Missing data analysis in longitudinal data. How to analyze it?

Jorge Curto García DNI: 08107073X

10 de enero de 2018

Jorge J. Curto García

M0.185 - TFM-Estadística y Bioinformática

Máster Universitario en Bioinformática y Bioestadística UOC-UB

Nombre Consultor/a: Núria Pérez

Barcelona, 10 de enero 2018. Fase 3.

Índice:

- 6.9 Imputación por regresión simple (30%)
- 6.10 Imputación múltiple PMM (10%)
- 6.11 Imputación múltiple PMM (20%)
- 6.12 Imputación múltiple PMM (30%)

7. REPRODUCCIÓN DE LOS ANÁLISIS ORIGINALES
   - 7.1 Reproducción de los análisis con datos originales
     - 7.1.1 Reproducción de los análisis: Diseños únicos
     - 7.1.2 Reproducción de los análisis: Errores perseverantes
8. REPRODUCCIÓN DE LOS ANÁLISIS CON 10% DE DATOS FALTANTES
   - 8.2 Reproducción de los análisis: Eliminación de los casos (listwise) (10%)
     - 8.2.1 Reproducción de los análisis. Modelos nº diseños únicos: Eliminación de los casos (listwise) (10%)
     - 8.2.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Eliminación de los casos (listwise) (10%)
   - 8.3 Reproducción de los análisis: Media (10%)
     - 8.3.1 Reproducción de los análisis. Modelos nº diseños únicos: Media (10%)
     - 8.3.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Media (10%)
   - 8.4 Reproducción de los análisis: Regresión (10%)
     - 8.4.1 Reproducción de los análisis. Modelos nº diseños únicos: Regresión (10%)
     - 8.4.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Regresión (10%)
   - 8.5 Reproducción de los análisis: MI (PMM) (10%)
     - 8.5.1 Reproducción de los análisis. Modelos nº diseños únicos: MI (PMM) (10%)
     - 8.5.2 Reproducción de los análisis. Modelos nº de errores perseverantes: MI (PMM) (10%)
9. REPRODUCCIÓN DE LOS ANÁLISIS CON 20% DE DATOS FALTANTES
   - 9.1 Reproducción de los análisis: Eliminación de los casos (listwise) (20%)
     - 9.1.1 Reproducción de los análisis. Modelo nº diseños únicos: Eliminación de los casos (listwise) (20%)
     - 9.1.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Eliminación de los casos (listwise) (20%)
   - 9.2 Reproducción de los análisis: Media (20%)
     - 9.2.1 Reproducción de los análisis. Modelo nº diseños únicos: Media (20%)
     - 9.2.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Media (20%)

1. DATOS UTILIZADOS PARA EL ANÁLISIS

El objetivo del presente TFM, es la aplicación de diferentes técnicas de imputación de datos faltantes en un conjunto de datos longitudinales reales.
Una limitación que surgió a la hora de decidir qué base de datos escoger para la realización del presente TFM es que en la mayoría de los casos los investigadores publican bases de datos completas. Finalmente se ha optado por seleccionar una base de datos con un número significativo de registros y en caso de que no contuviese datos faltantes, generarlos de manera aleatoria.

Existen diversos repositorios online que recopilan bases de datos, uno de ellos es "The Dryad Digital Repository" "http://datadryad.org/" (último acceso 17-11-2017), que es un recurso comisariado que permite acceder libremente a los datos utilizados en diversas publicaciones científicas. A través de Dryad, se obtuvo acceso a una base de datos correspondiente a un estudio longitudinal que incluye resultados del test de Ruff Figural Fluency/ Test de Fluidez de diseños de Ruff (RFFT) en 2515 participantes del estudio Prevention of Renal and Vascular ENd-stage Disease (PREVEND), de la ciudad de Groningen, en los Países Bajos.

La cumplimentación del RFFT se recogió en tres ocasiones durante un periodo de seguimiento de 6 años. El RFFT es un prueba cognitiva que evalúa la función ejecutiva (FE) midiendo la programación visomotora y asociándose con la actividad frontal derecha. La FE engloba habilidades cognitivas propias de la corteza prefrontal, entre los cuales se incluyen la anticipación, la elección de objetivos, la planeación, la selección de la conducta, la autorregulación, el autocontrol y el uso de la retroalimentación. Además de los resultados del RFTT, la base de datos contiene la edad de los participantes en la primera cumplimentación del test, el género, el nivel de educación y el tiempo transcurrido entre cada cumplimentación del test.

Estudios longitudinales han demostrado que las pruebas repetidas mejoran el rendimiento en RFFT dificultando la interpretación de los resultados de la prueba en el entorno clínico. Por este motivo los autores investigaron el efecto de las mediciones consecutivas sobre el rendimiento en el RFFT mediante la utilización de un modelo de regresión lineal multivariado en el que incluyeron la edad, género, nivel de educación y el término de interacción entre el número de medición consecutivo y la edad como variables independientes.

El test RFFT consiste de cinco partes que a su vez contienen 35 patrones de cinco puntos y, de forma resumida, la tarea de los participantes es dibujar tantos diseños únicos en cada parte como les sea posible durante un minuto, conectando los puntos evitando repetir diseños. El rendimiento en el RFFT se expresa como el número total de diseños únicos en las cinco partes y el número total de repeticiones de los diseños

o errores perseverantes. En el estudio PREVEND los resultados (número de diseños únicos y errores perseverantes) fueron analizados por dos examinadores independientes y en caso de no coincidir, se introdujo a un tercer examinador y se promediaron los resultados de los dos examinadores con resultados más concordantes.

Los datos se recogen en una base de datos con formato .sav (SPSS) disponible en: "http://datadryad.org/bitstream/handle/10255/dryad.77302/RFFT%20longitudinal%20data%20Groningen%20the%20Netherlands.sav?sequence=1"

Análisis descriptivo de los datos mediante R.

Inicialmente se habían reclutado a 4158 participantes, pero solo un total de 2515 (61%) completaron las tres medidas del test RFFT y son los casos incluidos en la base de datos disponible a través de Dryad.

VARIABLE	Descripción	Tipo
Casenr	Nº de individuo	entero
Age	Edad en la 1ª medida (años)	continua
Gender	Género	categórica
Education	Nivel de educación	categórica
Measurement	Nº de medida consecutiva	ordinal
Unique	Nº de diseños únicos	entero
Perseverative	Nº de errores perseverantes	entero
Interval	Tiempo desde la medida previa (años)	continua

---

2. OBJETIVO PRINCIPAL

Ejemplificación de la aplicación de los métodos estudiados mediante el análisis de una base de datos longitudinales en el ámbito de la biomedicina, generando un informe estadístico dinámico (utilizando software de licencia libre: R y Markdown). Se generará un informe estadístico dinámico (utilizando R y Markdown) que contendrá aquellos métodos de tratamiento de datos faltantes identificados previamente y que sean aplicables a los datos contenidos en la base de datos disponible. El informe estadístico se incluirá como anexo de la memoria final.

---

3. MÉTODOS

Para realizar el análisis que permita responder a los objetivos planteados se van a seguir una serie de pasos que se describen a continuación:

- En primer lugar se importará la base de datos desde su ubicación en la web datadryad.org.

Análisis de los datos

- Análisis descriptivo de las variables incluidas en la muestra mediante técnicas descriptivas: medidas de tendencia central y dispersión para las variables cuantitativas y frecuencias absolutas y relativas para las variables cualitativas. También se utilizarán representaciones gráficas como son los histogramas para variables cuantitativas y gráficos de barras para las variables cualitativas.

- Generación de los datos perdidos:

La base de datos está compuesta por todos los registros de los resultados del test RFFT para 2515 participantes, por lo que no hay datos perdidos. Se generaran nuevas bases de datos que contendrán un 10%, 20% y 30% de datos faltantes respectivamente, en las variables que contienen los resultados del test RFFT (número de diseños únicos y número de errores perseverantes).

- Análisis de los patrones de los datos faltantes: Análisis comparativo de las variables basales que contienen datos faltantes frente a los que no.

- Tratamiento de los datos faltantes: Se aplicarán 4 métodos distintos para el tratamiento de datos faltantes.

- Reproducción de los análisis originales:
  - Se reproducirán los análisis llevados a cabo por los autores publicados en el artículo original en la base de datos original y en las bases de datos resultantes del tratamiento de datos faltnates.
- Comparación de los resultados originales con los de las bases de datos generadas con 10%, 20% y 30% de datos perdidos.

Nota aclaratoria Se adjuntará un archivo con todo el código en R, que permite rehacer el análisis que se presenta en este informe.

---

4. ANÁLISIS DE RESULTADOS

Se cargan los paquetes que se utilizarán para el análisis estadístico: tratamiento y depuración de datos, descriptivo, inferencia y datos perdidos.

- Paquetes para análisis descriptivo e inferencial (modelos multivariados): knitr', 'ROCR', 'ggplot2', 'aod', 'Rcpp', 'dplyr', 'plyr', 'lattice', 'car', 'foreign', 'samplingbook', 'pps', 'pwr', 'Rcmdr', 'nlme', 'nlmeU', 'plyr', 'reshape', 'RLRsim', 'WWGbook', 'graphics', 'stats', 'lme4'.

- Paquetes para análisis de valores faltantes: 'MissingDataGUI', 'VIM', 'VIMGUI', 'MICE', 'Amelia', 'Hmisc', 'mi', 'BaylorEdPsych'.

```
##    Casenr Age Gender Education Measurement
## 1       1  74      0         1           1
## 2       1  74      0         1           2
## 3       1  74      0         1           3
## 4       2  48      0         2           1
## 5       2  48      0         2           2
## 6       2  48      0         2           3
## 7       3  54      0         2           1
## 8       3  54      0         2           2
## 9       3  54      0         2           3
## 10      4  70      0         2           1
```

```
##  [1] "Casenr"
##  [2] "Age"
##  [3] "Gender"
##  [4] "Education"
##  [5] "Unique.1ª medida (2003-06)"
##  [6] "Perseverative.1ª medida (2003-06)"
##  [7] "Interval.1ª medida (2003-06)"
##  [8] "Unique.2ª medida (2006-08)"
##  [9] "Perseverative.2ª medida (2006-08)"
## [10] "Interval.2ª medida (2006-08)"
## [11] "Unique.3ª medida (2008-12)"
## [12] "Perseverative.3ª medida (2008-12)"
## [13] "Interval.3ª medida (2008-12)"
```

Tras importar la base de datos desde su ubicación en la web, etiquetar los valores y variables y transponerla a formato *"wide"* se muestran los valores en todas las variables para los 5 primeros casos.

```
##    Casenr Age Gender           Education Unique_1_2003_06
## 1       1  74 Hombre            Primaria               32
## 4       2  48 Hombre  Secundaria inicial               26
## 7       3  54 Hombre  Secundaria inicial               91
## 10      4  70 Hombre  Secundaria inicial               59
## 13      5  52 Hombre Secundaria superior               60
##    Perseverative_1_2003_06 Interval_1_2003_06 Unique_2_2006_08
## 1                        1                -2             35.0
## 4                        3                -2             47.0
## 7                       11                -2             83.5
## 10                      49                -2             54.0
## 13                      14                -2             68.0
```

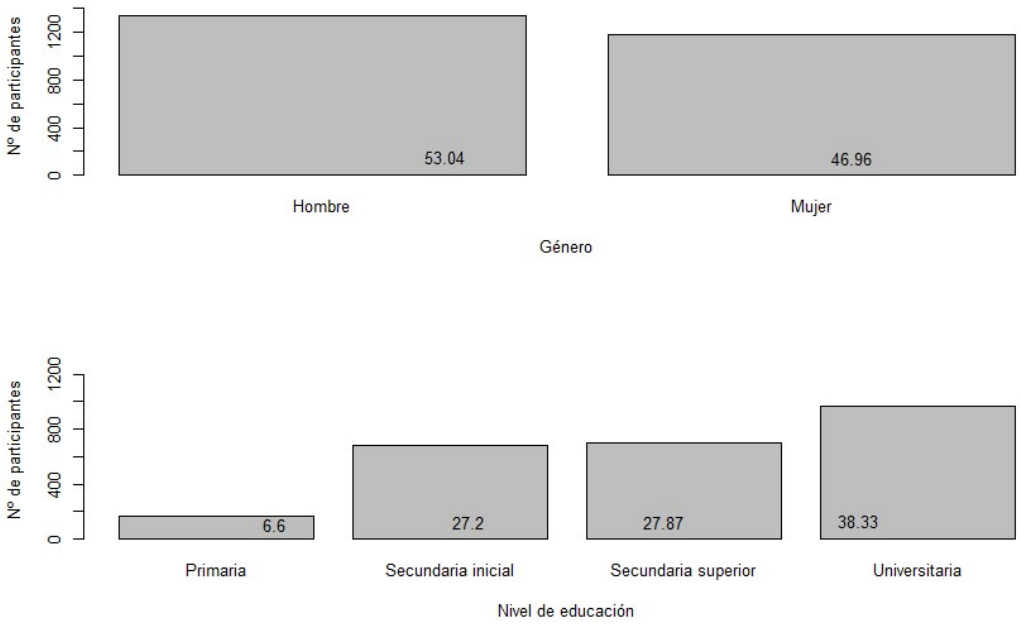
```
##    Perseverative_2_2006_08 Interval_2_2006_08 Unique_3_2008_12
## 1                        0           2.842123             50.5
## 4                        3           2.762718             71.0
## 7                       21           2.628553             77.0
## 10                      39           2.762718             64.0
## 13                      16           2.844861             21.0
##    Perseverative_3_2008_12 Interval_3_2008_12
## 1                     14.5           2.201232
## 4                      0.5           2.376454
## 7                     17.0           3.392197
## 10                    11.0           2.650240
## 13                     0.0           2.247775
```

4.1 Características basales

Se analizan a modo descriptivo todas las variables cualitativas basales: Género y nivel de educación.

```
##
## counts:
##
## Hombre  Mujer
##   1334   1181
##
## percentages:
##
## Hombre  Mujer
##  53.04  46.96

##
## counts:
##
##            Primaria  Secundaria inicial Secundaria superior
##                 166                 684                 701
##       Universitaria
##                 964
##
## percentages:
##
##            Primaria  Secundaria inicial Secundaria superior
##                6.60               27.20               27.87
##       Universitaria
##                38.33
```

Figura 1. Frecuencias variables cualitativas (I)

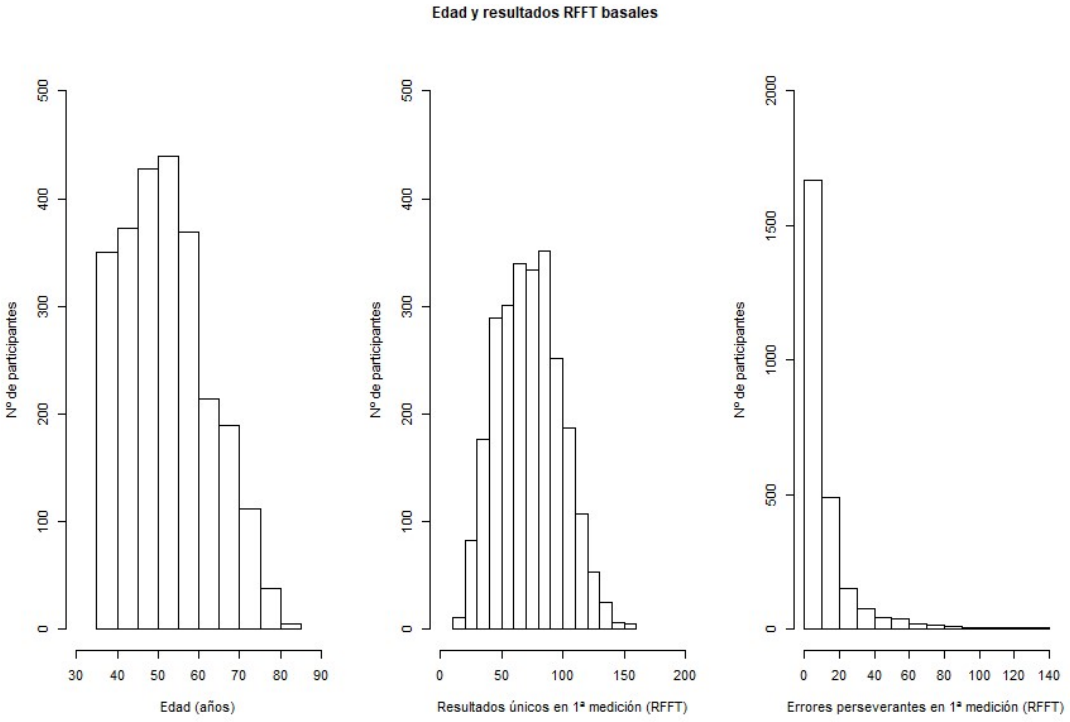


La muestra está compuesta en un 46.96% de mujeres y el 38.33% del total de participantes tenían educación universitaria.

A continuación se presentan los resultados descriptivos de las variables cuantitativas edad y los resultados basales del test RFFT: Nº de diseños únicos y Nº de errores perseverantes.

```
##                          mean   sd IQR    cv skewness kurtosis 0% 50% 1
00%
## Age                     52.6 10.4  15 0.198    0.413   -0.557 35  52
82
## Unique_1_2003_06        72.9 25.6  38 0.351    0.212   -0.468 14  72
157
## Perseverative_1_2003_06 11.9 15.8  10 1.329    3.501   15.899  0   7
132
```

Figura 2. Histograma variables cuantitativas



La media de edad de la muestra fue 52.6 años (mediana: 52) y la media de los resultados únicos en el RFFT basales fue de 72.94 (mediana: 72), con una media de errores perseverantes en el RFFT basal de 11.91(mediana: 7).

NOTA

Para elaborar las secciones que vienen a continuación se ha utilizado información recogida en distintas referencias y páginas web disponibles en los siguientes enlaces:

- MissingDataLab
- rstudio publications
- VIM-Imputation
- VIM
- Multiple-imputacion.com

- Multiple imputation with mice
- Imputing missing data with R; MICE package
- Linear Mixed Effects Modeling using R
- Análisis bioestadístico con modelos de regresión en R
- Package 'nlme'

---

5. GENERACIÓN DE DATOS FALTANTES

Tanto para la variable que contiene el *Nº de diseños únicos* y el *Nº de errores perseverantes*, se generan nuevas bases de datos con un 10%, 20% y 30% de datos faltantes en cada una de las 3 mediciones, es decir con el dataset en formato *"wide"*. Para asegurar la reproducibilidad de la generación de los datos faltantes se fija una semila (*seed*) para cada caso.

- *seed=456789 para Nº de diseños únicos*
- *seed=456987 para Nº de errores perseverantes*

5.1 Generación de datos fatantes: 10%

```
##    Casenr Age Gender           Education Interval_2_2006_08
## 1       1  74 Hombre            Primaria           2.842123
## 4       2  48 Hombre  Secundaria inicial           2.762718
## 7       3  54 Hombre  Secundaria inicial           2.628553
## 10      4  70 Hombre  Secundaria inicial           2.762718
## 13      5  52 Hombre Secundaria superior           2.844861
## 16      6  73 Hombre Secundaria superior           2.801051
## 19      7  50  Mujer       Universitaria           2.779147
## 22      8  58  Mujer Secundaria superior           2.861289
## 25      9  53  Mujer Secundaria superior           2.543672
## 28     10  49 Hombre       Universitaria           2.858551
##    Interval_3_2008_12 Unique.mcar1 Unique.mcar2 Unique.mcar3
## 1            2.201232           32         35.0         50.5
## 4            2.376454           26         47.0         71.0
## 7            3.392197           91           NA           NA
## 10           2.650240           NA         54.0         64.0
## 13           2.247775           60           NA         21.0
## 16           2.294319           63         68.5         60.5
## 19           2.409309           42         50.0         91.0
## 22           2.277892           83         86.0        104.5
## 25           3.088296           78         49.5         86.0
```

```
## 28           2.291581           96        115.0        127.5
##    Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
## 1                    1                 0.0                14.5
## 4                   NA                 3.0                 0.5
## 7                   11                21.0                17.0
## 10                  49                39.0                11.0
## 13                  14                16.0                 0.0
## 16                  NA                18.0                13.5
## 19                  25                36.0                30.0
## 22                   7                 9.0                12.5
## 25                   7                 4.5                 8.0
## 28                  57                43.0                25.5

##
## counts:
##
## FALSE  TRUE
##  2260   255
##
## percentages:
##
## FALSE  TRUE
## 89.86 10.14

##
## counts:
##
## FALSE  TRUE
##  2266   249
##
## percentages:
##
## FALSE  TRUE
##  90.1   9.9

##
## counts:
##
## FALSE  TRUE
##  2240   275
##
## percentages:
##
## FALSE  TRUE
## 89.07 10.93

##
## counts:
##
## FALSE  TRUE
##  2257   258
```

```
##
## percentages:
##
## FALSE  TRUE
## 89.74 10.26

##
## counts:
##
## FALSE  TRUE
##  2263   252
##
## percentages:
##
## FALSE  TRUE
## 89.98 10.02

##
## counts:
##
## FALSE  TRUE
##  2293   222
##
## percentages:
##
## FALSE  TRUE
## 91.17  8.83
```

- Para las variables ***Nº de diseños únicos*** se han generado un total de 10.14%, 9.9% y 10.93% de datos faltantes, respectivamente en cada una de las 3 mediciones.

- Para las variables ***Nº de errores perseverantes*** se han generado un total de 10.26%, 10.02% y 8.83% de datos faltantes, respectivamente en cada una de las 3 mediciones.

5.2 Generación de datos fatantes: 20%

```
##    Casenr Age Gender           Education Interval_2_2006_08
## 1       1  74 Hombre            Primaria           2.842123
## 4       2  48 Hombre  Secundaria inicial           2.762718
## 7       3  54 Hombre  Secundaria inicial           2.628553
## 10      4  70 Hombre  Secundaria inicial           2.762718
## 13      5  52 Hombre Secundaria superior           2.844861
## 16      6  73 Hombre Secundaria superior           2.801051
## 19      7  50  Mujer       Universitaria           2.779147
## 22      8  58  Mujer Secundaria superior           2.861289
## 25      9  53  Mujer Secundaria superior           2.543672
## 28     10  49 Hombre       Universitaria           2.858551
##    Interval_3_2008_12 Unique.mcar1_20 Unique.mcar2_20 Unique.mcar3_20
## 1            2.201232              32            35.0            50.5
```

```
## 4            2.376454              26            47.0            71.0
## 7            3.392197              91              NA              NA
## 10           2.650240              NA            54.0            64.0
## 13           2.247775              60              NA            21.0
## 16           2.294319              NA              NA            60.5
## 19           2.409309              NA            50.0            91.0
## 22           2.277892              83            86.0           104.5
## 25           3.088296              78            49.5            86.0
## 28           2.291581              96           115.0           127.5
##    Perseverative.mcar1_20 Perseverative.mcar2_20 Perseverative.mcar3_2
0
## 1                       1                    0.0                      N
A
## 4                      NA                    3.0                      N
A
## 7                      11                   21.0                    17.
0
## 10                     49                   39.0                    11.
0
## 13                     14                   16.0                     0.
0
## 16                     NA                   18.0                    13.
5
## 19                     25                   36.0                    30.
0
## 22                     NA                     NA                    12.
5
## 25                      7                    4.5                     8.
0
## 28                     57                   43.0                    25.
5

##
## counts:
##
## FALSE  TRUE
##  2009   506
##
## percentages:
##
## FALSE  TRUE
## 79.88 20.12

##
## counts:
##
## FALSE  TRUE
##  2002   513
##
## percentages:
```

```
##
## FALSE  TRUE
##  79.6  20.4

##
## counts:
##
## FALSE  TRUE
##  1985   530
##
## percentages:
##
## FALSE  TRUE
## 78.93 21.07

##
## counts:
##
## FALSE  TRUE
##  2020   495
##
## percentages:
##
## FALSE  TRUE
## 80.32 19.68

##
## counts:
##
## FALSE  TRUE
##  1982   533
##
## percentages:
##
## FALSE  TRUE
## 78.81 21.19

##
## counts:
##
## FALSE  TRUE
##  2047   468
##
## percentages:
##
## FALSE  TRUE
## 81.39 18.61
```

- Para las variables ***Nº de diseños únicos*** se han generado un total de 20.12%, 20.4% y 21.07% de datos faltantes, respectivamente en cada una de las 3 mediciones.

- Para las variables ***Nº de errores perseverantes*** se han generado un total de 19.68%, 21.19% y 18.61% de datos faltantes, respectivamente en cada una de las 3 mediciones.

