BCB744: Intro R Test

Smit, A. J.

2025-03-17

ABOUT THE TEST

The Intro R Test will start at 8:30 on 17 March, 2025 and you have until 08:30 on 18 March to complete it. The Theory Test must be conducted on campus, and the Practical Test at home or anywhere you are comfortable working. The test constitutes a key component of Continuous Assessment (CA) and are designed to prepare you for the final exam.

The test consists of two parts:

Theory Test (30%)

This is a written, closed-book assessment where you will be tested on theoretical concepts. The only resource available during this test is RStudio, the R help system, your memory, and your mind.

Practical Test (70%)

In this open-book coding assessment, you will apply your theoretical knowledge to real data problems. While you may reference online materials (including ChatGPT), collaboration with peers is strictly prohibited.

ASSESSMENT POLICY

The marks indicated for each section reflect the relative weight (and hence depth expected in your response) rather than a rigid checklist of individual points. Your answer should demonstrate a comprehensive understanding of the concepts and techniques required, showing thoughtful integration of multiple R skills. Higher marks will be awarded for solutions that demonstrate not only technical correctness but also elegant code, insightful analysis, and clear communication of findings. We are assessing your ability to think systematically through complex data problems, make appropriate methodological choices, and present your findings in a coherent narrative that reveals meaningful patterns in the data. Your code should be well-structured, adequately commented, and reflect good programming practices.

Please refer to the Assessment Policy for more information on the test format and rules.

THEORY TEST

This is the closed book assessment.

Below is a set of questions to answer. You must answer all questions in the allocated time of 3-hr. Please

write your answers in a neatly formatted Word document and submit it to the iKamva platform.

Clearly indicate the question number and provide detailed explanations for your answers. Use Word's headings and subheadings facility to structure your document logically.

Naming convention: `Intro_R_Test_Theory_YourSurname.docx`

Question 1

You are a research assistant who have just been given your first job. You are asked to analyse a dataset about patterns of extreme heat in the ocean and the possible role that ocean currents (specifically, eddies) might play in modulating the patterns of extreme sea surface temperature extremes in space and time.

Being naive and relatively inexperienced, and misguided by your exaggerated sense of preparedness as young people tend to do, you gladly accept the task and start by exploring the data. You notice that the dataset is quite large, and you have no idea what's happening, what you are doing, why you are doing it, or what you are looking for. Ten minutes into the job you start to question your life choices. Your feeling of bewilderment is compounded by the fact that, when you examine the data (the output of the `head()` and `tail()` commands is shown below), the entries seem confusing.

```
fpath <- "/Volumes/OceanData/spatial/processed/WBC/misc_results"
fname <- "KC-MCA-data-2013-01-01-2022-12-31-bbox-v1_ma_14day_detrended.csv"
data <- read.csv(file.path(fpath, fname))
```

```
> nrow(data)
[1] 53253434

> head(data)
           t     lon    lat      ex    ke
1 2013-01-01 121.875 34.625 -0.7141 2e-04
2 2013-01-01 121.875 34.625 -0.8027 2e-04
3 2013-01-02 121.875 34.625 -0.8916 2e-04
4 2013-01-02 121.875 34.625 -0.9751 2e-04
5 2013-01-03 121.875 34.625 -1.0589 3e-04
6 2013-01-03 121.875 34.625 -1.1406 3e-04

> tail(data)
                  t     lon    lat     ex      ke
53253429 2022-12-29 174.375 44.875 0.4742 -0.0049
53253430 2022-12-29 174.375 44.875 0.4856 -0.0049
53253431 2022-12-30 174.375 44.875 0.4969 -0.0050
53253432 2022-12-30 174.375 44.875 0.5169 -0.0050
53253433 2022-12-31 174.375 44.875 0.5367 -0.0051
53253434 2022-12-31 174.375 44.875 0.5465 -0.0051
```

You resign yourself to admitting that you don't understand much, but at the risk of sounding like a fool when you go to your professor, you decide to do as much of the preparation you can do so that you at least have something to show for your time.

a. What will you take back to your professor to show that you have prepared yourself as fully as possible? For example:

- What is in your ability to understand about the study and the nature of the data?
- What will you do for yourself to better understand the task at hand?
- What do you understand about the data?
- What will you do to aid your understanding of the data?
- What will your next steps be going forward?

b. What will you need from your professor to help you understand the data and the task at hand so that you are well equipped to tackle the problem?

