Name:
DNI/Passport:

GRAU D'ENGINYERIA INFORMÀTICA (UPC).
CURS 19-20 Q2 –QUIZ 2
Anàlisi de Dades i Explotació de la Informació (ADEI) .
(Data: 29/5/2020 10:00-12:00 h On-line

Professor:	Lídia Montero Mercadé
Rules for the quiz:	Emailing and chatting is strictly forbidden. Mobile phones should be switched off. PC camara should be turned on to invigilate you. You have to deliver 1 Name.FamilyName.pdf file containing answers to the questions, used commands and R output results needed to justify your answers.
Duration:	1h 45 min
Marks:	Before 5/6/20 Subject ATENEA website.
Open Office- online:	5/6/20 10:000

Problem 1: All questions account for 1 point

1793 choices by 561 individuals of a transport mode from/to Freetown airport (Sierra Leone) to downtown. This problem exploits an unusual transportation setting to generate some of the first revealed preference value of a statistical life (VSL) estimates from a low-income setting. Four alternatives are available: ferry, helicopter, water-taxi and hovercraft. A striking characteristic of the study is that all these alternatives experienced fatal accidents in recent years, so that the fatality risk is non-negligible and differs much from an alternative to another. For example, the probabilities of dying using the water taxi and the helicopter are respectively of 2.55 and 18.41 out of 100,000 passenger-trips.

Variable	Description
id	Individual id *(not to be used in this exercise)*
choice	1 for the chosen mode
mode	One of Helicopter, *(not to be used in this exercise)*WaterTaxi (a small craft for 12 to 18 pax), Ferry, and Hovercraft
cost	the generalised cost of the transport mode (US$) – *numeric target*
risk	The fatality rate, numbers of death per 100,000 trips for the selected mode
weight	Weights *(not to be used in this exercise)*
seats	Level of seat availability - comfort (Likert scale 1 to 5, transformed to 0 to 1 scale)
noise	Level for less noise disturbance (Likert scale 1 to 5, transformed to 0 to 1 scale)
crowdness	Level for less crowdedness (Likert scale 1 to 5, transformed to 0 to 1 scale)
convloc	Level of convenience location for the transfer (Likert scale 1 to 5, transformed to 0 to 1 scale)
clientele	Level of quality of 'trip makers' (Likert scale 1 to 5, transformed to 0 to 1 scale)
chid	Choice situation id *(not to be used in this exercise)*
african	yes if born in Africa, no otherwise
lifeExp	declared life expectancy
dwage	declared hourly wage
iwage	imputed hourly wage
educ	level of education, one of low and high
fatalism	self-ranking of the degree of fatalism
gender	gender, one of female and male
age	age
haveChildren	yes if the traveler has children, no otherwise
swim	yes if the traveler knows how to swim, 'no', otherwise
noalt	Number of available alternatives for the selected choice

Name:
DNI/Passport:

The trade-offs that individuals are willing to make between mortality risk and cost as they travel to and from the international airport in Sierra Leone are estimated. The setting and original dataset allow us to address some typical variable concerns, and also to compare VSL estimates for travelers from different countries, all facing the same choice situation. The average VSL estimate for African travelers in the sample is US$ 577,000 compared to US$ 924,000 for non-Africans. The two covariates of interest are cost (the generalized cost in $PPP unit, not *leones*) and risk (mortality per 100,000 passenger-trips). The risk variable being purely alternative specific, intercepts for the alternatives cannot therefore be estimated. To avoid endogeneity problems, the authors introduce as covariates marks the individuals gave to 5 attributes of the alternatives: comfort, noise level, crowdedness, convenience and transfer location and the "quality" of the clientele.

Source

`data("RiskyTransport")` # mlogit package – long format dataset
American Economic Association data archive.

References

León, Gianmarco, and Miguel, Edward. *Risky Transportation Choices and the Value of a Statistical Life*. Nashville, TN: American Economic Association [publisher], 2017. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2019-10-12. https://doi.org/10.3886/E113686V1.

Let us focus on travel cost (cost variable). Firstly, restrict your active data set to observations involving 4 available alternatives (noalt=4) and actual choice (choice=1). Secondly, define a new binary factor containing WaterTaxi choice versus Others.

1. Indicate by data exploration tools which are globally the most associated variables with the **response variable (cost)**.
2. Calculate the linear model that explains the cost of the transfer from the imputed wage (iwage) and factor mode: interpret the regression lines and assess its global quality. What is the percentage of the cost variability that is explained by the transportation mode?
3. Calculate a linear model for the target cost using all available numeric variables. Are there any collinearity issues in the model? Justify the solution to remove collinearity.
4. Once the best model for target cost using explanatory numeric variables has been proposed, are there any significant main factor effects to be included? And interactions? Justify your answer.
5. Select the best model available so far. Let us assume an observation on the median of numeric variables and reference levels for the factors. Estimate a 90% confidence interval for predicted transfer cost.
6. Graphically assess the best model obtained so far. Assess the presence of outliers in the studentized residuals at 95% confidence level. Indicate which those observations are and why they are showing lack of fit.
7. Study the presence of *a priori and a posteriori* influential data observations. Indicate thresholds to be applied to the statistic involved in the diagnostic.
8. **WaterTaxi binary choice factor is the new target to be addressed.** Estimate a logit model including seats, crowdness, convloc covariates and educ and swim factors. Discuss model fit taking into account marginal trends and residual plots.
9. Interpret model equations and the effects in the odds scale of involved factors.
10. What would be the expected probability of using a 'WaterTaxi' for a high education and swimmer trip maker when numeric explanatory variables are set to their sample minimum?