Trabajo datos faltantes

Juan David Castillo Garza

2022-10-06

La base de datos personal de justicia tiene la informacion de la tasa por 100 000 habitantes de policias, personal penitenciario, jueces profesionales o magistrados, y personal de la fiscalia en base al pais, la region-continente y subregion (Ademas un codigo IS03), pero pose datos faltantes en las tasas. El objetivo de este trabajo es a partir de la base de datos se debe realizar un ejercicio de imputación de los valores faltante, empleeando las funciones vistas diferentes a las de la librería mice y para ello se propone los siguientes puntos:

1) Describir la base de datos empleando Casos Completos y calculando medidas descriptivas y calculando la matriz de correlaciones entre las variables (Antes de ello, describir los valores faltantes de la base)
2) Usando la variable Region como variable de Clase de Imputación realizar un proceso de imputación mediante el método de Hot Deck. (Calcularnuevamente las medidas descriptivas del punto 1)
3) Aplique el método de imputación del vecino más cercano y calcule nuevamente las medidas descriptivas del punto 1
4) Aplique el método de imputación pmm y calcule nuevamente las medidas descriptivas del punto 1
5) Compare, de forma gráfica (ud decide cómo), los resultados de las medidas descriptivas calculadas en cada punto

Librerias utilizadas

Desarrollo

Para la region, la subregion, el codigo del pais, y la ciudad, no hay ningun valor faltante como se menciono arriba, pero para la tasa por 100 000 habitantes de: prison (personal penitenciario), police (personal de policia), profesional (jueces profesionales o magistrados) y prosecution (personal de la fiscalia) si hay valores faltantes. Por variable el porcentaje de valores faltantes con respecto a la totalidad de la base :

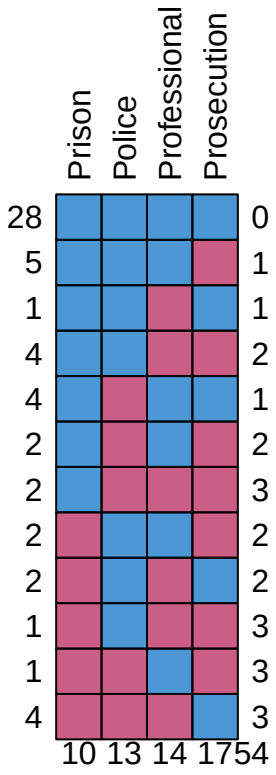
Tasa (100 00 habitantes)	% faltantes
Police	2.91%
Prision	2.23%
Profesional	3.13%
Prosecution	3.79%
Total	12.05%

De esta tabla, se puede observar que el 12.05% de las celdas totales de la base son faltantes, y el porcentaje de faltantes de cada tasa no supera el 4% de la base total, por lo que en principio no es un alto porcentaje de faltantes, algo a proponer aqui, seria ver el porcentaje de faltantes sin tener en cuenta el pais y el Iso3,

ya que realmente estas dos columnas son para identificar los paises, y si se tiene el pais se tiene el Iso3 y en sentido contrario

Con respecto a los patrones de valores faltantes, se puede ver lo siguiente:

- Hay 11 patrones de valores faltantes, de 56 paises 28 tienen informacion completa
- A 5 paises no se les tiene la informacion solamente de prosecution, a 1 de profesional ,4 de police y no hay paises a los que les falte solamente prision
- A 10 paises no se les tiene informacion en una pareja de tasas, de los cuales falta las 3 combinaciones de tasas con prosecution (4 paises les falta prosecution-profesional, a 2 prosecution-police y a 2 prosecution-prison) y a 2 les falta prison-profesional
- 8 paises se les tiene la informacion de una sola tasa, a 4 paises solamente se tiene la informacion de prosecution, a 2 de prision a 1 de police y 1 de profesional
- No hay ningun pais del cual no se tenga al menos la informacion de una tasa



Con respecto a la region, hay 4 las cuales : son Africa (con un pais), Americas (con 16 paises), Asia (con 6 paises) y Europa (con 33 paises)

Casos completos

Como se indico en la seccion anterior, hay 28 paises de los cuales se tiene informacion completa, lo que equivale al 50% de los paises de la base, asi que las siguientes estadisticas estan calculadas en base a ello

Resumen	Police	Prison	Professional	Prosecution
min	67.61	35.06	2.102	3.136
Q1	260.05	54.72	7.157	6.756
Q2	372.92	70.67	17.077	11.638
media	384.96	84.94	18.961	12.882
Q3	447.30	102.13	27.800	16.519
max	1117.97	240.38	48.886	29.886

Con casos completos se observa que la tasa x 100 000 habitantes de police es la que presenta los mayores valores como se esperaria y se mueve entre 67.61 a 1117.97 , algo a revisar con respecto a esto es tanto los paises que tienen las menores tasas y los que tienen las mayores, y cuanto se esperaria de esta tasa, Con respecto a Prison, es la segunda tasa con mayores valores moviendose entre 35.06 y 240.38, despues Professional moviendose entre 2.1 y 48.86 y finalmente la tasa con menores valores x 100 000 habitantes es prosecution moviendose entre 3.13 y 29.88. Para prison Profesional y prosecution vale la pena revisar lo mismo que con la tasa prison

La matriz de correlaciones entre cada una de las tasas es la siguiente:

	Police	Prison	Professional	Prosecution
Police	1.0000000	0.0285157	0.1064405	-0.0367905
Prison	0.0285157	1.0000000	-0.2210093	0.3389968
Professional	0.1064405	-0.2210093	1.0000000	0.4996661
Prosecution	-0.0367905	0.3389968	0.4996661	1.0000000

