A1: Preproceso de datos

Enunciado

2019.1

Índice

1. Carga del fichero **2**

2. Breve inspección de los datos **2**

3. Formato de las variables cuantitativas **3**

4. Valores extremos **3**

5. Valores del resto de variables cuantitativas **3**

6. Inconsistencias **3**

7. Formato de las variables cualitativas **3**

8. Imputación **3**

9. Tabla resumen de las variables cualitativas **4**

10. Tabla resumen de las variables cuantitativas **4**

11. Grabar a un archivo **4**

Introducción

La obesidad es un factor de riesgo para el cáncer de mama en mujeres premenopáusicas. La literatura existente ha encontrado que la obesidad es un factor de riesgo, porque puede aumentar los niveles de estrógenos en las mujeres. Concretamente, un biomarcador del nivel de estrógeno, serum estradiol, es un factor de riesgo identificado para el cáncer de mama. Para evaluar este tipo de relaciones, unos investigadores estudiaron un grupo de 211 mujeres en edad premenonáusica en USA, 151 de las cuales eran negras (African American) y 60 blancas (Caucasian). La adiposidad se cuantificó con dos medidas diferentes: BMI y WHR. Se midió también el nivel de estradiol a partir de un análisis hormonal. También se incluyen otros factores de riesgo como el número de hijos y edad de la mujer. Concretamente, el archivo incluye los siguientes atributos:

- Id: identificador.
- Estrad (serum estradiol): medida del estradiol (analítica hormonal).
- Ethnic (ethnicity): African American o Caucasian.
- Entage: edad de la persona.
- NumChild: número de hijos.
- Agefbo: edad a la que la persona ha tenido el primer hijo.
- Anykids: 1 si ha tenido hijos, 0 si no ha tenido.
- Agemenar: edad de la menarquia (primera menstruación).

- BMI: medida de la adiposidad general. Corresponde al ratio $peso(kg)/Altura^2(m)$.
- WHR: medida de la adiposidad abdominal. Corresponde al ratio $cintura/cadera$.
- Area: rural (1) o urbana (0).

El archivo se denomina *DataEstradiol.csv*, y contiene 211 registros y 11 variables.

Preprocesamiento de los datos

El objetivo concreto de esta actividad es preparar el archivo para su posterior análisis. Para guiar la actividad, sugerimos qué tipos de preprocesamiento hay que hacer para resolver satisfactoriamente esta actividad. En primer lugar, se presentan los criterios a seguir para realizar el preprocesamiento. A continuación, explicamos cuáles son los pasos necesarios que se deben seguir.

Criterios de preprocesamiento

Cuando se realiza un preprocesamiento, el analista decide como estandarizar/normalizar las variables. Por ello, establece unos criterios, que idealmente se escriben para homogeneizar versiones o para posteriores preprocesados que aparezcan sobre nuevos datos. A continuación, indicamos los criterios a seguir:

- El punto (.) es el separador decimal de cualquier variable numérica.
- Los valores de la variable Ethnic se estandarizan en "African American" y "Caucasian".
- Los valores de la variable Anykids deben ser "Yes" (valor 1 en el archivo), y "No" (valor 0).
- Los valores de la variable Area son "Rural" (1) o "Urban" (0).
- Hay que tener en cuenta las inconsistencias entre los valores Anykids y Numchild. En caso de inconsistencia, prevalece el valor de Numchild. Y por lo tanto, hay que ajustar Anykids adecuadamente.
- Hay que tener en cuenta las posibles inconsistencias entre Entage y Agemenarq. Si Agemenarq es superior a EntAge, probablemente se debe a un error al introducir los datos y por tanto, hay que intercambiar los valores.
- Agefbo debe ser superior a Agemenarq si la mujer ha tenido algún hijo. Toma el valor 0 en caso contrario.
- En caso de realizar imputaciones de valores, se debe realizar la imputación con los 3 vecinos más cercanos, usando sólo las variables cuantitativas y haciendo la imputación a partir de datos de la misma etnia.

