TP1 Bioestdística

Irisarri - Landa

TP 1: Validación de datos

Alumnos: Malena Irisarri, Román Landa

UNIVERSIDAD NACIONAL
DE ROSARIO
FACULTAD DE CIENCIAS
ECONÓMICAS Y ESTADÍSTICA

UNR

Rosario, Argentina

17 de Abril de 2025

Introducción

Se llevó a cabo un estudio de investigación clínica multicéntrico con el fin de implementar estándares mundiales de crecimiento fetal que faciliten la detección temprana de alteraciones en el desarrollo del feto dentro del últero, y de esta manera reducir la morbi-mortalidad perinatal asociada con el crecimiento.

Durante el período de reclutamiento, las mujeres admitidas en alguna de las clínicas de salud participantes cursando su primer trimestre de embarazo fueron invitadas a formar parte del estudio. Aquellas que cumplieron con los criterios de elegibilidad y dieron su consentimiento fueron seguidas según un esquema de visitas programado a las 14, 18, 24, 28, 32, 36 y 40 semanas de gestación. En cada visita, se tomaron medidas antropométricas del feto por medio de un ultrasonido.

La información necesaria para llevar a cabo el análisis se recolectó a lo largo de 17 formularios en papel. Particularmente, en el formulario de admisión se registraron algunas características de la mujeres al momento de ingresar en el estudio. A las mujeres que cumplieron todos los criterios les fue asignado un código identificador único (Subject Number) compuesto por su código de páis, código del médico responsable, y el orden de ingreso. Sólo se entrevistaron mujeres mayores de edad (18 años o más al momento de la entrevista).

Datos

Contamos con una base de 1000 pacientes, con las siguentes variables:

```
## # A tibble: 8 x 5
##   variable      description                                    type  format values
##   <chr>         <chr>                                          <chr> <chr>  <chr>
## 1 countrycode   a) Country code                                nume~ <NA>   4, 11~
## 2 patientid     b) Patient ID                                  char~ dd/mm~ <NA>
## 3 interview     c) Date of interview                           date  dd/mm~ <NA>
## 4 ethnicgroup   d) Ethnic group                                inte~ <NA>   1 = '~
## 5 scr           1. The woman is eligible according to CLIN-~ inte~ <NA>   1 = '~
## 6 usscr         2. The woman is eligible according to US-SC~ inte~ <NA>   1 = '~
## 7 consent       3. The woman agreed to participate and sign~ inte~ <NA>   1 = '~
## 8 subjectnumber Subject number                                 char~ <NA>   <NA>
```

Los primeros registros son:

```
## # A tibble: 6 x 10
##   countrycode patientid interview  ethnicgroup   scr usscr consent subjectnumber
##         <int> <chr>     <date>           <int> <int> <int>   <int> <chr>
## 1          14 15/03/19~ 2014-02-03           3     2     2       2 014003017
## 2          31 19/03/19~ 2011-03-29           1     2     2       2 031001011
## 3          23 26/11/19~ 2014-06-25           1     2     2       2 023009024
## 4          11 18/04/19~ 2014-03-18           3     2     2       2 011002135
## 5          11 01/01/19~ 2014-02-25           3     2     2       2 011002120
## 6          31 13/12/19~ 2011-04-01           1     2     2       2 031002012
## # i 2 more variables: fechaid <date>, inicialesid <chr>
```

Reglas

Se plantean las siguientes reglas

```
## # A tibble: 27 x 3
##    id    desc                         cond
##    <chr> <chr>                        <chr>
##  1 r1    (patientid) es faltante      !is.na(patientid)
##  2 r2    (countrycode) es faltante    !is.na(countrycode)
##  3 r3    (countrycode) es entero      is.integer(countrycode)
##  4 r4    (countrycode) fuera de rango countrycode %in% c(4,11,14,23,31,48,54,65~
##  5 r5    (fechaid) es faltante        !is.na(fechaid)
##  6 r6    (fechaid) es fecha           is.Date(fechaid)
##  7 r7    (inicialesid) es faltante    !is.na(inicialesid)
##  8 r8    (inicialesid) es caracter    is.character(inicialesid)
##  9 r9    (inicialesid) es mayuscula   grepl('^[A-Z]{2}$', inicialesid)
## 10 r10   (interview) es faltante      !is.na(interview)
## # i 17 more rows
```

NOTA: se adjunta en la entrega un archvio *reglas.xlsx* con la especificaciín de todas las reglas.

