**SEGUNDO EJERCICIO OBLIGATORIO PARA SEMINARIO IMPUTACIÓN DE RESPUESTAS: MÉTODOS CLÁSICOS**

**Responde a las preguntas en este archivo de Word (o similar) y anexa el script de R que hayas creado**

**Métodos clásicos de imputación. Abrir el archivo Ejercicio2.dat desde R**

El objetivo de este ejercicio es probar los métodos clásicos de imputación y comprobar el sesgo que se produce ante una situación de perdida de datos MAR.

Tenemos dos variables (además del ID), IQ = Cociente intelectual y REND = Rendimiento. La media poblacional de Rendimiento es μ = 4,89, su desviación típica poblacional es σ = 2,34, la correlación entre ambas variables (poblacional) es ρ = 0,50.

Vamos a generar un 50% de valores perdidos en la variable REND bajo el mecanismo MAR (la pérdida de datos en REND dependerá en IQ). Para ello, tendremos aproximadamente un 80% de casos válidos en REND para los casos por encima de la mediana de IQ (MDNIQ = 101). En cambio, tendremos únicamente un 20% (aprox.) de caso válidos en REND para los casos por debajo de la mediana de IQ. A esta variable con un 50% de los casos válidos de REND podemos llamarla REND\_MAR.

El propósito es ir imputando datos en esta variable REND\_MAR con los siguientes métodos clásicos: remplazamiento con la media, método Hot-deck, imputación por regresión e imputación por regresión estocástica (puedes ayudarte del script visto en clase).

Rellenar la siguiente tabla:[[1]](#footnote-1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Parámetro estimado por métodos clásicos (media, desviación típica y correlación): , y | | | |
|  | Parámetro | Imputación media | Hot-deck | Regresión | Regresión estocástica |
| Media | 4.89 | 5.430 | 5.431 | 4.885 | 4.883 |
| D.T. | 2.34 | 1.624 | 2.258 | 1.867 | 2.339 |
| Corr | 0.50 | 0.271 | 0.194 | 0.623 | 0.496 |

¿A qué conclusiones llegas respecto a los métodos clásicos de imputación en cuanto a sesgo (consistencia) en los tres parámetros estimados: , y ? ¿Qué métodos son peores y cuáles mejores?

Se realizó el ingreso del mecanismo MAR de acuerdo con las indicaciones y se ejecutó mil replicaciones que fueron resumidas en la tabla anterior. Se tiene lo siguiente en cada parámetro:

* **Media (μ̂)**: La imputación por regresión y regresión estocástica recuperan muy bien la media real (4.88 vs. 4.89). En cambio, la imputación por la media y la versión de hot-deck utilizada sobrestiman la media, probablemente porque los valores imputados no consideran la relación entre REND e IQ.
* **Desviación típica (σ̂)**: Solo la regresión estocástica recupera bien la variabilidad original (2.34), ya que incorpora error aleatorio. La imputación por la media y por regresión determinista subestiman la variabilidad, al introducir valores poco dispersos. El hot-deck logra una desviación más cercana, pero con un sesgo importante en otros aspectos.
* **Correlación (ρ̂)**: La regresión determinista y la estocástica son los únicos métodos que preservan o incluso amplifican la relación entre REND e IQ (especialmente la determinista: 0.623). Los otros métodos reducen drásticamente la correlación, mostrando que ignoran completamente la estructura conjunta entre las variables.

Los métodos más precisos y consistentes son la regresión estocástica, que recupera fielmente los tres parámetros poblacionales, y la regresión determinista, que aunque subestima la variabilidad, preserva la relación con IQ. En contraste, la imputación por la media y el hot-deck aleatorio presentan sesgos considerables, especialmente en la media y la correlación.

