Problem Set 3: Linear & Logistic Regression

The National Health and Nutrition Examination Survey (NHANES) is a survey conducted annually by the US Centers for Disease Control and Prevention (CDC) involving about 10,000 Americans aged 21 and older. In this problem set, you will use a data set obtained by randomly sampling 500 cases from the 10,000 NHANES survey participants. This new data set is called **nhanes_samp_adult_500** and is available online at https://www.openintro.org/data/csv/nhanes.samp.adult.500.csv.

Importing the Data

Up to this point in this course we have been using data frames that we either created in R or loaded from some package. In this problem set, you will use data from an online source. You will import this data in your workspace as a tibble. Tibbles are same as data frames but with several advantages such as

- They load faster than data frames.
- Allow use of non-standard variable names (i.e., variables can have sapces and can start with a number)
- Allow you to have columns as lists.

To import the data set into your work space, use the command `read_csv()` which imports the data as a tibble. So you know, the command `read.csv()` loads data as a data frame. As noted earlier, the data is stored on a website and we have the link (URL). Use the command below to load the data into your workspace. Note that we are naming the data **nhanes_samp_adult_500**.

If you data loads properly, it is highly recommended that you take some time to study it so you understand its context and variables. Doing this will make your analyses easier and interpretations will be meaningful.

Problem 1 (Linear Regression)

Several studies point out that maintaining a healthy weight is an important part of well being. The goal in this problem is to identify, via linear regression, the most important factors that can determine someone's weight. Here is a list of the variables you will consider in this problem:

PhysActive: A categorical variable coded as **Yes** if the participant does moderate or vigorous-intensity sports, fitness, or recreational activities, and **No** if otherwise.

`Height`: A numerical variable measuring height in cm.

`SleepHrsNight`: A numerical variable representing the average sleep time per night in hours.

`AgeMonths`: A numerical variable giving someone's age in months.

`MaritalStatus`: A categorical variable with levels Divorced, LivePartner, Married, NeverMarried, Separated, and Widowed.

`Education`: A categorical variable with levels 8th Grade 9 - 11th Grade, High School, Some College, and College Grad.

`Gender`: a categorical variable with levels female, male.

As we have learned in this course, having a lot of predictors does not necessarily mean you have a great model. You are going to start by fitting a model with all predictors listed above and then perform model selection (backward elimination) to come up with the ''best model'' for predicting weight.

a) Create a model for predicting weight based on all variables in the list above. Based on this model, what factor has the most influence on weight? What factor has the least influence?

b) What is the R-squared value of the model in part (a) above? Interpret the value in the context of the data.

c) Perform model selection (backward elimination) to come up with the ''best model'' for predicting weight. Write down the equation of the model. You do not need to show all the steps of the backward elimination process but you should explain these steps here in brief.

d) Interpret the intercept and all slopes in the final model from part (c).

e) Examine the first case in the data set. Does your model underestimate or overestimate the weight of the individual? By how much?

Problem 2 (Logistic Regression)

Many studies show that being overweight increases one's risk of heart disease, stroke, type 2 diabetes, and other conditions. A common measure for determining whether someone is overweight is the Body Mass Index (BMI). BMI is a measure that accounts for both weight and height of an individual and is computed as follows:

$$BMI = \frac{Weight}{(Height)^2}$$

Where `weight` is measured in **Kilograms** and `Height` is measured in **meters**.

In this problem, you will use the same data set used in problem 1 to run a model for the likelihood of diabetes based on `BMI`, `Age`, and `Gender`. Note that the variable `Diabetes` is categorical with levels **Yes** and **No**.

a) Create a regression model (name it `Diabetes_Model`) for predicting whether someone has diabetes or not based on their BMI, Age, and Gender. (Note: You may have to first mutate the Diabetes variable to use 1's and 0's. Check out the labs.

b) Write down the equation of the model and interpret the coefficient for `Gender` and `BMI`.

c) James (male) is 57 years old, weighs 87.9 Kg, and is 156 cm tall. How would the `Diabetes_Model` classify this individual (Diabetes or No diabetes)?

d) Create a confusion matrix to determine the accuracy of the `Diabetes_Model`.