Microcredential Assignment

Jasmine Hughes

11/04/2022

Overview

This data set is a synthetic data set similar to real-world electronic health record data. It describes patients treated at two different hospitals ("F1" and "F2"). In this assignment, you will demonstrate your ability to use R to perform typical data manipulations and to calculate clinical parameters with the end goal of assessing acute kidney injury.

Please use R for all analysis and graphing. You should be able to complete the assignment using only packages covered in class, but if you find a package that helps you accomplish a task you are welcome to use it.

Make sure your R code is readable. Use variable names that are easily understood. Use comments where appropriate to explain how your code works.

You may write your answers in this R Markdown Notebook, or you may submit your assignment as a separate .R or .Rmd file.

Please email your solution to `wk@buffalo.edu` by October 8th. To receive the microcredential badge, you will need to receive a grade of at least 75% (partial credit will be awarded for all questions).

Question 1: Explore Data

1(a): Load Data

Two CSV files have been provided:

1. "patients.csv" contains basic demographic data about patients treated at hospital F1 and hospital F2. This data set includes patient age (in years), patient sex ("M" for male, "F" for female), patient weight (in lbs), and patient height (in cm). The column "ID" contains an identifier unique to each patient.
2. "lab_tests.csv" contains the results of blood tests for these patients. The compound analyzed in this data set is serum creatinine, which is measured in units of mg/dl. There is a column for ID, which corresponds to the column ID in "patients.csv". There is a column for the lab result, and a column for the measurement time.

Using `read.csv()` or another appropriate function, read these two files and save them as variables.

1(b) Unit conversion

Patient weight has been provided in pounds, but in a later question, we need weight to be in kilograms. Write a function called `lb_to_kg` that takes one value in pounds, or a vector of values in pounds, and converts it to kilograms. Use the conversion 1 lb = 0.454 kg.

If your function works correctly, the code below should return TRUE!

```
identical(round(lb_to_kg(c(1, 150)), 3), c(0.454, 68.100))
```

1(c) Convert patient weights from pounds to kilograms

Using the function you defined, convert patient weights from pounds to kilograms, saving this data frame as a new variable.

1(d) Convert measurement time to a datetime class

Depending on how you imported in the data in "lab_test.csv", the column "measurement_time" is probably of class "character" or "factor". Convert it to be of class POSIXct, saving this data frame as a new variable.

Question 2: Estimating Glomerular Filtration Rate

Glomerular filtration rate (GFR) is a measure of kidney function. One common way of estimating GFR is with the Cockcroft-Gault formula.

For male patients, the Cockcroft-Gault formula is:

$$eGFR = (140 - AGE) * WT/(SCR * 72)$$

For female patients, the Cockcroft-Gault formula is:

$$eGFR = 0.85 * (140 - AGE) * WT/(SCR * 72)$$

where AGE is age in years, WT is weight in kilograms, and SCR is serum creatinine in mg/dl.

2(a) eGFR function

Write a function called `calc_egfr` that takes age, sex, weight and serum creatinine as arguments, and returns eGFR.

If your function works, the code below should return TRUE for both lines.

```
egfr1 <- calc_egfr(sex = "F", age = 70, creat = 0.5, weight = 70)
egfr2 <- calc_egfr(
  sex = c("F", "M"),
  age = c(70, 54),
  creat = c(0.5, 1.0),
  weight = c(70, 80)
)

identical(round(egfr1), 116)
identical(round(egfr2), c(116, 96))
```

2(b) Calculate eGFR

Using the function you created above, calculate eGFR for each patient.

Serum creatinine is in the "lab_test.csv" data set, while the other variables you need are in "patients.csv". You will need to combine the two data frames. Here is an example to help you. For more, see Workbook 8!

```
pt_char <- data.frame(
  pt = c("M.H.", "P.J.", "K.G.", "A.C."),
  age = c(77, 35, 56, 48),
  weight = c(65, 82, 70, 100),
  stringsAsFactors = F
)
pt_doses <- data.frame(
  pt = c("M.H.", "P.J.", "K.G.", "M.H.", "S.C."),
  mg_per_kg = c(15, 15, 10, 20, 15),
  interval = c("Q8", "Q12", "Q6", "Q8", "Q8"),
  stringsAsFactors = F
)

combined_data <- dplyr::left_join(pt_char, pt_doses, by = "pt")
```

- Many patients have more than one serum creatinine lab value. For these patients, you should calculate an eGFR for each serum creatinine value.

2(c) Graph eGFR

Using ggplot, create a histogram of eGFRs, with separate distributions shown for male patients and for female patients. Add a title to your graph and label your axis.

Question 3: Calculate Acute Kidney Injury

Acute kidney injury (or AKI) is a medical syndrome involving sudden damage to the kidney, resulting in reduced kidney function. A reduction in kidney function causes serum creatinine levels to rise.

We will use a simplified metric to diagnosis kidney injury: (1) an increase in serum creatinine from baseline of at least 0.3 mg/dl OR (2) an increase in serum creatinine of 1.5 times baseline.

For example:

- Patient 1 comes into the hospital, and their serum creatinine levels are 0.31 mg/dl. Later during treatment, their serum creatinine lab result measures 0.55 mg/dl. Since $(0.31 * 1.5) < 0.55$, they should be classified as having an AKI because of condition (2) in our definition above.
- Patient 2 comes into the hospital with a serum creatinine level of 0.70 mg/dl. Later during treatment, their serum creatinine lab result measures 1.03 mg/dl. Since $(0.7 + 0.3) < 1.03$, they should be classified as having an AKI because of condition (1) in our definition about.
- Patient 3 comes into the hospital with a serum creatinine level of 0.50 mg/dl. Later during treatment, their serum creatinine lab results fall to 0.15 mg/dl. They should NOT be classified as having an AKI because $(0.5 * 1.5) > 0.15$ mg/dl and $(0.5 + 0.3) > 0.15$ mg/dl.

3(a): Baseline serum creatinine

For this analysis, assume that baseline serum creatinine is the earliest (by datetime) measurement of serum creatinine. Make a new column in your lab result data frame that contains the first serum creatinine measured for that patient. (Hint: the measurement_time column isn't sorted chronologically.)

3(b): Maximum serum creatinine

Add a column to your data frame containing the maximum serum creatinine for each patient. Your data frame should have (at a minimum) columns containing the following data: ID, facility, baseline serum creatinine, maximum serum creatinine.

3(c): Determine AKI

Create a new column in your lab result data frame indicating whether that patient experienced an AKI using the definitions above, the baseline serum creatinine value you calculated in 3(a), and the maximum serum creatinine value you calculated in 3(b).

3(d): AKI rates by hospital

What proportion of patients at facility F1 had an AKI? What proportion of patients at facility F2 had an AKI? (Make sure you count each patient only once!). Do these facilities have different rates of AKI? Use a graph, or the results of `prop.test()` to justify your answer.