Assignment 1

STAT462 2024-S2

Before you start, here's some guidance on what is being expected of you.

Marking Rubric

On a high level:

- C grade (50%-65%): You have attempted the assignment, but in a limited way that might not show demonstrate full understanding of the concepts involved.
- B grade: (65%-79%) You have done what you were asked to do, and you did it correctly
- A grade (80%-100%): In addition to satisfying the requirements for a B grade, you have done something beyond, like:
    - exploring an additional avenue of research,
    - including work beyond the lecture content,
    - discussing current debates in the subject area,
    - demonstrating extraordinary investigative research or
    - writing an exceptional report,
    - …

You are usually not expected to write any long essays, so keep it brief, legible, and understandable. Support your arguments with figures, mathematical arguments, or references, where suitable.

In slightly more detail: Your assignment will be marked according to the following criteria. The relative weightings are only an approximate guide and will vary between questions.

Skills	Expectation for a C grade	(additional) Expectation for a B grade	(additional) Expectation for an A grade	approximate relative weighting
Communicating clearly via text	Text is in understandable English. For most questions, a few short sentences will be totally adequate, unless stated otherwise.	Uses adequate technical jargon, and concise language.	"Generic garbage" as sometimes produced by GenAI will lead to detractions here.	**necessary prerequisite.** Anything else cannot be marked if it is not communicated clear enough

R Markdown and format of submission	Submission is a single file in either .html or .pdf format. File is generated by R Markdown.	—	—	**necessary prerequisite.**
R programming skills	Syntax correct, code runs without errors.	Code does what it is intended to do. Code is readable and uses code comments and suitable naming conventions to clarify what is being done. `dplyr`, `ggplot`, and statistical learning modules are employed in a suitable way.	Computations are reasonably efficient.	25%
Knowing and applying statistical learning algorithms	Basic statistical learning task is addressed.	Correct choice (within constraints of admissible modules) and application of algorithm.	—	30%
Interpreting and evaluating algorithm results and output	Some interpretation of the findings is provided, even if not correct.	Algorithm results are (when required) put into context, correctly explained, and implications as relevant for the assignment are correctly identified.	—	25%
Communicating clearly via figures	An attempt at visualisation (if required) is made.	Axes are labelled, figure is suitably scaled, intention is clearly communicated. Interpretation of the figure is given in either an expressive caption or in the accompanying full text.	—	10%
Extra Effort	—	A good report will use the models and statistical methods to the extent developed in class.	A very good assignment (e.g., A- or better) will extend to something not covered, and/or show reflection beyond the	10%

narrow research question. Sometimes, question parts marked with “extra effort” give an indication for possible avenues to extend your work.

Braking distance

In this question, do not use the `lm` function, or a module that provides an implementation of k-NN. You are allowed to use elementary statistical objects like `mean`, `var`, etc.

We will be predicting the distance that a car takes to get from driving at a certain speed to a full stop. Note that the dataset we are using is from 1930, so this might be a bit outdated. Units are miles per hour (for `speed`) and distance in feet (for `dist`).

1. Load the dataset `braking.csv`. In order to make the results more legible for a New Zealand audience, convert the units into metric, i.e. kilometers per hour (for `speed`) and meters (for `dist`). Split the dataset randomly into a training set (80%) and test set (20%).
2. Conduct a simple linear regression (without using `lm`): You will need to compute the slope and intercept parameters.
    a. In this linear regression model: If you increase your speed by 5km/h, how many more meters of braking distance do you expect?
    b. How much of the variation in the data can be explained by your linear regression model?
    c. Is `speed` a significant predictor for `dist` at the 95% confidence level?
    d. Using your linear regression model, predict the braking distance for a car going at 30 km/h, and include an 80% prediction interval.
3. Fit a k-NN model to the training set. Predict the braking distance for a car going at 30km/h using this model.
4. Visualise both models in a graph that also shows the dataset. Compare the performance of both the k-NN model and the linear regression model on the test set by computing both their test set MSE. Which one is performing better?

