EJERCICIOS OBLIGATORIOS PARA SUPERAR EL SEMINARIO

Cópiate este ejercicio en un Word independiente (o editor de textos que quieras) y entrega también anexado el script de R.

Abrir el archivo Ejercicio1.dat desde R

El objetivo de este ejercicio es familiarizarte con cómo inspeccionar los valores perdidos de un archivo antes de empezar a trabajar con él con técnicas inferenciales (regresiones, ANOVAS, SEM, etc.). Utilizaremos las herramientas de R que hemos visto en clase. Además, también se pretende con este ejercicio que sepas falsar que la pérdida de tus datos se ocasiona por mecanismos MCAR.

En este archivo encontrarás 779 estudiantes seleccionados al azar de carreras técnicas en el que recogemos 8 variables. ID es el número de identificador del sujeto. REND es la nota media en la carrera (primera convocatoria a la que se presentaron al examen). HESTUDIO contiene el percentil de horas de estudio, MOT el percentil en una escala de motivación por los contenidos aprendidos en su grado, SEXO (0 = mujer y 1 = varón), SELEC = la nota de selectividad, STUAULA son los estudiantes/aula de su grado y GASTO es el gasto en miles de euros por año (matrícula, piso, materiales, etc.).

1.1) Con la función md.pattern del paquete mice, contesta a la siguientes preguntas. ¿Cuántos casos tenemos sin valores perdidos (casos completos)?

172 casos

1.2) ¿Cuál es el patrón más común de valores perdidos que tenemos (sin contar el de casos completos)? ¿Cuántos casos tiene ese patrón y qué variable/es tienen en ese patrón valores perdidos? Si quitásemos o eliminásemos del archivo a esa/as variable/es que implican a ese patrón, ¿cuántos casos completos tendríamos?

El patrón más común es el que tiene solo como caso perdido la variable MOT. Son 103 casos. Si lo quitásemos, habría 275 casos completos.

1.3) ¿Consideras que se puede asumir que la pérdida de datos para la variable REND en este archivo es MCAR? Justifica tu respuesta (nota: conviene que generes primero una variable binaria con códigos 1 si tenemos valor válido en REND y 0 si tenemos valor perdido en REND. Te puedes ayudar del script visto en clase)

Existen diferencias estadísticamente significativas para la primera variable probada (HESTUDIO), por lo que no podemos mantener la hipótesis de pérdida MCAR en REND.

1.4) ¿Cuál es la media de horas de estudio (medida en percentiles) para los casos con un valor

válido en REND? ¿Y para los casos con valores perdidos en REND? ¿Hay diferencias

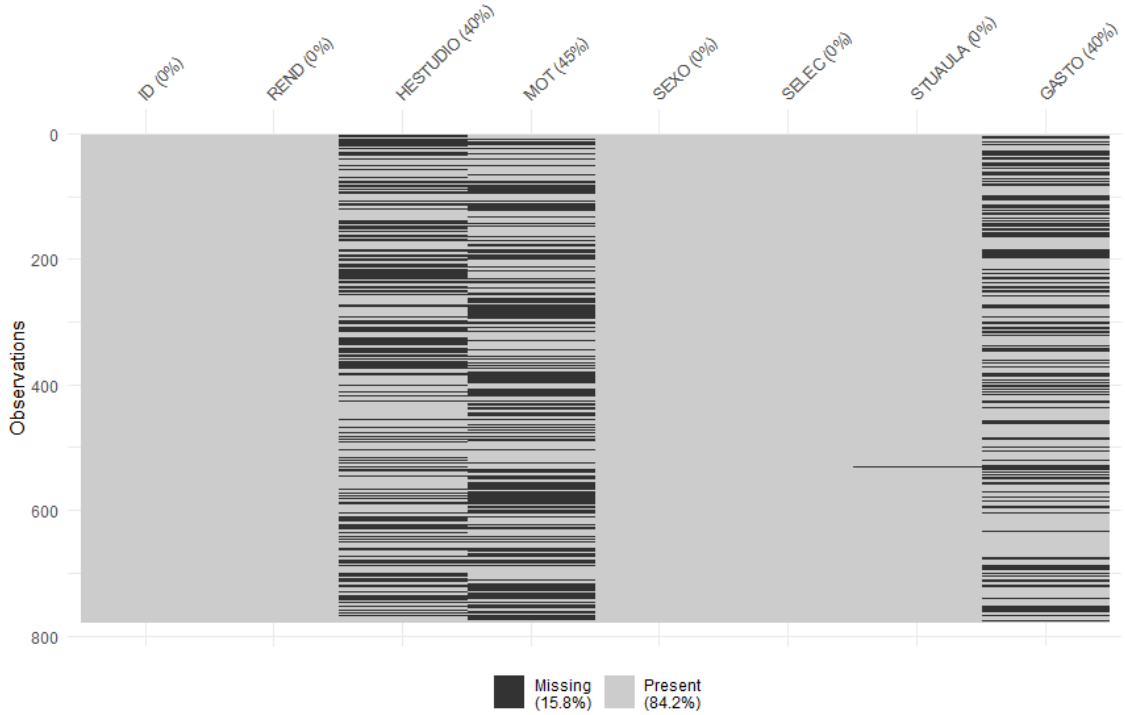
significativas en las horas de estudio entre los dos grupos?