5.3 Generación de datos fatantes: 30%

```
##    Casenr Age Gender           Education Interval_2_2006_08
## 1       1  74 Hombre            Primaria           2.842123
## 4       2  48 Hombre  Secundaria inicial           2.762718
## 7       3  54 Hombre  Secundaria inicial           2.628553
## 10      4  70 Hombre  Secundaria inicial           2.762718
## 13      5  52 Hombre Secundaria superior           2.844861
## 16      6  73 Hombre Secundaria superior           2.801051
## 19      7  50  Mujer       Universitaria           2.779147
## 22      8  58  Mujer Secundaria superior           2.861289
## 25      9  53  Mujer Secundaria superior           2.543672
## 28     10  49 Hombre       Universitaria           2.858551
##    Interval_3_2008_12 Unique.mcar1_30 Unique.mcar2_30 Unique.mcar3_30
## 1            2.201232              32            35.0            50.5
## 4            2.376454              26            47.0              NA
## 7            3.392197              NA              NA              NA
## 10           2.650240              NA            54.0            64.0
## 13           2.247775              60              NA            21.0
## 16           2.294319              NA              NA              NA
## 19           2.409309              NA            50.0            91.0
## 22           2.277892              83            86.0           104.5
## 25           3.088296              78            49.5              NA
## 28           2.291581              96           115.0           127.5
##    Perseverative.mcar1_30 Perseverative.mcar2_30 Perseverative.mcar3_3
0
## 1                       1                     NA                      N
A
## 4                      NA                    3.0                      N
A
## 7                      11                   21.0                    17.
0
## 10                     49                   39.0                    11.
0
## 13                     NA                   16.0                     0.
0
## 16                     NA                   18.0                    13.
5
## 19                     NA                   36.0                      N
A
## 22                     NA                     NA                    12.
5
## 25                      7                    4.5                     8.
0
## 28                     57                   43.0                    25.
5
```

```
##
## counts:
##
## FALSE  TRUE
##  1764   751
##
## percentages:
##
## FALSE  TRUE
## 70.14 29.86

##
## counts:
##
## FALSE  TRUE
##  1760   755
##
## percentages:
##
## FALSE  TRUE
## 69.98 30.02

##
## counts:
##
## FALSE  TRUE
##  1728   787
##
## percentages:
##
## FALSE  TRUE
## 68.71 31.29

##
## counts:
##
## FALSE  TRUE
##  1763   752
##
## percentages:
##
## FALSE  TRUE
##  70.1  29.9

##
## counts:
##
## FALSE  TRUE
##  1716   799
##
## percentages:
```

```
##
## FALSE  TRUE
## 68.23 31.77

##
## counts:
##
## FALSE  TRUE
##  1800   715
##
## percentages:
##
## FALSE  TRUE
## 71.57 28.43
```

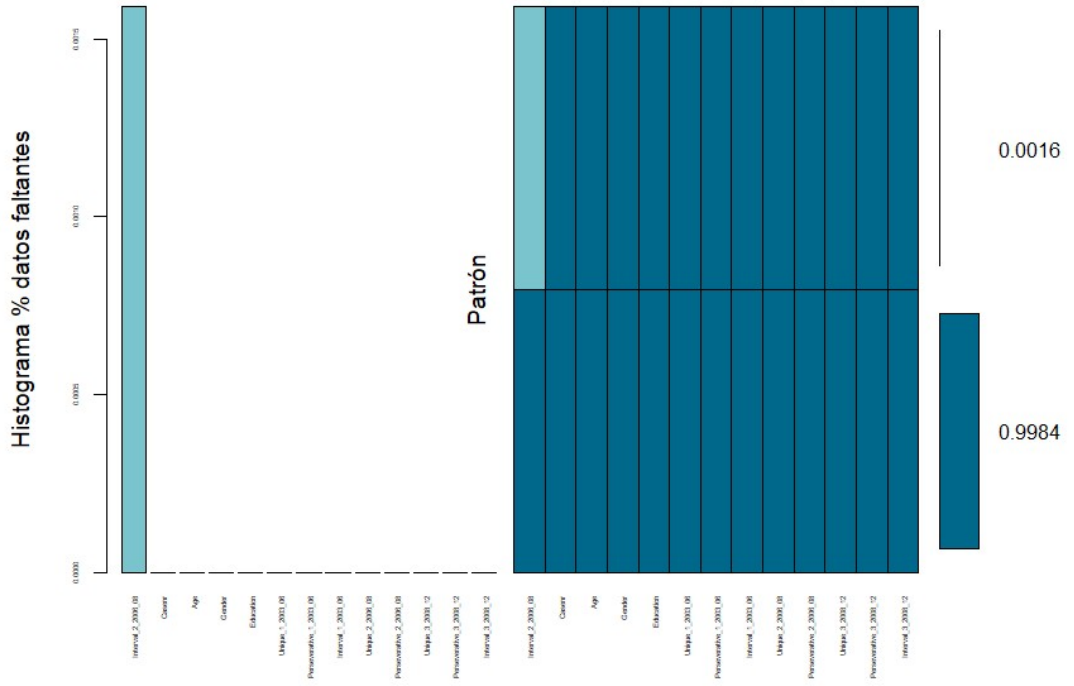
- Para las variables ***Nº de diseños únicos*** se han generado un total de 29.86%, 30.02% y 31.29% de datos faltantes, respectivamente en cada una de las 3 mediciones.
- Para las variables ***Nº de errores perseverantes*** se han generado un total de 29.9%, 31.77% y 28.43% de datos faltantes, respectivamente en cada una de las 3 mediciones.

5.4 Análisis de los patrones de los datos faltantes (10%)

Patrón datos faltantes en dataset original

Se utiliza la función ***agrr*** del paquete **VIM** para visualizar los datos faltantes y obtener el patrón de dichos datos faltantes en los ***datos iniciales***.

Figura 3. Patrón datos faltantes en dataset original




```
##
##  Variables sorted by number of missings:
##                  Variable       Count
##       Interval_2_2006_08 0.001590457
##                   Casenr 0.000000000
##                      Age 0.000000000
##                   Gender 0.000000000
##                Education 0.000000000
##         Unique_1_2003_06 0.000000000
##  Perseverative_1_2003_06 0.000000000
##       Interval_1_2003_06 0.000000000
##         Unique_2_2006_08 0.000000000
##  Perseverative_2_2006_08 0.000000000
##         Unique_3_2008_12 0.000000000
##  Perseverative_3_2008_12 0.000000000
##       Interval_3_2008_12 0.000000000
```

Se obtiene que la variable Interval_2_2006_08 contiene 4 casos con datos faltantes, lo que supone 0.16% respecto al total de casos. Se tendrá que tener en cuenta de cara a la imputación. Obviamente no presenta ningún tipo de patrón relevante.

Patrón datos faltantes en dataset modificado (10%)

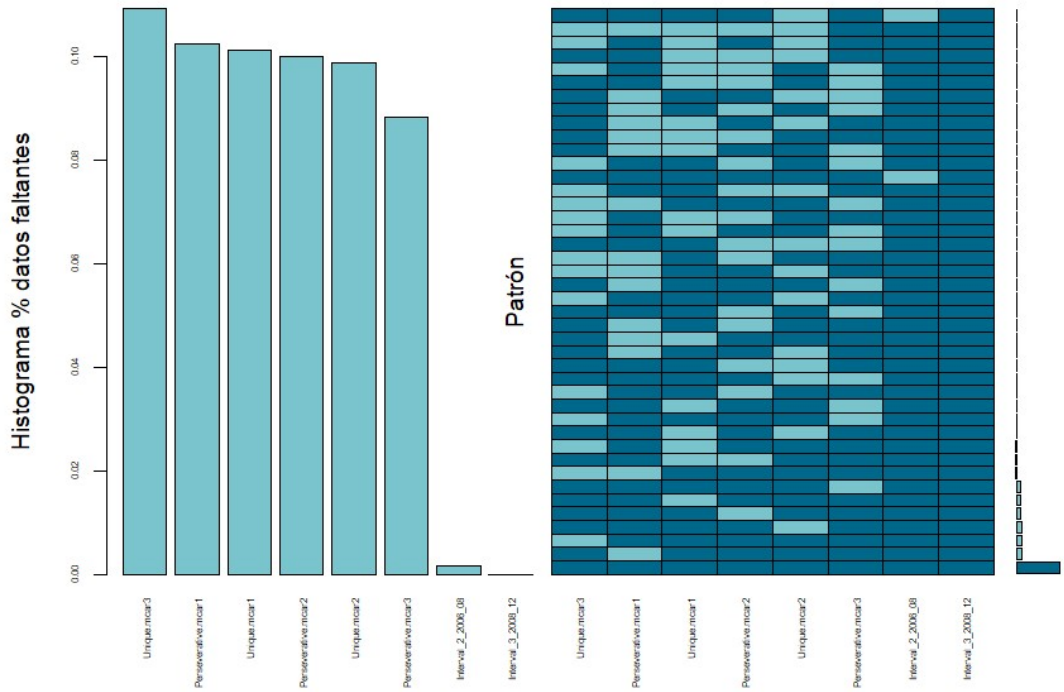
Se identifican los individuos con al menos 1 caso con dato faltante en global (incuyendo los casos de la variable) y separado para las 3 mediciones de **Unique** y **Perseverative** respectivamente.

```
##
## counts:
##
## Sin NA Con NA
##   1342   1173
##
## percentages:
##
## Sin NA Con NA
##  53.36  46.64

##
## counts:
##
## Sin NA Con NA
##   1815    700
##
## percentages:
##
## Sin NA Con NA
##  72.17  27.83

##
## counts:
##
## Sin NA Con NA
##   1846    669
##
## percentages:
##
## Sin NA Con NA
##   73.4   26.6
```

- Finalmente del total de pacientes el 46.64% presenta al menos un valor perdido en alguna de los valores del data set, mientras que en las 3 mediciones de unique es el 27.83% y en las mediciones de persevaritive es el 26.6%.

De nuevo se utiliza _ agrr_ para obtener el patrón de datos faltantes en el dataset que contiene los datos faltantes generados de manera aleatoria en las 6 variables que contienen las mediciones de unique y perseverative (3 mediciones en cada una de ellas). Solo se muestran los valores faltantes en dichas variables y la variable *Interval 2 2006 08* puesto que para el resto de variables se dispuso de los datos en todos los individuos.

Figura 4. Patrón datos faltantes en dataset modificado (I) (10%)




```
##
##  Variables sorted by number of missings:
##              Variable       Count
##         Unique.mcar3 0.109343936
##  Perseverative.mcar1 0.102584493
##         Unique.mcar1 0.101391650
##  Perseverative.mcar2 0.100198807
##         Unique.mcar2 0.099005964
##  Perseverative.mcar3 0.088270378
##   Interval_2_2006_08 0.001590457
##   Interval_3_2008_12 0.000000000
```

No se aprecia ningún tipo de patrón que presente una frecuencia superior al resto.

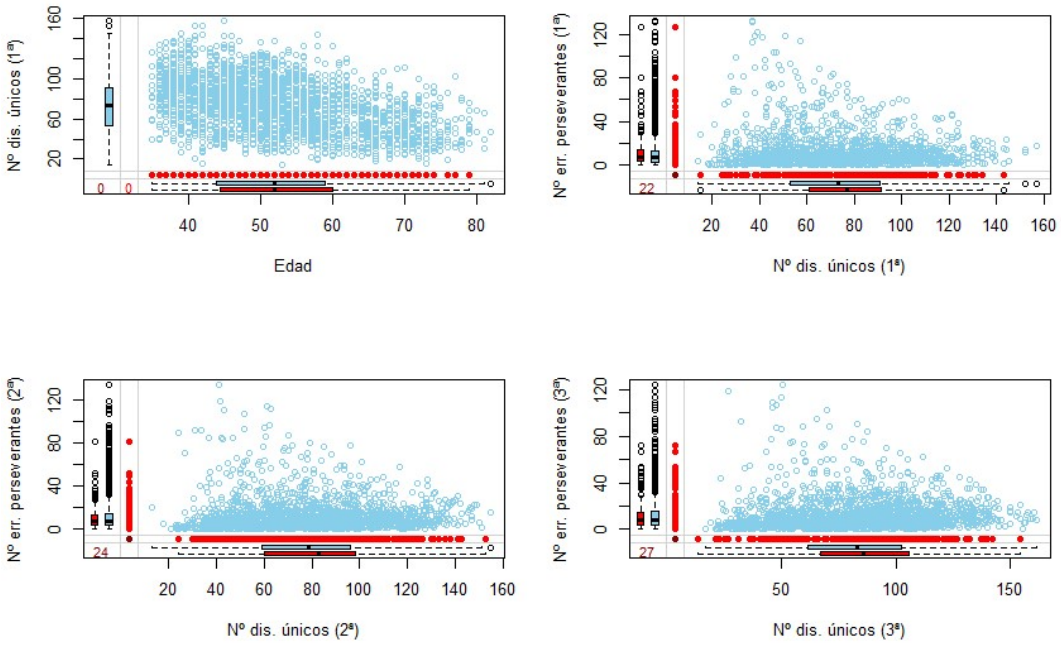
El paquete ***BaylorEdPsych***, contiene el una función (***LittleMCAR***)que permite aplicar el test *Little´s MCAR-test*, para analizar la hipótesis nula de que *los datos faltantes son MCAR*, donde un resultado significativo indicaría que se debe rechazar dicha hipótesis. Se aplica sobre el dataset con datos faltantes generados de manera aleatoria (*se aplica el test sobre las variables de interés para el estudio: edad, género, nivel de educación, 1ª medición únicos, 2ª medición únicos, 3ª medición únicos, 1ª medición perseverantes, 2ª medición perseverantes, 3ª medición perseverantes*).

```
## this could take a while

## [1] 0.9442149
```

A partir del resultado de test de Little's, no se puede rechazar la hipótesis nula (*p-valor*=0.94421>α=0.05), y por lo tanto, como era de esperar al haber sido generados de manera aleatoria, los datos faltantes cumplen las características de **MCAR**.

A su vez el paquete VIM también dispone de más gráficos que permiten comparar la distribución de los datos faltantes por pares de variables. A modo de ejemplo se plantean 4 comparaciones:

Figura 5. Patrón datos faltantes en dataset modificado (II) (10%)



Los gráficos anteriores permiten comparar la distribución de los datos faltantes por pares de variables. De modo que en el primer gráfico, la fila de puntos rojos indicarían como se distribuirían los datos faltantes de la variable *nº de diseños únicos 1ª* en función de los valores de la variable *edad*, y los puntos azules serían el resto de valores. El hecho de que los boxplots representados en el eje x (*edad*) sean similares indica que los valores de edad con y sin valores faltantes en el nº de diseños únicos, son similares y es un indicativo de que los datos cumplen las premisas de **MCAR**. Igualmente entre el resto de pares de variables representadas, visualmente tampoco se han hallado diferencias entre la presencia o ausencia de datos faltantes.

5.4.1 Características basales Vs datos faltantes todas las variables (10%)

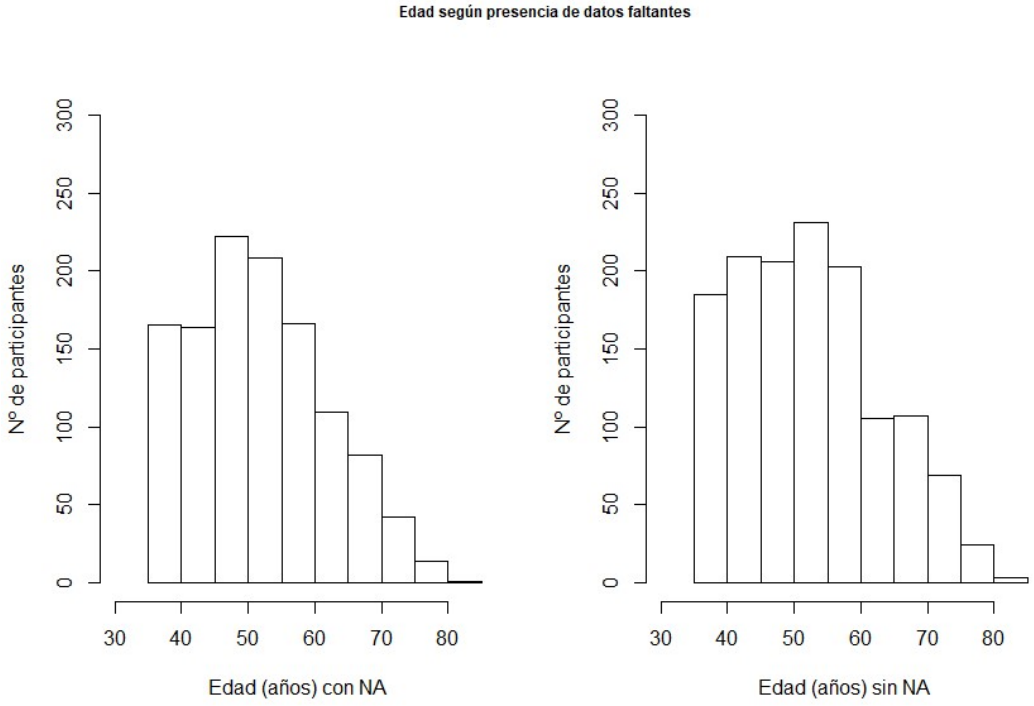
A continuación se analiza en mayor detalle las posibles diferencias entre la presencia y ausencia de datos faltantes en función de las características basales de los individuos.

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en alguna de las variables (10%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados del test RFFT. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 6. Edad en función de la presencia de datos faltantes (10%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA	0

Casos sin NA	0				
Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA	Media Edad
T-Test para muestras independientes	0.108	-0.146	1.478	Con NA	52.870
Test de Wilcoxon rank sum test with continuity correction	0.192	0.000	1.000	Sin NA	52.204

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, *p-valor*=0<α=0.05. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes *p-valor*=0.192> α=0.05.

Género y nivel de educación vs datos faltantes (10%)

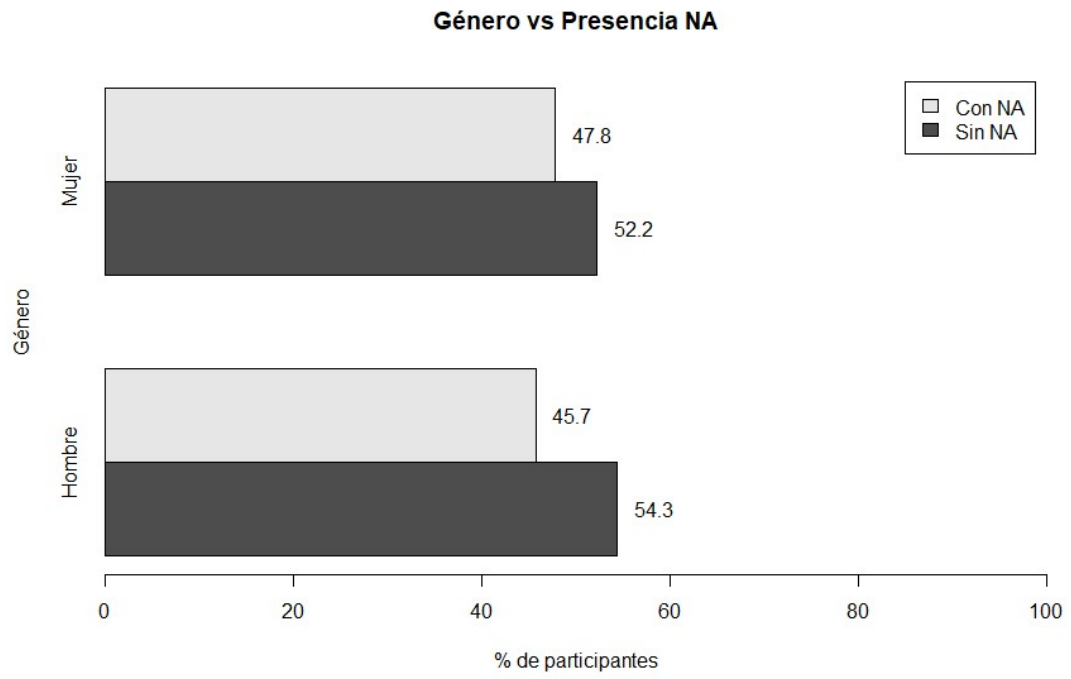
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (10%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 7. Presencia de datos faltantes según género (10%)

```
##
## Column percentages:
##              Gender
## Missing_data Hombre  Mujer
##       Sin NA   54.3   52.2
##       Con NA   45.7   47.8
##       Total   100.0  100.0
##       Count  1334.0 1181.0
```




```
##         Missing_data
## Gender   Sin NA Con NA
##   Hombre    725    609
##   Mujer     617    564
```

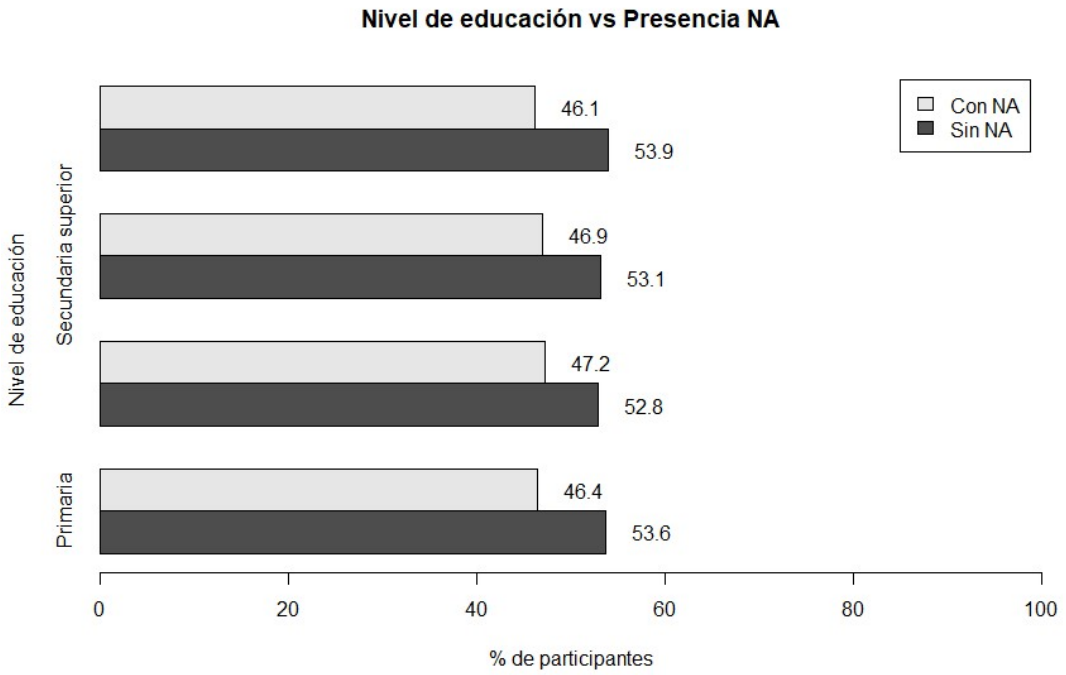
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.291	0.298

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.291>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes no varía entre géneros.