Con casos completos la matriz de correlaciones presenta valores principalmente bajos para las relaciones de Police con las otras 3 tasas, y para las 3 tasas hay valores bajos-medios en las correlaciones. La mayor correlacion es cercana al 50% y la menor al 3%

Hot deck

Ahora empleando Hot deck como metodo de imputacion, el metodo no funciona para solo un valor faltante, y es que de la region de Africa solo se tiene un pais y la tasa Police es faltante, por lo que no hay un candidato para donante utilizando la region como variable auxiliar, y esto se observa en el siguiente resumen con la base imputada

Resumen	Police	Prison	Professional	Prosecution
min	67.61	13.92	2.102	2.300
Q1	260.82	47.81	6.318	6.035
Q2	376.91	65.98	10.275	9.931
media	402.81	86.02	15.863	10.423
Q3	472.63	108.84	24.029	13.473
max	1117.97	318.36	48.886	29.886
NA	1	0	0	0

Sigue existiendo un NA para police, Ahora la escala de las variables no cambio y se siguen moviendo en los mismos rangos cada una, lo que si cambio es el valor de la media siendo mayor utilizando Hot Deck para Police junto con prision y menor para Professional junto con prosecution, y cada uno de los cuartiles

Con respecto a la matriz de correlaciones, al existir un valor faltante se puede tomar 2 opciones, utilizar casos completos osea que elimine la fila donde se tiene el valor todavia faltante o realizarla con las parejas de

las cuales se tiene la informacion completa o "eliminacion inteligente". En este caso presentamos la matriz de de correlaciones por eliminacion inteligente, ya que sigue conservando las propiedades de una matriz de correlaciones (es definida positiva, simetrica y su determinante es positivo), pero tambien se podria omitir la fila de la cual existe el faltante

	Police	Prison	Professional	Prosecution
Police	1.0000000	0.1097009	0.1016223	-0.0038584
Prison	0.1097009	1.0000000	-0.1600042	0.1027525
Professional	0.1016223	-0.1600042	1.0000000	0.4476751
Prosecution	-0.0038584	0.1027525	0.4476751	1.0000000

Los cuatro valores propios de la matriz de correlaciones son: 1.4580353, 1.1811272, 0.9439747 y 0.4168628 y su determinante es 0.6776699

Si se compara con casos completos, la relacion entre prision-police y professional-police cambia de signo, lo que significa que la asociacion esta cambiando, antes positiva debil, ahora negativa debil

Vecino mas cercano

Para el vecino mas cercano, para las 4 tasas que se necesita imputar, se usaron como variables regresoras las mismas 4 tasas y por ende la medida de distancia que calcula tiene en cuenta Police, Prison, Professional y Prosecution, pero ademas no sera con respecto a toda la base que calculara el vecino mas cercano que le pediremos, sino que tendra en cuenta la region, por lo que al igual que Hot Deck, no va a imputar ningun valor para Africa ya que para la distancia de Gower, no hay un valor real con cual comparar para imputar, y el K es igual a 1 por que en asia para prosecution solo hay 3 donantes, e incrementando a k=2 o k=3, consideramos que cambia mucho dependiendo que donante escoja la imputacion

Despues de realizar la imputacion, un resumen descriptivo de la imputacion es el siguiente:

Resumen	Police	Prison	Professional	Prosecution
min	67.61	13.92	2.102	2.300
Q1	245.76	46.33	5.949	6.148
Q2	368.93	67.17	11.554	10.151
media	388.01	90.14	17.021	11.039
Q3	456.41	108.84	26.278	14.185
max	1117.97	318.36	48.886	29.886
NA	1	0	0	0

Se observa que sigue existiendo un valor faltante para Police, el rango de las variables no cambia, y comparando con casos completos difiere menos que Hot deck para este caso. Con respecto a la matriz de correlaciones, al igual que en hot deck se emplea "eliminacion inteligente" donde la matriz de correlaciones conserva las propiedades que debe de tener, pero hay un cambio comparando con casos completos, y es que la relacion entre prison y police cambia de una asociacion positiva debil, a una asociacion negativa debil. Cuando solo se tiene como variable regresora la region, la matriz de correlaciones coincide en signos con casos completos

	Police	Prison	Professional	Prosecution
Police	1.0000000	-0.0047820	0.0462030	-0.0485711
Prison	-0.0047820	1.0000000	-0.2583031	0.2589768
Professional	0.0462030	-0.2583031	1.0000000	0.4574527
Prosecution	-0.0485711	0.2589768	0.4574527	1.0000000

Pmm

Para pmm, se imputo usando la Region como variable auxiliar para que el modelo que realice impute con valores de la misma region, por lo que igual que en los otros casos para africa sigue existiendo el valor faltante, y como medidas resumen se tiene lo siguiente

Resumen	Police	Prison	Professional	Prosecution
min	67.61	13.92	2.102	2.300
Q1	275.09	48.04	6.440	6.456
Q2	376.91	70.67	11.858	10.064
media	393.71	84.65	16.611	10.989
Q3	456.41	117.97	24.369	14.185
max	1117.97	318.36	48.886	29.886
NA	1	0	0	0

Se conserva el rango de las variables, y a difrerencia de con hot deck o Knn, de esta forma concervo los signos de asociacion entre las variables

	Police	Prison	Professional	Prosecution
Police	1.0000000	0.0885042	-0.0538085	-0.1050893
Prison	0.0885042	1.0000000	-0.1940851	0.0587390
Professional	-0.0538085	-0.1940851	1.0000000	0.5253888
Prosecution	-0.1050893	0.0587390	0.5253888	1.0000000

Parte grafica comparacion

Link pmm

https://github.com/DavidCastillo13/Faltantes/commit/a6044afd77b1215fb7ac7bbacb6c83c6109c3f2b