ICT513 *Data Analytics*
Semester 2, 2024
Project

Due: Friday, 31 October 2024, 11:00 PM

The ICT513 project can be completed in groups of one to three. Students are strongly encouraged to work as a pair so that they have another person to share ideas with as they go through the process of considering the data, formulating research questions, analysis and communication of the project in the final report. Either way you must select yourself into a project group prior to submission and submit one copy of the project. Those working in pairs must also complete the group evaluation form and include it with your report (preferably as the final page), this page is over and above the 10 page limit (see below).

The ICT513 project includes a thorough data analysis related to your choice of **one** of two datasets. For your chosen dataset, you can investigate two questions using two distinct methods covered in class or prepare your two own questions and answer using two distinct methods covered in class. **You should present your project in the last session of the semester and also submit your report through the Project link on the LMS webpage**.

While this is a data analysis project and not a research project (*i.e.*, you are not expected to do an in-depth investigation of the particular topic), you should imagine that this report is being written for publication, so the style of the report should reflect that degree of formality and not be written as an assignment. Additionally, if considering linear regression models, although it is important that you use diagnostic plots to determine an appropriate model, these need not be included in the main body of the report. Rather you can simply make mention of any transformations made to ensure greater compliance with the assumptions of linear regression. Finally, `R` code and summary output should not be pasted into the document, but instead relevant results should be presented in nicely formatted tables and/or visualisations. `R` code is to be submitted in a separate script file. To assist in your writing of the report, an example report following a similar format is presented under the "Projects" link on the unit webpage.

Final report

The report must not exceed 10 pages in length and should include the following sections:

1. **Introduction:** provides background to the research, clearly lays out the research questions of interest, and motivates why these questions are of interest.

2. **Methods and Analysis:** provides a description of the data and the statistical analyses that will be performed. For example if a linear regression is to be carried out, this section should provide an explanation of and motivation for the variables you have included in your model.

3. **Results:** provides a thorough description of the results of the analyses you described in the previous section. This section should first include some carefully chosen descriptive statistics (statistics, tables, data visualisations) that are useful in describing the data and providing a glimpse of what you might expect from your statistical analyses. You must explain what you can derive from these. Then you report the results of your application of statistical techniques; include tables and/or data visualisations with relevant output. If analyses are carried out that involve the estimation of parameters, this should include an interpretation of the parameters for the variables of interest. Any requirements/assumptions needed to perform the analyses carried out must be addressed. Additionally, if outliers are present, you may want to determine whether or not the results change by removing these outliers. Finally, remember that even if a result is statistically significant, this does not always mean that it is practically significant.

4. **Discussion:** concisely summarizes your conclusions to the research questions of interest as well as any supplementary analyses carried out. This section also should include a brief description of the limitations of your analyses as well as other research questions that may be worth exploring in response to any exploratory analyses you have carried out.

5. **R code (not included in the 10 page limit):** your R script file, appropriately named and tested. It should be able to be run error-free from start to finish and should include all R code used to produce your results.

6. **The data (not included in the 10 page limit).**

We take seriously academic integrity, and reports will be closely scrutinised for plagiarism or collusion. If you are unsure of what constitutes plagiarism or collusion, carefully read:

https://www.murdoch.edu.au/TNE/Student-Information/Academic-Integrity/

If in doubt, you should discuss with the unit coordinator.

1. **Breast Cancer Prognosis**

   Breast cancer is one of the most common forms of cancer, affecting up to 1 in 7 females in Australia alone. In recent years, advances in detection and treatment of breast cancer have improved survival rates significantly. The average 5-year survival rate for women with invasive breast cancer is 90%, however this drops to only 77% for patients with so-called triple negative breast cancer (patients have tested negative for estrogen and progesterone receptors and excess HER2 protein). Researchers are interested in determining which patients have the more agressive form of cancer and would like to predict patient prognosis (whether or not they will still be alive 5 yrs after initial diagnosis).

   When a patient is diagnosed with breast cancer, clinicians traditionally have collected many clinical and pathological measurements such as tumour grade and stage. In the last 20 years, new biotechnological advances have led to the ability to detect the expression (mRNA) of thousands of genes at the same time using gene expression microarrays. This technology has been used to assess the expression profiles of hundreds of tumours of different types. These profiles can then potentially be used to predict the prognosis of patients.

   In 2002, a group of researchers published a key paper that identified a set of 70 genes whose gene expression (mRNA) levels could predict prognosis in breast cancer patients. This finding was replicated in another study published the New England Journal of Medicine (NEJM) in the same year. The NEJM study by Marc van de Vijver and colleagues at the Netherlands Cancer Institute profiled (measured) the gene-expression levels of genes in breast cancer tumours from 295 women. Traditional criteria that are routinely used to predict a patient's prognosis, such as the number of lymph nodes that the cancer has spread to and the size of the tumour (in mm), were also measured.

   We will consider data on these 295 individuals who were diagnosed and treated for breast cancer at the Netherlands Cancer Institute. These data include more than 20 variables that reflect the traditional clinico-histopathological characteristics of the patient as well as the gene expression data from their tumours.

For this project, there are three primary questions to be investigated:

(a) Gene expression data can be expensive to collect. In the non-research world, this cost must be born by the patient. How well can common clinico-pathological variables help to classify an individual's chance of being distant metastasis free after 5 years?

(b) Do the data show a clear link between breast cancer survival and hormone receptor status (ER/PR/HER2)?

(c) Does combining the traditional variables with the gene expression data improve the ability to predict patient prognosis?