Figura 8. Presencia de datos faltantes según Nivel de educación (10%)

```
##
## Column percentages:
##             Education
## Missing_data Primaria Secundaria inicial Secundaria superior Universit
aria
##       Sin NA     53.6               52.8                53.1
53.9
##       Con NA     46.4               47.2                46.9
46.1
##       Total     100.0              100.0               100.0         1
```

```
00.0
##       Count     166.0              684.0              701.0          9
64.0
```




```
##                       Missing_data
## Education             Sin NA Con NA
##   Primaria               89     77
##   Secundaria inicial    361    323
##   Secundaria superior   372    329
##   Universitaria         520    444
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.969	0.968

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.969>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes no varía entre los niveles de educación.

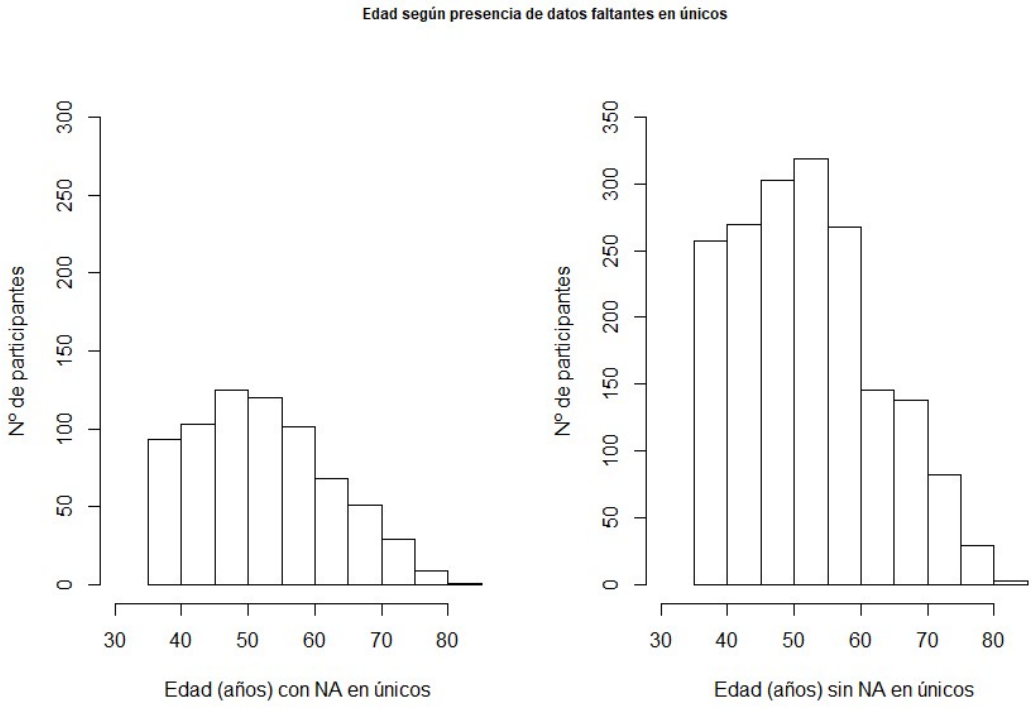
5.4.2 Características basales Vs datos faltantes en *diseños únicos* (10%)

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en diseños únicos (10%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados de las mediciones de la variable Unique. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 9. Edad en función de la presencia de datos faltantes en diseños únicos (10%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA en únicos	0
Casos sin NA en únicos	0

Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA en únicos	Media Edad
T-Test para muestras independientes	0.912	-0.953	0.851	Con NA	52.545
Test de Wilcoxon rank sum test with continuity correction	0.849	-1.000	1.000	Sin NA	52.596

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, $p\text{-}valor=0<\alpha=0.05$. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes $p\text{-}valor=0.849>\alpha=0.05$.

Género y nivel de educación vs datos faltantes (10%)

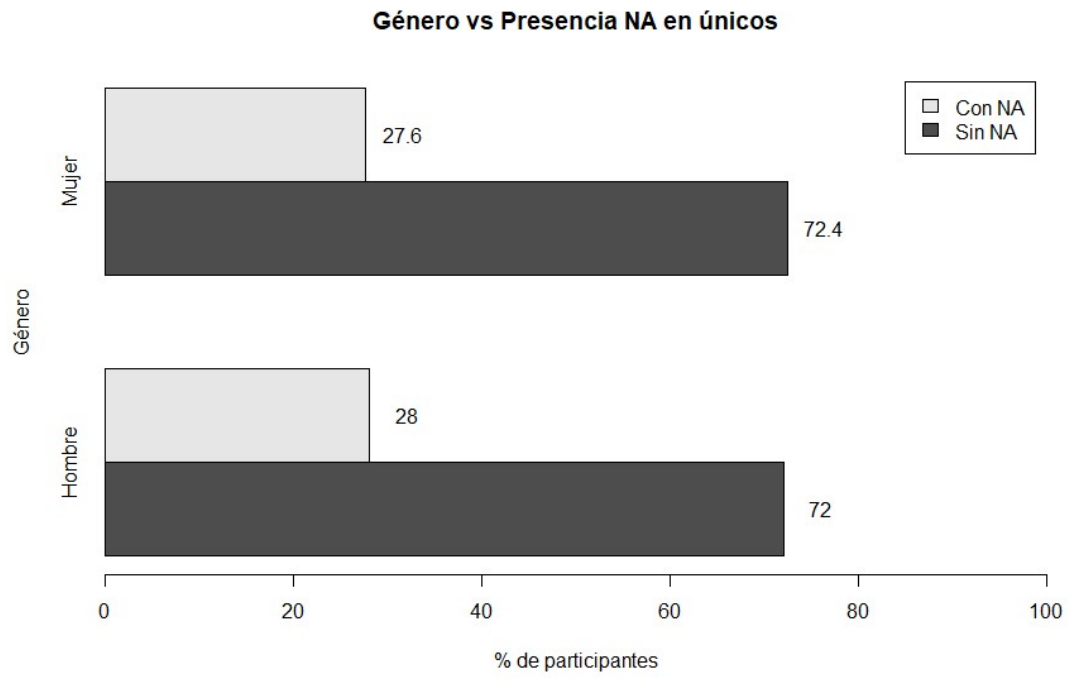
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (10%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 10. Presencia de datos faltantes en diseños únicos según género (10%)

```
##
## Column percentages:
##                Gender
## Missing_Unique Hombre  Mujer
##         Sin NA     72   72.4
##         Con NA     28   27.6
##         Total     100  100.0
##         Count    1334 1181.0
```




```
##         Missing_Unique
## Gender   Sin NA Con NA
##   Hombre    960    374
##   Mujer     855    326
```

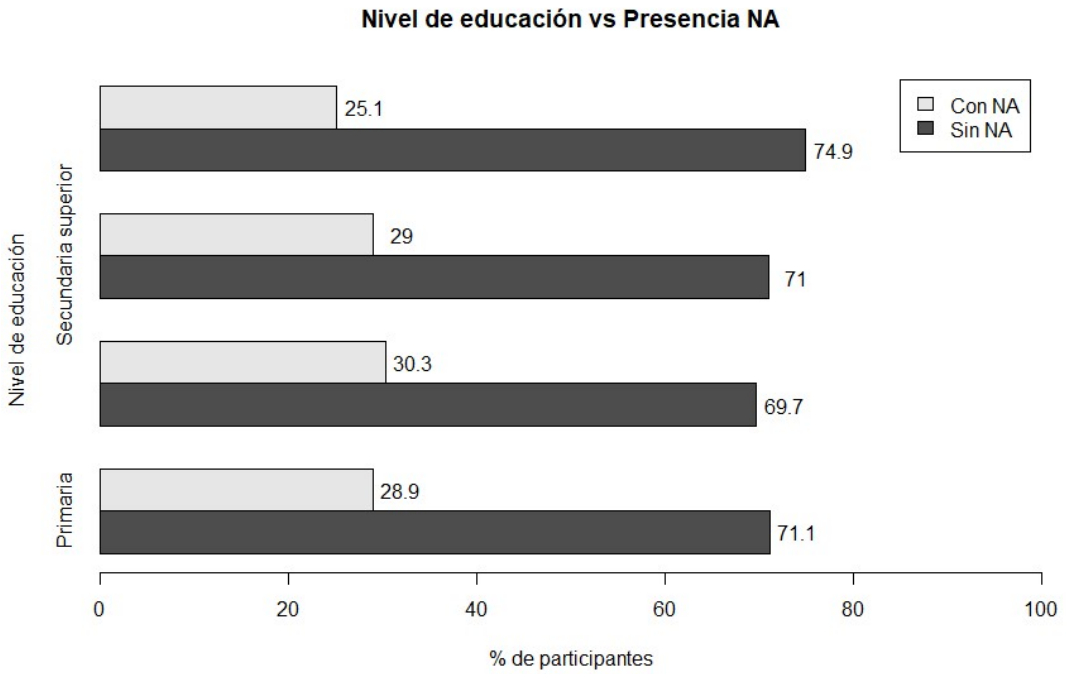
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.809	0.824

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.809>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en Unique no varía entre géneros.

Figura 11. Presencia de datos faltantes en diseños únicos según Nivel de educación (10%)

```
##
## Column percentages:
##                Education
## Missing_Unique Primaria Secundaria inicial Secundaria superior
##         Sin NA     71.1                69.7                  71
##         Con NA     28.9                30.3                  29
##         Total     100.0               100.0                 100
##         Count     166.0               684.0                 701
##                Education
## Missing_Unique Universitaria
```

```
##          Sin NA          74.9
##          Con NA          25.1
##          Total          100.0
##          Count          964.0
```




```
##                       Missing_Unique
## Education              Sin NA Con NA
##   Primaria                118     48
##   Secundaria inicial      477    207
##   Secundaria superior     498    203
##   Universitaria           722    242
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.106	0.102

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.106>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en diseños únicos no varía entre los niveles de educación.

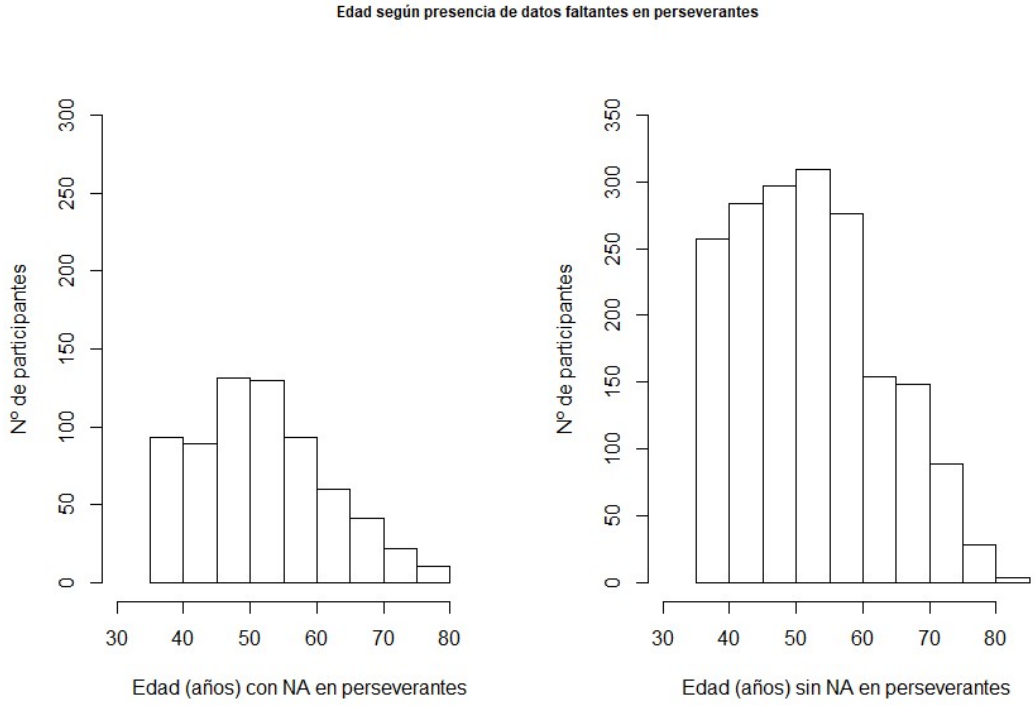
5.4.3 Características basales Vs datos faltantes en *errores perseverantes* (10%)

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en errores perseverantes (10%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados de las mediciones de los errores perseverantes. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 12. Edad en función de la presencia de datos faltantes en errores perseverantes (10%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA en perseverantes	0
Casos sin NA en perseverantes	0

Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA en perseverantes	Media Edad
T-Test para muestras independientes	0.122	-0.187	1.588	Con NA	52.745
Test de Wilcoxon rank sum test with continuity correction	0.227	0.000	1.000	Sin NA	52.045

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, $p\text{-}valor=0<\alpha=0.05$. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes $p\text{-}valor=0.227> \alpha=0.05$.

Género y nivel de educación vs datos faltantes (10%)

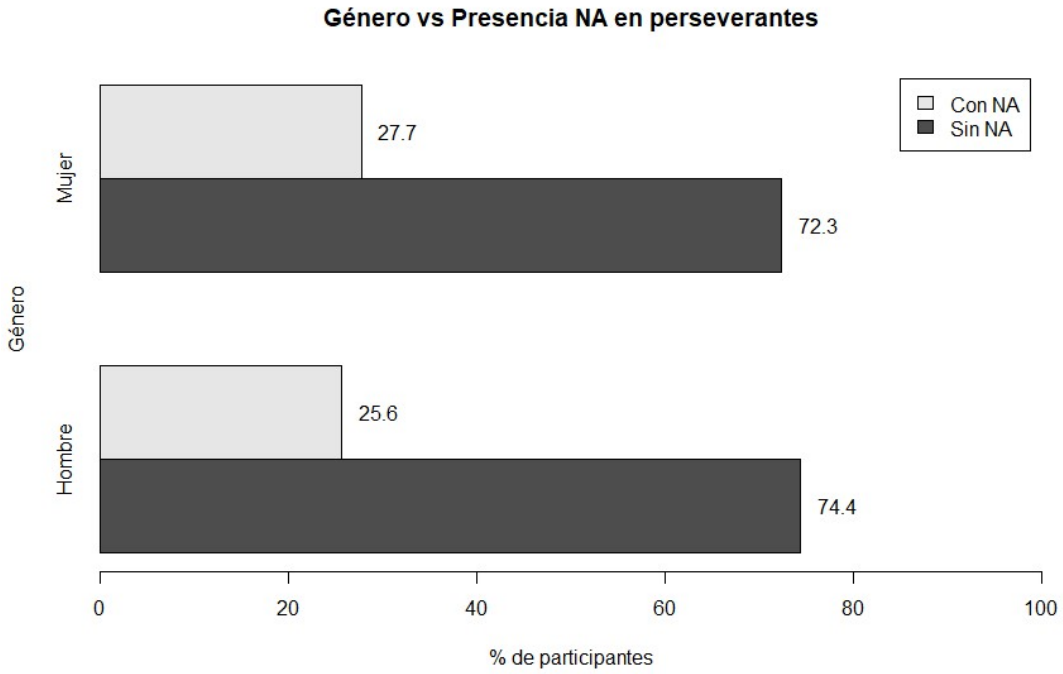
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (10%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 13. Presencia de datos faltantes en errores perseverantes según género (10%)

```
##
## Column percentages:
##                         Gender
## Missing_Perseverative Hombre  Mujer
##                Sin NA   74.4   72.3
##                Con NA   25.6   27.7
##                Total   100.0  100.0
##                Count  1334.0 1181.0
```




```
##         Missing_Perseverative
## Gender   Sin NA Con NA
##   Hombre    992    342
##   Mujer     854    327
```

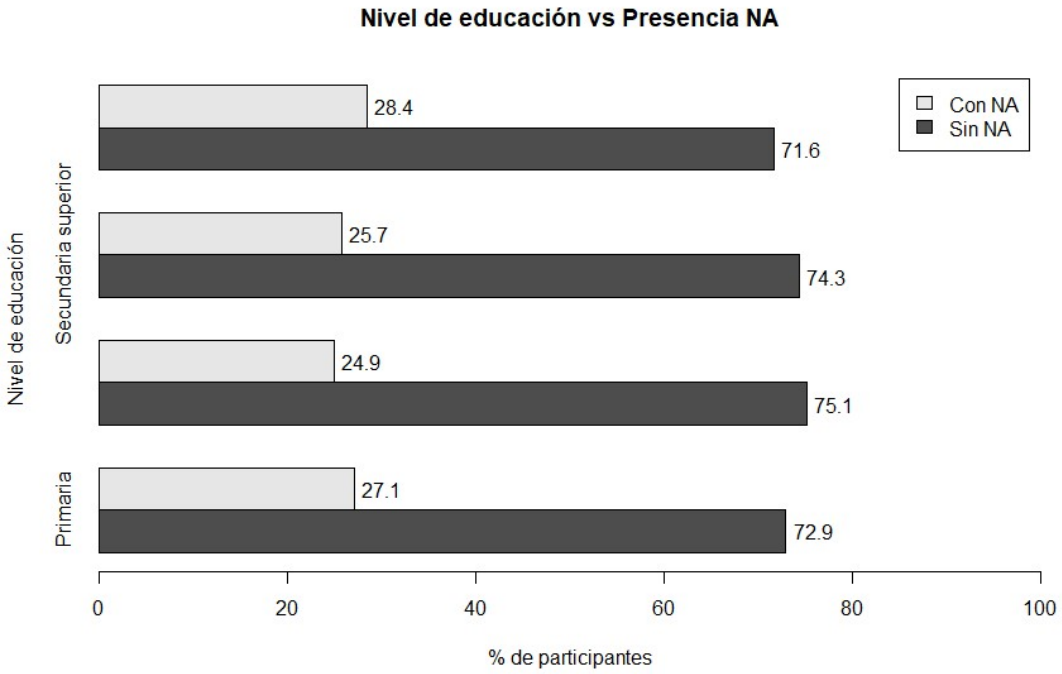
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.245	0.258

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.245>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en perseverantes no varía entre géneros.

Figura 14. Presencia de datos faltantes en errores perseverantes según Nivel de educación (10%)

```
##
## Column percentages:
##                         Education
## Missing_Perseverative Primaria Secundaria inicial Secundaria superior
##                Sin NA     72.9               75.1                74.3
##                Con NA     27.1               24.9                25.7
##                Total     100.0              100.0               100.0
##                Count     166.0              684.0               701.0
##                      Education
```

```
## Missing_Perseverative Universitaria
##                 Sin NA          71.6
##                 Con NA          28.4
##                 Total          100.0
##                 Count          964.0
```




```
##                       Missing_Perseverative
## Education              Sin NA Con NA
##   Primaria                121     45
##   Secundaria inicial      514    170
##   Secundaria superior     521    180
##   Universitaria           690    274
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.386	0.388

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.386>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en perseverantes no varía entre los niveles de educación.

En función de todos los resultados anteriores: el test ***Little´s MCAR-test***, las representaciones gráficas de los patrones de los datos faltantes y la ausencia de relación entre la presencia de datos faltantes y las características basales, tanto para el

nº de diseños únicos como para el nº de errores perseverantes, permite afirmar que los datos cumplen las características de **MCAR**.

5.5 Análisis de los patrones de los datos faltantes (20%)

Patrón datos faltantes en dataset modificado (20%)

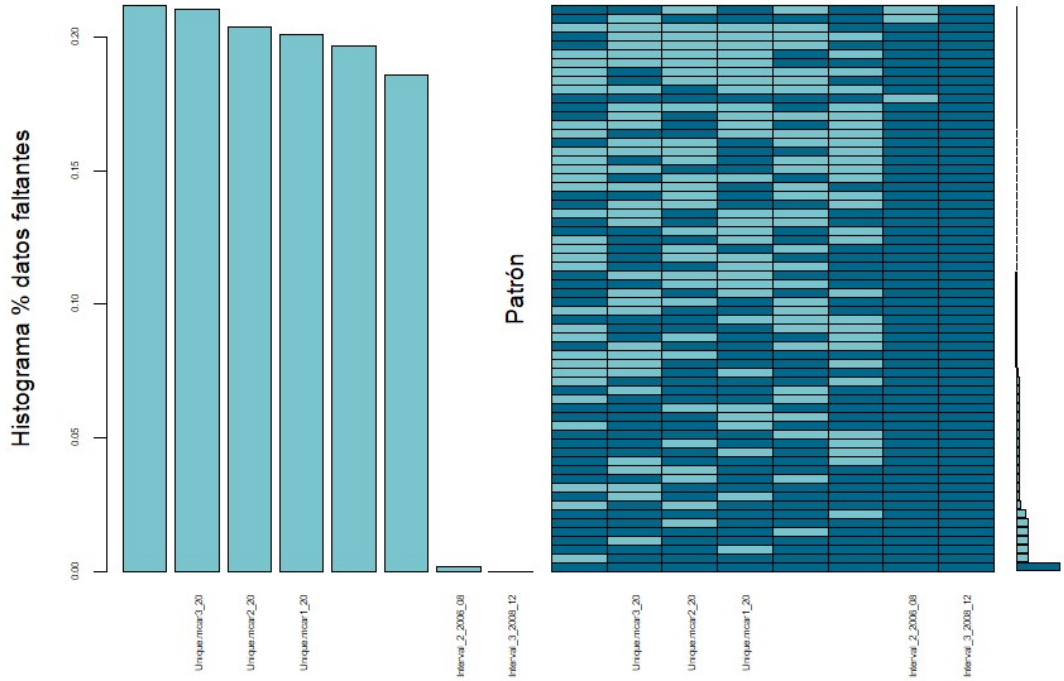
Se identifican los individuos con al menos 1 caso con dato faltante en global (incuyendo los casos de la variable) y separado para las 3 mediciones de **Unique** y **Perseverative** respectivamente.

```
##
## counts:
##
## Sin NA Con NA
##    636   1879
##
## percentages:
##
## Sin NA Con NA
##  25.29  74.71

##
## counts:
##
## Sin NA Con NA
##   1245   1270
##
## percentages:
##
## Sin NA Con NA
##   49.5   50.5

##
## counts:
##
## Sin NA Con NA
##   1297   1218
##
## percentages:
##
## Sin NA Con NA
##  51.57  48.43
```

- Finalmente del total de pacientes el 74.71% presenta al menos un valor perdido en alguna de los valores del data set, mientras que en las 3 mediciones de unique es el 50.5% y en las mediciones de persevaritive es el 48.43%.

De nuevo se utiliza _ agrr_ para obtener el patrón de datos faltantes en el dataset que contiene los datos faltantes generados de manera aleatoria en las 6 variables que contienen las mediciones de unique y perseverative (3 mediciones en cada una de ellas). Solo se muestran los valores faltantes en dichas variables y la variable *Interval 2 2006 08* puesto que para el resto de variables se dispuso de los datos en todos los individuos.

Figura 15. Patrón datos faltantes en dataset modificado (I) (20%)




```
##
##  Variables sorted by number of missings:
##                Variable       Count
##  Perseverative.mcar2_20 0.211928429
##         Unique.mcar3_20 0.210735586
##         Unique.mcar2_20 0.203976143
##         Unique.mcar1_20 0.201192843
##  Perseverative.mcar1_20 0.196819085
##  Perseverative.mcar3_20 0.186083499
##      Interval_2_2006_08 0.001590457
##      Interval_3_2008_12 0.000000000
```

No se aprecia ningún tipo de patrón que presente una frecuencia superior al resto.

El paquete ***BaylorEdPsych***, contiene el una función (***LittleMCAR***)que permite aplicar el test *Little´s MCAR-test*, para analizar la hipótesis nula de que *los datos faltantes son MCAR*, donde un resultado significativo indicaría que se debe rechazar dicha hipótesis. Se aplica sobre el dataset con datos faltantes generados de manera aleatoria (*se aplica*

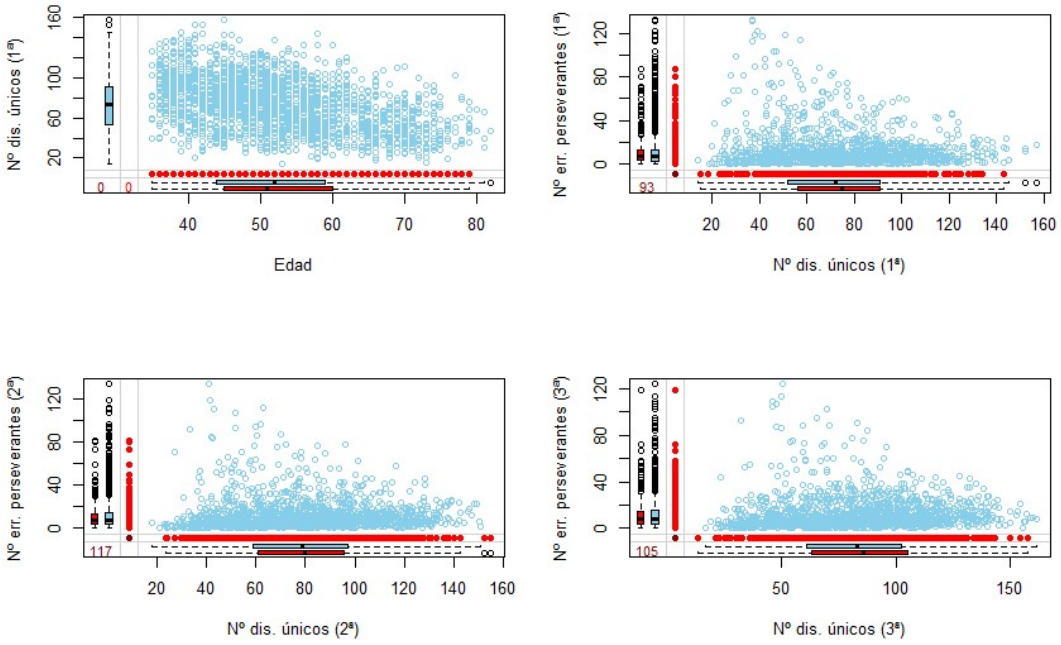
el test sobre las variables de interés para el estudio: edad, género, nivel de educación, 1ª medición únicos, 2ª medición únicos, 3ª medición únicos, 1ª medición perseverantes, 2ª medición perseverantes, 3ª medición perseverantes).

