Tema 24: Observaciones perdidas

The best solution to handle missing data is to have none.
R.A. Fisher[1]

□ Peter B. Mandeville

Las observaciones perdidas son parte de casi toda la investigación, por lo cual de vez en cuando todos tenemos que decidir cómo manejarlas. En estadística, la imputación sustituye algún valor para una observación perdida.

En las encuestas en las que se anticipan problemas con observaciones perdidas se deben utilizar variables altamente correlacionadas con las variables, en las que se espera encontrar observaciones perdidas, porque entonces es relativamente fácil predecir las observaciones perdidas.[2-4]

Hay varios tipos de observaciones faltantes:

- Faltantes completamente al azar (*Missing Completely at Random*, MCAR). MCAR se refiere a los datos, cuando el proceso de la pérdida no depende de las otras variables explicativas en el conjunto de datos. Con MCAR, cualquier observación tiene igual probabilidad de perderse. Esto significa que los datos se recolectaron al azar, y no dependen de ninguna otra variable en el conjunto de datos. Ejemplos: las observaciones que pueden perderse, porque el equipo no funcionaba correctamente, el clima era terrible, la gente se enfermó, o los datos no se capturaron correctamente.

  Los datos sobre los ingresos familiares no se considerarían MCAR, si las personas con bajos sueldos fueran menos propensas a reportarlos que las personas con mayores ganancias, debido a que la probabilidad de la pérdida de observaciones se correlaciona con el ingreso. Con los datos MCAR se puede perder poder, pero los parámetros estimados no están sesgados.

- Faltantes al azar (*Missing at Random*, MAR). MAR significa que las observaciones faltantes están condicionadas por otras variables explicativas en el conjunto de datos, aunque no con la variable de respuesta.[2-4] Los valores faltantes no se distribu-

yen al azar con todas las observaciones, pero se distribuyen al azar dentro de una o más submuestras. MAR es mucho más común que MCAR.

Por ejemplo, las personas que sufren depresión podrían estar menos inclinadas a reportar sus ingresos, por lo que reportarlos se relaciona con la depresión. Las personas deprimidas también pueden tener ingresos más bajos y, por lo tanto, una alta tasa de observaciones perdidas, el ingreso medio calculado podría ser menor de lo que es en un grupo sin observaciones perdidas. Sin embargo, si dentro de los pacientes con depresión la probabilidad de reportar ingreso no se relaciona con el nivel de ingresos, se considera que los datos son MAR, no MCAR.

- Faltantes no al azar (*Missing not at random*, MNAR). Si los datos no son MAR o MCAR, entonces se clasifican como MNAR. Por ejemplo, si está estudiando la salud mental de personas diagnosticadas como deprimidas y éstas son menos propensas que otras a informar de su estado mental, no hay una pérdida de las observaciones al azar. Es evidente que la puntuación media del estado mental de las observaciones disponibles no será una estimación sin sesgo de la media que habría obtenido con los datos completos. Lo mismo ocurre cuando las personas con bajos ingresos son menos propensas a reportar sus ingresos. Cuando se tienen datos MNAR, la única manera de obtener una estimación de los parámetros sin sesgo es modelar el proceso de perder de observaciones.

Hay una serie de formas alternas de tratar con observaciones perdidas.

- Supresión de la lista: cuando el número de casos con observaciones perdidas es pequeño (menos de 5% en muestras grandes), es común simplemente suprimir estos casos, y ejecutar el análisis sobre las observaciones que quedan. Este enfoque se denomina *supresión de la lista o análisis de los casos completos*. Aunque supresión de la lista puede resultar en una disminución sustancial en el tamaño de la muestra disponible para el análisis, tiene ventajas importantes. Bajo el supuesto de que los datos son MCAR, el procedimiento es válido y da lugar a estimaciones de los parámetros sin sesgo, pero con una pérdida de poder. Cuando las observaciones perdidas no son MCAR, y las observaciones perdidas muy pocas veces son MCAR, los resultados están sesgados.[2-4]
- Sustitución de la media: un viejo procedimiento es sustituir las observaciones perdidas con la media. No añade ninguna información nueva, la media global es la misma y se subestima el error.
- Imputación múltiple (*Multiple Imputation*, MI). En MI se generan valores imputados sobre la base de los datos existente. En lugar de utilizar un solo valor para cada observación perdida (imputación simple o sencilla), el procedimiento MI reemplaza cada observación perdida con un conjunto de valores plausibles que representan la incertidumbre sobre el valor apropiado a imputar.[5,6] Con MI se toman los valores pronosticados, y se agrega un componente de error al azar de la distribución de los residuos. Esto subestima los errores estándar, y se repite la imputación varias veces, cinco por defecto, y genera varios conjuntos de datos cuyos coeficientes varían de un conjunto a otro. Los conjuntos de datos imputados se analizan mediante el procedimiento habitual para los datos completos. Finalmente, se combinan los resultados de estos análisis, al producir inferencias estadísticas válidas que reflejen de manera adecuada la incertidumbre, debido a las observaciones perdidas.