Filipino household income

In this question, you are allowed to use `lm`.

Load the dataset `income.csv`, which contains more than 40000 entries of data about households in the Philippines. There is a lot to learn in this dataset, but in this assignment we are going to focus on predicting household income (`Total.Household.Income`) from the number of underage children living in the household (`Members.with.age.5...17.years.old`).

1. For this assignment we will just be looking at the two features mentioned, so feel free to remove unnecessary columns and rename the feature names if that makes your life easier (I will just refer to these two features by `income` and `children` from now on).
2. Split the dataset randomly into a training set (80%) and test set (20%).
3. Perform a linear regression of type `income = b0 + b1 * children` for analysing the influence on `children` on `income`.
    a. What is the specific form of the affine-linear model, i.e. what are `b0` and `b1`.?
    b. What is the predicted mean income of a household with n children, for $n \in \{0, 1, \dots, 8\}$? What are the associated 90% prediction intervals? Summarise all of this in a table.
    c. Using your test set, check how many percent of datapoints lie within the 90% prediction intervals.
4. Do all steps of part 3. again, but this time you will be predicting `log_income = log(income)` instead of `income`.
5. *Opportunity for showing extra effort:* Reflect on your results. What did you observe, and what do you think are the reasons for that?

Predicting Possum age

Note: In this question, you are allowed to use `lm`. Do not use any "sophisticated" statistical learning modules apart from `lm` for this question, i.e. don't use `stepAIC`, `ffs`, `regsubsets`, or other implementations of forward feature selection. You will have to implement this yourself.

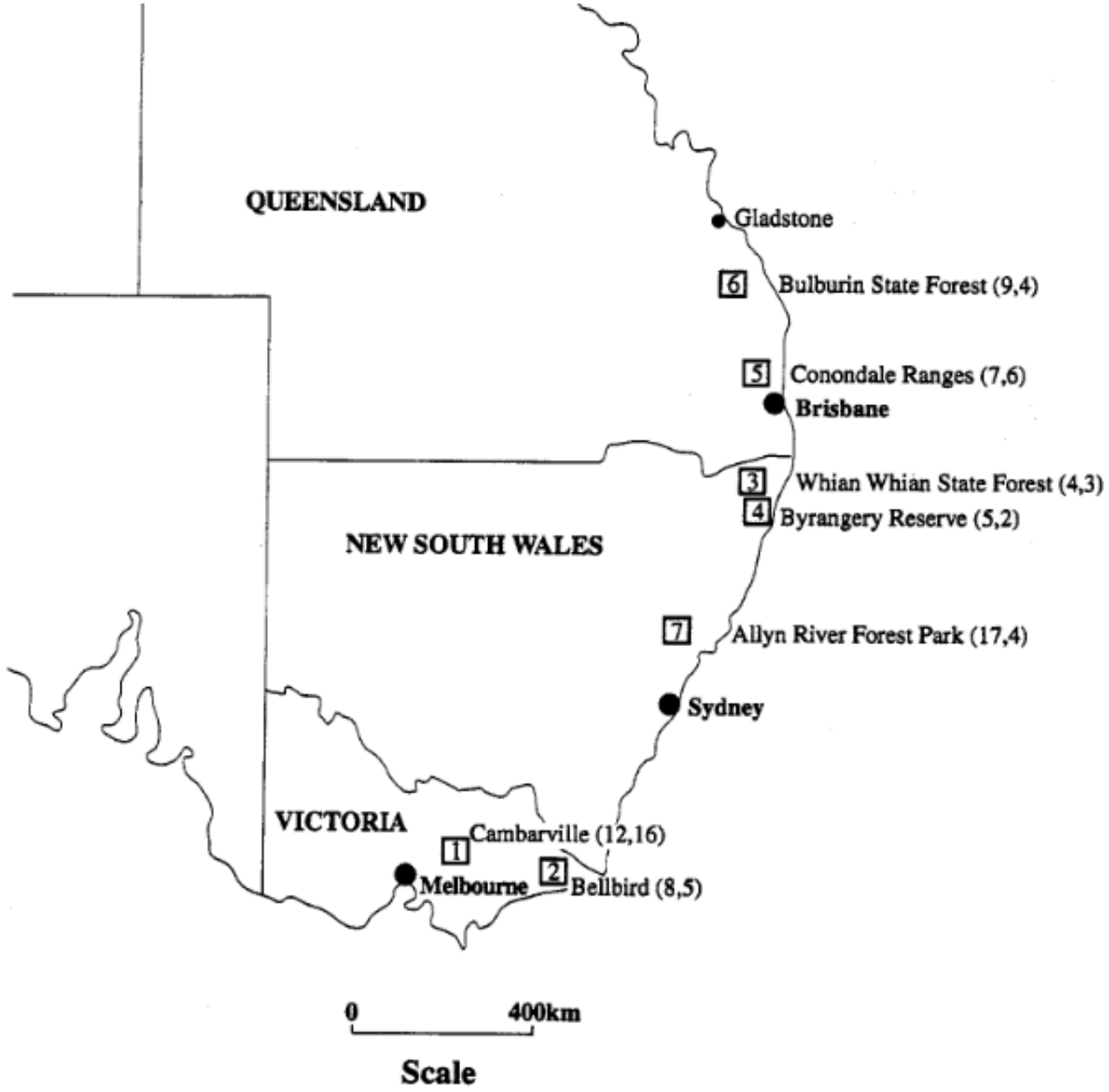
We will look at a dataset of trapped and measured possums in 7 different sites in Australia (spread over Queensland, New South Wales, and Victoria, see figure below). Our goal will be to predict the age of a possum from its body measurements.

