Tratamiento de datos faltantes

Causas por las que pueden faltar datos

Errores en la toma de datos

Por ejemplo, se dejan campos en blanco que deberían contener información

Rechazo a responder

Algunas preguntas en un cuestionario pueden resultar muy personales o quien responde puede no tener los conocimientos suficientes

Respuestas inaplicables

Tipos de datos faltantes

Outliers tratados como datos faltantes

Cuando se conocen los límites de las diferentes variables del dataset, los datos que caen fueran del rango definido se deben etiquetar como faltantes.

Datos faltantes completamente al azar

La probabilidad de que la variable Y tenga un dato faltante es independiente de X.

Los datos existentes en Y son una muestra al azar de los valores de Y

Datos faltantes al azar

La probabilidad de que la variable Y tenga un dato faltante depende de X, pero no de Y.

Es decir, el patrón de los datos faltantes se puede predecir a partir de otras variables de la base de datos.

Tipos de datos faltantes

Datos faltantes que dependen de un predictor no observado

El dato faltante se podría estimar a partir de otra variable, pero que no fue registrada.

Se puede intertar modelar la variable con datos faltantes a partir de las presentes en el dataset.

Datos faltantes que dependen de la misma variable

Por ejemplo, las personas con altos ingresos es más probable que no quieran informar sus ingresos.

Métodos para tratar con datos faltantes

Utilizar únicamente los casos completos

Eso sólo se puede utilizar cuando los datos faltantes son del tipo competamente al azar.

Se puede considerar también cuando la proporción de casos con datos faltantes es chica.

En cualquier otro caso se introducirá un sesgo.

Eliminar de a pares

Eliminar casos o variables seleccionados

Si los datos faltantes se concentran en una clase, puede eliminarse la clase completa, o si una variable tienen una alta proporción de datos faltantes puede eliminarse la variable

Generalmente no es una práctica razonable. Puede haber una razón para que una clase o variable concentre datos faltantes.

Métodos para tratar con datos faltantes

Imputar los datos faltantes

Se estiman los datos faltantes a partir de los valores válidos de otras variables

Sustitución de casos

Sustitución por la media / mediana

Imputación “cold deck”

Imputación “hot deck”

Imputación múltiple

Procedimientos basados en modelos

Imputación por regresión / regresión estocástica

Métodos para tratar con datos faltantes

Sustitución de casos

Se reemplaza un caso que tiene datos faltantes, por otro que no estaba incluído en la muestra o que se había dejado fuera del estudio.

Sustitución por la media / mediana

Se reemplaza el dato faltante por la media de los datos presentes, o una muestra de ellas.

Es una técnica de uso frecuente pero se subestima la varianza, se puede modificar la distribución de los datos, y al repetir valores se deprime la correlación entre variables.

Una alternativa que a veces se puede considerar es la sustitución por la media/mediana de una clase (pero los problemas mencionados persisten)

Métodos para tratar con datos faltantes

Imputación “cold deck”

Se toman valores de otras bases de datos o se calculan a partir de relaciones también obtenidos de otras fuentes.

Hay que determinar si son datasets comparables.

Puede introducir distorsiones importantes, casi no se usan aislados, sino en combinación con otros métodos.

Imputación “hot deck”

Se reemplazan los valores faltantes por valores de casos similares, por ejemplo, utilizando KNN (K vecinos más próximos)

Métodos para tratar con datos faltantes

Imputación múltiple

El registro conteniendo data faltantes se reemplaza por varias imputaciones, con el mismo o diferentes métodos. En el primer caso es necesario introducir ruido aleatorio para no tener casos idénticos

Imputación por regresión / regresión estocástica

Es un método simple, pero tiene algunas desventajas.

Se asume que hay una correlación significativa entre las variables que se usan para la estimación de los datos faltantes.

Se refuerzan las relaciones que ya existen entre los datos.

El valor estimado puede escaparse de los rangos establecido para las variable con datos faltantes.

El uso de la regresión estocástica soluciona algunos de estos problemas

Herramientas (muy) básicas en R

```
# primero creamos una matriz con datos faltantes
a <- 1:50 + runif(50,-5,5)
b <- a*0.3 + 4 + runif(50, -5,5)
c <- 25:-24 + rnorm(50,0,6)
a[sample(1:50,5)] <- NA
b[sample(1:50,5)] <- NA
c[sample(1:50,10)] <- NA
abc <- cbind(a,b,c)

# Podemos obtener un dataset "limpio" (eliminación de casos completos)
na.omit(abc)
attr(na.omit(abc), "na.action")
# Información sobre datos faltantes
is.na(abc)
which(is.na(abc))
apply(abc,2,function(x) which(is.na(x)))
apply(abc,2,function(x) table(is.na(x)))

cor(abc)
cor(abc, use="complete.cases")
cor(abc, use="pairwise.complete.obs")
```

Paquetes de R para manejo de datos faltantes: imputation, mitools, Amelia II