PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Tarea 1

Taller de análisis de datos I

Fecha de entrega: 17/07/2023 hasta las 23:59 hrs.

Aspectos formales:

- Deberá entregar una carpeta comprimida (`.zip`) que contenga: a) su reporte con las respuestas (`.pdf` o `.docx`), b) su proyecto (`.Rproj`) y c) el código de análisis (`.R`) **Este código debe ser reproducible**.
- Los aspectos de formato son flexibles, pero respetando la formalidad de un documento de trabajo.
- Todo el procesamiento, análisis, tablas y visualizaciones de datos deben ser realizado en `R`. En el código debe comentar el paso a paso de su tarea.
- Fecha de entrega: 17/07/2023 hasta las 23:59 hrs. **No se recibirán evaluaciones después de la fecha de entrega**
- Debe enviar su evaluación al profesor (jdconejeros@uc.cl) y a la ayudante del curso (samadariaga@uc.cl)
- La tarea deberá ser realizada en parejas o individual.

Tutoría el lunes 10/07/2023 (19:00 - 20:30 hrs): tendremos una tutoría voluntaria en que la ayudante entregara tips y resolvera dudas de la tarea. Dudas al correo del profesor y/o la ayudante. También pueden recurrir a cualquier apoyo disponible en el web, además de la literatura del curso.

Descripción

El objetivo de la tarea 1 es practicar las principales herramientas de `R` y `RStudio` para la manipulación de objetos, lectura de tablas y análisis. Se espera que puedan utilzar las herramientas vistas en el laboratorio 1 y 2 del curso.

PONTIFICIA UNIVERSIDAD CATÓLICA DE CHILE

Ejercicio 1

a. Genere en su `enviroment` los siguientes vectores:

`vector_1`: Contiene una secuencia de 9 numeros que inicia en 1 y termina en 5.

`vector_2`: Contiene una secuencia de números impares que van desde el 1 hasta el 20 de dos en dos.

`vector_3`: Contiene el número `2023` repetido 10 veces.

`vector_4`: Contiene 5 repeticiones del número `1` y 5 repeticiones del número `0`.

`vector_5`: Contiene la suma del `vector_1` y del `vector_2`.

Para el `vector_5` revise el `Warning message` que indica R y explique cuál la coerción implícita aplicada. Además, explique la diferencia entre un vector, una matris de datos y una lista. **(1 punto)**

b. Genere una matriz con los vectores definidos en **a**. (*Hint:* recuerde que la información se pega por columnas). Debiera llegar al siguiente resultado:

```
      vector_1 vector_2 vector_3 vector_4 vector_5
 [1,]      1.0        1     2023        1      2.0
 [2,]      1.5        3     2023        1      4.5
 [3,]      2.0        5     2023        1      7.0
 [4,]      2.5        7     2023        1      9.5
 [5,]      3.0        9     2023        1     12.0
 [6,]      3.5       11     2023        0     14.5
 [7,]      4.0       13     2023        0     17.0
 [8,]      4.5       15     2023        0     19.5
 [9,]      5.0       17     2023        0     22.0
[10,]      1.0       19     2023        0     20.0
```

Explique la coerción implícita que se aplica en la columna `vector_1`. ¿Por qué sería importante saber esto? ¿Cuál es la diferencia entre mantener este objeto como una matriz a transformarlo en un data-frame? **(1 punto)**

c. El estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$ es una especificación lineal para el vector de respuesta $\boldsymbol{Y}$ que tiene una distribución normal multivariada ($\boldsymbol{Y} \sim Normal(\boldsymbol{X\beta}, \sigma^2\boldsymbol{I})$). Dicho estimador es de la forma:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^t\boldsymbol{X})^{-1}\boldsymbol{X}^t\boldsymbol{Y} \sim Normal(\boldsymbol{\beta}, (\boldsymbol{X}^t\boldsymbol{X})^{-1}\sigma^2)$$

Donde $\boldsymbol{X}$ es el vector que predice $\boldsymbol{Y}$[1]. En otras palabras, el estimador $\hat{\boldsymbol{\beta}}$ corresponde a la multiplicación entre la inversa del vector $\boldsymbol{X}$ traspuesto multiplicada por el vector $\boldsymbol{X}$, y el vector $\boldsymbol{X}$ traspuesto multiplicado por el vector $\boldsymbol{Y}$. El estimador $\hat{\boldsymbol{\beta}}$ se interpretará como la asociación entre el predictor $\boldsymbol{X}$ y mi vector de respuesta $\boldsymbol{Y}$. Este estimador será fundamental a la hora de evaluar el impacto en políticas públicas.

