Regression Lab

You have been provided with three medium-sized data sets. A description of each data set is provided below. Your task, using Excel, is to:

1. Decide whether regression is appropriate, and if so, decide on what type of regression to use. Explain your decision
2. Implement the appropriate regression:
   1. Report the model parameters
   2. Write out the full mathematical model. E.g. in the case of linear regression, in the form Y=mx+c
3. Are the models any good?
   1. How would you assess the ‘goodness’ of a linear regression model (info in the lecture notes)?
   2. How would you assess the ‘goodness’ of a logistic regression model. This was not covered in the lectures. You might want to think about:
      1. If the model predicts a ‘true’ value, but the real value is ‘false’ – this is a false positive.
      2. If the model predicts a ‘false’ value, but the real value is ‘true’ – this is a false negative.
      3. How can you use these two metrics to help assess the model? Which is more important, and why?

The expected outputs of this lab are:

1. Excel spreadsheet(s) showing appropriate regressions for the data sets.
2. Short answer questions (a few sentences) to answer Q2 and Q3

# Extension

Repeat Q2 using R. You will need to use the glm function. GLM (short for generalized linear model), as the name suggests, is a way of unifying linear and logistic regression (as well as other types of regression) into one model. Further information on the glm function can be found within R’s help pages.

## Data Description

**Regression1.csv**: This dataset contains three variables, a unique participant ID, their level of education in years – this should be a value between 10 and 20, and their spending per month in GBP. This data comes from a generative model – i.e. no participants were actually used, and there is a definitive ‘right’ answer.

**Regression2.csv (this is the trickiest one – you may want to do Regression3 first!) –** The data show, for a set of medical appointments, whether the patient turned up to the appointment, and a set of information about the patient and the appointment:

**Regression3.csv –** The data (X1 -> X5) are by city:

X1 = death rate per 1000 residents

X2 = doctor availability per 100,000 residents

X3 = hospital availability per 100,000 residents

X4 = annual per capita income in thousands of dollars

X5 = population density (people per square mile)

Reference *Life in America’s small cities – G.S. Thomas*