```
## this could take a while

## [1] 0.9930425
```

A partir del resultado de test de Little's, no se puede rechazar la hipótesis nula (*p-valor*=0.99304>α=0.05), y por lo tanto, como era de esperar al haber sido generados de manera aleatoria, los datos faltantes cumplen las características de **MCAR**.

A su vez el paquete VIM también dispone de más gráficos que permiten comparar la distribución de los datos faltantes por pares de variables. A modo de ejemplo se plantean 4 comparaciones:

Figura 16. Patrón datos faltantes en dataset modificado (II) (20%)



Los gráficos anteriores permiten comparar la distribución de los datos faltantes por pares de variables. De modo que en el primer gráfico, la fila de puntos rojos indicarían como se distribuirían los datos faltantes de la variable *nº de diseños únicos 1ª* en función de los valores de la variable *edad*, y los puntos azules serían el resto de valores. El hecho de que los boxplots representados en el eje x (*edad*) sean similares indica que los valores de edad con y sin valores faltantes en el nº de diseños únicos, son similares y es un indicativo de que los datos cumplen las premisas de **MCAR**. Igualmente entre el resto de pares de variables representadas, visualmente tampoco se han hallado diferencias entre la presencia o ausencia de datos faltantes.

5.5.1 Características basales Vs datos faltantes todas las variables (20%)

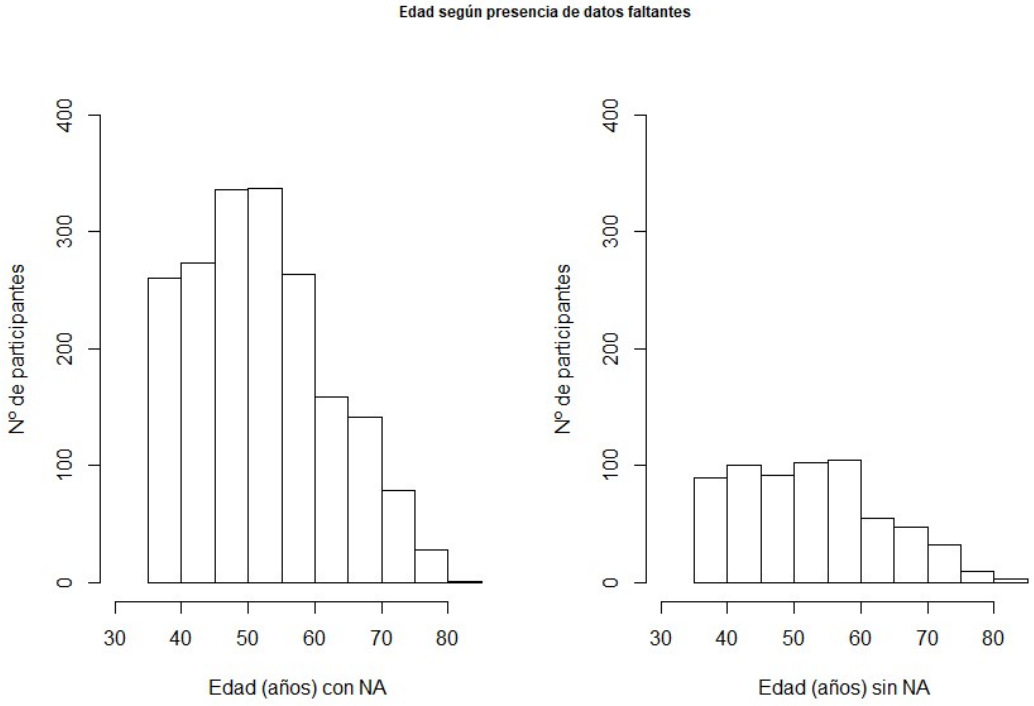
A continuación se analiza en mayor detalle las posibles diferencias entre la presencia y ausencia de datos faltantes en función de las características basales de los individuos.

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en alguna de las variables (20%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados del test RFFT. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 17. Edad en función de la presencia de datos faltantes (20%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA	0

Casos sin NA	0				
Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA	Media Edad
T-Test para muestras independientes	0.285	-0.436	1.481	Con NA	52.950
Test de Wilcoxon rank sum test with continuity correction	0.361	-1.000	1.000	Sin NA	52.427

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, *p-valor*=0<α=0.05. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes *p-valor*=0.361> α=0.05.

Género y nivel de educación vs datos faltantes (20%)

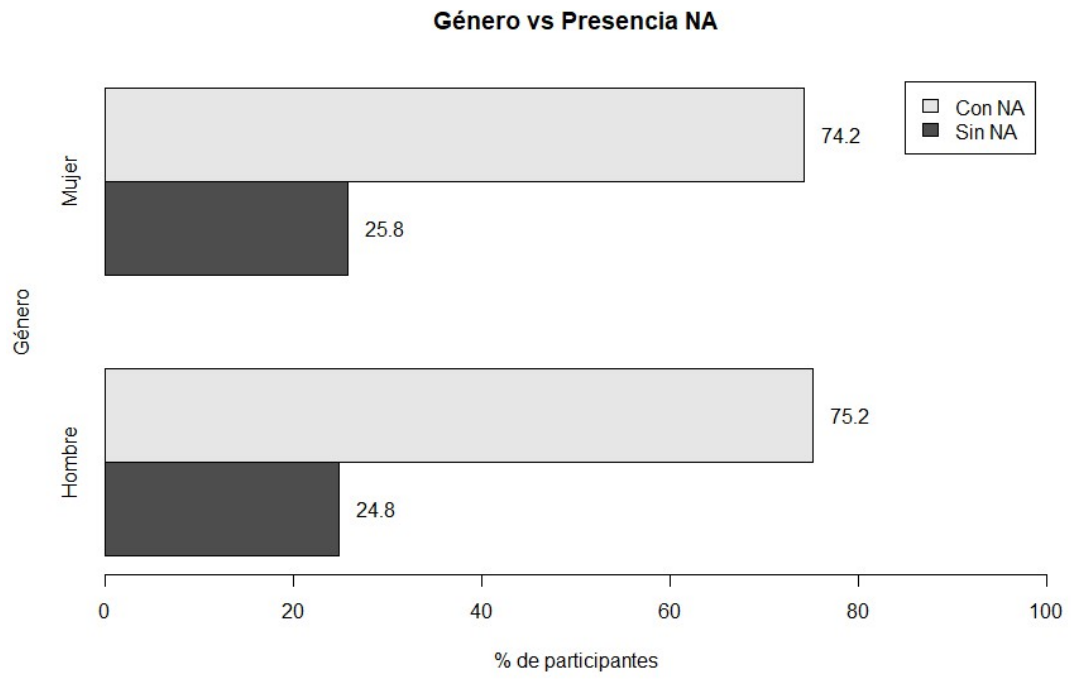
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (20%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 18. Presencia de datos faltantes según género (20%)

```
##
## Column percentages:
##              Gender
## Missing_data Hombre  Mujer
##       Sin NA   24.8   25.8
##       Con NA   75.2   74.2
##       Total   100.0  100.0
##       Count  1334.0 1181.0
```




```
##         Missing_data
## Gender   Sin NA Con NA
##   Hombre    331   1003
##   Mujer     305    876
```

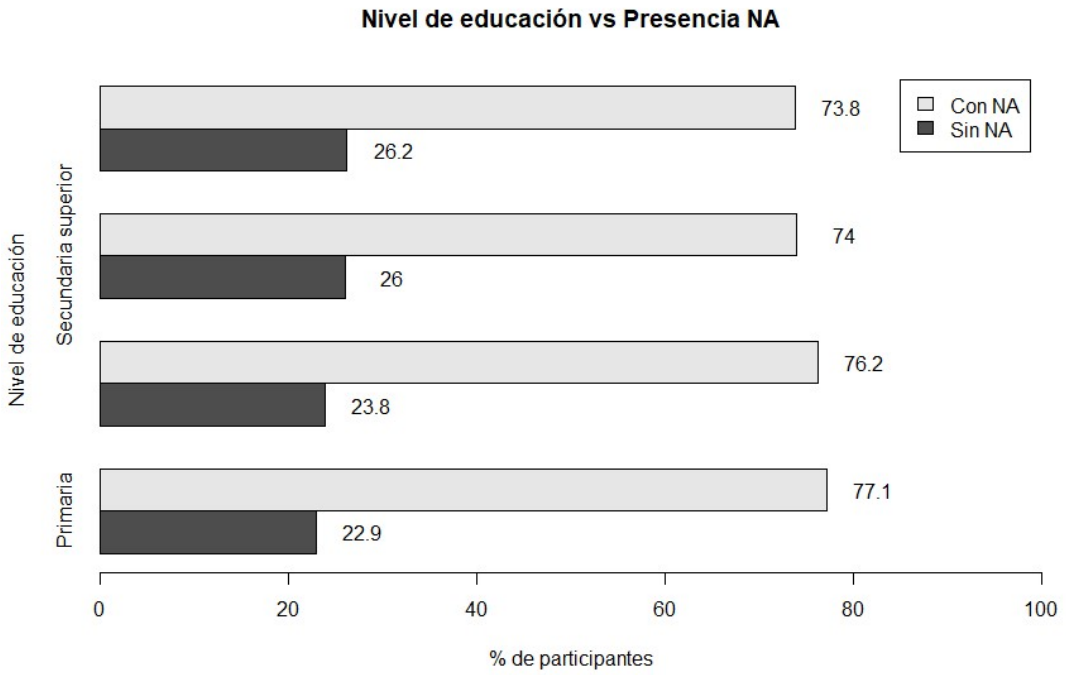
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.56	0.581

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.56>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes no varía entre géneros.

Figura 19. Presencia de datos faltantes según Nivel de educación (20%)

```
##
## Column percentages:
##             Education
## Missing_data Primaria Secundaria inicial Secundaria superior Universit
aria
##       Sin NA     22.9               23.8                 26
26.2
##       Con NA     77.1               76.2                 74
73.8
##       Total     100.0              100.0                100         1
```

```
00.0
##       Count     166.0              684.0                 701          9
64.0
```




```
##                       Missing_data
## Education             Sin NA Con NA
##   Primaria                38    128
##   Secundaria inicial     163    521
##   Secundaria superior    182    519
##   Universitaria          253    711
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.591	0.601

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.591>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes no varía entre los niveles de educación.

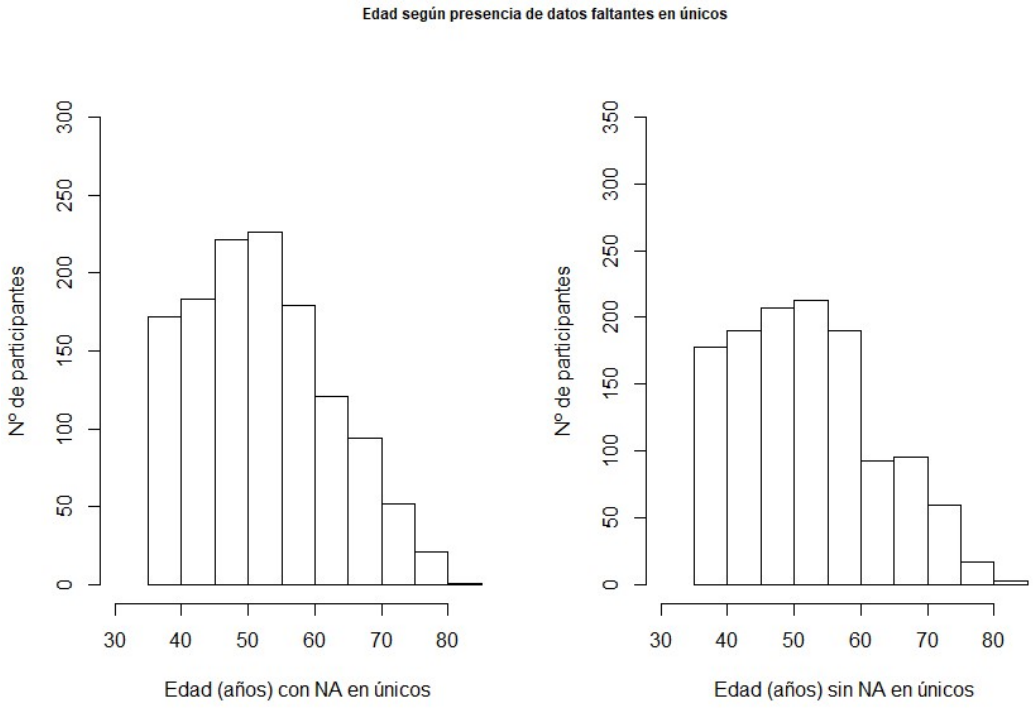
5.5.2 Características basales Vs datos faltantes en *diseños únicos* (20%)

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en diseños únicos (20%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados de las mediciones de la variable Unique. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 20. Edad en función de la presencia de datos faltantes en diseños únicos (20%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA en únicos	0
Casos sin NA en únicos	0

Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA en únicos	Media Edad
T-Test para muestras independientes	0.673	-0.989	0.639	Con NA	52.471
Test de Wilcoxon rank sum test with continuity correction	0.585	-1.000	1.000	Sin NA	52.646

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, *p-valor*=0<α=0.05. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes *p-valor*=0.585> α=0.05.

Género y nivel de educación vs datos faltantes (20%)

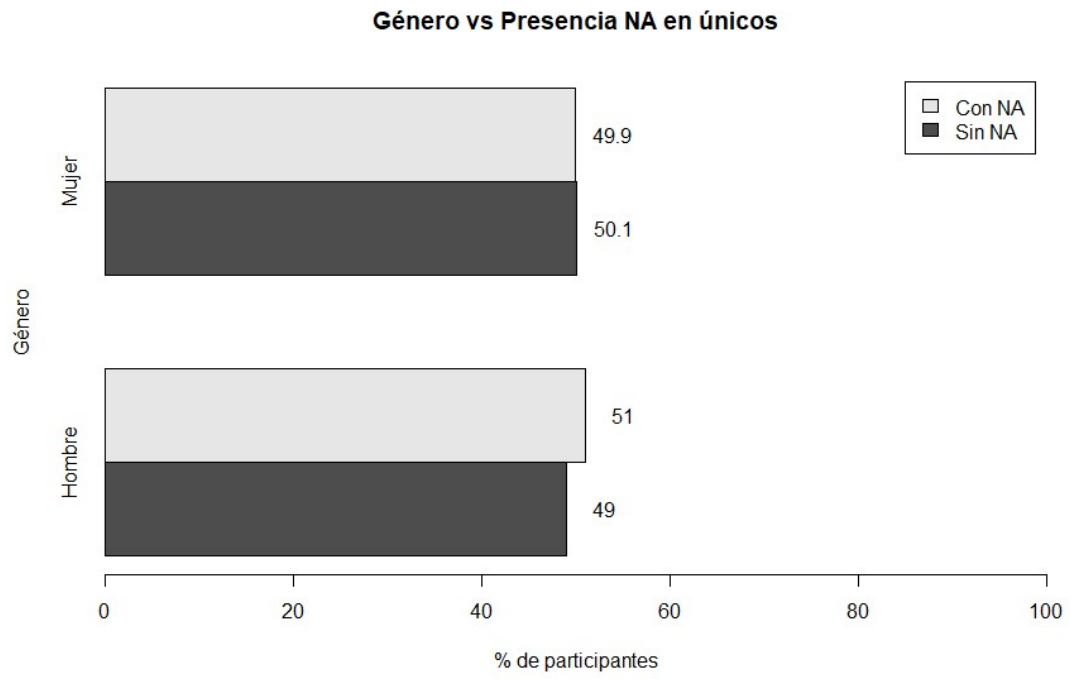
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (20%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 21. Presencia de datos faltantes en diseños únicos según género (20%)

```
##
## Column percentages:
##                Gender
## Missing_Unique Hombre  Mujer
##         Sin NA     49   50.1
##         Con NA     51   49.9
##         Total     100  100.0
##         Count    1334 1181.0
```




```
##         Missing_Unique
## Gender   Sin NA Con NA
##   Hombre    653    681
##   Mujer     592    589
```

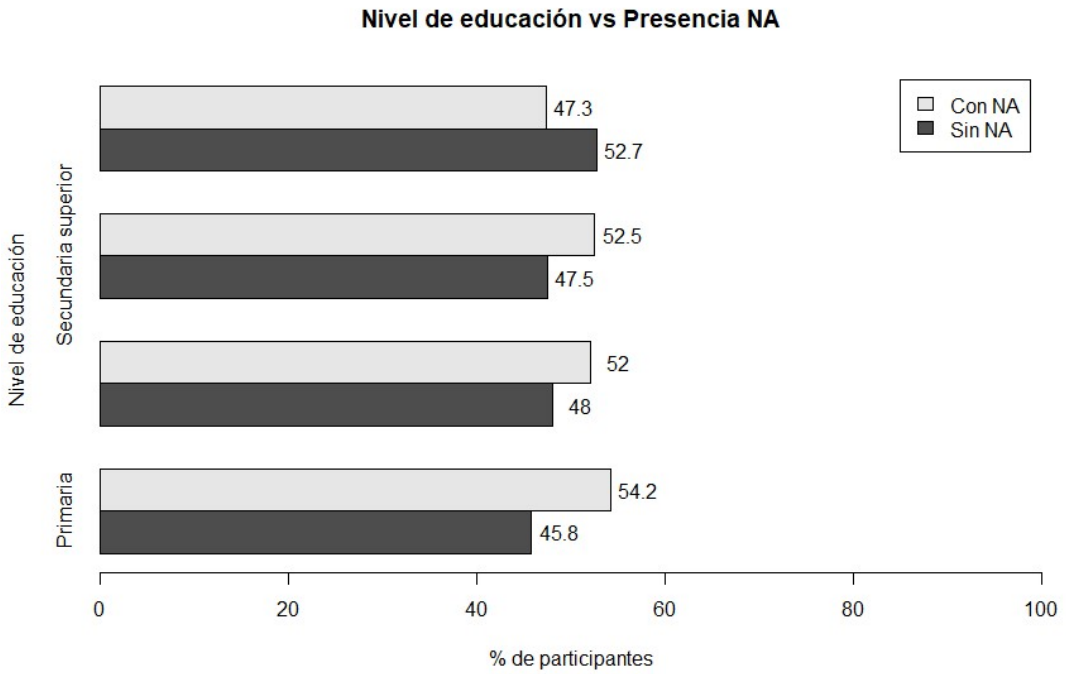
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.556	0.576

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.556>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en Unique no varía entre géneros.

Figura 22. Presencia de datos faltantes en diseños únicos según Nivel de educación (20%)

```
##
## Column percentages:
##                Education
## Missing_Unique Primaria Secundaria inicial Secundaria superior
##         Sin NA     45.8                 48                47.5
##         Con NA     54.2                 52                52.5
##         Total     100.0                100               100.0
##         Count     166.0                684               701.0
##                Education
## Missing_Unique Universitaria
```

```
##          Sin NA          52.7
##          Con NA          47.3
##          Total          100.0
##          Count          964.0
```




```
##                       Missing_Unique
## Education              Sin NA Con NA
##   Primaria                 76     90
##   Secundaria inicial      328    356
##   Secundaria superior     333    368
##   Universitaria           508    456
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.085	0.085

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.085>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en diseños únicos no varía entre los niveles de educación.

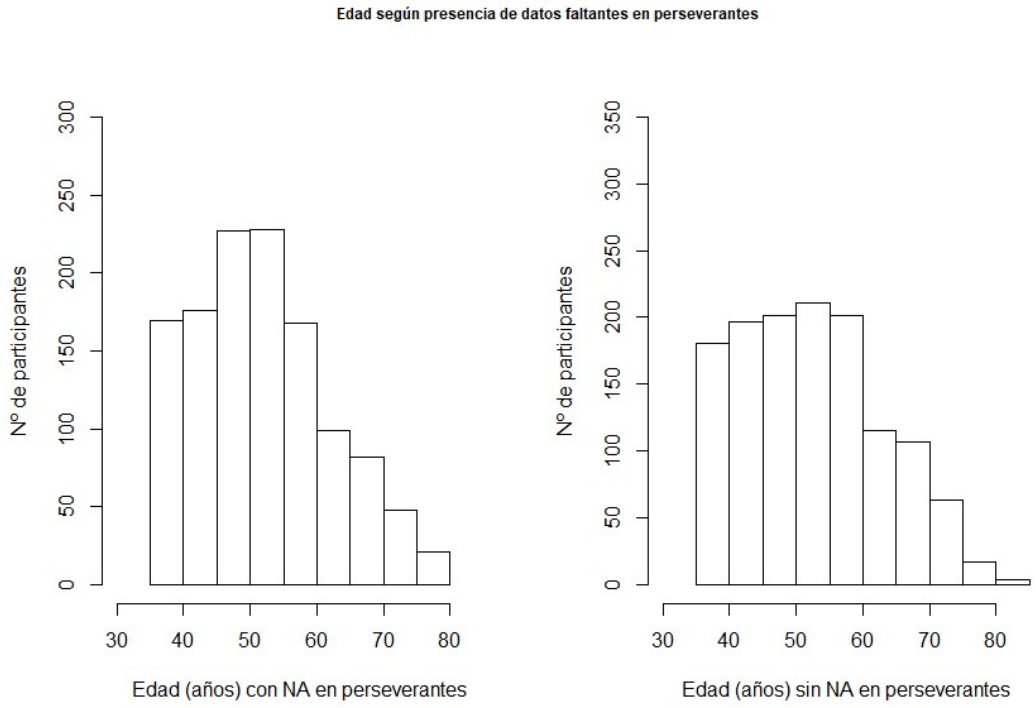
5.5.3 Características basales Vs datos faltantes en *errores perseverantes* (20%)

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en errores perseverantes (20%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados de las mediciones de los errores perseverantes. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 23. Edad en función de la presencia de datos faltantes en errores perseverantes (20%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA en perseverantes	0
Casos sin NA en perseverantes	0

Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA en perseverantes	Media Edad
T-Test para muestras independientes	0.080	-0.088	1.536	Con NA	52.910
Test de Wilcoxon rank sum test with continuity correction	0.105	0.000	2.000	Sin NA	52.186

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, $p\text{-}valor=0<\alpha=0.05$. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes $p\text{-}valor=0.105> \alpha=0.05$.

Género y nivel de educación vs datos faltantes (20%)

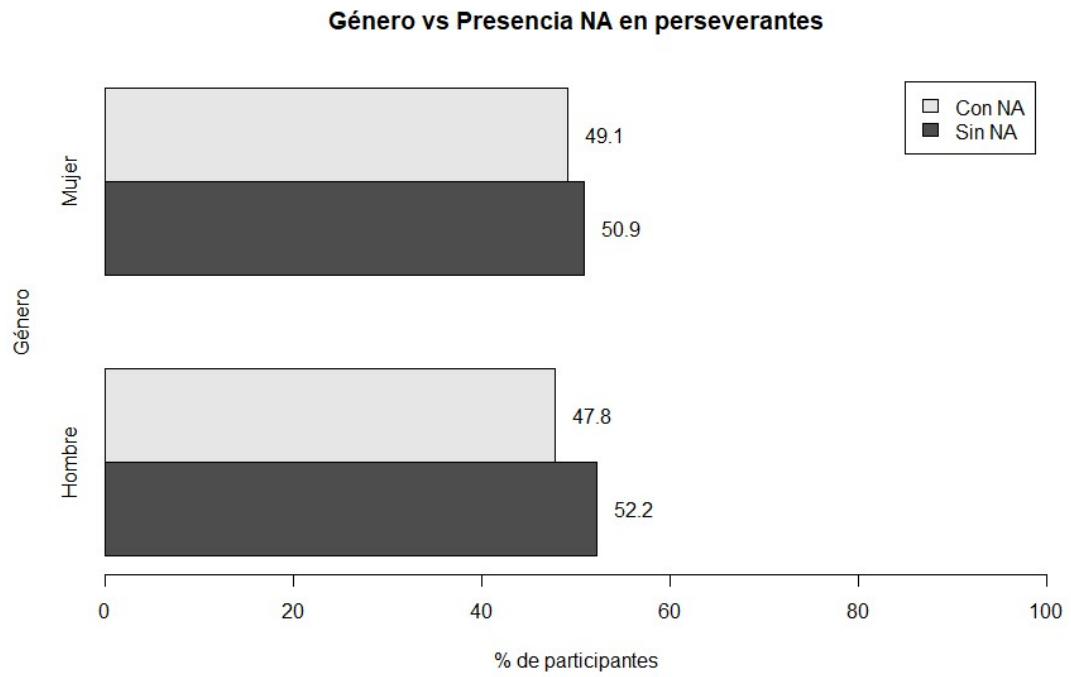
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (20%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 24. Presencia de datos faltantes en errores perseverantes según género (20%)

```
##
## Column percentages:
##                         Gender
## Missing_Perseverative Hombre  Mujer
##                Sin NA   52.2   50.9
##                Con NA   47.8   49.1
##                Total   100.0  100.0
##                Count  1334.0 1181.0
```




```
##         Missing_Perseverative
## Gender   Sin NA Con NA
##   Hombre    696    638
##   Mujer     601    580
```

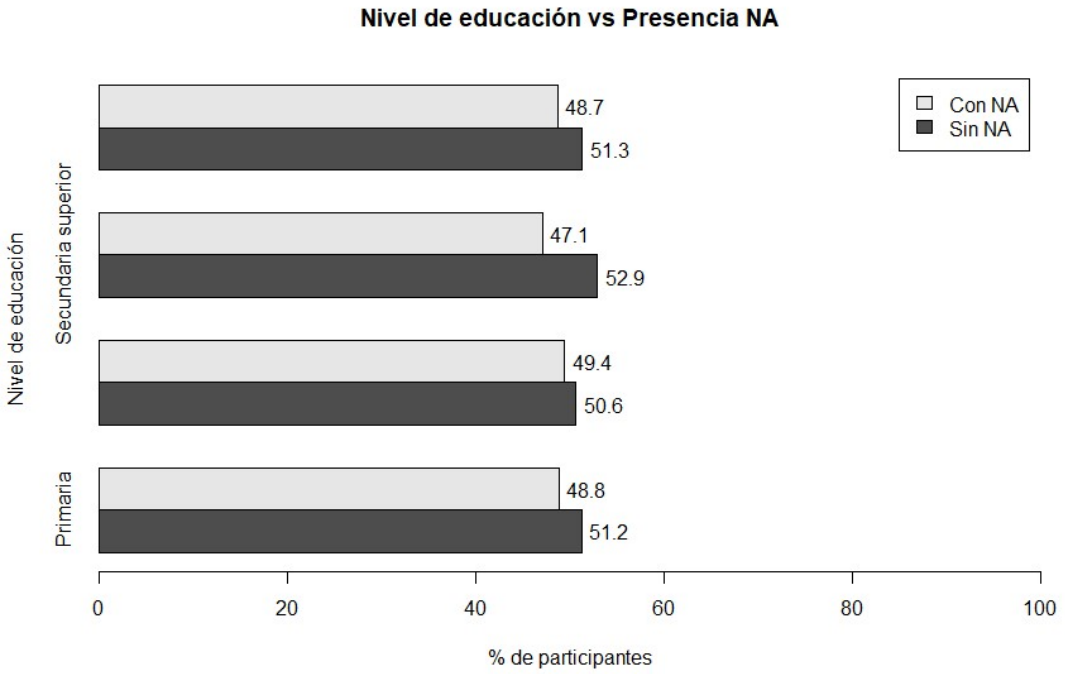
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.52	0.523

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.52>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en perseverantes no varía entre géneros.

Figura 25. Presencia de datos faltantes en errores perseverantes según Nivel de educación (20%)

```
##
## Column percentages:
##                         Education
## Missing_Perseverative Primaria Secundaria inicial Secundaria superior
##                Sin NA     51.2               50.6                52.9
##                Con NA     48.8               49.4                47.1
##                Total     100.0              100.0               100.0
##                Count     166.0              684.0               701.0
##                      Education
```

```
## Missing_Perseverative Universitaria
##                 Sin NA          51.3
##                 Con NA          48.7
##                 Total          100.0
##                 Count          964.0
```




```
##                       Missing_Perseverative
## Education              Sin NA Con NA
##   Primaria                 85     81
##   Secundaria inicial      346    338
##   Secundaria superior     371    330
##   Universitaria           495    469
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.847	0.848

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.847>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en perseverantes no varía entre los niveles de educación.