The features are as follows (we are going to ignore the index variable `case` and the `Pop` variable).

Features

Feature name	Explanation
`site`	index of geographical trapping site
`sex`	male or female

age	age in years
hdlngth	length of head
skullw	length of skull
totlngth	total body length
taill	tail length
footlgth	foot length
earconch	length of the ear conch
eye	eye size
chest	chest girth
belly	belly girth



Location of the seven sites where possums were trapped. Numbers in brackets record number of animals (male,female)

1. Load the dataset `possums.csv`. Using the `ggplot` module, plot `age` (on the y-axis) against `totlngth`. What does this tell you about the relationship between total length and age?

2. Data clean up and preparation:

    a. remove the `case` and `Pop` columns from the dataset since they will not be used for the regression problem.

    b. One-hot encode the `site` variable into six binary variables `site1`, `site2`, ..., `site6`. Why don't we need a seventh variable `site7`? Remove the `site` variable afterwards.

    c. Turn the `sex` variable into a 0-1 variable (instead of containing `"m"` and `"f"`). The best way to not get confused is to turn this into a new feature `female` which is `1` if the animal is female, and `0` otherwise. Remove the feature `sex` afterwards.

    d. Split the dataset into a training, a test, and a validation set (roughly 80%-10%-10%).

3. This leaves us with a set of columns that we are going to use for predicting `age`. How many possible multiple linear regression models can we fit? (For example, we could just use the “empty” model predicting always the mean age, or we could just look at `totlngth`, etc.)
4. Perform stepwise forward feature selection to find a suitable subset of features for a multiple linear regression model. Document your approach, and provide short justifications and helpful graphics to illustrate the decisions you are making.
5. Summarising your work in the previous step, which are the most important predictors for `age`, in your opinion?
6. Train a multiple linear regression model on the subset of features you have picked in the previous step. Compute the mean square error.
7. *Opportunity for showing extra effort:* Formulate a meaningful and interesting research question on this dataset, explore it thoroughly, and document your process and findings in an excellent way.