Valores faltantes

Definición y tipos

Leticia Gracia Medrano

IIMAS -UNAM

Septiembre 2020

¿Qué son?

Los valores faltantes son observaciones que en un principio se tenía la intención de hacerlas, pero por distintas razones no se obtuvieron.

Puede ser que un individuo no responda todas las preguntas, o que en un estudio longitudinal un individuo no siga siendo captado hasta al final.

Consecuencias

Actuar como si no faltaran observaciones, lleva a errores en la inferencia.
La pérdida de información puede afectar:

- las propiedades de los estimadores (sesgos)
- la potencia de las pruebas
- las longitudes de los intervalos de confianza

¿Qué se hace?

Se necesitará hacer suposiciones adicionales, que permitan **inferir sólo con la información observada** (incompleta), estas suposiciones serán en términos de la relación de los valores faltantes y los valores que hubieran tomado y del comportamiento (patrón) de los datos no-observados.

Determinar si estas suposiciones son plausibles, no puede hacerse en función de los datos que se tienen. Se necesitará conocer muy bien el contexto de la investigación.

(No es como en regresión, donde sí se puede verificar el supuesto de homoscedasticidad a partir de los datos que se tienen, se hace a través de los residuales).

Es muy conveniente hacer un análisis de sensibilidad para explorar que pasaría si se cambian las suposiciones.

Notación

Denotando a Y como la información con intención de ser recoletada.

$$Y = \{Y_o, Y_m\}.$$

Dónde Y_o es la información observada y Y_m la información perdida. El indicador de información perdida M se define como:

$$M = [M_{ij}] = \begin{cases} 1 & \text{si} \quad Y_{ij} \quad \text{no se observa} \\ 0 & \text{si} \quad Y_{ij} \quad \text{sí se observa} \end{cases}$$

M está en relación a Y.

Probabilidad de pérdida

La validez del análisis dependerá del mecanismo de pérdida, que está definido por

$$P(M|y_o, y_m)$$

Observaciones Perdidas Completamente al Azar (MCAR)

Cuando la probabilidad de pérdida no depende de lo observado ni de lo no observado

$$P(m|y_o, y_m) = Pr(m)$$

En este caso puede hacerse la inferencia como si no hubiese faltantes.

Ejemplos MCAR

- Ejemplo una muestra de laboratorio accidentalmente cae y se rompe.
- Descompostura de los aparatos de medición, que impide hacer el registro ese día.
- Una persona puede desconocer si ha está vacunada contra influenza o no.
- la letra del encuestador puede ser ilegible.

No es tan fácil

Algunas pareden MCAR pero analizándolas bien, podrían no serlo. Aquí dos ejemplos:

- En un estudio longitudinal alguien no prosigue pues tiene un accidente ("caida bajo el autobús"), si el estudio fuera de avance en aprendizaje de un idioma parece que no continuar en el estudio, no tiene nada que ver con la variable de aprendizaje, pero si se trata de un estudio psiquiátrico sobre depresión, podría ser que el sujeto no esté respondiendo al tratamiento y entonces la no respuesta (accidente) si tiene que ver con la variable respuesta.
- Un cuestionario acerca de seguridad no es respondido pues es robado en la oficina de correos, esto NO es aleatorio, pues probablemente esté relacionado con la variable zona dónde la oficina está localizada.

Observaciones Perdidas al Azar (MAR)

En este caso el mecanismo de pérdida no depende de la información perdida, pero si de la observada.

$$P(m|y_o, y_m) = P(m|y_o).$$

Viendo la siguiente tabla:

	variables					
unidad	1	2	3	4	5	6
1	1	3	4.3	3.5	1	4.46
2	1	3	?	3.5	?	?

Las unidades 1 y 2 tienen los mismos valores observados, dados estos valores observados, bajo el supuesto de mecanismo de pérdida MAR, las variables 3,5 y 6 de la unidad 2 puede considerarse que tienen la misma distribución que las variables 3,5 y 6 de la unidad 1.

Ejemplos MAR

Si en un estudio se tienen datos de ingreso y categoría de pago de impuestos. Es común que los que ganan más se niegan a revelar su ingreso.
Si se conoce la categoría de pago de impuestos de todos los individuos, la no respuesta acerca del ingreso es MAR, el mecanismo de pérdida depende de la categoría de pago de impuestos (que si es observada), así que la pérdida del dato de ingreso NO depende de valor del ingreso mismo (que no se observó).

Promediar los datos completos llevaría a subestimar el ingreso. Para estimar convendría primero calcular el promedio en cada categoría de impuestos, dado que dentro de cada categoría la pérdida es aleatoria, el promedio es un estimador válido y luego para estimar el ingreso medio de toda la población, se combinan estos estimadores en un promedio ponderando de manera proporcional al tamaño de las categorías de impuestos.