En función de todos los resultados anteriores: el test ***Little´s MCAR-test***, las representaciones gráficas de los patrones de los datos faltantes y la ausencia de relación entre la presencia de datos faltantes y las características basales, tanto para el

nº de diseños únicos como para el nº de errores perseverantes, permite afirmar que los datos cumplen las características de **MCAR**.

5.6 Análisis de los patrones de los datos faltantes (30%)

Patrón datos faltantes en dataset modificado (30%)

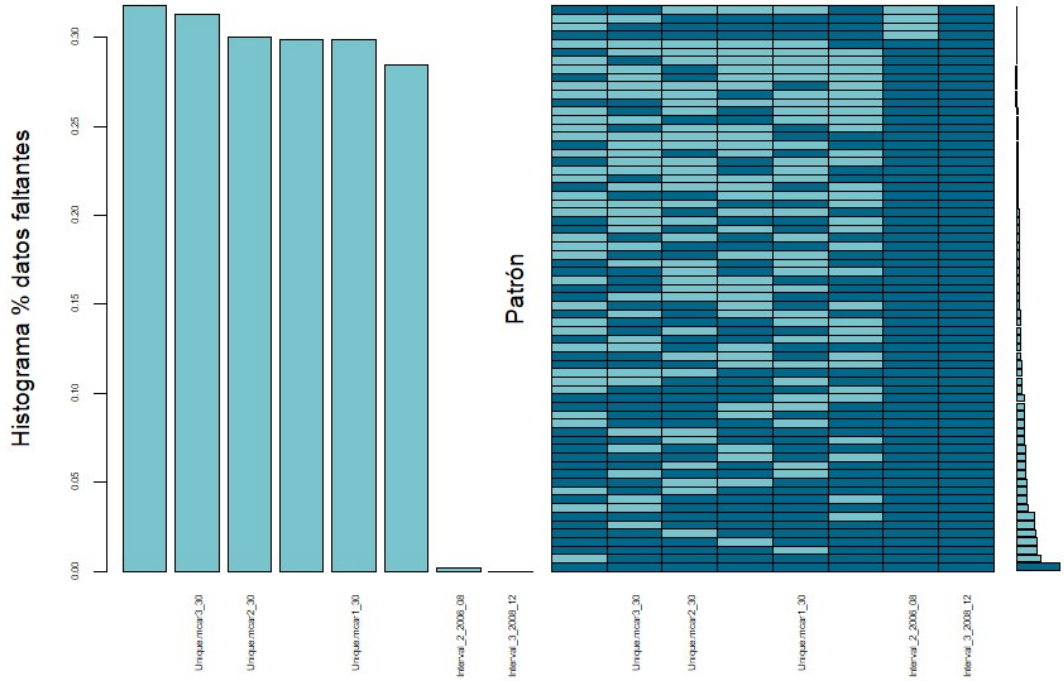
Se identifican los individuos con al menos 1 caso con dato faltante en global (incuyendo los casos de la variable) y separado para las 3 mediciones de **Unique** y **Perseverative** respectivamente.

```
##
## counts:
##
## Sin NA Con NA
##    275   2240
##
## percentages:
##
## Sin NA Con NA
##  10.93  89.07

##
## counts:
##
## Sin NA Con NA
##    832   1683
##
## percentages:
##
## Sin NA Con NA
##  33.08  66.92

##
## counts:
##
## Sin NA Con NA
##    833   1682
##
## percentages:
##
## Sin NA Con NA
##  33.12  66.88
```

- Finalmente del total de pacientes el 89.07% presenta al menos un valor perdido en alguna de los valores del data set, mientras que en las 3 mediciones de unique es el 66.92% y en las mediciones de persevaritive es el 66.88%.

De nuevo se utiliza _ agrr_ para obtener el patrón de datos faltantes en el dataset que contiene los datos faltantes generados de manera aleatoria en las 6 variables que contienen las mediciones de unique y perseverative (3 mediciones en cada una de ellas). Solo se muestran los valores faltantes en dichas variables y la variable *Interval 2 2006 08* puesto que para el resto de variables se dispuso de los datos en todos los individuos.

Figura 26. Patrón datos faltantes en dataset modificado (I) (30%)




```
##
##  Variables sorted by number of missings:
##                Variable       Count
##  Perseverative.mcar2_30 0.317693837
##         Unique.mcar3_30 0.312922465
##         Unique.mcar2_30 0.300198807
##  Perseverative.mcar1_30 0.299005964
##         Unique.mcar1_30 0.298608350
##  Perseverative.mcar3_30 0.284294235
##      Interval_2_2006_08 0.001590457
##      Interval_3_2008_12 0.000000000
```

No se aprecia ningún tipo de patrón que presente una frecuencia superior al resto.

El paquete ***BaylorEdPsych***, contiene el una función (***LittleMCAR***)que permite aplicar el test *Little´s MCAR-test*, para analizar la hipótesis nula de que *los datos faltantes son MCAR*, donde un resultado significativo indicaría que se debe rechazar dicha hipótesis. Se aplica sobre el dataset con datos faltantes generados de manera aleatoria (*se aplica*

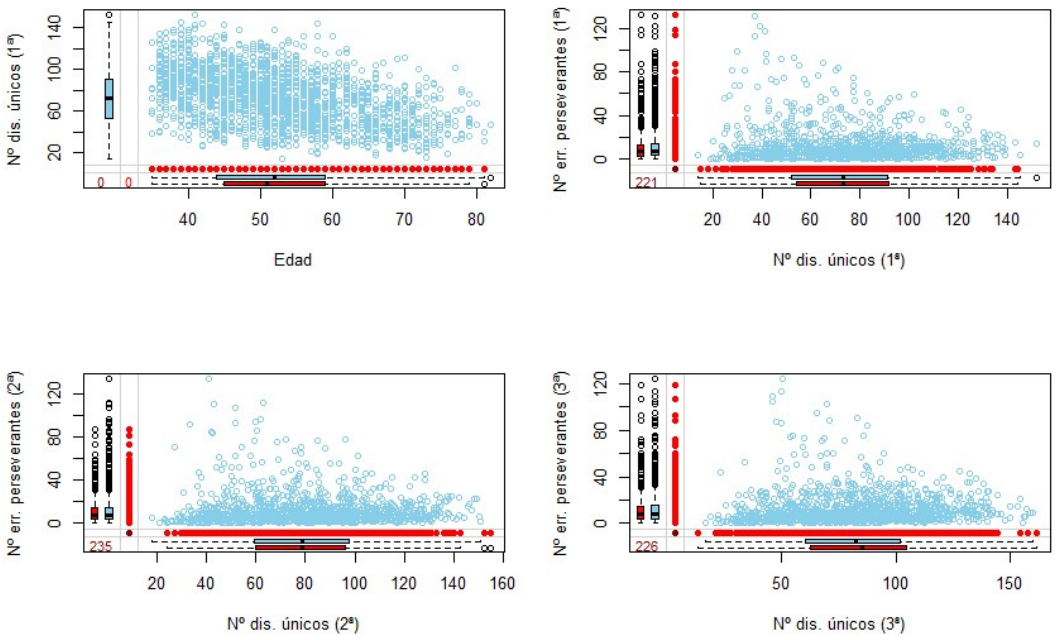
el test sobre las variables de interés para el estudio: edad, género, nivel de educación, 1ª medición únicos, 2ª medición únicos, 3ª medición únicos, 1ª medición perseverantes, 2ª medición perseverantes, 3ª medición perseverantes).

```
## this could take a while

## [1] 0.5108541
```

A partir del resultado de test de Little's, no se puede rechazar la hipótesis nula (*p-valor*=0.51085>α=0.05), y por lo tanto, como era de esperar al haber sido generados de manera aleatoria, los datos faltantes cumplen las características de **MCAR**.

A su vez el paquete VIM también dispone de más gráficos que permiten comparar la distribución de los datos faltantes por pares de variables. A modo de ejemplo se plantean 4 comparaciones:

Figura 27. Patrón datos faltantes en dataset modificado (II) (30%)



Los gráficos anteriores permiten comparar la distribución de los datos faltantes por pares de variables. De modo que en el primer gráfico, la fila de puntos rojos indicarían como se distribuirían los datos faltantes de la variable *nº de diseños únicos 1ª* en función de los valores de la variable *edad*, y los puntos azules serían el resto de valores. El hecho de que los boxplots representados en el eje x (*edad*) sean similares indica que los valores de edad con y sin valores faltantes en el nº de diseños únicos, son similares y es un indicativo de que los datos cumplen las premisas de **MCAR**. Igualmente entre el resto de pares de variables representadas, visualmente tampoco se han hallado diferencias entre la presencia o ausencia de datos faltantes.

5.6.1 Características basales Vs datos faltantes todas las variables (30%)

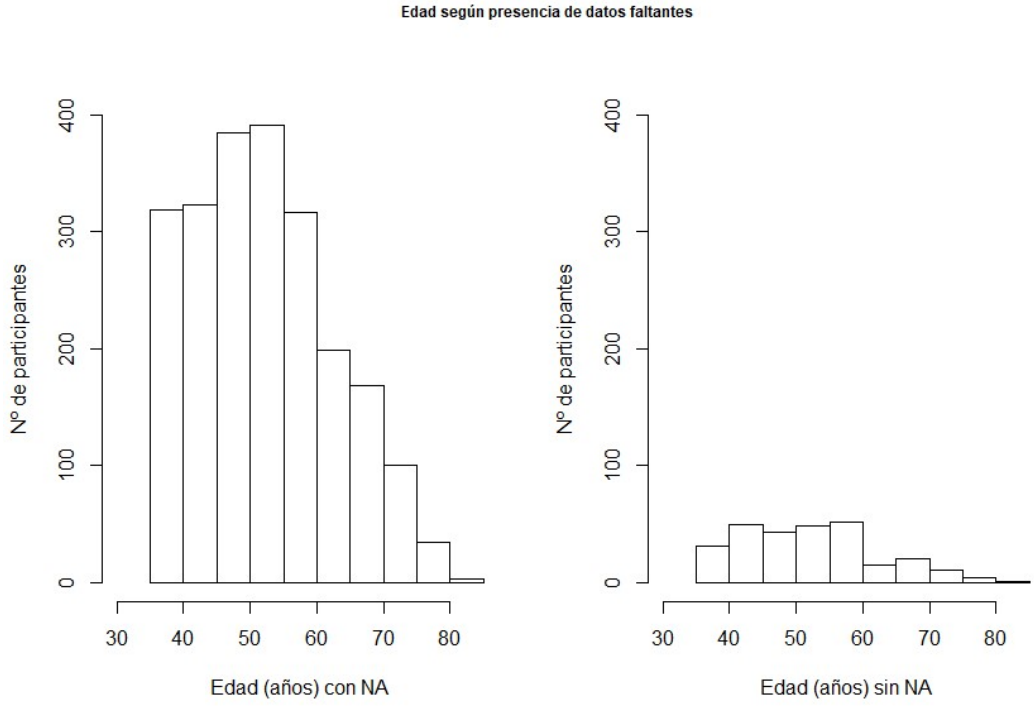
A continuación se analiza en mayor detalle las posibles diferencias entre la presencia y ausencia de datos faltantes en función de las características basales de los individuos.

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en alguna de las variables (30%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados del test RFFT. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 28. Edad en función de la presencia de datos faltantes (30%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA	0

Casos sin NA	0				
Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA	Media Edad
T-Test para muestras independientes	0.890	-1.197	1.379	Con NA	52.640
Test de Wilcoxon rank sum test with continuity correction	0.851	-1.000	1.000	Sin NA	52.549

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, *p-valor*=0<α=0.05. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes *p-valor*=0.851> α=0.05.

Género y nivel de educación vs datos faltantes (30%)

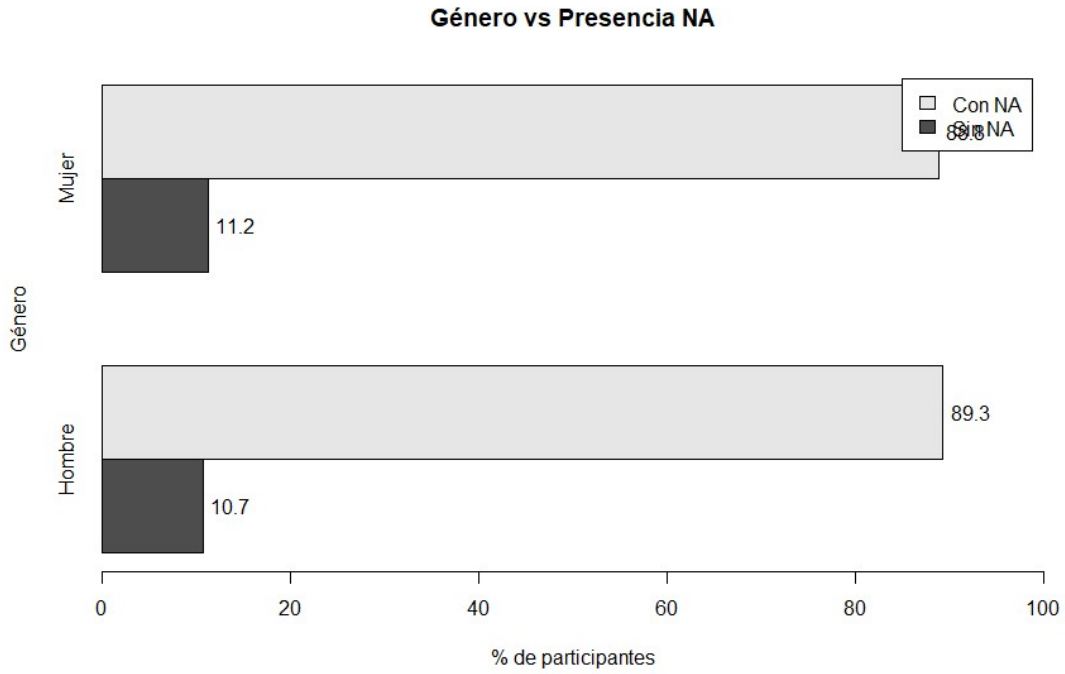
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (30%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 29. Presencia de datos faltantes según género (30%)

```
##
## Column percentages:
##              Gender
## Missing_data Hombre  Mujer
##       Sin NA   10.7   11.2
##       Con NA   89.3   88.8
##       Total   100.0  100.0
##       Count  1334.0 1181.0
```




```
##         Missing_data
## Gender   Sin NA Con NA
##   Hombre    143   1191
##   Mujer     132   1049
```

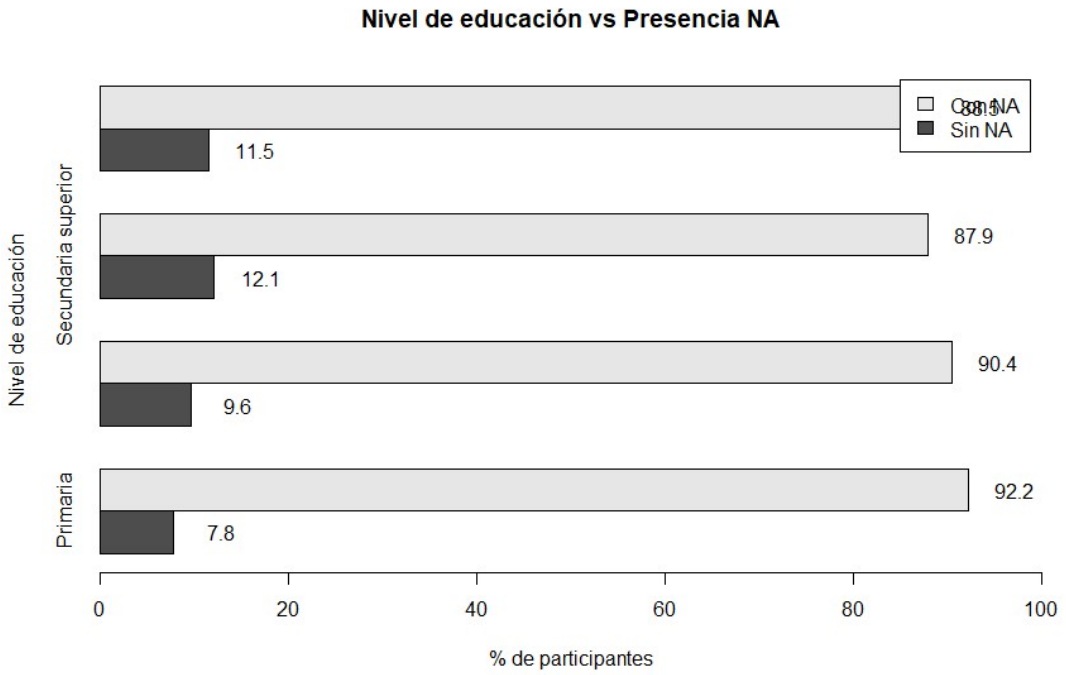
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.714	0.749

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.714>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes no varía entre géneros.

Figura 30. Presencia de datos faltantes según Nivel de educación (30%)

```
##
## Column percentages:
##             Education
## Missing_data Primaria Secundaria inicial Secundaria superior Universit
aria
##       Sin NA      7.8                9.6                12.1
11.5
##       Con NA     92.2               90.4                87.9
88.5
##       Total     100.0              100.0               100.0         1
```

```
00.0
##       Count      166.0              684.0              701.0          9
64.0
```




```
##                       Missing_data
## Education             Sin NA Con NA
##   Primaria                13    153
##   Secundaria inicial      66    618
##   Secundaria superior     85    616
##   Universitaria          111    853
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.245	0.253

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.245>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes no varía entre los niveles de educación.

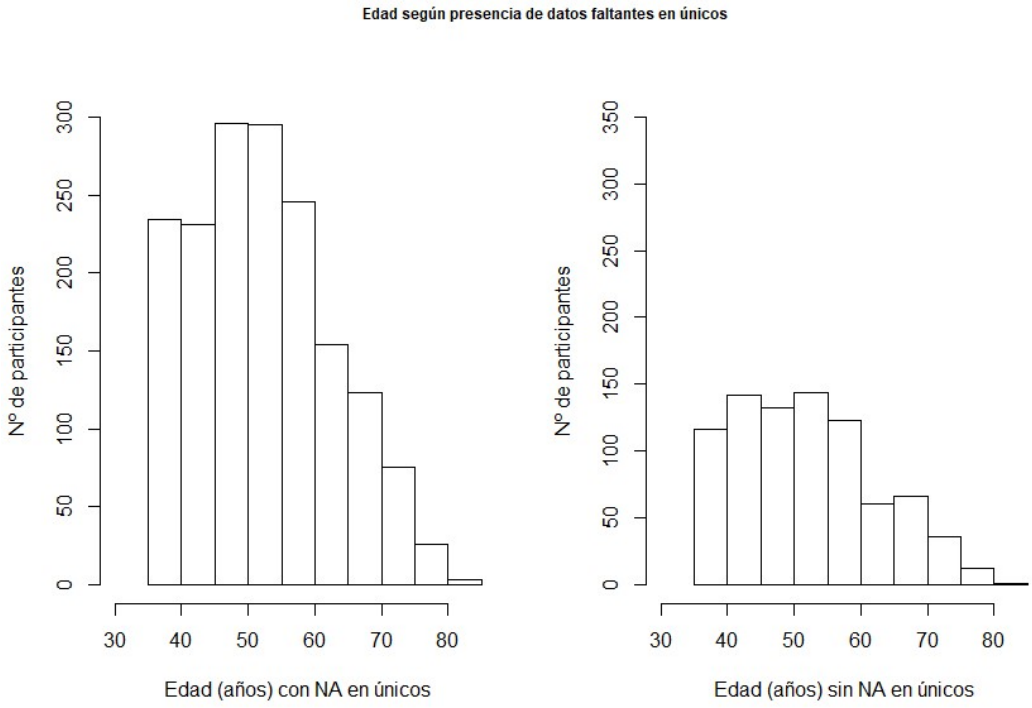
5.6.2 Características basales Vs datos faltantes en *diseños únicos* (30%)

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en diseños únicos (30%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados de las mediciones de la variable Unique. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 31. Edad en función de la presencia de datos faltantes en diseños únicos (30%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA en únicos	0
Casos sin NA en únicos	0

Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA en únicos	Media Edad
T-Test para muestras independientes	0.337	-1.29	0.441	Con NA	52.275
Test de Wilcoxon rank sum test with continuity correction	0.297	-1.00	0.000	Sin NA	52.699

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, $p\text{-}valor=0<\alpha=0.05$. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes $p\text{-}valor=0.297> \alpha=0.05$.

Género y nivel de educación vs datos faltantes (30%)

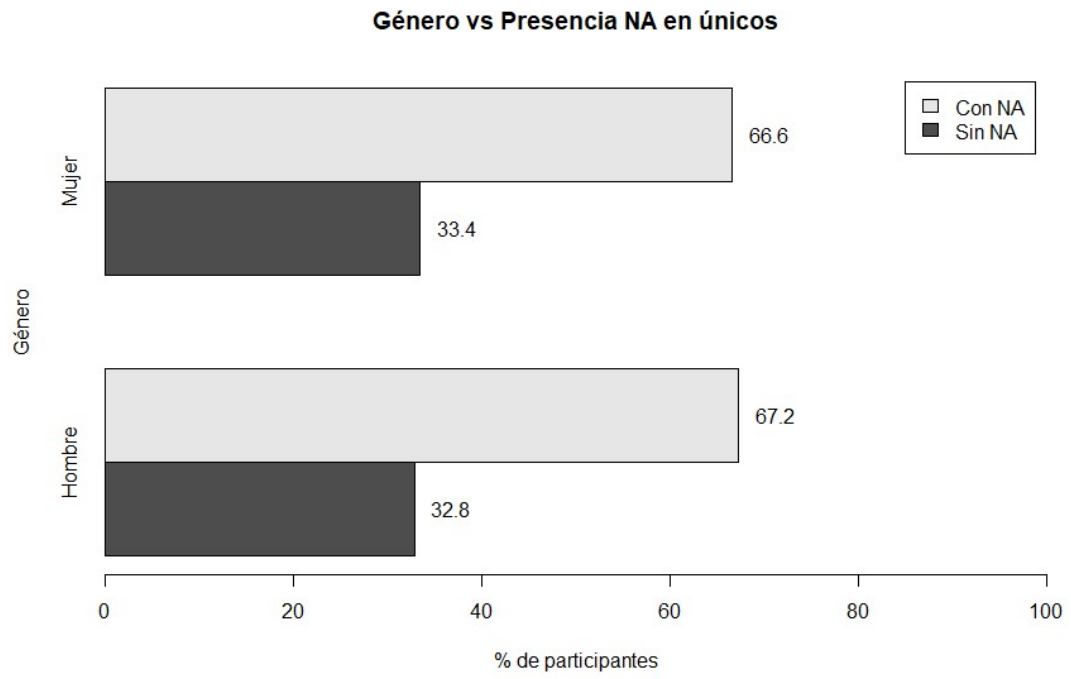
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (30%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 32. Presencia de datos faltantes en diseños únicos según género (30%)

```
##
## Column percentages:
##                Gender
## Missing_Unique Hombre  Mujer
##         Sin NA   32.8   33.4
##         Con NA   67.2   66.6
##         Total   100.0  100.0
##         Count  1334.0 1181.0
```




```
##         Missing_Unique
## Gender   Sin NA Con NA
##   Hombre    437    897
##   Mujer     395    786
```

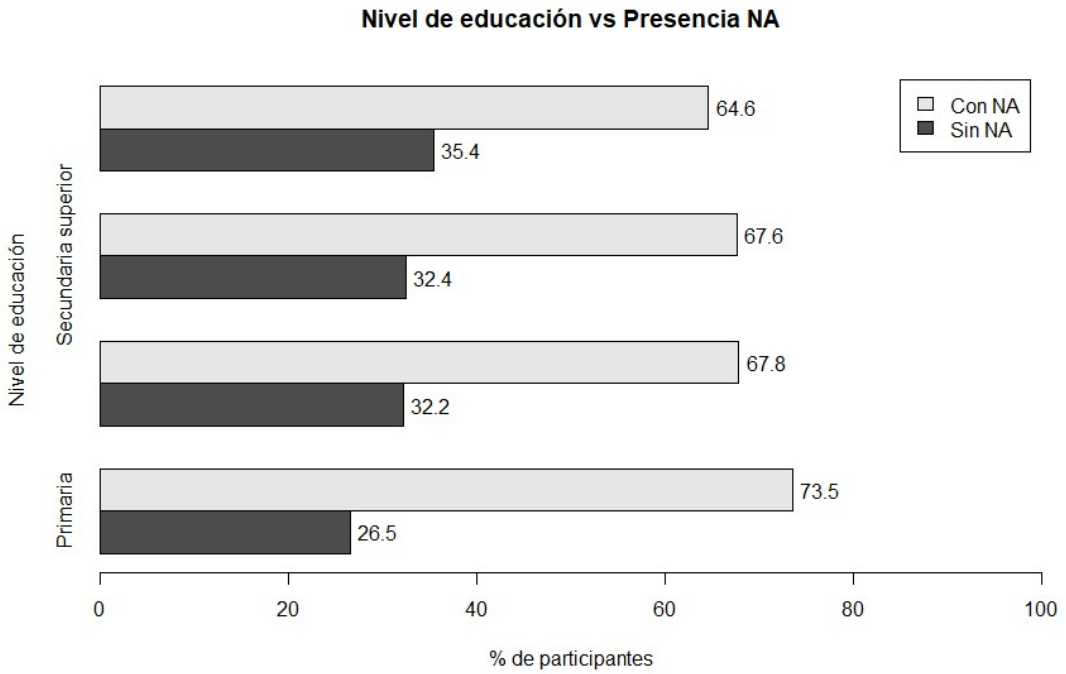
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.715	0.734

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.715>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en Unique no varía entre géneros.

Figura 33. Presencia de datos faltantes en diseños únicos según Nivel de educación (30%)

```
##
## Column percentages:
##                Education
## Missing_Unique Primaria Secundaria inicial Secundaria superior
##         Sin NA     26.5                32.2                32.4
##         Con NA     73.5                67.8                67.6
##         Total     100.0               100.0               100.0
##         Count     166.0               684.0               701.0
##                Education
## Missing_Unique Universitaria
```

```
##          Sin NA          35.4
##          Con NA          64.6
##          Total          100.0
##          Count          964.0
```




```
##                       Missing_Unique
## Education              Sin NA Con NA
##   Primaria                 44    122
##   Secundaria inicial      220    464
##   Secundaria superior     227    474
##   Universitaria           341    623
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.114	0.115

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.114>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en diseños únicos no varía entre los niveles de educación.

