Assignment 2

Download the data from the following link.

Database kept confidential. It cannot be shared.

NOTE: The data is confidential and can be used just for educational purpose in this course. You should not upload it for sharing.

The data is from a well-known online firm, which provides a matchmaking service for dating. To become a member of it, you are expected to input your physical characteristics such as height, weight, picture, etc., and background information such as marital status, education level, monthly income, wealth level, etc. There is a total of 30 items that you need to fill in. But you can choose just a few of them to provide your information, and strategically, you can provide wrong information as well.

The site investigated its 203 members' information provision practice (file name= "eum_assignment.csv"). There are nine variables in the data.

Var	Description
ID	User ID number
know	Whether the user has the knowledge on this site regarding how it works (Y=1, N=2)
expr	Whether the user has the direct or indirect experience of online matchmaking (Y=1, N=2)
itr	Internet trust (Likert scale, 5=trust much, 1=I don't believe internet much)
ip	Internet privacy concern (Likert scale, 5=high , 1=low)
sex	1=male, 2=female
item	Information item such as 1=name, 2=age, 3=height, …
prov	1=correct information provided, 2=not provided, 3=incorrect information provided

The site also conducted a survey to other users to collect information regarding the information item (file name= "eum_sensrelv.csv"). There are three variables in the data.

Var	Description
item	Information item such as 1=name, 2=age, 3=height, …
sen	How much the item is sensitive to provide (0=not at all, 100=extremely sensitive) (mean value of 63 respondents)
relv	How much the item is relevant to find a good matching (0=not at all, 100=extremely relevant) (mean value of 63 respondents)

Q1. Merge the two data tables into a single one. Explore the data. Can you find any interesting pattern or relationship between variables?

Generate a new dummy variable `D` where 1=correct info and 0=otherwise. We want to find a relationship between `D` and `sen`, and `D` and `relv`.

Q2. Use a pooling model to regress `D` on `sen`, `relv` and other control variables which seem to be appropriate. Criticize this pooling model. Apply the first difference model, fixed effects model and

random effects model to the data. Do you find any different results between models? Which model will you choose?

Q3. Provide the interpretation on your model.

Q4. Check whether there is a moderating effect of `relv` on the relationship between `D` and `sen`.

Q5. Do we need more data to establish the relationship between D and `sen`, `relv?` Which control variables do we need to incorporate more into the model to find correct correlations between variables?

Guideline for Assignment 2:

Submit your answer and R-codes (or RapidMiner Screenshots) used for the analysis to the course Blackboard. Use *plm* command as you encounter an error with *pglm* command for the fixed effect model. Please include your student number and name in the header of the document. Your document should be in the same format as the final report (See the syllabus) and should not exceed three A4 pages. For the model comparison, please make a good-looking table like Table 2 of Pitt (JPE, 1998). Append all R-codes or RapidMiner screenshots used for the analysis. There is no page limit for the appendix. The due is Nov 1, 23:59PM.