En valores válidos, la media es de 51.28966; en perdidos, 30.75862. Sí hay diferencias significativas ($p<0.05$).

1.5) Para visualizar el patrón de valores perdidos univariado utiliza la función vis_miss()del

paquete naniar. Pega el grafico y comenta cuál es la variable del archivo que tiene más

valores perdidos (informa del porcentaje)



Como podemos observar, la variable MOT es la que cuenta con más valores perdidos (45%)

Mecanismos de valores perdidos. Abrir el archivo Ejercicio2.dat desde R

Este ejercicio tiene como objetivo poner en práctica tu comprensión sobre los mecanismos que generan los valores perdidos: MCAR, MAR y MNAR.

Tenemos dos variables, IQ = Cociente intelectual y REND = Rendimiento. Supongamos que 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 ver la recuperación de estos parámetros bajo MCAR, MAR y MNAR haciendo nuevamente cinco réplicas (o, si lo deseas, puedes hacer muchas más utilizando la estructura for) con cada uno de estos mecanismos. El objetivo es generar pérdida de datos en la variable rendimiento.

Condiciones: en todas las condiciones queremos un 50% de valores perdidos en Rendimiento.

Para MCAR generamos valores perdidos en Rendimiento a partir de una variable aleatoria uniforme (0,1) de forma que si el valor es < 0,50 tenemos un valor perdido en rendimiento y, en caso contrario, un valor válido.

Para MAR generamos valores perdidos en rendimiento de la siguiente manera. Para el primer cuartil de CI la probabilidad de valor perdido en Rendimiento = 0,80, para el segundo cuartil de CI la probabilidad de valor perdido en Rendimiento = 0,60, tercer cuartil de CI una probabilidad de 0,40 y para el cuarto cuartil de CI una probabilidad de 0,20.

Para MNAR generamos valores perdidos en rendimiento de la siguiente manera. Para el primer cuartil de rendimiento la probabilidad de valor perdido será de 0,80, probabilidad de 0,60 para el segundo cuartil de rendimiento, probabilidad de 0,40 para el tercer cuartil y probabilidad de 0,20 para el cuarto cuartil.

2.1) Justifica brevemente que esta forma de generar valores perdidos es, efectivamente, MCAR, MAR y MNAR.

MCAR: Sí se trata de este mecanismo de pérdida, dado que es completamente aleatorio, es decir, no dependiente de otras variables ni la misma.

MAR: Sí se trata de este mecanismo de pérdida, dado que no depende de la propia variable pero sí de las observadas (CI).

MNAR: Sí se trata de este mecanismo de pérdida, dado que depende de la propia variable.

2.2) Rellena la siguiente tabla (o si has utilizado más réplicas, pon directamente las medias obtenidas para el conjunto de réplicas):

MEDIA RENDIMIENTO					
	R1	R2	R3	R4	R5
MCAR	4.77	4.86	4.91	4.82	4.78
MAR	5.15	5.2	5.3	5.3	5.26
MNAR	5.53	5.56	5.69	5.59	5.61

Nota: el valor poblacional es $\mu = 4{,}89$

DESVIACIÓN TÍPICA RENDIMIENTO					
	R1	R2	R3	R4	R5
MCAR	2.3	2.35	2.22	2.3	2.24
MAR	2.34	2.4	2.31	2.29	2.27
MNAR	2.27	2.27	2.22	2.23	2.21

Nota: el valor poblacional es σ = 2,34

CORRELACIÓN REND-CI					
	R1	R2	R3	R4	R5
MCAR	0.57	0.5	0.46	0.48	0.44
MAR	0.54	0.5	0.48	0.47	0.44
MNAR	0.5	0.47	0.46	0.45	0.46

Nota: el valor poblacional es ρ = 0,50

2.3) Comenta brevemente los resultados encontrados.

En términos de consistencia, las simulaciones MCAR muestran una ausencia de sesgo para los 3 parámetros.

En MAR, se produce una sobreestimación de la media, aún más acusada en MNAR.

En cuanto a la desviación típica, observamos infraestimación sistemática en MNAR; este fenómeno quizá se deba a la influencia de la propia variable en la pérdida, tratando dichos datos como medidas repetidas (menor variabilidad).