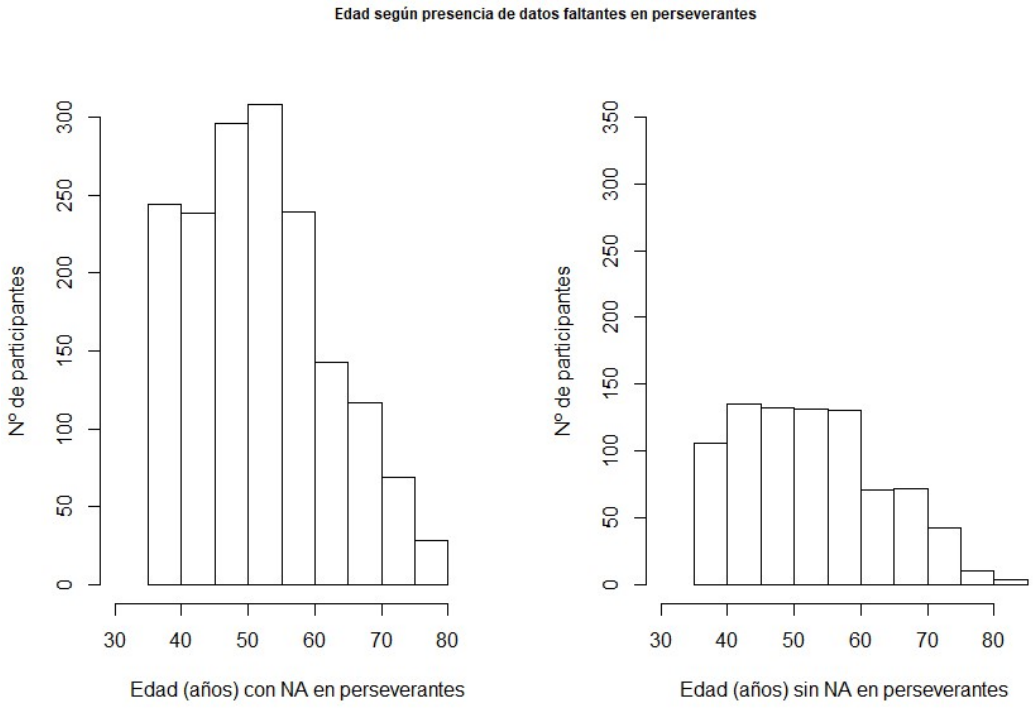
5.6.3 Características basales Vs datos faltantes en *errores perseverantes* (30%)

- Edad (variable cuantiativa)
- Género y Nivel de Educación (variable cualitativas)

Edad vs datos faltantes en errores perseverantes (30%)

Se realiza una prueba de normalidad de la variable *edad* en cada grupo según la presencia o no de datos faltantes en los resultados de las mediciones de los errores perseverantes. Para comparar la edad media entre ambos grupos se utiliza un test paramétrico y uno no paramétrico. Se fija el nivel de significación en α=0.05.

Figura 34. Edad en función de la presencia de datos faltantes en errores perseverantes (30%)



Edad	p-valor (Shapiro-Wilk normality test)
Casos con NA en perseverantes	0
Casos sin NA en perseverantes	0

Tipo de test	p-valor	IC 95% diferencia inf	IC 95% diferencia sup	Presencia NA en perseverantes	Media Edad
T-Test para muestras independientes	0.102	-0.144	1.606	Con NA	53.048
Test de Wilcoxon rank sum test with continuity correction	0.139	0.000	2.000	Sin NA	52.317

Observando los resultados obtenidos, no se pudo asumir la normalidad de los datos, ambos p-valores en el test saphiro-wilks son significativos, $p\text{-}valor=0<\alpha=0.05$. Por lo tanto para analizar si hay diferencias en las edades entre ambos grupos de pacientes, se tendrá que utilizar el resultado obtenido en el test de Wilcoxon y se puede afirmar que no se han hallado diferencias estadísticamente significativas en las edades entre los grupos definidos en función de la presencia/ausencia de datos faltantes $p\text{-}valor=0.139> \alpha=0.05$.

Género y nivel de educación vs datos faltantes (30%)

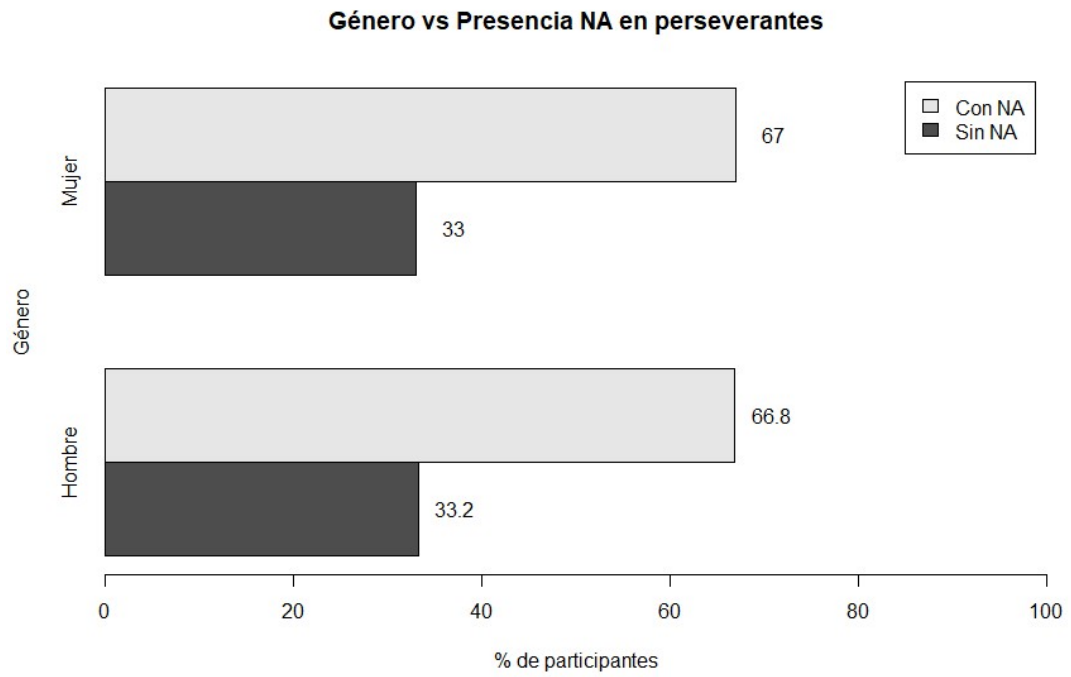
Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Género vs Datos faltantes (30%)

Tanto para analizar si hay diferencias en la proporción de datos faltantes entre géneros y entre los grupos del nivel de educación, se utiliza el test de Chi-cuadrado o en su defecto el test de Fisher.

Figura 35. Presencia de datos faltantes en errores perseverantes según género (30%)

```
##
## Column percentages:
##                         Gender
## Missing_Perseverative Hombre Mujer
##                Sin NA   33.2    33
##                Con NA   66.8    67
##                Total   100.0   100
##                Count  1334.0  1181
```




```
##         Missing_Perseverative
## Gender   Sin NA Con NA
##   Hombre    443    891
##   Mujer     390    791
```

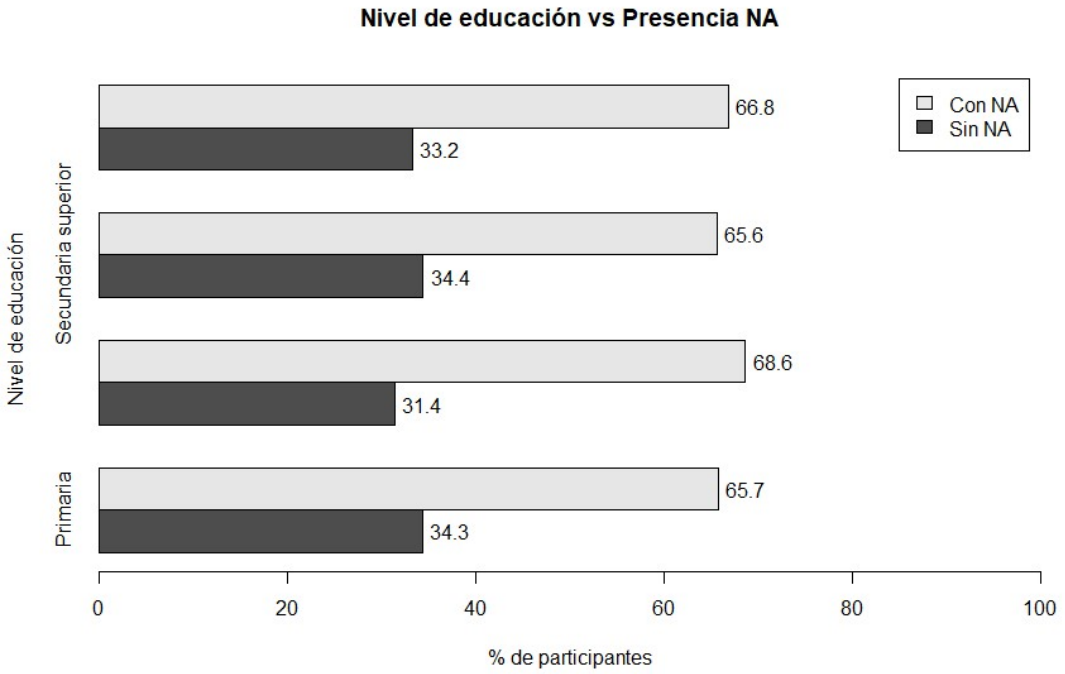
	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Género vs Presencia de NA	0	0.921	0.932

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.921>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en perseverantes no varía entre géneros.

Figura 36. Presencia de datos faltantes en errores perseverantes según Nivel de educación (30%)

```
##
## Column percentages:
##                         Education
## Missing_Perseverative Primaria Secundaria inicial Secundaria superior
##                Sin NA     34.3               31.4                34.4
##                Con NA     65.7               68.6                65.6
##                Total     100.0              100.0               100.0
##                Count     166.0              684.0               701.0
##                      Education
```

```
## Missing_Perseverative Universitaria
##                 Sin NA          33.2
##                 Con NA          66.8
##                 Total          100.0
##                 Count          964.0
```




```
##                       Missing_Perseverative
## Education              Sin NA Con NA
##   Primaria                 57    109
##   Secundaria inicial      215    469
##   Secundaria superior     241    460
##   Universitaria           320    644
```

	% de celdas con valores esperados <5%	P-valor Chi-cuadrado	P-valor Test exacto de Fisher
Nivel de educación vs Presencia de NA	0	0.684	0.679

Como se puede observar en la tabla de resultados que se presenta a continuación, el % de celdas de la tabla de contingencia con valor esperado <5 es 0% y por lo tanto el test que utilizaremos para contrastar las hipótesis es el Test de Chi-cuadrado y como el p-valor=0.684>α=0.05, no podemos rechazar la hipótesis nula y por lo tanto el % de datos faltantes en perseverantes no varía entre los niveles de educación.

En función de todos los resultados anteriores: el test ***Little´s MCAR-test***, las representaciones gráficas de los patrones de los datos faltantes y la ausencia de relación entre la presencia de datos faltantes y las características basales, tanto para el

nº de diseños únicos como para el nº de errores perseverantes, permite afirmar que los datos cumplen las características de **MCAR**.

---

6. TRATAMIENTO DE DATOS FALTANTES

Como ya se ha detallado previamente en la sección 10.1 de la memoria, existen diversas técnicas para el tratamiento de datos faltantes. El objetivo de este TFM, además de detallar las diversas técnicas existentes, es la aplicación de los métodos estudiados mediante el análisis de una base de datos biomédicos. A modo de ejemplo se han seleccionado algunos de los métodos más habitualmente utilizados por la comunidad científica:

- En primer lugar se propone uno de los métodos de eliminación, *Eliminación de los casos (listwise)*.
- En segundo lugar se propone utilizar dos métodos de imputación simples: *Imputación reemplazando por la media* e *Imputación por regresión simple (media condicional)* .
- Finalmente se propone uno de los métodos modernos sobre los que más investigación se esta llevando a cabo en los últimos años como es el *Algoritmo de imputación múltiple (MI) (método Predictive Mean Matching o Equiparación de media predictiva, PMM)*.

Los 4 métodos se han aplicado a cada uno de los 3 casos: 10%, 20% y 30% de datos faltantes.

6.1 Eliminación de los casos (listwise) (10%)

En este primer caso se generan dos dataframes en los que se excluyen todos los casos para los que hay algún dato faltante (por separado para los casos de *Nº de diseños únicos* y *Errores perseverantes*), para el dataframe con 10% de datos faltantes.

En el escenario del 10% de datos faltantes:
- El nº de participantes disponibles para el análisis del nº de diseños únicos fue de 1815 (nº de participantes con las 3 mediciones completas).
- El nº de participantes disponibles para el análisis del nº de errores perseverantes fue de 1846 (nº de participantes con las 3 mediciones completas).

6.2 Eliminación de los casos (listwise) (20%)

En este primer caso se generan dos dataframes en los que se excluyen todos los casos para los que hay algún dato faltante (por separado para los casos de *Nº de diseños únicos* y *Errores perseverantes*), para el dataframe con 20% de datos faltantes.

En el escenario del 20% de datos faltantes:
- El nº de participantes disponibles para el análisis del nº de diseños únicos fue de 1245 (nº de participantes con las 3 mediciones completas).
- El nº de participantes disponibles para el análisis del nº de errores perseverantes fue de 1297 (nº de participantes con las 3 mediciones completas).

6.3 Eliminación de los casos (listwise) (30%)

En este primer caso se generan dos dataframes en los que se excluyen todos los casos para los que hay algún dato faltante (por separado para los casos de *Nº de diseños únicos* y *Errores perseverantes*), para el dataframe con 30% de datos faltantes.

En el escenario del 30% de datos faltantes:
- El nº de participantes disponibles para el análisis del nº de diseños únicos fue de 832 (nº de participantes con las 3 mediciones completas).
- El nº de participantes disponibles para el análisis del nº de errores perseverantes fue de 833 (nº de participantes con las 3 mediciones completas).

6.4 Imputación por media (10%)

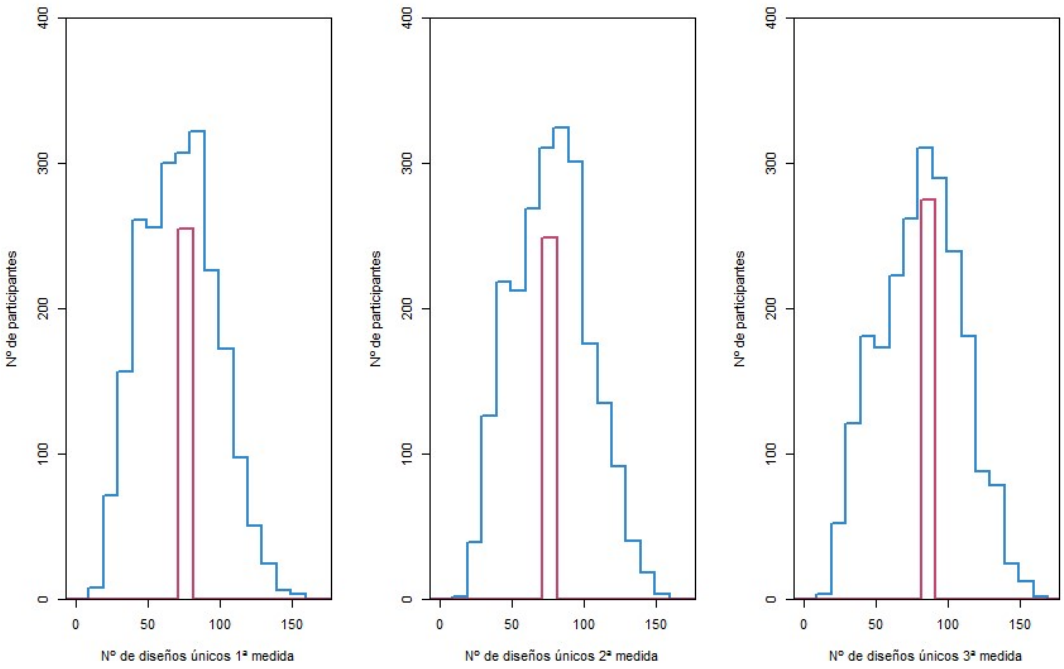
Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el método de *Imputación por media* , que reemplaza los datos faltantes por el valor medio del resto de datos conocidos en de la variable, en el caso con 10% de datos faltantes.

La función *"mice"* permite calcular la media en cada variable y reemplzar los datos faltantes de manera sencilla. Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 37. Imputación datos faltantes por media (MICE): Nº de diseños únicos (10%)

```
##
##  iter imp variable
##   1   1  Unique.mcar1  Unique.mcar2  Unique.mcar3

## Único 1ª medida 2003-06 (con NA)
## [1] 73.34336
```

```
## Único 2ª medida 2006-08 (con NA)
## [1] 78.81443

## Único 3ª medida 2008-12 (con NA)
## [1] 82.62411
```

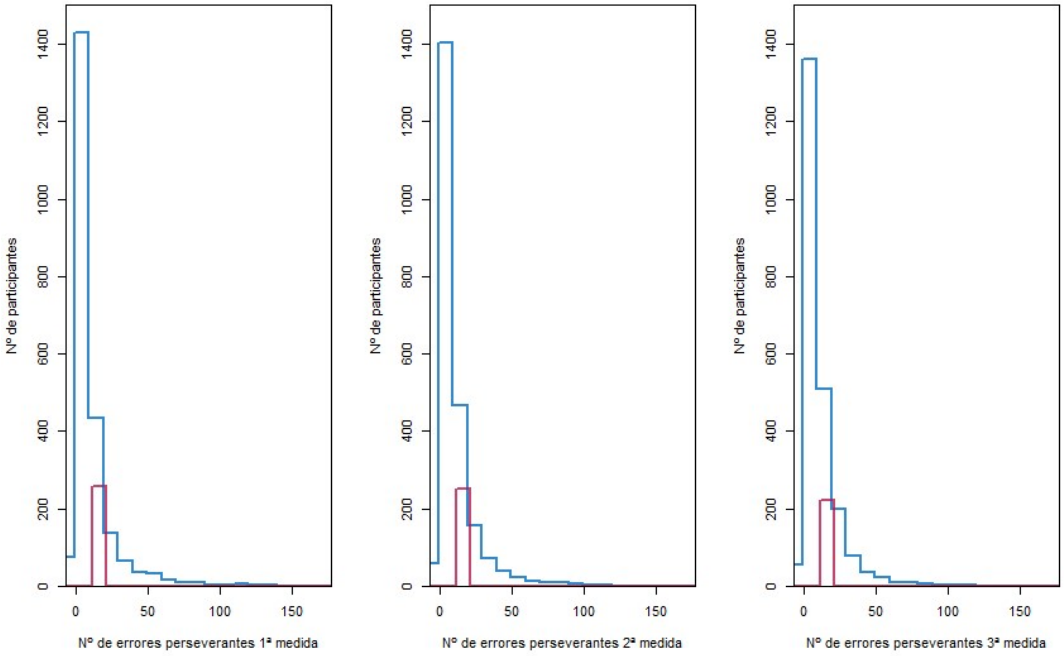


La media del Nº de diseños únicos en la 1ª medida fue 73.34, en la 2ª medida fue 78.81 y en la 3ª medida 82.62.

De modo similar se imputan los datos faltantes para el ***nº de errores perseverantes*** en cada una de las 3 mediciones.

Figura 38. Imputación datos faltantes por media (MICE): Nº de errores perseverantes (10%)

```
##
##  iter imp variable
##   1   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3

## Perseverante 1ª medida 2003-06 (con NA)
## [1] 11.75809

## Perseverante 2ª medida 2006-08 (con NA)
## [1] 11.7373

## Perseverante 3ª medida 2008-12 (con NA)
## [1] 11.98561
```



La media del Nº de errores perseverantes en la 1ª medida fue 11.76, en la 2ª medida fue 11.74 y en la 3ª medida 11.99.

6.5 Imputación por media (20%)

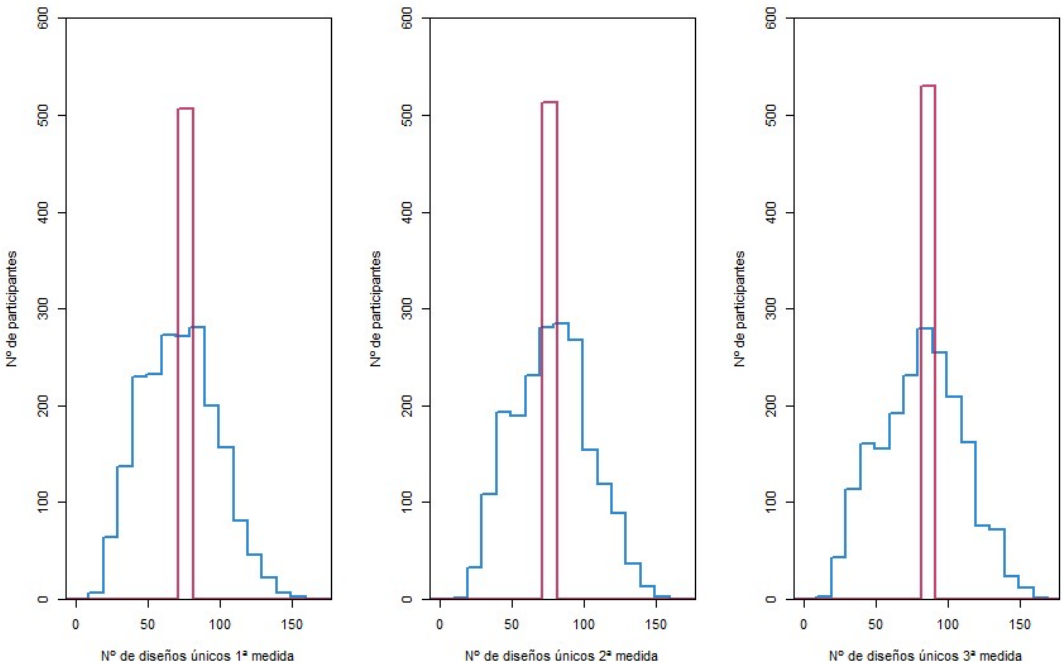
Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el método de *Imputación por media* , que reemplaza los datos faltantes por el valor medio del resto de datos conocidos en de la variable, en el caso con 20% de datos faltantes.

La función *"mice"* permite calcular la media en cada variable y reemplzar los datos faltantes de manera sencilla. Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 39. Imputación datos faltantes por media (MICE): Nº de diseños únicos (20%)

```
##
##  iter imp variable
##   1   1  Unique.mcar1  Unique.mcar2  Unique.mcar3

## Único 1ª medida 2003-06 (con NA)
## [1] 73.29517
```

```
## Único 2ª medida 2006-08 (con NA)
## [1] 79.06943

## Único 3ª medida 2008-12 (con NA)
## [1] 82.66801
```

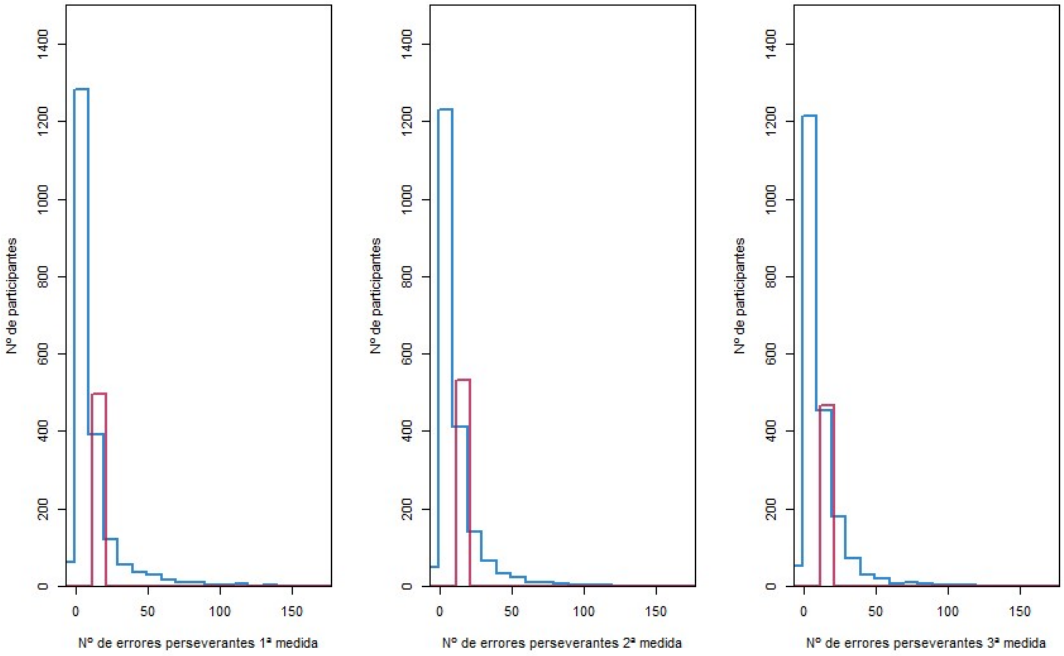


La media del Nº de diseños únicos en la 1ª medida fue 73.3, en la 2ª medida fue 79.07 y en la 3ª medida 82.67.