Pasos a seguir:

1. Carga del fichero

Abrir el fichero datos en R y realizar una breve descripción del mismo, donde se indiquen el número de registros, el número de variables y el nombre de las mismas. Se recuerda que antes de cargar el archivo, hay que inspeccionar qué tipo de formato csv se trata para que su lectura sea apropiada.

2. Breve inspección de los datos

Mostrar el tipo y valores de los datos que se han leído del fichero.

3. Formato de las variables cuantitativas

Revisar el tipo de dato y el formato de las variables que deben ser cuantitativas. Convertid a tipo numérico si las variables no se han cargado con este tipo. Antes, sin embargo, es necesario corregir las posibles inconsistencias en el punto decimal.

4. Valores extremos

Analizar la presencia de posibles valores extremos (outliers) en la variable Estradiol. Si se observan valores muy extremos, eliminar los registros correspondientes del fichero. Mostrar un diagrama de caja (boxplot) para visualizar la posible presencia de valores extremos. Al finalizar la eliminación (si es el caso), mostrar un diagrama de caja con los valores finales.

5. Valores del resto de variables cuantitativas

Revisar mediante diagramas de caja o histogramas los valores del resto de variables cuantitativas. Idenficar si hay algún caso con valores anómalos. En caso de valores anómalos, se deben seguir las instrucciones del apartado 8. En este apartado, simplemente se debe identificar y explicar.

6. Inconsistencias

Corregir las posibles inconsistencias entre los siguientes conjuntos de variables, y aplicar las correcciones necesarias, siguiendo los criterios especificados:

- las variables Numchild y Anykids.
- las variables Entage y Agemenarq.
- las variables Agefbo, Agemenarq y Numchild.

7. Formato de las variables cualitativas

Estandarizar/normalizar las variables cualitativas, según los criterios establecidos.

8. Imputación

En el caso de detectar algún valor anómalo en las variables cuantitativas edad, edad de menarquia o número de hijos (no resuelto anteriormente) realizar una imputación de valores en estas variables.

En caso de realizar imputaciones de valores, la imputación debe hacerse con los 3 vecinos más cercanos usando la distancia de Gower, usando sólo la información de las variables cuantitativas y los datos de la misma etnia.

Después de realizar la imputación es necesario verificar que los valores asignados se han copiado sobre el conjunto de datos originales y que los valores resultantes tienen sentido y coherencia con el resto de datos del conjunto. Visualizar el resultado de las imputaciones realizadas (para evitar mostrar todo el conjunto de datos, sólo se deben mostrar los registros del conjunto de datos que contienen la imputación realizada).

9. Tabla resumen de las variables cualitativas

Realitzar un resumen descriptivo de los valores de las variables cualitativas.

10. Tabla resumen de las variables cuantitativas

Realitzar una tabla de la tendencia central y dispersión de las variables cuantitativas. Usad medidas robustas y no robustas.

11. Grabar a un archivo

Grabar los datos preprocesados a un fichero final denominado: "ESTRADL_clean.csv".

Comentarios importantes

1. **No se puede inspeccionar ni corregir de manera manual** el fichero de datos. Por ejemplo, **no** se pueden realizar instrucciones de este tipo:

   ```
   data[1,5] <- 32.5
   ```

   Este tipo de transformaciones se deben hacer con funcionalidades de búsqueda (buscar los registros que tienen errores o inconsistencias) y luego hacer las correcciones oportunas con funcionalidades de R. Así el procedimiento de limpieza es útil, independientemente del fichero de datos y de la posición y valores concretos del archivo.

2. **No se pueden hacer listados completos de los datos del fichero a pantalla**, porque generan archivos de salida excesivamente grandes. Si se desea validar el resultado de una instrucción sobre los datos, se puede usar la función **head** que muestra las primeras filas de la tabla de datos o **tail** que muestra las últimas.

Puntuaciones de los apartados

- Apartados 1,2,3 (10 %)
- Apartado 4 (10 %)
- Apartado 5 (10 %)
- Apartado 6 (10 %)
- Apartado 7 (10 %)
- Apartado 8 (20 %)
- Apartados 9,10,11 (10 %)
- Calidad del informe dinámico (20 %)