Assignment 1

Biomedical Data Science (MATH11174), 22/23, Semester 2

March 9, 2023

Due on Thursday, 9th of March 2023, 5:00pm

> **!** Pay Attention
>
> The assignment is marked out of 100 points, and will contribute to ***20%*** of your final mark. The aim of this assignment is to produce a precise report in biomedical studies with the help of statistics and machine learning. Please complete this assignment using **Quarto/Rmarkdown file and render/knit this document only in PDF format** and submit using the **gradescope link on Learn**. You can simply click render on the top left of Rstudio (`Ctrl+Shift+K`). If you cannot render/knit to PDF directly, open **Terminal** in your RStudio (`Alt+Shift+R`) and type `quarto tools install tinytex`, otherwise please follow this link. If you have any code that does not run you will not be able to render nor knit the document so comment it as you might still get some grades for partial code.
> **Clear and reusable code will be rewarded**. Codes without proper indentation, choice of variable identifiers, **comments**, error checking, etc will be penalised. An initial code chunk is provided after each subquestion but **create as many chunks as you feel is necessary** to make a clear report. Add plain text explanations in between the chunks when required to make it easier to follow your code and reasoning. Ensure that all answers containing multiple values should be presented and formatted with `kable()` and `kable_styling()` or using Markdown syntax. All plots must be displayed with clear title, label and legend.

Problem 1 (25 points)

Files `longegfr1.csv` and `longegfr2.csv` (available on Assessment > Assignment 1) contain information regarding a longitudinal dataset containing records on 250 patients. For each

subject, eGFR (**estimated glomerular filtration rate, a measure of kidney function**) was collected at irregularly spaced time points: variable `fu.years` contains the follow-up time (that is, the distance from baseline to the date when each eGFR measurement was taken, expressed in years).

Problem 1.a (4 points)

- Convert the files to data table format and merge in an appropriate way into a single data table.
- Order the observations according to subject identifier and follow-up time.
- Print first 10 values of the new dataset using `head()`.

```
#Answer in this chunk
```

Problem 1.b (6 points)

- Compute the average eGFR and length of follow-up for each patient.
- Print first 10 values of the new dataset using `head()`.
- Tabulate the number of patients with average eGFR in the following ranges: $(0, 15]$, $(15, 30]$, $(30, 60]$, $(60, 90]$, $(90, \texttt{max(eGFR)})$.
- Count and report the number of patients with missing average eGFR.

```
#Answer in this chunk
```

Problem 1.c (6 points)

- For patients with average eGFR in the $(90, \texttt{max(eGFR)})$ range, collect their identifier, sex, age at baseline, average eGFR, time of last eGFR reading and number of eGFR measurements taken in a data table.
- Print the summary of the new dataset.

```
#Answer in this chunk
```

Problem 1.d (9 points)

For patients 3, 37, 162 and 223:

- Plot the patient's eGFR measurements as a function of time.
- Fit a linear regression model and add the regression line to the plot.

- Report the 95% confidence interval for the regression coefficients of the fitted model.
- Using a different colour, plot a second regression line computed after removing the extreme eGFR values (one each of the highest and the lowest value).

(All plots should be displayed in the same figure. The plots should be appropriately labelled and the results should be accompanied by some explanation as you would communicate it to a colleague with a medical background with a very little statistical knowledge.)

```
#Answer in this chunk
```

Problem 2 (25 points)

The MDRD4 and CKD-EPI equations are two different ways of estimating the glomerular filtration rate (eGFR) in adults:

$$\texttt{MDRD4} = 175 \times (\texttt{SCR})^{-1.154} \times \texttt{AGE}^{-0.203}[\times 0.742 \text{ if female}][\times 1.212 \text{ if black}]$$

, and

$$\texttt{CKD-EPI} = 141 \times \min(\texttt{SCR}/\kappa, 1)^{\alpha} \times \max(\texttt{SCR}/\kappa, 1)^{-1.209} \times 0.993^{\texttt{AGE}}[\times 1.018 \text{ if female}][\times 1.159 \text{ if black}]$$

, where:

- `SCR` is serum creatinine (in mg/dL)
- κ is 0.7 for females and 0.9 for males
- α is -0.329 for females and -0.411 for males

Problem 2.a (7 points)