De modo similar se imputan los datos faltantes para el ***nº de errores perseverantes*** en cada una de las 3 mediciones.

Figura 40. Imputación datos faltantes por media (MICE): Nº de errores perseverantes (20%)

```
##
##  iter imp variable
##   1   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3

## Perseverante 1ª medida 2003-06 (con NA)
## [1] 11.9297

## Perseverante 2ª medida 2006-08 (con NA)
## [1] 11.66246

## Perseverante 3ª medida 2008-12 (con NA)
## [1] 11.94968
```



La media del Nº de errores perseverantes en la 1ª medida fue 11.93, en la 2ª medida fue 11.66 y en la 3ª medida 11.95.

6.6 Imputación por media (30%)

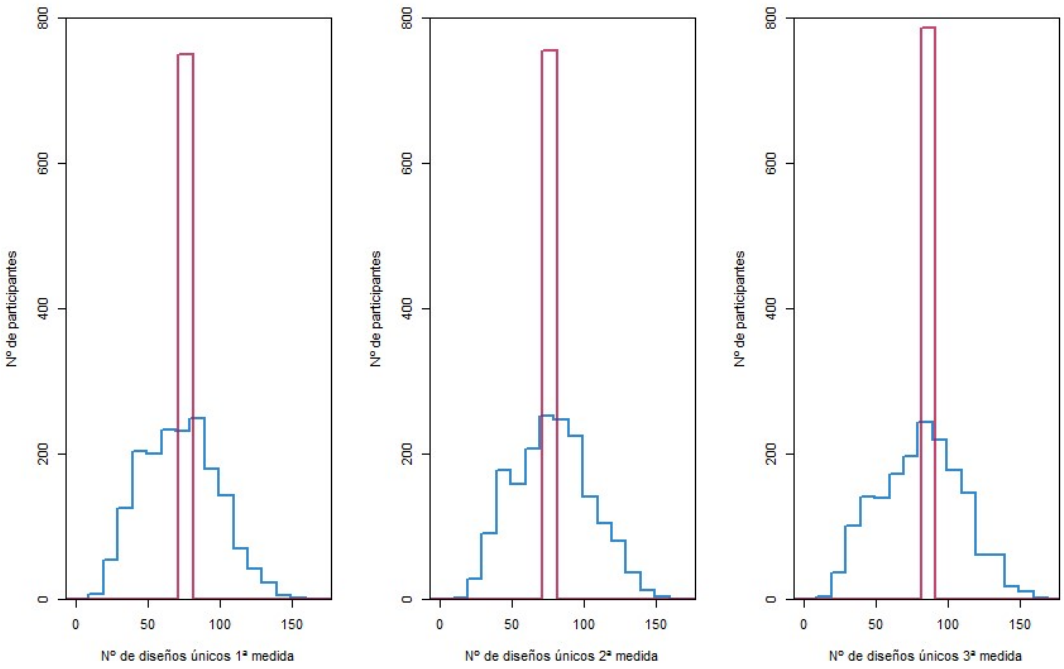
Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el método de *Imputación por media* , que reemplaza los datos faltantes por el valor medio del resto de datos conocidos en de la variable, en el caso con 30% de datos faltantes.

La función *"mice"* permite calcular la media en cada variable y reemplzar los datos faltantes de manera sencilla. Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 41. Imputación datos faltantes por media (MICE): Nº de diseños únicos (30%)

```
##
##  iter imp variable
##   1   1  Unique.mcar1  Unique.mcar2  Unique.mcar3

## Único 1ª medida 2003-06 (con NA)
## [1] 73.45692
```

```
## Único 2ª medida 2006-08 (con NA)
## [1] 79.29233

## Único 3ª medida 2008-12 (con NA)
## [1] 82.18576
```

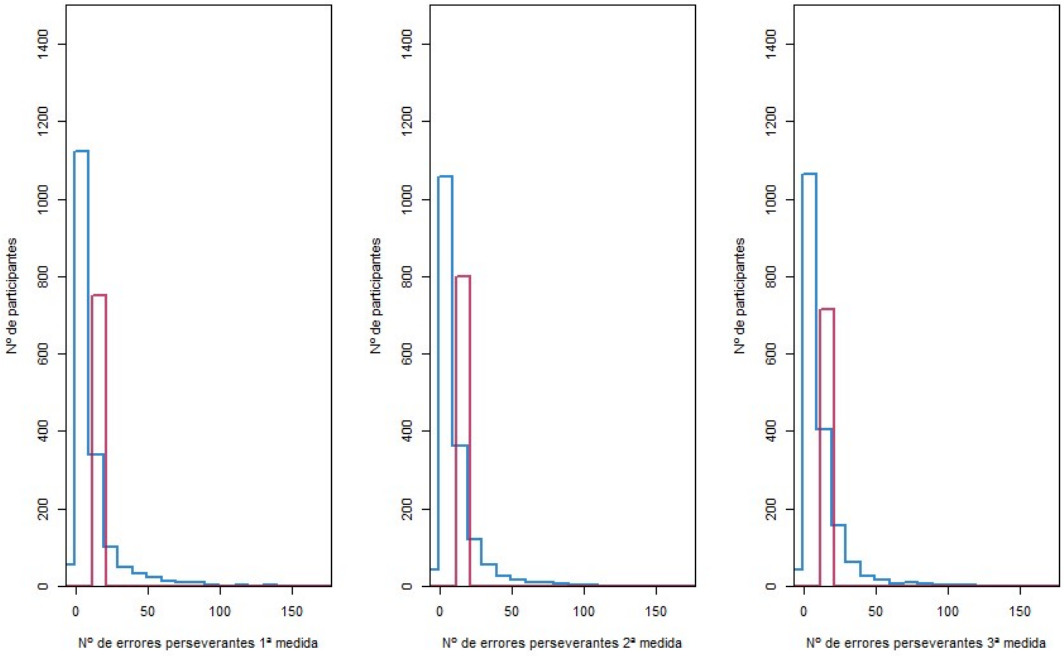


La media del Nº de diseños únicos en la 1ª medida fue 73.46, en la 2ª medida fue 79.29 y en la 3ª medida 82.19.

De modo similar se imputan los datos faltantes para el ***nº de errores perseverantes*** en cada una de las 3 mediciones.

Figura 42. Imputación datos faltantes por media (MICE): Nº de errores perseverantes (30%)

```
##
##  iter imp variable
##   1   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3

## Perseverante 1ª medida 2003-06 (con NA)
## [1] 11.76347

## Perseverante 2ª medida 2006-08 (con NA)
## [1] 11.72523

## Perseverante 3ª medida 2008-12 (con NA)
## [1] 12.12472
```



La media del Nº de errores perseverantes en la 1ª medida fue 11.76, en la 2ª medida fue 11.73 y en la 3ª medida 12.12.

6.7 Imputación por regresión simple (10%)

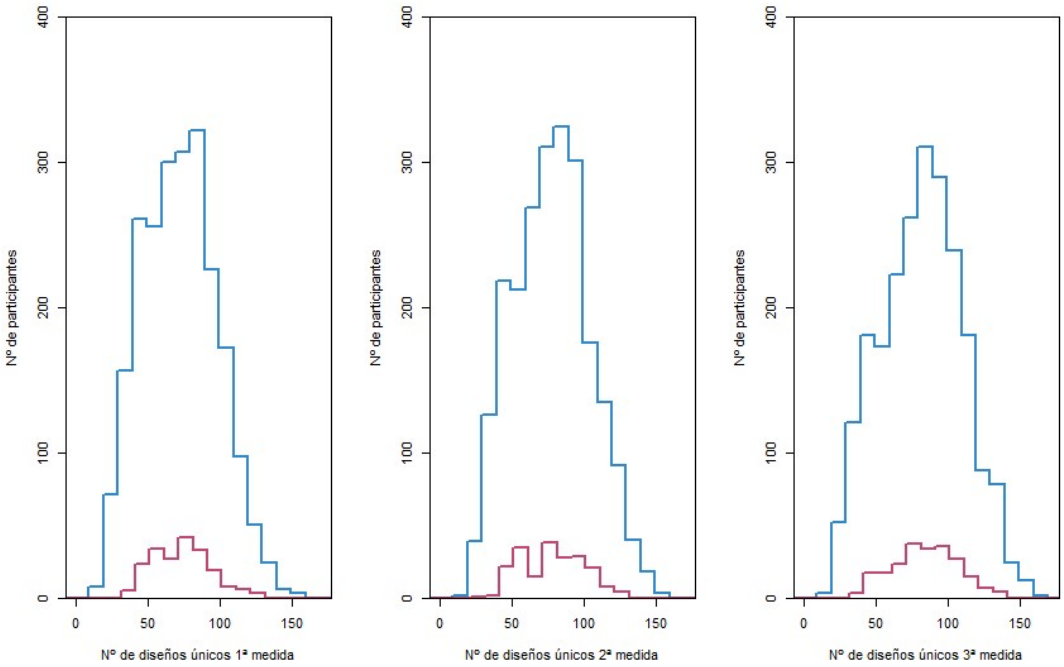
Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el método de *Imputación por regresión simple (media condicional)*, que reemplaza los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto, en el caso con 10% de datos faltantes.

Se utiliza la opción *"norm.predict"* en la función *"mice"* para imputar datos faltantes por regresión simple. Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 43. Imputación datos faltantes por regresión (MICE): Nº de diseños únicos (10%)

```
##
##  iter imp variable
##   1   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   2   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   3   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
```

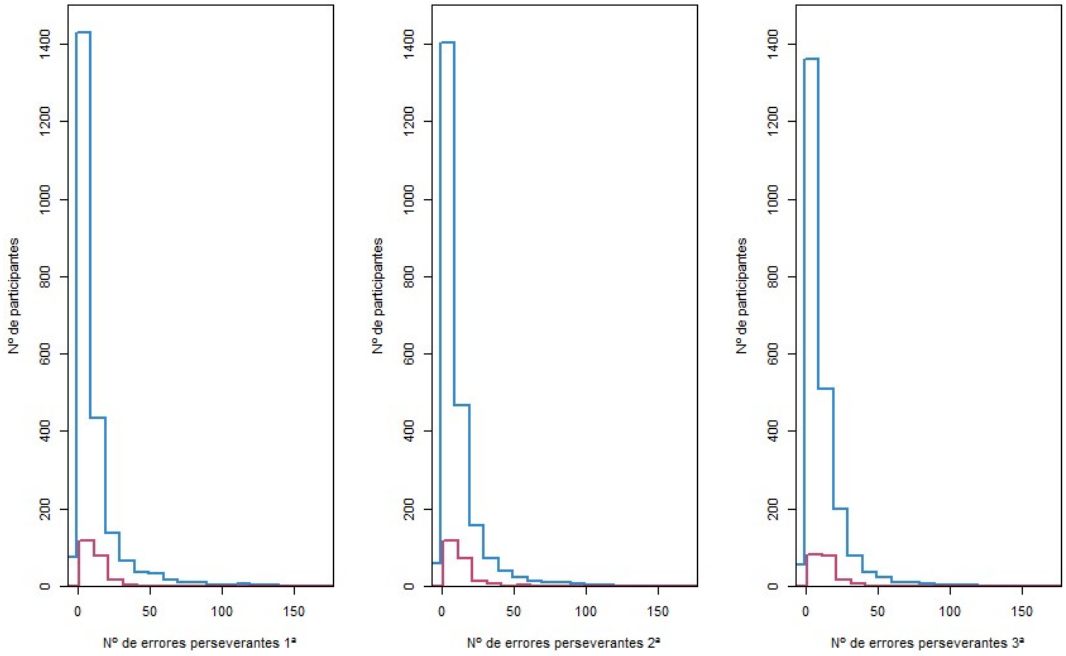
```
##   4   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   5   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
```



De modo similar se imputan los datos faltantes para el nº de errores perseverantes en cada una de las 3 mediciones.

Figura 44. Imputación datos faltantes por regresión (MICE): Nº de errores perseverantes (10%)

```
##
##  iter imp variable
##   1   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   2   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   3   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   4   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   5   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
```



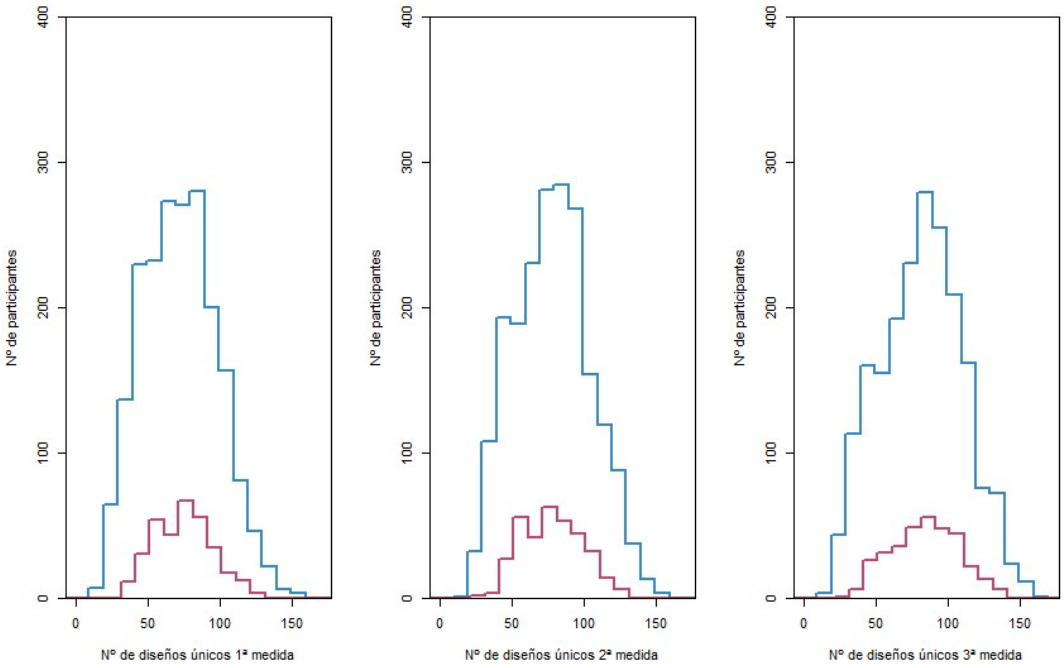
6.8 Imputación por regresión simple (20%)

Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el método de *Imputación por regresión simple (media condicional)*, que reemplaza los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto, en el caso con 20% de datos faltantes.

Se utiliza la opción *"norm.predict"* en la función *"mice"* para imputar datos faltantes por regresión simple. Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 45. Imputación datos faltantes por regresión (MICE): Nº de diseños únicos (20%)

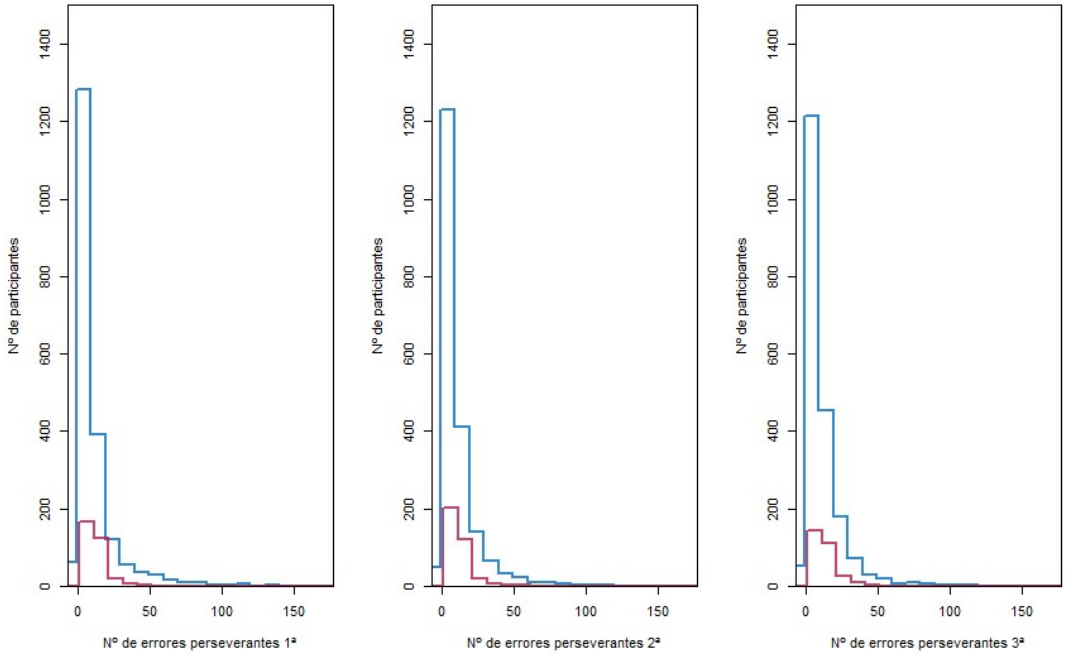
```
##
##  iter imp variable
##   1   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   2   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   3   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   4   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   5   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
```



De modo similar se imputan los datos faltantes para el nº de errores perseverantes en cada una de las 3 mediciones.

Figura 46. Imputación datos faltantes por regresión (MICE): Nº de errores perseverantes (20%)

```
##
##  iter imp variable
##   1   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   2   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   3   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   4   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   5   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
```



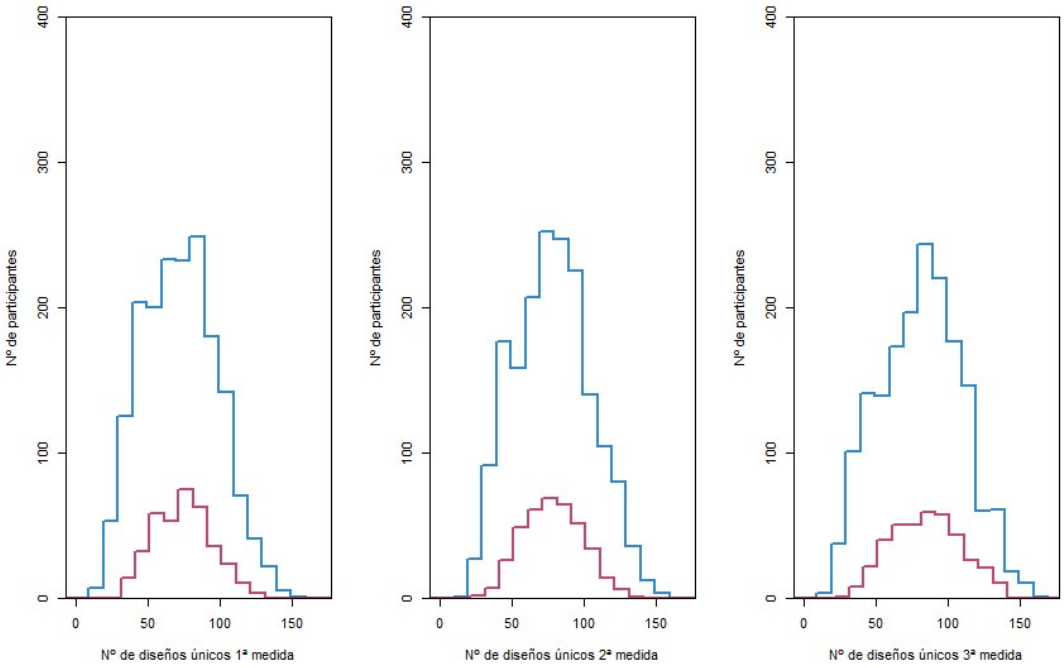
6.9 Imputación por regresión simple (30%)

Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el método de *Imputación por regresión simple (media condicional)*, que reemplaza los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto, en el caso con 30% de datos faltantes.

Se utiliza la opción *"norm.predict"* en la función *"mice"* para imputar datos faltantes por regresión simple. Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 47. Imputación datos faltantes por regresión (MICE): Nº de diseños únicos (30%)

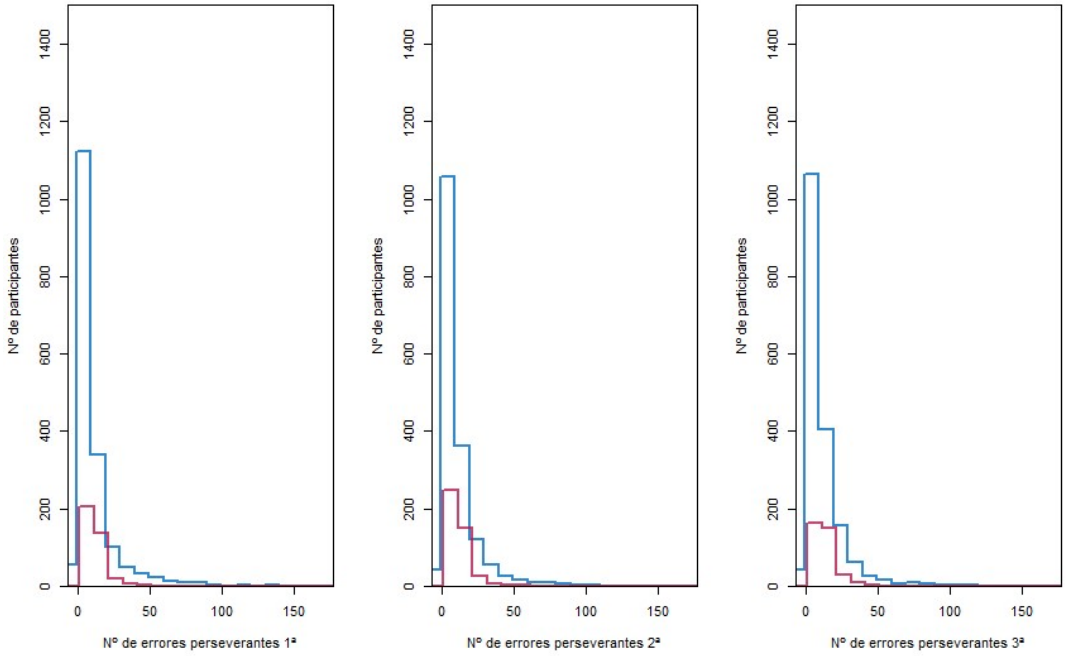
```
##
##  iter imp variable
##   1   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   2   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   3   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   4   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
##   5   1  Unique.mcar1  Unique.mcar2  Unique.mcar3
```



De modo similar se imputan los datos faltantes para el nº de errores perseverantes en cada una de las 3 mediciones.

Figura 48. Imputación datos faltantes por regresión (MICE): Nº de errores perseverantes (30%)

```
##
##  iter imp variable
##   1   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   2   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   3   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   4   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
##   5   1  Perseverative.mcar1  Perseverative.mcar2  Perseverative.mcar3
```



6.10 Imputación múltiple (método PMM) (10%)

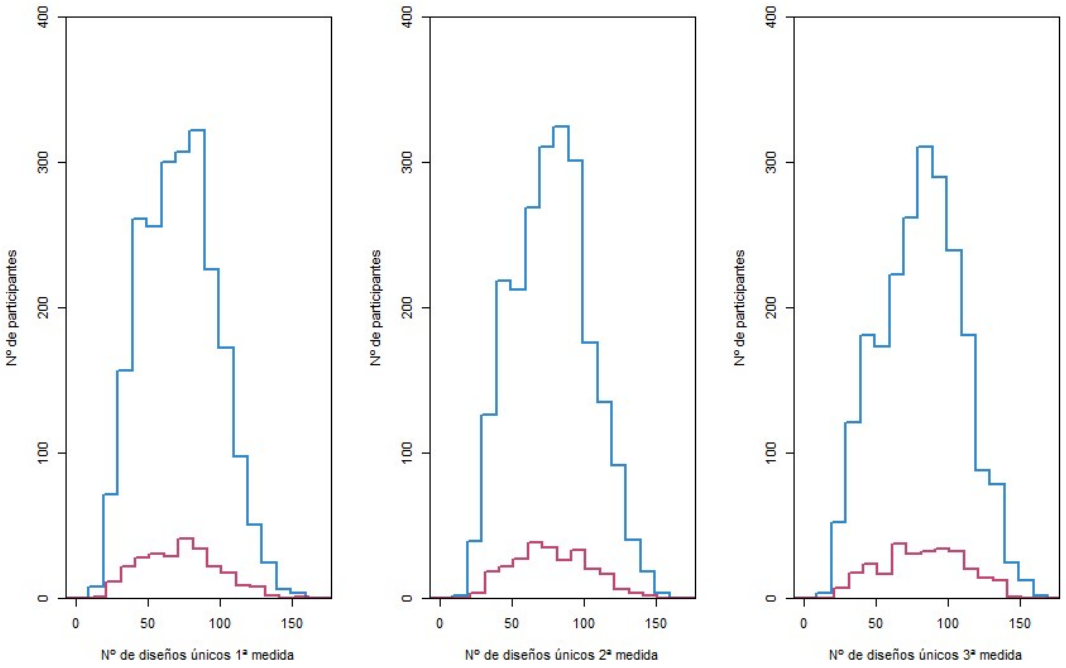
Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*, en el caso con 10% de datos faltantes.. Este método obtiene buenos resultados tanto para variables continuas como categóricas (binarias o más categorías) sin necesidad de calcular los errores (residuals) ni ajustar por máxima verosimilitud.

