Problem 1
Length of hospital stay [mp]

Problem

Find out whether the duration of hospitalization varies according to the type of medical insurance the patient has.

There are two major types of medical insurance in the U.S.

- *Health Maintenance Organization (HMO)*, also called *managed care*. There is a fixed fee for each medical visit, prescription, day of hospitalization. There is a requirement to get treatment from a provider the HMO has a contract with. Most HMO's have their own hospitals and clinics.
- *Regular insurance*. The patient pays a deductible for each visit, prescription, hospitalization, expressed as a % of the real cost. The patient can freely choose a medical provider.

The concern is that the HMO's reduce the duration of hospital stay of their patients to save costs.

Overall specifications

1. Treat the response (length of stay) as a continuous variable.
2. Perform a two-sample test comparing the expected lengths of stay for the two types of insurance.
3. Consider the (classical) linear regression model for expected length of stay (possibly transformed), adjusted for insurance type and other potential confounders. Can you find a model that fits data reasonably well? Use this model and evaluate the effect of insurance type on the length of stay. Provide quantification of this effect in an interpretable way.
4. Consider generalized linear models for expected length of stay, adjusted for insurance type and other potential confounders. Try models suitable for continuous responses, assuming different distributions (normal, gamma, inverse Gaussian) and different link functions (identity, log, inverse).
5. Choose the model that, in your opinion, provides the best fit to the data. Explain and justify your choice.
6. When considering the model choice, pay attention to the size and direction of the estimated covariate effects. Do they agree with your prior expectations?
7. Use the selected model to test and evaluate the effect of insurance type on the length of stay. Interpret the regression parameter and answer the question. Is the answer the same as with a two-sample test? Is the answer the same as with the model from point 3? [Do not forget to calculate and report the confidence intervals.]

Table 1: Variable coding table

Variable Name	Variable Label	Variable Coding
los	Length of stay	Numeric
hmo	Insurance type	1 = *HMO*, 0 = *other*
agef	Age group	Factor
admit	Type of hospital admittance	Factor
died	Patient died in hospital	1 = *died*, 0 = *discharged alive*
white	Race of patient	1 = *white*, 0 = *non-white*

Requirements

Write a report (prepared by LaTeX, LibreOffice, MS Word, ...) summarizing your solution to the problems specified during the exercise classes.

More precise specification of what exactly should be (and also what should not be) included in the report will be provided during the exercise classes related to this assignment (February 28 and March 7). Pay attention to those instructions!

The report in the pdf format (file named as `Surname_Firstname_1.pdf`) and the related R script (file named as `Surname_Firstname_1.R`) have to be submitted in Moodle by ***Sunday March 12, 2023 [23:59 CET]***.

Population

The data were collected in a survey covering several hospitals within a certain area in the U.S.

Dataset

The dataset can be downloaded from
`https://www2.karlin.mff.cuni.cz/~komarek/vyuka/2022_23/nmst412/Problem_1/GLM_1_mp.RData`

The dataframe is called `mp`. It contains 1 495 rows (patients) and 6 variables.

Variable list: See Table 1.