[15 marks]

Question 2

Please translate the following code into English by providing an explanation for each line:

```
monthlyData <- dailyData %>%
    dplyr::mutate(t = asPOSIXct(t)) %>%
    dplyr::mutate(month = floor_date(t, unit = "month")) %>%
    dplyr::group_by(lon, lat, month) %>%
    dplyr::summarise(temp = mean(temp, na.rm = TRUE)) %>%
    dplyr::mutate(year = year(month)) %>%
    dplyr::group_by(lon, lat) %>%
    dplyr::mutate(num = seq(1:length(temp))) %>%
    dplyr::ungroup()
```

In your answer, simply refer to the line numbers (1-9) before each line of code and provide an explanation for each line.

[10 marks]

Question 3

What is 'Occam's Razor'?

[5 marks]

Question 4

Explain the difference between R and RStudio.

[5 marks]

Question 5

By way of example, please explain some key aspects of R code conventions. For each line of code, explain also in English what aspects of the code are being adhered to.

For example:

1. `a ← b` is not the same as `a < -b`. The former is correct because there is a space preceding and following the assignment operator (←, a less-than sign immediately followed by a dash to form an

arrow); this has a different meaning from the latter, which is incorrect because there is no space between the less-than sign and the dash, reading as “a is less than negative b”.

Hint: In your Word document, use a fixed-width font to indicate the code as a separate block which is distinct from the rest of the text.

[10 marks]

Question 6

a. Explain why one typically prefers working with CSV files over Excel files in R.
b. What are the properties of a CSV file that make it more suitable for data analysis in R?
c. What are the properties of an Excel file that make it less suitable for data analysis in R?

[15 marks]

Question 7

Explain each of the following in the context of their use in R. For each, provide an example of how you would construct them in R:

a. A vector
b. A matrix
c. A dataframe
d. A list

Hint: See my hint under Question 5.

[20 marks]

Question 8

a. Write a 150 to 200 word abstract about your Honours research project. In your abstract, draw attention to the types of data you will be expected to generate, and mention how these will be used to address your research question.
b. Explain which of the R data classes will be most useful in your research and why.
c. With reference to the abstract you wrote in Question 8.a, explain how you would visualise (or display your finding in tabular format) your research findings. Provide an example of how you would do this in R. Which of your research questions would be best answered using a visualisations or tables? What do you expect your visualisations or tables to show?
d. Provide an example of how you would create a plot or table in R. Generate mock code (it does not need to run) that you would use to create the plot or table.

Note 1: In the unlikely event that your research will not require visualisations or tables, please explain why this is the case and how you would communicate your findings.

Note 2: If you haven’t defined your research project yet, describe a hypothetical project in your field of interest.

[30 marks]

TOTAL MARKS: 110

PRACTICAL TEST

This is the open book assessment.

Below is a set of scripting problems to solve. You have 21 hours from the end of the Theory Test to complete this section Please write your code in an R script file and submit it to the iKamva platform by no later than 8:30 on Tuesday, 18 March 2025.

Please follow a clear structure (appropriate, clearly numbered headings and subheadings) in your code, including comments and explanations.

Ensure that all code runs without errors before submitting it – serious penalties will apply to non-functional scripts.

Naming convention: `Intro_R_Test_Practical_YourSurname.R`

Question 1

Download the `fertiliser_crop_data.csv` data.

The data represent an experiment designed to test whether or not fertiliser type and the density of planting have an effect on the yield of wheat. The dataset contains the following variables:

- Final yield (kg per acre) – make sure to convert this to the most suitable SI unit before continuing with your analysis
- Type of fertiliser (fertiliser type A, B, or C)
- Planting density (1 = low density, 2 = high density)
- Block in the field (north, east, south, west)

Undertake a full visual assessment of the dataset and establish which of the influential variables are most likely to have an effect on crop yield. Provide a detailed explanation of your findings.

[25 marks]

Question 2

The Bullfrog Occupancy and Common Reed Invasion data are here: `AICcmodavg::bullfrog` (i.e. the `bullfrogs` dataset resides within the **AICcmodavg** package, which you might have to install).

Create a tidy dataframe from the bullfrog data.

