EJERCICIOS OBLIGATORIOS PARA SUPERAR EL SEMINARIO. Parte II – IMPUTACIÓN MÚLTIPLE con mice

Recuerda entregar este Word y el script de R que hayas creado

El archivo de datos que vamos a abrir es *Regresion estudiantes TDAH.txt*. El archivo contiene una muestra de 70 niños diagnosticados de TDHA de cuarto, quinto y sexto de primaria. La VD es el aprendizaje (variable *aprendizaje*). Esta variable es la calificación sacada a partir de la valoración que da el tutor del niño al final del año académico, de su rendimiento escolar y de la opinión del orientador del centro (se escala de 0 a 10 la variable *aprendizaje*). Aproximadamente la mitad de los niños han aplicado durante un año un programa extraescolar de modificación de conducta en el aula y de habilidades sociales.

Los investigadores han recogido variables relevantes a partir de una revisión teórica exhaustiva que afectan al aprendizaje: la motivación por el estudio (escala 0 – 9), horas de estudio semanales (de 0 a 15), el salario de los padres (escala 0 = bajo a 4 = alto) y el nivel educativo de los padres (0 = bajo y 4 = alto). Estas variables se incluirán en el modelo. También está recogida la variable *programa* (0 = no aplica, 1 = sí aplica) que es la más interesante para los investigadores.

Realiza dos modelos de regresión. Uno utilizando lm desde R (que utiliza el método de eliminación por lista) y otro con el paquete mice utilizando imputación múltiple. Para la imputación múltiple puedes utilizar el método que viene por defecto ("pmm"), pero genera un total de *m* = 50 bases de datos (si quieres replicar tus resultados más tarde, debes fijar una semilla de aleatorización).

Rellena la tabla: Variable dependiente *Aprendizaje*

	Regresión con R función lm			
	B	S.E.	*p*	β
Intersección	-1.09	1.17	0.36	
Motivación	0.54	0.15	0	0.37
Horas est.	0.38	0.11	0	0.46
Salario padres	0.20	0.22	0.35	0.1
Estudios padr.	0.44	0.20	0.04	0.25
Programa	0.96	0.48	0.053	0.21
σ_e (residual)	1.46			
R^2	0.64			
Adj(R^2)	0.59			

	Regresión con mice			
	B	S.E.	*p*	β
Intersección	-1.38	0.86	<0.05	
Motivación	0.57	0.11	<0.05	***
Horas est.	0.42	0.08	<0.05	***
Salario padres	0.2	0.19	<0.05	***
Estudios padr.	0.28	0.18	<0.05	***
Programa	1.16	0.43	<0.05	***
σ_e (residual)	***			
R^2	0.73			
Adj(R^2)	0.71			

Comenta las principales diferencias que encuentras entre la regresión llevada a cabo con lm (eliminación por lista) y la que se calcula con mice (Imputación Multiple). Por ejemplo, comenta lo que ocurre en relación a los predictores significativos (usa un nivel de significación, α, de 0,05), el tamaño muestral que se utiliza en cada modelo, diferencias en las estimaciones de los coeficientes de regresión, diferencias en los errores típicos…

En lm, solo resultan ser significativos la motivación, las horas de estudio, y los estudios de los padres, mientras que con mice, son todos los predictores.

Por otra parte, los errores típicos son menores en mice, dado que se emplea un tamaño muestral menor en lm (70-27 = 43) frente a los 70 de casos completos.

En general, los coeficientes son mayores, por lo que podemos intuir que se ha capturado en mayor medida las contribuciones de cada variable en el modelo imputado.

Finalmente, el modelo mice presenta mayor ajuste.

Interpreta el coeficiente de regresión de la variable Programa en el modelo de regresión calculado por mice.

Se estima que los niños a los que se les ha aplicado el programa, obtendrán mayores puntuaciones en la valoración del aprendizaje por parte del tutor (aprendizaje).

¿Qué conclusión se hubiese sacado sobre el programa en el caso de haber trabajado con el método de regresión clásico?

La calificación en la valoración del aprendizaje está relacionada directamente proporcional, y en orden descendente de impacto, con las horas de estudio, la motivación y los estudios de los padres.

Establece un método diferente al que viene por defecto para validar los resultados (por defecto, mice usa "pmm"), pero puedes utilizar la regresión bayesiana (método "norm"). ¿Se replican los resultados? Haz que la secuencia de imputación sea en orden inverso al de la proporción de valores perdidos. Es decir, imputa primero las variables que tengan menos valores perdidos e imputa al final las que más tienen (esta es una recomendación general).

Sí se replican los resultados en "norm", incluida la magnitud de las betas. Mejora el ajuste ligeramente.

También se replican los resultados al introducir por orden las variables.

Por último, vamos a ajustar el modelo de regresión lineal con el paquete lavaan utilizando *full maximum likelihood*.

	Regresión con lavaan utilizando Full Maximum Likelihood			
	B	S.E.	*p*	β
Intersección	-1.29	0.77	0.1	-0.48
Motivación	0.55	0.11	0	0.36
Horas est.	0.42	0.07	0	0.51
Salario padres	0.2	0.17	0.2	0.08
Estudios padr.	0.27	0.17	0.1	0.13
Programa	1.07	0.41	0.01	0.20
σ_e (residual)	4.72e-7			
R^2	0.75			
Adj(R^2)	***			

Comenta la salida que te ofrece lavaan y los parecidos o diferencias que encuentras con el otro método modernos (imputación múltiple) que has ejecutado previamente con mice.

Recuerda que la teoría dice ambos métodos son equivalentes (es decir, deben arrojar resultados similares, puesto que ambos, con muestras suficientemente grandes, convergen), de modo que puedes enfocar tus comentarios un poco en relación a si crees que encuentras o no resultados parecidos. Por ejemplo, puedes comentar qué coeficientes son significativos con ambos métodos, cuáles no o cómo son los errores típicos aportados por ambos métodos.

En este nuevo modelo, obtenemos que, las variables que influyen en la predicción en orden descendiente de importancia son horas de estudio, motivación y programa. El hecho de que dichos resultados varíen con respecto a mice puede deberse al tamaño de muestra que manejamos, que impide que converjan ambos métodos.

A su vez, los errores típicos son menores con ml, y obtenemos un ligero mejor ajuste.