so 3 out of 10 for the part taught by me).

Note that this is not a paper, you should answer directly to the questions. Aim for 5 to 10 pages but the less you need to answer the questions, the better.

Questions:

1) Provide a table with the main variables in the dataset and provide some descriptive evidence about the role of observable characteristics in choices.
2) Estimate the model assuming students are not forward looking, i.e. a model that predicts the schooling decision in each period and the probability to obtain a high school degree, but assume students only care about current period utility.
   a. Explain every step and the assumptions you make.
   b. Explain how you estimate the model
   c. Show the parameter estimates and discuss the signs
   d. What's the impact on the utility of going to college of not having a parent who went to college? Answer in kilometers.
   e. Counterfactual 1: college available everywhere (i.e. distance=0). What is the impact on total college enrollment? What is the impact on high school enrollment?
   f. Counterfactual 2: only 50% of students with a high school degree are (randomly) allowed to enter college. What's the impact on total college enrollment? What is the impact on high school enrollment?
3) Repeat every step of question 2, but now with forward looking students that use a discount factor of 0.95. Why are estimates different? Why are counterfactual simulations different?
4) For the model with discount factor 0.95: estimate the parameters of high school utility without estimating a structural model for entering college using a CCP estimator.
5) Now assume this model continues after period 2 as follows: as long as students are in college, they can choose between staying there or leaving college. If they leave college, they cannot re-enter. Try again to estimate the parameters of high school utility using a CCP estimator. What does this estimator look like? Do you need more data? Why (not)?