VALIDACIÓN DE DATOS FALTANTES EN ENCUESTAS AGROPECUARIAS

Jaime Pinto A.

La generación de Bases de datos, para los análisis a realizarse, requiere considerar los muchos criterios que se plantean para depurar la base, que en algunos casos debido a diversos factores como una no adecuada supervisión o control de relevamiento de datos, critica y trascripción, hacen que haya datos faltantes en la base de datos.
Esta omisión se debe subsanar, mediante procedimiento de imputación que sugieren o indican cual pudiese ser el dato faltante.

La imputación generalmente se utiliza para asignar valores a los elementos faltantes. Frecuentemente se asigna un valor de reemplazo al valor faltante mediante un valor de otra persona en la encuesta, similar a la que no responde al elemento con respecto a otras variables. Al usar la imputación, hay que crear una variable adicional en el conjunto de datos que indique si la respuesta fue medida o imputada.

Algunos métodos de imputación son los siguientes:

IMPUTACIÓN DE LA MEDIA POR CELDA

Los productores que responden se dividen en clases (celdas) con base en variables conocidas como en los ajustes de clases de ponderación. Entonces, sustituimos el promedio de los valores de las unidades que responden y que están en la celda c, en cada valor faltante.
La imputación de la media por celda, supone que los elementos faltantes son faltantes completamente al azar dentro de las celdas.

Para este extractamos parte de la boleta y los datos obtenidos en una encuesta agropecuaria, realizada en una zona de producción.
Parte de la Boleta: *ENCUESTA AGROPECUARIA*

IDENTIFICACIÓN.-
P1.- Uso: Agrícola (1) Pecuaria (2)
P2.- Unidad de Observación:Grande (1) No grande (2)

UNIDAD DE PRODUCCIÓN.-
P3.- Cual es la Superficie total de su Unidad de Producción Agropecuaria (UPA)?

SUPERFICIE		
Cantidad	Unidad	Uso de oficina

P4.- Cual es el Numero de Parcelas de su Unidad de Producción Agropecuaria?

P5.- Numero de personas que trabajaron en la Unidad de Producción Agropecuaria.?

Conjunto de datos para explicar el método:

Productor	**P2**	**P3**	**P4**	**P5**
	Unidad de Observación	**Sup. Total de su UPA (Has.)**	**Numero de Parcelas de la UPA**	**Nro. de personas que trabajaron en la UPA**
1	2	0.25	1	2
2	2	0.90	4	
3	2	0.25	3	
4	2	0.25	3	2
5	2	0.8	3	1
6	1	4.00	9	10
7	1	6.00	7	14
8	2	0.30	6	1
9	2	0.25	4	2
10	2	0.10	5	
11	2	0.85	7	1
12		0.25	2	3
13	2	0.90	2	1
14	1	5.00	12	10
15	1	4.00	14	16

Para esto construimos cuatro celdas usando las variables Número de Parcelas y Unidad de Observación (Grande y No Grande)

NUMERO DE PARCELAS

UNIDAD DE OBSERVACIÓN	Menor e igual a 5	Mayor a 5
Grande	Productores	Productores 7, 6, 14, 15
No Grande	Productores 1, 12, 13, **3**, 4, 5, **2**, 9,**10**,	Productores 8,11

Revisando la Base de Datos los Productores 2, 3 y 10 no tienen el valor para los "Numero de personas que trabajaron en la Unidad de producción Agropecuaria", recibirán el promedio de los restantes productores donde se encuentran, es decir de la columna "Menor e igual a 5 Parcelas" La media para cada celda después de la imputación es igual a la media de quienes respondieron.

Realizando operaciones:

Promedio = (Suma de datos de la variable, de los que respondieron) / Total de personas que respondieron.

Promedio = (2+3+1+2+1+2) 1 6 = 11/ 6 = **1,83**

Los valores a colocar en la variable "Numero de personas que trabajan en la Unidad de Producción Agropecuaria" para los Productores 2, 3 y 10 son en cada caso " **2 Personas**".

Imputación Deductiva.-

Algunos valores se pueden asignar en la edición de datos, mediante las relaciones lógicas entre las variables.

En la base o conjunto de datos la persona 12 no respondió si es "Productor Grande", pero como respondió que su Superficie Total de Unidad de Producción Agropecuaria es 0,25, la respuesta de Unidad de Observación debe colocarse 2 .(por corresponder a Productor No Grande).

Hago mención a otros métodos de Imputación, que también se utilizan como:
Imputación Hot-deck

En esta Imputación, al igual que en la imputación de la media por celda y los métodos de ajuste con ponderación, las unidades de la muestra se dividen en clases. El valor de una de las unidades de la clase y que responde se sustituye en cada respuesta faltante. Con frecuencia, los valores de un conjunto de elementos faltantes relacionados entre si se toman del mismo donante para preservar algunas de las relaciones multivariadas.
El nombre hot deck proviene de los días en que los programas de computadora y los conjuntos de datos se perforaban en tarjetas.

Imputación cold-deck.

En la imputación cold-deck, los valores se asignan a partir de una encuesta anterior o de otras informaciones, como datos históricos. Hay poca teoría para este método
Como en el caso de la imputación de hot-deck, la imputación de cold-deck no garantiza la eliminación del sesgo de selección.

Imputación por regresión.

La imputación por regresión, predice el valor faltante usando una regresión del elemento de interés sobre las variables observadas para todos los casos. Una variante es la imputación por regresión estocástica, donde el valor faltante se reemplaza mediante el valor predicho a partir del modelo de regresión, más un término de error generado aleatoriamente.

[Kennedy] leía una de cada 50 cartas de las treinta mil que llegaban semanalmente a la Casa Blanca, al igual que un resumen estadístico de todo el lote, aunque él sabía que con frecuencia era tan organizado y no representativo como las estacas de Pensylvania Avenue.

Theodore Sorensen , Kennedy

Índice