Team Assignment – Part 4

The assignment is due to **Friday, May 26, 23:59**. You can work in groups of at most three students, and in your answers, clearly explain all the steps you take. The assignment should contain on its front page the names of the students in the team as well as their SNR 7-digit numbers.

Assuming your team is already registered, submit your solution on behalf of your team: go to section "Assignments", select Assignment 4, and submit your solution as a PDF file on Canvas (other types of files are not allowed).

Question 1

Consider the Pareto distribution, which was originally designed to describe the distribution of wealth in a society. The Pareto distribution function is given by

$$F(x) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha} & \text{if } x \geq x_m \\ 0 & \text{if } x < x_m \end{cases},$$

where $x_m > 0$ represents the minimum possible value of the Pareto distributed random variable and $\alpha > 0$ is the tail index. Both x_m and α can be unknown. To estimate these parameters, we assume here that the sample $x_1, \ldots, x_n$ follows the Pareto distribution with a particular x_m and α values.

1. Write down the likelihood function for sample $x_1, \ldots, x_n$ as a function of the parameters x_m and α.
2. Find the maximum likelihood estimator of the parameter x_m. *Hint:* you can obtain it without taking the logarithm or first-derivative of the objective function.
3. Assume now that x_m is know (it can of course be estimated using the method proposed in point 2). Write down the log-likelihood function, obtain the first-order conditions with respect to α, and the corresponding maximum likelihood estimator of α.
4. Let $\hat{\alpha}_n$ represent the finite-sample estimate proposed in point 3. Using the asymptotic distribution of the maximum likelihood estimator, derive the formula for the standard error of $\hat{\alpha}_n$ and propose its estimator.

Question 2

Consider the "docvis.csv" (see Canvas) containing information about the number of doctor visits (variable *docvis*) of a random sample of individuals. We will model and study the probability $P(y_i = 1|x_i)$

that an individual visits a doctor in a given year, $y_i = I(docvis_i > 0)$, given the following characteristics x_i: age (variable *age*), dummy indicating gender (variable *female*), dummy indicating a chronic disease (variable *chronic*), and dummy indicating marital status (variable *married*). *Note:* the label GLM in the text below stands for the generalized linear models. This is a broad class of models that can be written in the form $E(y_i|x_i) = g(x_i^\top \beta)$ and $Var(y_i|x_i) = c \cdot V(x_i^\top \beta)$, where y_i and x_i represent the dependent and explanatory variables and β are regression parameters; the function g is called the link function. Some binary-choice models form a special case of GLM with binomial response variable y_i.

1. Report summary statistics for the variables in the data set. How many people in the sample have not been to the doctor at all?

2. To predict the probability of visiting a doctor at least once a year, one can regress the binary variable $Any = I(docvis > 0)$ on variables *Age, Female, Chronic,* and *Married* using the probit model (e.g., by command *glm* in R using the option 'family=binomial(link="probit")'). Include in the model interaction effects between these four variables, and after selecting the significant ones, report the final model, the code used for estimation, and its output. Do both explanatory variables *Age* and *Chronic* have expected signs and are their effects significant?

3. To check for heteroscedasticity, you estimate also heteroscedastic probit (e.g., by command *hetglm* in package "glmx" using again the 'family=binomial(link="probit")'). Use the model constructed in point 2, but allow the heteroscedasticity to be a function of all variables *Age, Female, Chronic,* and *Married* (the interaction effects do not have to be employed in modelling the variance). Report the estimation output and the likelihood ratio statistics for the test of heteroscedasticity including the hypotheses, the value of the test statistics, the corresponding critical value or p-value, and conclusion. Given the results obtained in points 2 and 3, which model would you choose for modelling the probability of visiting a doctor?

4. One can also regress the binary variable *Any* on *Age, Female, Chronic,* and *Married* and the previously selected interaction terms in point 2 using the logit model (e.g., by command *glm* in R using the 'family = binomial(link="logit")'). Report the code and its output. Do both explanatory variables have the same signs and magnitude as in point 2?

5. To interpret the results of probit and logit, marginal effects have to be computed. Compute marginal effects at average for probit and logit and compare their signs and magnitudes (e.g., manually or using commands *logitmfx* and *probitmfx* from package "mfx"). Are the marginal effects for variables *Age* and *Chronic* computed in the same way (i.e., using the same formula)?

6. Interpret the marginal effects of *Age* and *Chronic* for men.

7. Finally, evaluate the fit of the probit model by computing the percentage of correctly predicted observations and by tabulating the actual against predicted number of individuals visiting a doctor; in both cases, use the cut-off point 0.5. Discuss the results.