EJERCICIOS OBLIGATORIOS PARA SUPERAR EL SEMINARIO

Cópiate el ejercicio en un 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 $\mu = 4{,}89$, su desviación típica poblacional es $\sigma = 2{,}34$, la correlación entre ambas variables (poblacional) es $\rho = 0{,}50$.

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:

	Parámetro	Imputación media	Hot-deck	Regresión	Regresión estocástica
Media	4.89	5.35	5.37	4.74	4.76
D.T.	2.34	1.61	2.36	1.91	2.41
Corr	0.50	0.3	0.19	0.68	0.52

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

La mayor consistencia se da en la imputación por regresión estocástica; este hecho se debe a la introducción del componente aleatorio, que tiene en cuenta la variabilidad de REND,

compensando el sesgo introducido por las restricciones del modelo lineal al que sometemos los datos por este procedimiento.

La regresión simple no introduce dicho componente, por lo que infraestimará la desviación típica (variabilidad), y consecuentemente, forzamos a una correlación sobreestimada, fruto del sobreajuste sin término error.

El método hot-deck, dado que funciona por "similaridad" de datos, captura la variabilidad, pero sesga media, dado que la ratio perdidos/válidos está desbalanceado, y correlaciones, dado que no tiene en cuenta la relación entre predictora y criterio.

Finalmente, la imputación por media es el peor método, ya que sustituye todos los valores por la media, resultando en una medida que falsea todos los parámetros.