Resultados

Luego de aplicar las reglas de validación a nuestra base, compartimos los siguientes resultados:

- Matriz de validación de tipo Registro (filas) X Regla (columnas)

```
## # A tibble: 6 x 28
##   registro     r1    r2    r3    r4    r5    r6    r7    r8    r9    r10   r11
##   <chr>        <lgl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl> <lgl>
## 1 15/03/1989-~ TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 2 19/03/1973-~ TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 3 26/11/1994-~ TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 4 18/04/1990-~ TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 5 01/01/1981-~ TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## 6 13/12/1981-~ TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
## # i 16 more variables: r12 <lgl>, r13 <lgl>, r14 <lgl>, r15 <lgl>, r16 <lgl>,
## #   r17 <lgl>, r18 <lgl>, r19 <lgl>, r20 <lgl>, r21 <lgl>, r22 <lgl>,
## #   r23 <lgl>, r24 <lgl>, r25 <lgl>, r26 <lgl>, r27 <lgl>
```

- Número de participantes limpios:

La cantidad de registros sin errores es 759, lo que equivale a un 75,9% de la base.

- Participantes con inconsistencias:

La lista de participantes con inconsistencias es

```
## # A tibble: 241 x 4
##    registro      n_errores reglas_fallidas desc
##    <chr>             <int> <chr>           <chr>
##  1 01/01/1989-KR         2 r21, r22        <NA>
##  2 01/04/1981-RT         2 r21, r22        <NA>
##  3 01/04/1981-ZT         2 r12, r14        <NA>
```

```
##  4 01/04/1987-CI         1 r25            (subjectnumber) no debería estar si ~
##  5 01/04/1987-GU         1 r26            (subjectnumber) no debería estar si ~
##  6 01/05/1982-UX         2 r21, r22       <NA>
##  7 01/06/1977-HF         2 r21, r22       <NA>
##  8 01/06/1992-EJ         1 r24            (subjectnumber) no debería estar si ~
##  9 01/09/1990-YH         1 r23            (subjectnumber) coincide con country~
## 10 01/09/1993-QZ         2 r12, r14       <NA>
## # i 231 more rows
```

- Inconsistencias más frecuentes:

Las inconsistencias mas frecuentes son:

1. En primer lugar, la regla 21, que indica *subject number* como dato faltante. El campo cuenta con 100 valores perdidos en nuestra base.
2. Lugo, la regla 22, que verifica que *subject number* es vacio siendo que debia estar completo.
3. Por último, la regla 14, que evalua si el campo *ethnic group* es uno de los valores admitidos.

Presentamos el resto de las reglas con errores en las siguiente tabla:

```
## # A tibble: 11 x 4
##    regla errores desc                                          cond
##    <chr>   <int> <chr>                                         <chr>
##  1 r21       100 (subjectnumber) es faltante                   !is.na(subjectnu~
##  2 r22        97 (subjectnumber) obligatorio si pasa screening ifelse(scr == 2 ~
##  3 r14        51 (ethnicgroup) fuera de rango                  ethnicgroup %in%~
##  4 r23        41 (subjectnumber) coincide con countrycode      as.integer(subst~
##  5 r26        30 (subjectnumber) no debería estar si consent=1 ifelse(consent =~
##  6 r12        26 (ethnicgroup) es faltante                     !is.na(ethnicgro~
##  7 r25        18 (subjectnumber) no debería estar si usscr=1   ifelse(usscr == ~
##  8 r24        10 (subjectnumber) no debería estar si scr=1     ifelse(scr == 1,~
##  9 r27         3 paciente mayor de edad                        !is.na(fechaid) ~
## 10 r19         1 (consent) es faltante                         !is.na(consent)
## 11 r20         1 (consent) dentro de rango                     consent %in% c(1~
```

- Campos con más errores:

El campo con mas errores de la base es *subject number* con 97 errores.