A full list of variables available for analysis is provided in Table 1.

Note that variables beginning with "I" represent immunohistochemistry data. The values of these variables are the percentage of positive cells for the relevant gene. The last 105 columns of the dataset contain the gene expression levels of specific genes from the microarray, these are the log (base 2) ratios of the tumor level to normal breast tissue levels. The column names are the relevant genes.

Variable	Description
B01_Profile_70_genes	70 gene prognosis signature; 0=poor; 1=good (Nature 2002 van 't Veer, NEJM 2002 van de Vijver)
E02_Event_DMFS_2005	Distant metastasis as first event(MCR) 2005 1=yes, 0=no
Cl01_Age	age at diagnosis
Cl02_pN_pos	Number of positive lymph nodes
Cl03_Size_(mm)	size of the tumor in mm
H02A_Type2007	IDC=1; ILC=2; Mucinous=3; Metaplastic=4; tubular.lobular=5; 6=apocrine; 10=IDC/ILC; 8=mucA; 9=mucB; 7=micropapillar
H03_ANGIOINV2001	Lymphangio invasion data; 0= none, 1= moderate, 2= extensive
H04_centralcoll	central collegen (0=absent) (1=present)
H05_matrix	matrix formation (0=absent) (1=present)
H06_necrosis	necrosis (0=negative; 1=some; 2=moderate; 3=extensive)
H08D_grade_07	A tumor with a final sum of Mitoses, nucl pleiomorphy and tubuli formation. 3-5 Grade I (well-differentiated). 6-7 Grade 2; 8-9 Grade 3 tumor
H09A_lymphinf2005	lymph infiltration (0=negative; 1=some; 2=moderate; 3=extensive)
M01B_PIK3CA	mutation (1) or wildtype (0) for the PIK3CA gene
M02G_TP53	mutation status for TP53 (1=silent/wildtype; 2=missense non-DBM, 3=missense DBM, 4=non-missense)
I01_ER_2007	Immunohistochemsitry (IHC) for estrogen receptor alpha
I02_PR_2007	Immunohistochemsitry for progesteron PR
I02B_ER/PR_clinical	ER/PR_Group: 1=ER+/PR+, 2=ER+/PR-, 3=ER-/PR-, 4=ER-/PR+
I03_HER2_2007	HER2 IHC 1= positive and 0 = negative
I06_CMYC_2007	C-myc antibody
I11_P53_2007	P53_%_positive

Table 1: Descriptions of variables contained in the dataset nkiproject.csv.

2. Bird Data Overview

This dataset captures the avian diversity of a specific area in Australia, featuring a remarkable collection of 98 different bird species across eight distinct zones. Observations detail a variety of flight behaviors, including gliding, flapping, and diving, offering insights into the ecological adaptations of these birds. Weather conditions, ranging from sunny to stormy, are recorded to understand their impact on bird activity. The dataset also includes information on seven habitat types, such as wetlands and forests, alongside geographic coordinates for precise location mapping.

For this project, there are three primary questions to be investigated:

(a) Predict bird species based on flight height and habitat.

(b) Do the data show a clear link between maximum flight, species and habitat ?

(c) Does combining the geographic coordinates variables with the maximum flight and site ID variables improve the ability to predict species?

A full list of variables available for analysis is provided in Table 2.

Variable	Description
SPECIES	The dataset includes 98 different bird species observed in the study area, aiding in assessing biodiversity and species-specific behaviors $(S_1, S_2, \cdots, \S_{98})$.
SITE_ID	Each observation is linked to one of the eight Site IDs, enabling comparisons across 8 different zones and facilitating localized ecological studies..
Behaviour	This variable captures six different flight behaviors, such as $B_1, B_2, \cdots, B_6$.
WEATHER_CONDITION	The dataset categorizes weather conditions into several categories.
N_INDIVIDUALS	This quantifies the number of individual birds observed for each species.
HABITAT	The dataset includes eight distinct habitat types, $X_1 = River, X_2 = Woodland, \cdots, X_8$.
Height_min	The lowest altitude at which birds were observed flying.
Hieght_max	The highest altitude at which birds were observed.
Longitude	Geographic coordinates pinpointing the exact location of observations.
Latitude	Geographic coordinates pinpointing the exact location of observations.
SEASON	The data is categorized by three distinct seasons (A, B, C), providing essential context for understanding migratory patterns and seasonal behaviors.

Table 2: Descriptions of variables contained in the dataset bird.csv.

Group Evaluation Form

For those working in pairs the following form must be completed and signed by both group members:

Student (Student No.)	% Contribution	Signature
Total	100%	

In most cases, the percentage contributions corresponding to individual group members would be expected to be equal (i.e. 50%), in which case both group members will receive exactly the same mark (corresponding to the mark given to the project). Where it is determined that percentage contributions are not equal, individual project marks will be assigned as follows:

- The group member with the highest percentage allocation will receive the mark given to the project.
- The group member with a lower percentage allocation will be given a mark commensurate with

$$\text{Project mark} \times \frac{\text{Their percentage contribution}}{\text{Higher group member percentage allocation}}$$

Where there is disagreement within a group as to appropriate percentage contributions and the group is unable to resolve this disagreement:

- **Do not sign the form.** (Students who sign the form are considered to be in agreement with the stated percentage allocations.)
- Contact the unit coordinator in regard to the dispute. Those with opposing views on the percentage allocation should explain why they believe a particular percentage allocation is warranted, including steps taken to try to ensure that all students had sufficient opportunity for equal contribution to the project.