*Anexo Código R*

##%######################################################%##  
# #  
#### Ejercicio 2 ###  
#### Brian Norman Peña Calero ###  
#### Seminario de Valores Perdidos ###  
# #  
##%######################################################%##  
  
# Carga de Paquetes -------------------------------------------------------  
  
library(tidyverse)  
library(mice)  
  
# Importar Datos Ejercicio 2 ----------------------------------------------  
  
rendimiento <- read\_delim("Ejercicio2.dat")  
  
# Función de una réplica: genera REND\_MAR y aplica imputaciones  
una\_replicacion <- function(df) {  
 set.seed(NULL) # para aleatoriedad real  
   
 mediana\_iq <- median(df$IQ)  
   
 # Generar REND\_MAR con MAR  
 df <- df |>  
 mutate(  
 REND\_MAR = case\_when(  
 IQ < mediana\_iq ~ ifelse(runif(n()) < 0.2, REND, NA),  
 IQ >= mediana\_iq ~ ifelse(runif(n()) < 0.8, REND, NA)  
 )  
 )  
   
 # Imputación por media  
 df <- df |>  
 mutate(REND\_media = ifelse(is.na(REND\_MAR), mean(REND\_MAR, na.rm = TRUE), REND\_MAR))  
   
 # Hot-deck (versión con media y desviación típica + error aleatorio)  
 media\_REND <- mean(df$REND\_MAR, na.rm = TRUE)  
 sd\_REND <- sd(df$REND\_MAR, na.rm = TRUE)  
   
 df <- df |>   
 mutate(REND\_hotdeck = ifelse(  
 is.na(REND\_MAR),  
 media\_REND + rnorm(n(), mean = 0, sd = sd\_REND),  
 REND\_MAR  
 ))  
   
 # Regresión determinista  
 modelo\_reg <- lm(REND\_MAR ~ IQ, data = df)  
 predicciones <- predict(modelo\_reg, newdata = df)  
 resid\_sd <- sd(modelo\_reg$residuals, na.rm = TRUE)  
   
 df <- df |>  
 mutate(  
 REND\_reg = ifelse(is.na(REND\_MAR), predicciones, REND\_MAR),  
 REND\_reg\_stoc = ifelse(  
 is.na(REND\_MAR),  
 predicciones + rnorm(n(), 0, resid\_sd),  
 REND\_MAR  
 )  
 )  
   
 # Calcular estadísticas  
 tibble(  
 media\_media = mean(df$REND\_media),  
 sd\_media = sd(df$REND\_media),  
 cor\_media = cor(df$REND\_media, df$IQ),  
   
 media\_hot = mean(df$REND\_hotdeck),  
 sd\_hot = sd(df$REND\_hotdeck),  
 cor\_hot = cor(df$REND\_hotdeck, df$IQ),  
   
 media\_reg = mean(df$REND\_reg),  
 sd\_reg = sd(df$REND\_reg),  
 cor\_reg = cor(df$REND\_reg, df$IQ),  
   
 media\_regstoc = mean(df$REND\_reg\_stoc),  
 sd\_regstoc = sd(df$REND\_reg\_stoc),  
 cor\_regstoc = cor(df$REND\_reg\_stoc, df$IQ)  
 )  
}  
  
# Ejecutar 1000 réplicas  
set.seed(123)  
resultados\_sim <- map\_dfr(1:1000, ~ una\_replicacion(rendimiento))  
  
resultados\_sim |>   
 pivot\_longer(  
 cols = everything(),  
 names\_sep = "\_",  
 names\_to = c("stat", "method"),  
 values\_to = "Valor"  
 ) |>   
 group\_by(method, stat) |>   
 summarise(Media = mean(Valor), .groups = "drop") |>   
 pivot\_wider(  
 names\_from = method,  
 values\_from = Media  
 ) |>   
 mutate(  
 stat = factor(  
 stat,  
 levels = c("media", "sd", "cor")  
 )  
 ) |>   
 arrange(stat) |>   
 relocate(stat, media) |>   
 rename\_with(~ c("Estadístico", "Imputación media", "Hot-deck",   
 "Regresión", "Regresión estocástica"))

1. Puedes hacer una sola réplica o, si lo prefieres, hacer múltiples réplicas insertando los scripts en una estructura for [↑](#footnote-ref-1)