Utilizando la tabla de datos `simce2m2016_extracto.csv` calcule manualmente el estimador $\hat{\boldsymbol{\beta}}$ que nos permite validar la siguiente hipótesis: *"Los resultados en el SIMCE están negativamente asociados. Un resultado positivo en la prueba de matemáticas implica un resultado negativo en la prueba de lenguaje"*[2]. ¿Qué puede decir de su hipótesis? ¿La relación entre el resultado de ambas pruebas es negativo o positivo? Justifique. **(1 punto)**

Ejercicio 2

Un buen diagnóstico es fundamental a la hora de levantar problemáticas relevantes para la sociedad y diseñar políticas públicas. Es por esto que una ONG internacional lo ha contratado a usted como consultor/a para indagar acerca de los cambios que ha observado Latinoamérica y el Caribe en salud, vivienda, ingresos y educación desde el año 2015 al año 2020. Para esto usted dispone de una librería en `R` que se conecta a una API del Banco Mundial lo que permite extraer información desde la comodidad de su escritorio, esta librería se llama: `WDI` (World Development Indicators and Other World Bank Data)[3]. A partir de esto se le pide desarrollar los siguientes requerimientos:

a. A partir de la librería `WDI` realice la extracción de un indicador del banco mundial que tenga data disponible desde el año 2015 al año 2020 para cada país de latinoamérica. Es solo un indicador que usted considere relevante para alguna de estas áreas: salud, vivienda, ingresos o educación[4]. Indique el número de observaciones, ¿cuántos países y años son parte del estudio? ¿Cuántas variables pudo generar en su extracción? **(0.5 puntos)**

Por ejemplo, si quisiéramos obtener información sobre el ingreso per cápita (GDP) de los países de Latinoamérica y el Caribe podriamos construir la siguiente tabla de datos:

[1] $\boldsymbol{X}$ está definido como una matriz de dos vectores: el vector 1 es una matriz solo de 1 con un largo equivalente al vector 2. Por otro lado, el vector 2 se compone por la variable independiente o predictora de interés.

[2] puede validar su resultado usando el código `lm(ptje_mate2m_alu ~ ptje_lect2m_alu, data=simce)`. Para que el código funcione recuerde llamar a su objeto que contiene la tabla de datos como `simce`.

[3] Puede encontrar toda la documentación de la librería aquí: http://vincentarelbundock.github.io/WDI/ y aquí: https://cloud.r-project.org/web/packages/WDI/WDI.pdf

[4] Puede realizar busquedas de indicadores con `WDIsearch("literacy")`. Lo importante es buscar el término en inglés. También podría utilizar otra librería que se llama `library(wbstats)` para su extracción de datos.

```
Rows: 132
Columns: 13
$ country           <chr> "Argentina", "Argentina", "Argentina", "Argentina", ~
$ iso2c             <chr> "AR", "AR", "AR", "AR", "AR", "AR", "AW", "AW", "AW"~
$ iso3c             <chr> "ARG", "ARG", "ARG", "ARG", "ARG", "ARG", "ABW", "AB~
$ year              <int> 2015, 2019, 2018, 2020, 2017, 2016, 2018, 2017, 2016~
$ NY.GDP.PCAP.PP.KD <dbl> 23933.887, 22071.748, 22747.242, 19685.216, 23597.11~
$ status            <chr> "", "", "", "", "", "", "", "", "", "", "", "", "", ~
$ lastupdated       <chr> "2023-06-29", "2023-06-29", "2023-06-29", "2023-06-2~
$ region            <chr> "Latin America & Caribbean", "Latin America & Caribb~
$ capital           <chr> "Buenos Aires", "Buenos Aires", "Buenos Aires", "Bue~
$ longitude         <chr> "-58.4173", "-58.4173", "-58.4173", "-58.4173", "-58~
$ latitude          <chr> "-34.6118", "-34.6118", "-34.6118", "-34.6118", "-34~
$ income            <chr> "Upper middle income", "Upper middle income", "Upper~
$ lending           <chr> "IBRD", "IBRD", "IBRD", "IBRD", "IBRD", "IBRD", "Not~
```

b. Agrupando la información de todos los países de latinoamérica, calcule estadísticos descriptivos (centro, dispersión y posición) para su variable de interés según el año de estudio. Interprete sus resultados. **(1 punto)**

Un ejemplo para el GPD sería de este estilo:

Table 1: Ingreso per cápita de latinoamérica en el tiempo

Region	Año	N	Media	SD	Min	Mediana	Max
Latin America & Caribbean	2015	22	17624.1	10431.1	5180.2	14215.7	41852.4
Latin America & Caribbean	2016	22	17651.9	10295.8	5287.3	14358.2	40857.8
Latin America & Caribbean	2017	22	17678.5	10113.1	5447.7	14334.9	38865.2
Latin America & Caribbean	2018	22	17696.3	9990.8	5561.3	14426.4	40706.7
Latin America & Caribbean	2019	22	17846.3	10207.6	5398.1	14616.1	40780.5
Latin America & Caribbean	2020	22	16154.5	8573.5	5028.1	13805.8	33155.2

Fuente: Elaboración propia.

c. Explore si es que su variable de interés presenta valores `NA` (pérdidos) ¿Qué países no cuentan con información? ¿Por qué el *missing value* sería un problema? Proponga alguna estrategia para abordar el problema de los *missing values* para este u otro contexto de análisis. **(0.5 puntos)**

d. ¿Cómo se comporta su variable de análisis para el caso de Chile? Presente una tabla e interprete sus resultados. **(0.5 puntos)**

e. ¿Qué puede concluir de sus análisis? ¿Cuál es el diagnóstico que puede ofrecer como consultor/a? **(0.5 puntos)**

Bonus (+0.7)

Busque una manera de representar el indicador seleccionado en el ejercicio 2 a partir de un mapa de Latinoamérica y el Caribe según el año. En otras palabras, tendrá 5 mapas de latinoamerica (uno para cada año) bajo la capa de su indicador. Tip: puede trabajar con la librería `sf()`u otra para la generación de mapas. Interprete sus resultados.