Observaciones NO Perdidas al Azar (MNAR)

Se dice que el mecanismo de pérdida NO es ignorable si los valores perdidos dependen de los valores no observados.

$$P(m|y_o, y_m) = Pr(m|y_m).$$

Ejemplos MNAR

- Cuando se preguntan datos de ingreso, los que ganan mucho o muy poco tienden a no responder.
- Cuando preguntamos cuantas veces se sienten deprimidos a la semana, los de valores altos tienden a no contestar
- Cuantas ayudas recibe la familia del gobierno, tienden a no contestar los que tienen más ayudas.
- Se registra el peso sólamente de aquellos que tienen sobrepeso.

Reporte de valores faltantes

- Porcentaje de valores faltantes para cada renglón
- Porcentaje de valores faltantes para cada columna
- Porcentaje de valores faltantes en total
- Reportes de tablas, por ejemplo si el porcentaje de faltantes es el mismo en el grupo de hombres que en el grupo de mujeres, qué pasa por grupos de edad o por estrato socioeconómico, etc.
- Reportar si hay algún patrón sistemático en los casos completos y/o los casos incompletos

Tratamiento de valores faltantes

Delete cases Deshacerse de los casos con algún valor faltante. Si no se trata de un MCAR este método causa sesgo. Causa también pérdida de potencia de las pruebas. A veces se tira demasiada información.

Imputar El valor faltante se rellena con algún valor. Hay muchas formas de hacerlo.

media El valor faltante se rellena con la media de los casos completos. (Genera sesgo en cualquiera de los tres tipos de valores faltantes)

hot deck El valor faltante se rellena con los valores de otro caso similar, pero que no pertenece a la muestra, por lo que se debe tener un"montoncito" de información reservado por si acaso hay valores faltantes en la muestra. (Genera sesgo en cualquiera de los tres tipos de valores faltantes)

modelo predicción Se rellena con la predicción de algún modelo por ejemplo uno de regresión lineal, que toma en cuenta las variables observadas. (resulta insesgada bajo MCAR y MAR, pero potencialmente sesgada bajo NMAR)

Ejemplos de imputación con modelo de predicción

1. Se rellena con la predicción de la media de un modelo de regresión, por ejemplo $ingreso_i = \hat{\alpha} + \hat{\beta}_i * edad_i$
2. Se rellena con la predicción de una observación cualquiera del modelo de regresión, por ejemplo $ingreso_i = \hat{\alpha} + \hat{\beta}_i * edad_i + e_i$, con e_i es una observación de una normal $N(0, \sigma)$.
3. En los casos MNAR se puede modelar la función de distribución conjunta de M y Y, denotada por $F_{MY}(M_{ij}, Y_{ij})$.

Este modelo no es tan sencillo y se utiliza el método iterativo de Esperanza y Maximización EM*.
También puede usarse un modelo de patrones: se clasifican a los casos según sus valores faltantes y sus valores observados y se hace un modelo en cada grupo para hacer el rellenado grupo por grupo.

(*Método EM: Suponiendo que los datos Y, tienen por ejemplo una distribución $NMV(\mu, \Sigma)$, con la información completa estimo μ y Σ, uso el modelo de predicción para rellenar los datos, mismos que se utilizan para re-estimar a μ y Σ, y de nuevo cambio el rellenado con las estimaciones nuevas y así sucesivamente.)

Ver artículo Schafer y Graham 2002. "Missing Data: Our View of the State of the Art", en Psycological Methods vol 7, 2, 147-177.

En este caso se generan varios conjuntos de datos *rellenados*, usando algún método de imputación simple, se tienen ahora D conjuntos.

Para cada conjunto se estima el parámetro de interés digamos γ , se tienen entonces las estimaciones

$$\hat{\gamma}_1, \ldots, \hat{\gamma}_D,$$

cada uno con error estándar $\sqrt{U_i}$ con $i = 1, \ldots, D$. El estimador global que se usa es

$$\bar{\gamma} = \sum_{i=1}^{D} \gamma_i / D$$

cuya incertidumbre o varianza es

$$T_D = \bar{U}_D + (1 + 1/D)B_D$$

con

$$\bar{U}_D = \sum_{i=1}^{D} U_i/D \quad \text{(Promedio de varianzas)}$$

y

$$B_D = \sum_{i=1}^{D} (\hat{\gamma}_i - \bar{\gamma})^2/(D-1) \quad \text{(Varianza entre imputaciones)}$$