# MSc/PGDip in Data Analytics
# Statistics for Data Analytics CA1 (35%)

**Publication Date: Friday, 24th February 2023**
**Submission Date: Sunday, 26th March 2023**

## Outline

Your task is to investigate socio-economic factors influencing the cancer mortality in the US. For this purpose we have sourced a dataset with **average data** for most of the 3000+ US counties. Note that depending on the state, the size of counties varies a lot. The dataset is provided in the file 'cancer.csv'.

This dataset includes for each county the **name** of the county and its **population size** some summary medical data and socio-economic data.

The medical data are:

- the **incidence rate**: the number of newly diagnosed cancer cases per 100,000 of population
- the **death rate**: the number of death by cancer per 100,000 of population. Note: Some of the death occur in the same year as the diagnosis, some may occur many years later. Occasionally the death rate can therefore be higher than the incidence rate.

The available socio-economic data are:

- **Income related**: the median income, the percentage of unemployed, and the poverty rate
- **Age related**: Median Age across the population, and for male and female separately
- **Household related**: Average Household Size and percentage of Married Households
- **Education related**: Percentage of the highest educational level attained (No High School / High School / Bachelor Degree) in the age groups 18-24 and over 25.
- **Health Insurance related**: Percentage of Private Insurance, Private Insurance paid by Employer, Public Insurance and Public Insurance Only, and
- **Race related**: percentage of White/Black/Asian/Other.

Your task is to find (from your point of view) the best suited multiple linear regression model for the expected cancer related death rate per county using the incidence rate of cancer and the available socio-economic data for the county. Note that the model for the death rate you can't use the mortality rate and vice versa. However it is your choice if you build a model for the mortality rate (and derive the death rate by multiplying the mortality rate with the incidence rate), or if you build a model for the death rate (and derive the mortality rate by dividing the death rate by the incidence rate).

## Submission

The submission consists of two parts: a report of up to 6 pages in .pdf format using the IEEE conference template and a supporting code file.

In your report you should:

- Use descriptive statistics and appropriate visualisations to enhance understanding of the variables in the dataset.
- Describe the model building steps you undertook in the process of arriving at your final regression model. The rationale for rejecting intermediate models should be explained clearly and details provided on the rationale the for choosing predictors, transformations undertaken, treatment of outliers, etc.
- Provide details on diagnostics undertaken to verify that the Gauss Markov and other relevant assumptions of multiple regression have been satisfied.
- Provide a succinct summary of the parameters of your final model and details of model performance and fit.

The supporting code file should contain material required to reproduce the results of your report:

- If you used **Jupyter Notebook**, submit the notebook file with all the output produced included. Make sure that it works using the "Restart Kernel and run all" option. For any computer generated graphics you used in the report, insert in the Jupyter notebook a comment referring to the figure number or caption.
- If you used **R Studio** or similar, submit the source file and make sure that one can run the code sequentially. For any computer generated graphics you used in the report, insert in the source code a comment referring to the figure number or caption.
- f you used a software package like **SPSS**, provide a .pdf document with a detailed description of the steps you have taken to obtain the results in your report.

## Academic Integrity

- By submitting your work on Moodle you declare that this is your own work.
- Any material created by others must be properly referenced. Verbatim text copies should be included in quotes.
- Figures not created by yourself should include an acknowledgement detailing the name(s) of the creator(s) and proper references.
- Code and figures copied from class material or other sources should be clearly marked as such and properly referenced. In particular it should not be (directly or implicitly) claimed as your own. Instead a comment should be included in the source code indicating where you obtained it from.
- Students are strongly advised to familiarise themselves with the Guide to Academic Integrity. All submissions will be electronically screened for evidence of academic misconduct, e.g. plagiarism, collusion and misrepresentation.

## Marking Scheme

| Component | Indicative Breakdown | |
|---|---|---|
| Introduction (10%) | Outlining the general foundations of the multiple linear regression with a focus on aspects relevant for your project | 10% |
| Explorative Data Analysis (20%) | Descriptive Statistics | 10% |
| | Visualisation and discussion of out of range or N/A values | 10% |
| Model Development (40%) | Description of the methodology used | 5% |
| | Considering the population size | 5% |
| | Use of Transformation on independent variables | 5% |
| | Handling of outliers | 5% |
| | Diagnostic of intermediate models | 5% |
| | Evaluation of the proposed final model | 10% |
| | Interpretation of the parameters of the final model | 5% |
| Code (20%) | The code is executable and allows reproduction of results | 20% |
| Presentation(10%) | Writing style, use of template, and adherence to page limit | 10% |
| Total | | 100% |