STATS 201/8 Data Analysis
Assignment 2, NEFU, 2021

Instructions concerning this assignment:

We are providing you an R Markdown document called **`STATS201_2021_NEFU_A2.rmd`** which will have some answers already filled in. You will need to fill in and complete the rest of the document. The data files you will be using for the assignment are described in the questions and are available online. Make sure you put these datasets in the same place you put the R markdown document because it is going to look for them there. The first change you need to make to the markdown document is put your name and ID number at the top.

NOTE: When writing Executive Summaries, remember the Questions of Interest/Goals.

Question 1. [14 Marks]

We want to build a model to explain the sale price of houses using their annual city tax bill for houses in Albuquerque, New Mexico. In particular, we are interested in estimating the effect on sales price for houses which differ in city tax bills by 1% and 50%. Data was collected from a random sample of 104 houses sold in Albuquerque. The data was collated in the file `hometax.csv` and includes variables:

`Price`	the sales price of the house (in thousands of dollars).
`Tax`	the amount of annual city tax paid for the house in the year of sale.

- Look at the initial plots of the data and comment on them.
- Justify why a log-log (power) model is appropriate here.
- Fit a log-log model to the data. Check the model assumptions.
- Generate inference output required from the final model.
- Write a Method and Assumption Checks section.
- Write an Executive Summary.

Story for questions 2 and 3:

Alzheimer's Disease Neuroimaging Initiative (ADNI) is a multi-site longitudinal study for the prevention and treatment of Alzheimer's Disease (AD). It collects and utilizes various predictors of AD, including 3D brain imaging, cognitive measurements and genetic data. Magnetic resonance imaging (MRI) is a useful tool in early diagnosis of Alzheimer's disease (AD). In the ADNI study, based on MRI-derived volumes, it is of interest to study the relationship between the size of parts of the brain and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with some dementia related symptoms.

For this assignment we will look at two different aspects of the brain, the size of the hippocampus (in question 2) and the size of the lateral ventricular sub-regions of the brain (in question 3). The data we will use is a simplified set of data from the study based on the initial round of data collection. For the purposes of the assignment, treat the data as if it came from random samples of subjects. The measurements from the MRI scans don't give units, but larger values mean larger volumes. There are different data sets for the two questions as not all subjects had both regions of the brain measured.

The data collected was:

`Hippocampus`	the volume of the hippocampus of the brain (measured by MRI).
`Ventricles`	the volume of the lateral ventricular sub-regions of the brain (measured by MRI).
`Age`	the age of the subject at the time of the scan (in years).
`AD`	whether or not the subject had dementia (`D` for dementia or `H` for healthy).

Question 2. [18 Marks]

For this question, it is of interest to study the relationship between hippocampus size and age. In particular, we are interested in whether the relationship differs between healthy individuals and individuals with dementia related symptoms. The data is stored in `Hippocampus.csv`.

- Look at the initial plots of the data and comment on them.
- Fit a model to the data.
  - Check the model assumptions.
  - Change the model and repeat checks as needed. You may have to do this more than once. (**Note:** only drop ONE variable at a time.)
- Generate inference output required from the final model.
- Plot the data with your appropriate model superimposed over it.
- Write a Method and Assumption Checks section.
- Write an Executive Summary. (See Assignment 1 notes for more information.)

Question 3. [12 Marks]

For this question, it is of interest to study the relationship between ventricles size and age. In particular, we are interested in whether the relationship varies between healthy individuals and individuals with dementia related symptoms. The data is stored in `Ventricles.csv`.

For this question we have given you all relevant R output. You do not need any further output.

- Look at the plots of the data and comment on them.
- In terms of slopes and/or intercepts, explain what the coefficient of `Age:ADH` is estimating.
- For each of the following, either write a sentence interpreting a confidence interval to estimate the requested information or state why we cannot answer this from the R-output given:
  - in general, the difference in size of ventricles between healthy people and those exhibiting dementia symptoms.
  - the effect on the size of ventricles for each additional years aging on healthy people.
  - the effect on the size of ventricles for each additional years aging on people exhibiting dementia symptoms.
- Looking at the plot with the model superimposed, describe what seems to be happening.