En mi experiencia con SPSS, sin duda muy limitada, he descubierto que éste automáticamente utiliza supresión de

la lista, cuando hay observaciones perdidas, que se puede sustituir la media, y se necesita comprar el módulo de *Missing Values* para efectuar MI. R también suprime la lista automáticamente, cuando hay observaciones perdidas, y también se puede sustituir la media en R. Finalmente, MI está implementado en el paquete *mice,*[5] que es gratuito, y se demuestra el uso más básico de *mice* a continuación.

```
# crear semilla para la generación de números
aleatorios
> sample(0:9,3,replace=T)
[1] 9 7 9
# aleatoriamente crear un conjunto de datos
> set.seed(979)
> col1 <- rep(c(«M»,»H»),10)
> col2 <- rep(c(«A»,»B»,»C»,»D»),5)
> col3 <- rnorm(20,mean=15,sd=2.5)
> col4 <- rexp(20)
> dat0 <- data.frame(col1,col2,col3,col4)
> head(dat0)
  col1 col2  col3   col4
1    M    A 15.74 0.3615
2    H    B 16.33 4.0569
3    M    C 13.14 1.0289
4    H    D 12.53 2.5323
5    M    A 14.11 2.6804
6    H    B 16.63 0.1244
# declarar los factores
> dat1 <- dat0
> fac <- c(1,2)
> for(j in 1:length(fac)) dat1[,fac[j]] <-
as.factor(dat0[,fac[j]])
> str(dat1)
'data.frame':   20 obs. of  4 variables:
 $ col1: Factor w/ 2 levels «H»,»M»: 2 1 2 1 2 1 2 1 2 1 ...
 $ col2: Factor w/ 4 levels «A»,»B»,»C»,»D»: 1 2 3 4 1 2
3 4 1 2 ...
 $ col3: num  15.7 16.3 13.1 12.5 14.1 ...
 $ col4: num  0.361 4.057 1.029 2.532 2.68 ...
# aleatoriamente crear una lista de 15 celdas para
borrar
> set.seed(979)
> col <- sample(4,15,replace=T)
> row <- sample(20,15,replace=T)
> miss <- data.frame(row,col)
> head(miss)
  row col
1  18   3
2   8   1
3   3   3
4  12   2
5   4   1
6  11   4
# asignar NA a la celdas seleccionadas
> dat2 <- dat1
> for(j in 1:nrow(miss)) dat2[miss[j,1],miss[j,2]] <- NA
> head(dat2)
  col1 col2  col3   col4
1    M    A 15.74 0.3615
2    H    B 16.33     NA
3 <NA>    C    NA 1.0289
4 <NA>    D 12.53 2.5323
5    M    A 14.11 2.6804
6    H <NA> 16.63 0.1244
# número de celdas con datos faltantes en cada
columna
> na <- function(x){sum(is.na(x))}
> apply(dat2,2,na)
col1 col2 col3 col4
   4    4    4    3
# proporción de las filas con datos faltantes
> tmp1 <- apply(dat2,1,na)
> tmp2 <- ifelse(tmp1==0,0,1)
> sum(tmp2)/length(tmp2)
[1] 0.65
```

65% de las filas tiene datos faltantes, por lo cual se necesita utilizar imputación múltiple.

```
# imputación múltiple
> library(mice)
> dat3 <- mice(dat2,seed=979)
> dat4 <- complete(dat3)
> head(dat4)
  col1 col2  col3   col4
1    M    A 15.74 0.3615
2    H    B 16.33 0.3046
3    H    C 11.82 1.0289
4    M    D 12.53 2.5323
5    M    A 14.11 2.6804
6    H    B 16.63 0.1244

# número de datos faltantes en cada columna
> apply(dat4,2,na)
col1 col2 col3 col4
   0    0    0    0

# escribir los datos imputados al directorio de trabajo en formato CSV
> write.csv(dat4,file=»dat3.csv»,row.names=F)
```

Publicaciones introductorias sobre el tema de las observaciones perdidas incluyen esta información.[1,6]

Referencias

1. Patrick E. McKnight, Katherine M. McKnight, Souraya Sidani y Aurelio José Figueredo. 2007. Missing Data: A Gentle Introduction. Methodology in the Social Sciences. The Guilford Press, New York, NY, USA.
2. Schafer, J.L. 1997. Analysis of incomplete multivariate data. Chapman & Hall, London. Book No. 72, Chapman & Hall series Monographs on Statistics and Applied Probability.
3. Rubin, D.B. 1987. Multiple imputation for nonresponse in surveys. John Wiley & Sons, New York.
4. Little, R.J.A. and D.B. Rubin. 1987. *S*tatistical analysis with missing data. John Wiley & Sons, New York.
5. Van Buuren, S., Groothuis-Oudshoorn, K. 2009. MICE: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, forthcoming. http://www.stefvanbuuren.nl/publications/MICE in R - Draft.pdf
6. Allison, P.D. 2001. Missing data. Thousand Oaks, CA: Sage Publications.