[10 marks]

Question 3

The Growth Curves for Sitka Spruce Trees in 1988 and 1989 data are here: `MASS::Sitka` and `MASS::Sitka89`.

Combine the two datasets and provide an analysis of the growth curves for Sitka spruce trees in 1988 and 1989. Give graphical support for the idea that i) ozone affects the growth of Sitka spruce trees, and ii) the

growth of Sitka spruce trees is affected by the year of measurement. In addition to showing the overall response in each year x treatment, also ensure that the among tree variability is visible.

Explain your findings.

[20 marks]

Question 4

The Frog Dehydration Experiment on Three Substrate Types data can be accessed here: `AICcmodavg::dry.frog`.

a. Based on the dataset, what do you think was the purpose of this study? Provide a 200 word synopsis as your answer.
b. Create new columns in the dataframe showing:
   - the final mass;
   - the percent mass lost; and
   - the percent mass lost as a function of the initial mass of each frog.
c. Provide the R code that would have resulted in the data in the variables `cent_initial_mass` and `cent_Air`.
d. An analysis of the factors responsible for dehydration rates in frogs. In your analysis, consider the effects substrate type, initial mass, air temperature, and wind.
e. Provide a brief discussion of your findings.

[25 marks]

Question 5

Consider this script:

```
ggplot(points, aes(x = group, y = count)) +
  geom_boxplot(aes(colour = group), size = 1, outlier.colour = NA) +
  geom_point(position = position_jitter(width = 0.2), alpha = 0.3) +
  facet_grid(group ~ ., scales = "free") +
  labs(x = "", y = "Number of data points") +
  theme(legend.position = "none",
    strip.background = element_blank(),
    strip.text = element_blank())
```

a. Generate fictitious (random, normal) data that can be plotted using the code, above. Make sure to assemble these data into a dataframe suitable for plotting, complete with correct column titles.
b. Apply the script *exactly as stated* to the data to demonstate your understanding of the code and convince the examiner of your understanding of the correct data structure.

[10 marks]

Question 6

For this assessment, you will analyse the built-in R dataset `datasets::UKDriverDeaths`, which contains monthly totals of car drivers killed or seriously injured in road accidents in Great Britain from January 1969 to December 1984. This time series data allows for examination of long-term trends, seasonal patterns, and potential correlations with societal factors.

a. **Data Exploration and Preparation**

   i. Load the `UKDriverDeaths` dataset and examine its structure. Convert the time series data into a standard data frame format with separate columns for:

      - Year
      - Month (both as a number and as a factor with proper names)
      - Number of deaths/injuries

   ii. Create a new variable that classifies each month into seasons (Winter: Dec-Feb, Spring: Mar-May, Summer: Jun-Aug, Autumn: Sep-Nov).
   iii. Create another variable identifying whether each observation falls during a major energy crisis period (e.g., the oil crises of 1973-1974 and 1979-1980).
   iv. Identify and handle any potential inconsistencies or issues in the dataset that might affect subsequent analyses.

[20 marks]

b. **Temporal Trend Analysis**

   i. Create a comprehensive visualisation showing the full time series with:

      - Clear temporal trends
      - A smoothed trend line
      - Vertical lines or shading indicating major UK policy changes related to road safety (e.g., 1983 seat belt law)
      - Annotations for key events

   ii. Develop a visualisation examining monthly fatality averages across the entire period, ordered appropriately to show seasonal patterns.
   iii. Create a visualisation that compares annual patterns between the first half of the dataset (1969-1976) and the second half (1977-1984).
   iv. Using *tidy* data manipulation techniques, calculate and visualise the year-over-year percent change in fatalities for each month throughout the dataset.

[20 marks]

c. **Pattern Analysis and Decomposition**

   i. Calculate and visualise the average number of fatalities by season across all years.
   ii. Create a heatmap showing fatalities by month and year, with appropriate color scaling to highlight temporal clusters.
   iii. Implement a decomposition of the time series to separate: - The overall trend - Seasonal patterns - Remaining variation

iv. Visualise each component and discuss what factors might contribute to the patterns observed.

Note: Some of this will be new to you. But don't worry, use any means available to you to solve the problem.

[25 marks]

d. **Data manipulation**

Starting with the data as presented in the `UKDriverDeaths` dataset, create a new dataframe identical to the `Seatbelts` dataset.

[5 marks]

TOTAL MARKS: 160

– THE END –