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

Publication Date: **Tuesday, 27th February 2024**
Submission Date: **Monday, 25th March 2024**

Outline

Your task is to find the best suited multiple linear regression model for the weekly earnings in the US. For this purpose we have sourced[1] a dataset with **sample data** from just over 1.3 Million working people across all industries and regions of the US which is representative for about 140 Million working people within the age group from 17 to 65 years.

This dataset **`microwage.csv`** (162.2MB) includes the following independent variables:

Geographic data:

- *region*: integer encoding of the census region. The encoding and names of the regions can be found in the additional file **`regions.csv`**.
- *statefip*: IT standard encoding for US states. The encoding and names of the states can be found in the additional file **`fips.csv`**.
- *metaread*: code for the metropolitan area (if any). The encodings and names of the metropolitan areas can be found in the additional file **`metareas.csv`**.
- *puma*: public use micro area, originally introduced for the 1990 census, split the states in areas of at least 100,000 inhabitants.

Individual data:

- *age*: age of a person in years. For this study the age is limited to 17..65 years.
- *edyrs*: number of years in education
- *female*: binary encoding of legal gender (0=male, 1=female)
- *race_nonwhite*: binary encoding of race (0=white Caucasian, 1=other).
- *perwt*: indicates for how many people across the whole US the individual is representative. A very small value indicates that the individuum described in that line is probably a rare case, a large value indicates that the individuum is probably a common case.

Professional data:

- *industry*: industry in which the person works using the encoding of industry (1990) standard. A list of encodings and names of the industries can be found in the additional file **`industries.csv`**.

[1] https://economics.mit.edu/people/faculty/david-h-autor/data-archive

- *expyrs*: number of potential years of experience on the job
- mutually exclusive job classification:
  - *occ_managprof*: 1=managerial or professional role
  - *occ_techsalad*: 1=technical, salaried
  - *occ_service*: 1=service industry
  - *occ_farm*: 1=farming
  - *occ_operator*: 1=machine operator
  - *occ_product*: 1=general production
- additional job classification:
  - *occ_service_broad*: 1=new, broader sense of service
  - *occ_service_np*: 1=non-professional services

and the dependent variable:

- *wkwage*: average weekly wage, taking into account the number of hours worked per week, but only considering weeks worked.

As part of the exploratory data analysis investigate the distribution of each of the individual variables. Use suitable visualisation to decide about possible transformations. Further use visualisation and suitable statistical tests to investigate the pairwise relations between each of the independent variables and the dependent variable.

For the dependent variable you may need to consider a non-linear model taking into account Mincer's earnings function or a modern refinement of the same.

The additional files **`regions.csv`**, **`fips.csv`**, **`metareas.csv`**, and **`industries.csv`** are for your information only. There is no need to process the additional files. They allow you to interpret differences between locations and industries, and to cross-check if the differences you may have found are actually sensible.

Submission

The submission consists of two parts:

1. a report of up to 6 pages in .pdf format using the IEEE conference template and
2. any supporting files compressed into a single .zip file.

In your report you should:

- Use descriptive statistics and appropriate visualisations to demonstrate an understanding of the variables in the dataset.
- Document the relationship between each of the independent variables and the dependent variable.
- 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 should be provided on the rationale for the chosen predictors, transformations undertaken, treatment of outliers, etc.
- Provide details on the diagnostics undertaken to verify that the relevant assumptions of multiple regression have been satisfied, and
- Provide a succinct summary of the parameters of your final model, details of model performance and fit.

The supporting file should contain intermediate versions of your final report (numbered in sequence) and the material required to reproduce the final results of your report:

- The intermediate versions of your report document your working process. It is your defence in case your final report is flagged as potentially written by AI tools. It is ok to use such tools to improve your use of language as long as you provide the raw version of your submission as you have written it yourself. Make sure that the intermediate versions of your text still have the original timestamps when you created them.
- If you used **Jupyter Notebook**, submit the notebook file with all the output included. Make sure that it works by using the "Restart Kernel and run all" option and then save the file. 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.
- If 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. Make sure that the .pdf file contains screenshots of the relevant interactions.
- If you use code snippets from the classes or other sources, make sure to include in the code a reference to the source and a comment clarifying if and what parts of the code you have modified yourself.

Note:

1. Do not attempt to upload any data files to Moodle. The file upload capacity in Turn-it-in is limited to 100MB (uncompressed). Remember: The part two of the submission is a single .zip file containing all supporting file. You need to create a folder structure with all your file and upload the compressed version of that folder.
2. The data file provided is too big to be processed in Microsoft Excel. If you attempt to do so, you will lose a substantial amount of data.

Academic Integrity

- By submitting your work on Moodle you declare that this is your own work.
- Any material created by others (human or AI) 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 methods applied in your project	10%
Explorative Data Analysis (20%)	Descriptive Statistics of dependent and independent variables	5%
	Visualisation and discussion of out of range or N/A values	5%
	Documentation of the relation between individual independent variables and the dependent variable	10%
Model Development (40%)	Description of the methodology and the class of models used	5%
	Use of Transformation on dependent or independent variables	5%
	Handling of outliers and/or weight function	5%
	Diagnostic of intermediate models and process of incremental improvement of models	10%
	Evaluation of the proposed final model	10%
	Interpretation of the parameters of the final model	5%
Code (20%)	The early versions of the report demonstrate the working process. The code is executable and allows reproduction of results, or alternatively, the documentation of the development process allows the reproduction of the result	20%
Presentation(10%)	Writing style, use of template, and adherence to page limit	10%
Total		**100%**