Problem n.2

The CEO of *AppliedStatistics* knows that the linear dependence between what is monthly earned and spent could be simply captured by a linear model: the more is earned, the more is spent. For such reason, he rarely increases his employees' gross yearly income. The manager of the Econometrics team of *AppliedStatistics* contests the CEO's statement providing the following argumentation: if it is true that an employee earning less would spend a rather constant amount of money, it might not be true that an employee earning a lot always buys expensive goods or lives in expensive apartments. To prove his idea, he analyses the average monthly credit card expenditures [€] of the employees of *AppliedStatistics*.

The file `AvgExpenditure.txt` contains the monthly credit card expenditures (`AvgExp`, expressed in €), averaged over the year 2022, of 150 `Employees`. The dataset also reports, for each `Employee`, the gross `Income` [K€] in 2022, the `Age` in years and whether he owns or rents a house (`OwnsRents` $\in \{0 = \text{he owns}, 1 = \text{he rents}\}$).

a) Formulate the CEO's *classical* linear regression model (**M0**) for $\texttt{AvgExp}_i$ for $i \in \texttt{Employee}$, as a function of all the other variables, with no interaction. Report the model and its parameterization, together with the estimates of the parameters of the model and the standard deviation σ of the error term $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

b) Given **M0**, report the mean difference of the average monthly expenditures of the employees who rent an house with respect to the ones who don't. Is such difference significant at 5%? Report the correspondent p-value.

c) Analyze the standardized residuals of **M0** and report the plot. Does such plot bring evidence in favor of the CEO's statement? Highlight possible weaknesses of the model.

d) Propose two different models (**M1** and **M2**) in which the error term assumes the expression $\epsilon_i \sim \mathcal{N}(0, \sigma_i^2)$ with in turn:

M1) $\sigma_i = \sigma \cdot |\texttt{Income}_i|^a$

M2) $\sigma_i = \begin{cases} \sigma \cdot |\texttt{Income}_i|^b & \text{if } \texttt{OwnsRents} = 1 \\ \sigma \cdot |\texttt{Income}_i|^c & \text{if } \texttt{OwnsRents} = 0 \end{cases}$

Report the estimates of a, b, c and, for both models, the estimate of σ and the Akaike Information Criterion (AIC).

e) Identify the best model by looking at the AICs. If possible, perform a test for the identification of the best model, specify H0 and report the p-value.

Upload your solution **here**