For the `scr.csv` dataset,

- Examine a summary of the distribution of serum creatinine and report the inter-quartile range.
- If you suspect that some serum creatinine values may have been reported in µmol/L convert them to mg/dL by dividing by 88.42.
- Justify your choice of values to convert and examine the distribution of serum creatinine following any changes you have made.

```
#Answer in this chunk
```

Problem 2.b (11 points)

- Compute the eGFR according to the two equations using the newly converted `SCR` values.
- Report (rounded to the second decimal place) mean and standard deviation of the two eGFR vectors and their Pearson correlation coefficient.
- Report the same quantities according to strata of MDRD4 eGFR: $(0 - 60)$, $(60 - 90)$ and (> 90).
- Print first 15 values for both datasets using `head()`.

```
#Answer in this chunk
```

Problem 2.c (7 points)

- Produce a scatter plot of the two eGFR vectors, and add vertical and horizontal lines (i.e.) corresponding to median, first and third quantiles.
- Is the relationship between the two eGFR equations linear? Justify your answer.

```
#Answer in this chunk
```

Problem 3 (31 points)

You have been provided with electronic health record data from a study cohort. Three CSV (Comma Separated Variable) files are provided on learn.

The first file is a cohort description file `cohort.csv` file with fields:

- `id` = study identifier
- `yob` = year of birth
- `age` = age at measurement
- `bp` = systolic blood pressure
- `albumin` = last known albuminuric status (categorical)
- `diabetes` = diabetes status

The second file `lab1.csv` is provided by a laboratory after measuring various biochemistry levels in the cohort blood samples. Notice that a separate lab identifier is used to anonymise results from the cohort. The year of birth is also provided as a check that the year of birth aligns between the two merged sets.

- `LABID` = lab identifier
- `yob` = year of birth
- `urea` = blood urea
- `creatinine` = serum creatinine
- `glucose` = random blood glucose

To link the two data files together, a third linker file `linker.csv` is provided. The linker file includes a `LABID` identifier and the corresponding cohort `id` for each person in the cohort.

Problem 3.a (6 points)

- Using all three files provided on learn, load and merge to create a single data table based dataset `cohort.dt`. This will be used in your analysis.
- Perform assertion checks to ensure that all identifiers in `cohort.csv` have been accounted for in the final table and that any validation fields are consistent between sets.
- After the checks are complete, drop the identifier that originated from `lab1.csv` dataset `LABID`.
- Ensure that a single `yob` field remains and rename it to `yob`.
- Ensure that the `albumin` field is converted to a factor and the ordering of the factor is `1="normo"`, `2="micro"`, `3="macro"`.
- Print first 10 values of the new dataset using `head()`.

```
#Answer in this chunk
```

Problem 3.b (10 points)

- Create a copy of the dataset where you will impute all missing values.
- Update any missing age fields using the year of birth.
- Perform mean imputation for all other continuous variables by writing a single function called `impute.to.mean()` and impute to mean, impute any categorical variable to the mode.
- Print first 15 values of the new dataset using `head()`.
- Compare each distribution of the imputed and non-imputed variables and decide which ones to keep for further analysis. Justify your answer.

```
#Answer in this chunk
```

Problem 3.c (6 points)

- Plot a single figure containing boxplots of potential predictors for `diabetes` grouped by cases and controls. (Hint : `par(mfrow=c(1,5))`)
- Use these to decide which predictors to keep for future analysis.
- For any categorical variables create a table instead. Justify your answers.

```
#Answer in this chunk
```

Problem 3.d (9 points)

- Use your findings from the previous exercise and fit an appropriate model of `diabetes` with two predictors.
- Print a summary and explain the results as you would communicate it to a colleague with a medical background with a very little statistical knowledge.

```
#Answer in this chunk
```

Problem 4 (19 points)

Problem 4.a. (9 points)

- Add a third predictor to the final model from **problem 3**, perform a likelihood ratio test to compare both models and report the p-value for the test.
- Is there any support for the additional term?
- Plot a ROC curve for both models and report the AUC, explain the results as you would communicate it to a colleague with a medical background with a very little statistical knowledge.
- Print a summary and explain the results as you would communicate it to a colleague with a medical background with a very little statistical knowledge.

```
#Answer in this chunk
```

Problem 4.b (10 points)

- Perform 10-folds cross validation for your chosen model based on the above answers.
- Report the mean cross-validated AUCs in 3 significant figures.

```
#Answer in this chunk
```