PMM es por defecto el algoritmo que utiliza la función *"mice"* para imputar datos faltantes, por lo que en la opción *meth* no sería necesario especificar que el método a utilizar es *pmm*. Se recomienda que la imputación múltiple (IM) se efectue con al menos 5 imputaciones (también es el nº que *mice* emplea por defecto) y el nº de iteraciones también por defecto es de 5. Una vez imputados, mediante la función *pool* se obtiene el valor medio para todas las imputaciones Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

Figura 49. Imputación datos faltantes por IM (MICE-PMM): Nº de diseños únicos (10%)

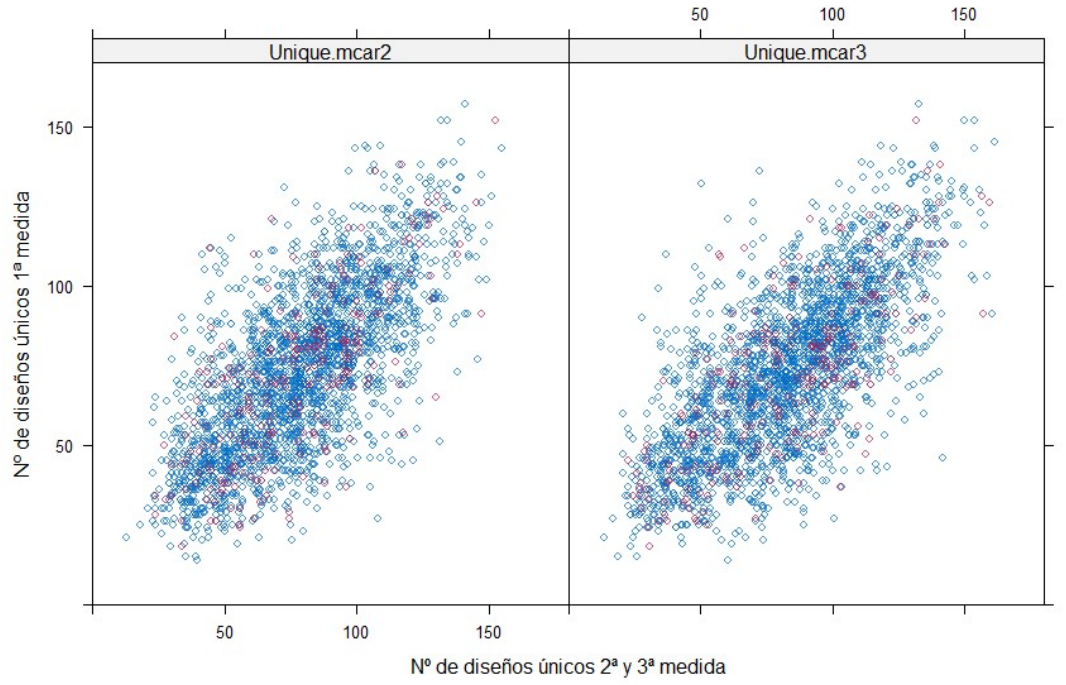
```
## Multiply imputed data set
## Call:
```

```
## mice(data = RFFT_wide_unique[, 2:7], printFlag = FALSE, seed = 123456)
## Number of multiple imputations:  5
## Missing cells per column:
##          Age       Gender    Education Unique.mcar1 Unique.mcar2
##            0            0            0          255          249
## Unique.mcar3
##          275
## Imputation methods:
##          Age       Gender    Education Unique.mcar1 Unique.mcar2
##           ""           ""           ""        "pmm"        "pmm"
## Unique.mcar3
##        "pmm"
## VisitSequence:
## Unique.mcar1 Unique.mcar2 Unique.mcar3
##            4            5            6
## PredictorMatrix:
##              Age Gender Education Unique.mcar1 Unique.mcar2 Unique.mca
r3
## Age            0      0         0            0            0
0
## Gender         0      0         0            0            0
0
## Education      0      0         0            0            0
0
## Unique.mcar1   1      1         1            0            1
1
## Unique.mcar2   1      1         1            1            0
1
## Unique.mcar3   1      1         1            1            1
0
## Random generator seed value:  123456
```



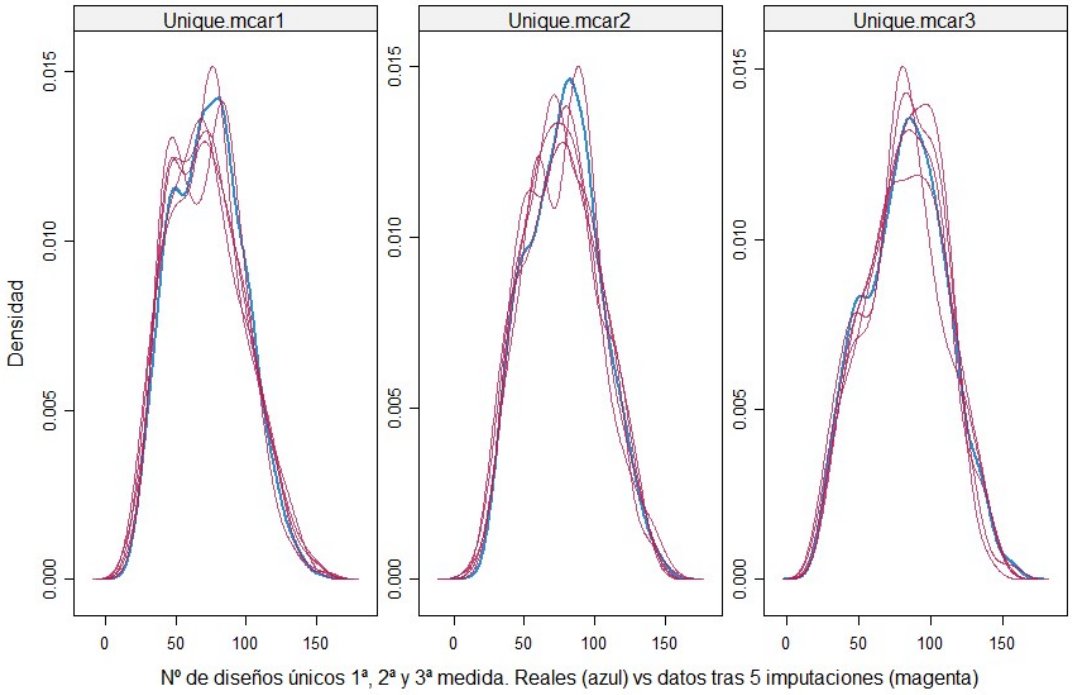
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las variables de medición del nº de diseños únicos (dos a dos).

Figura 50. Valores reales vs imputados IM (MICE-PMM): Nº de diseños únicos (10%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de las imputaciones de los datos faltantes.

Figura 51. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de diseños únicos (10%)

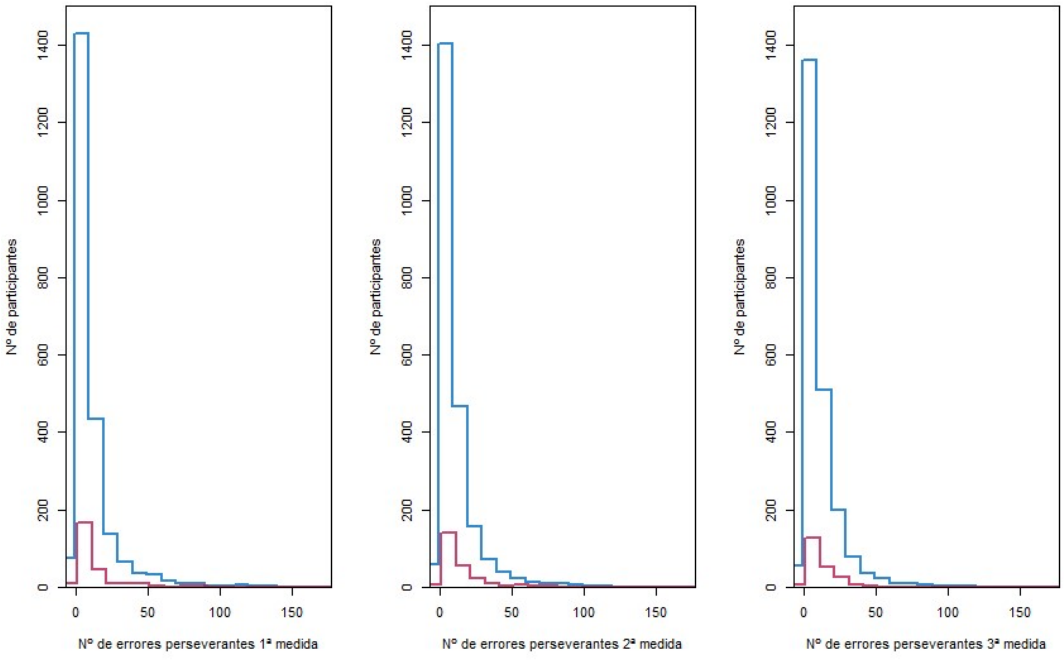


De modo similar se imputan los datos faltantes para el nº de errores perseverantes en cada una de las 3 mediciones, utilizando el método de imputación múltiple (IM).

Figura 52. Imputación datos faltantes por IM (MICE-PMM): Nº de errores perseverantes (10%)

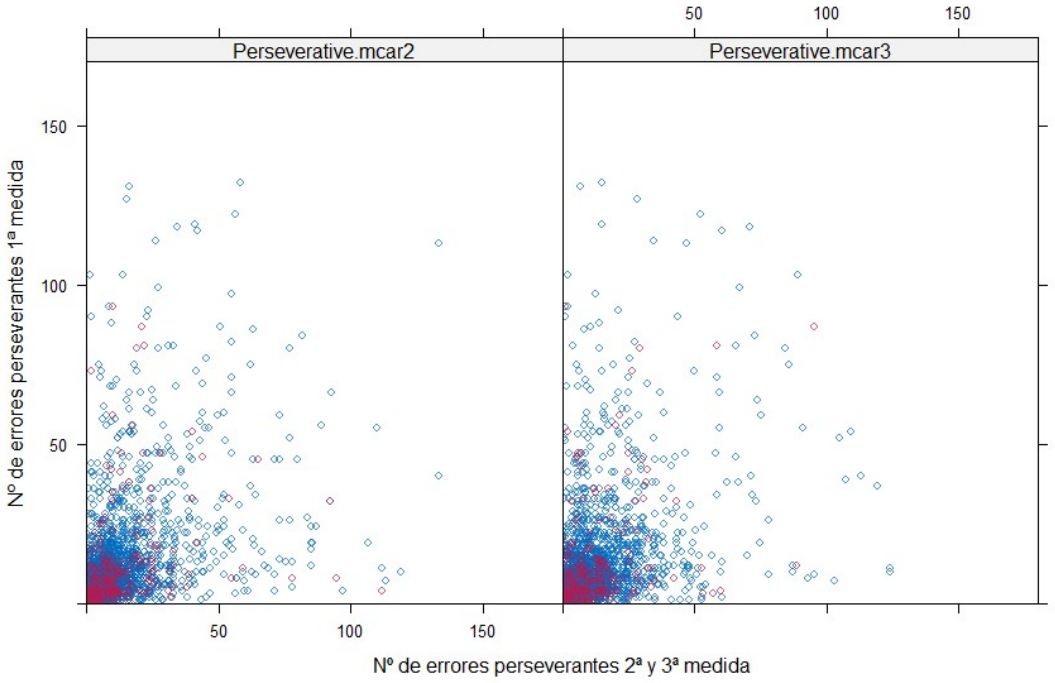
```
## Multiply imputed data set
## Call:
## mice(data = RFFT_wide_perseverative[, 2:7], printFlag = FALSE,
##     seed = 123456)
## Number of multiple imputations:  5
## Missing cells per column:
##                  Age               Gender            Education
##                    0                    0                    0
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##                  258                  252                  222
## Imputation methods:
##                  Age               Gender            Education
```

```
##                   ""                   ""                   ""
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##               "pmm"                "pmm"                "pmm"
## VisitSequence:
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##                   4                    5                    6
## PredictorMatrix:
##                     Age Gender Education Perseverative.mcar1
## Age                   0      0         0                    0
## Gender                0      0         0                    0
## Education             0      0         0                    0
## Perseverative.mcar1   1      1         1                    0
## Perseverative.mcar2   1      1         1                    1
## Perseverative.mcar3   1      1         1                    1
##                     Perseverative.mcar2 Perseverative.mcar3
## Age                                   0                    0
## Gender                                0                    0
## Education                             0                    0
## Perseverative.mcar1                   1                    1
## Perseverative.mcar2                   0                    1
## Perseverative.mcar3                   1                    0
## Random generator seed value:  123456
```



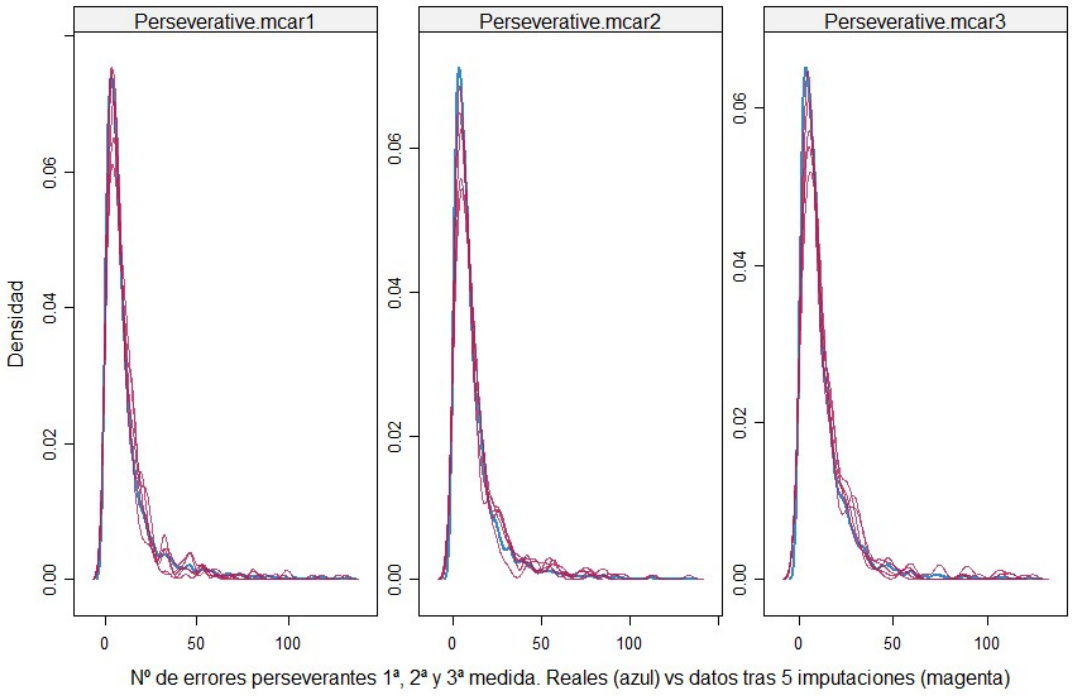
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las variables de medición del nº de errores perseverantes (dos a dos).

Figura 53. Valores reales vs imputados IM (MICE-PMM): Nº de errores perseverantes (10%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de las imputaciones de los datos faltantes.

Figura 54. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de errores perseverantes (10%)



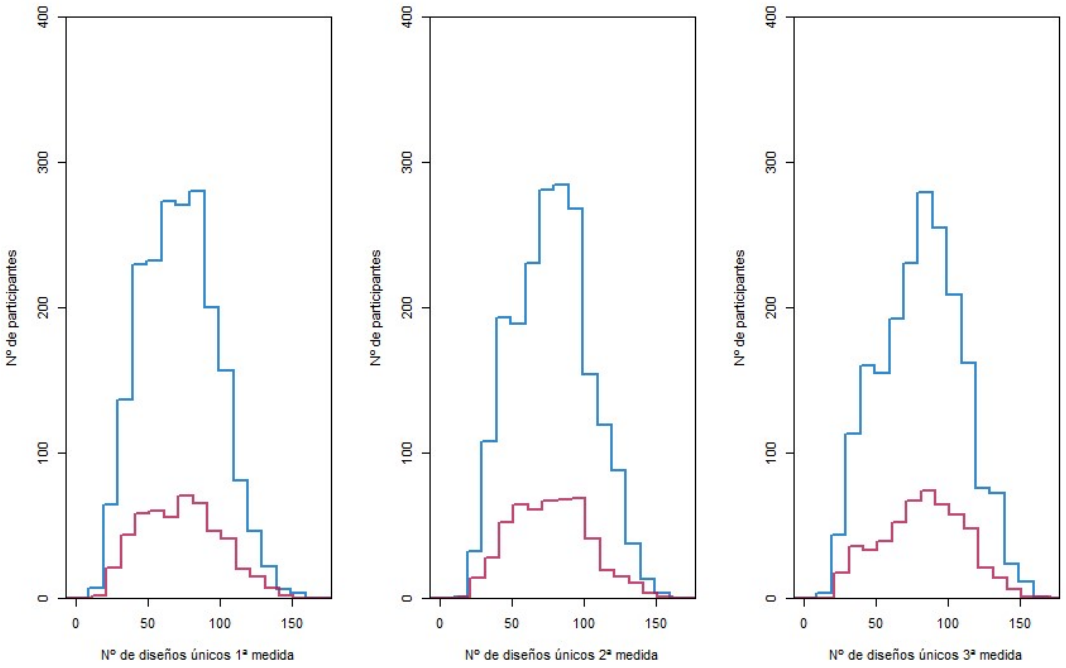
6.11 Imputación múltiple (método PMM) (20%)

Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*, en el caso con 20% de datos faltantes.. Este método obtiene buenos resultados tanto para variables continuas como categóricas (binarias o más categorías) sin necesidad de calcular los errores (residuals) ni ajustar por máxima verosimilitud.

PMM es por defecto el algoritmo que utiliza la función *"mice"* para imputar datos faltantes, por lo que en la opción *meth* no sería necesario especificar que el método a utilizar es *pmm*. Se recomienda que la imputación múltiple (IM) se efectue con al menos 5 imputaciones (también es el nº que *mice* emplea por defecto) y el nº de iteraciones también por defecto es de 5. Una vez imputados, mediante la función *pool* se obtiene el valor medio para todas las imputaciones Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

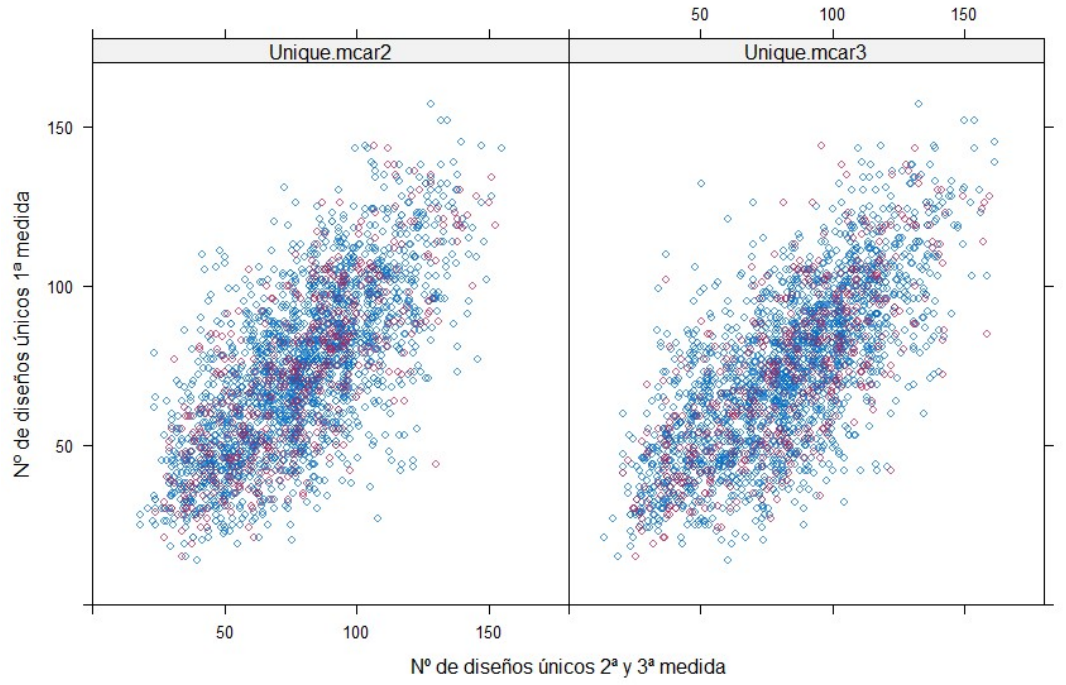
Figura 55. Imputación datos faltantes por IM (MICE-PMM): Nº de diseños únicos (20%)

```
## Multiply imputed data set
## Call:
## mice(data = RFFT_wide_unique_20[, 2:7], printFlag = FALSE, seed = 1234
56)
## Number of multiple imputations:  5
## Missing cells per column:
##          Age       Gender    Education Unique.mcar1 Unique.mcar2
##            0            0            0          506          513
## Unique.mcar3
##          530
## Imputation methods:
##          Age       Gender    Education Unique.mcar1 Unique.mcar2
##           ""           ""           ""        "pmm"        "pmm"
## Unique.mcar3
##        "pmm"
## VisitSequence:
## Unique.mcar1 Unique.mcar2 Unique.mcar3
##            4            5            6
## PredictorMatrix:
##              Age Gender Education Unique.mcar1 Unique.mcar2 Unique.mca
r3
## Age            0      0         0            0            0
0
## Gender         0      0         0            0            0
0
## Education      0      0         0            0            0
0
## Unique.mcar1   1      1         1            0            1
1
## Unique.mcar2   1      1         1            1            0
1
## Unique.mcar3   1      1         1            1            1
0
## Random generator seed value:  123456
```



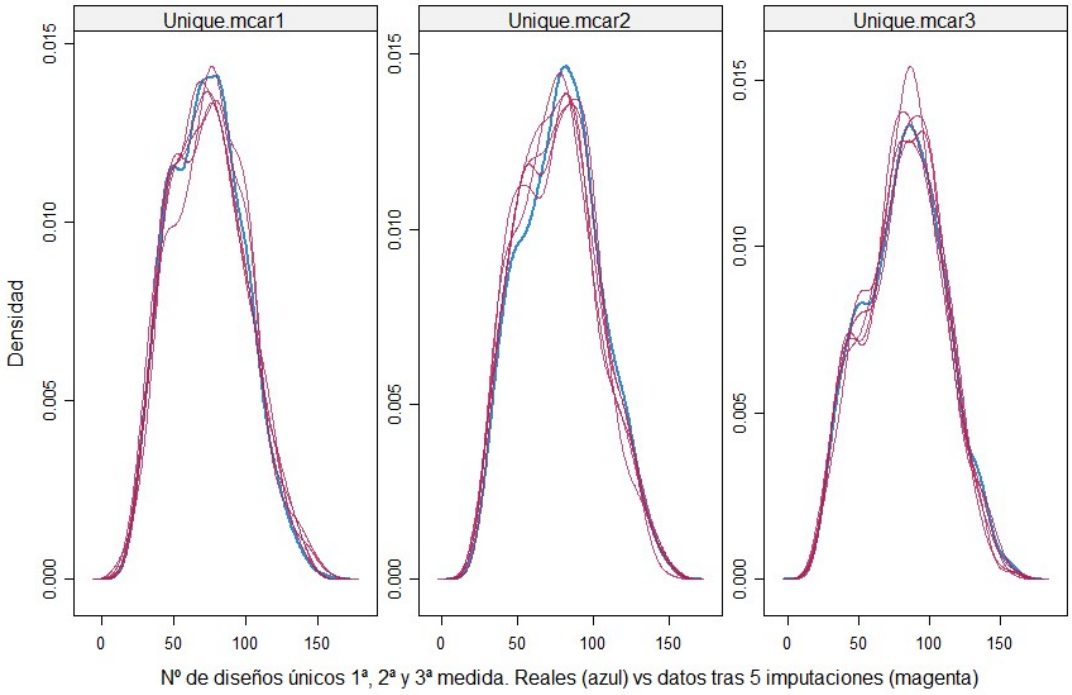
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las variables de medición del nº de diseños únicos (dos a dos).

Figura 56. Valores reales vs imputados IM (MICE-PMM): Nº de diseños únicos (20%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de las imputaciones de los datos faltantes.

Figura 57. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de diseños únicos (20%)

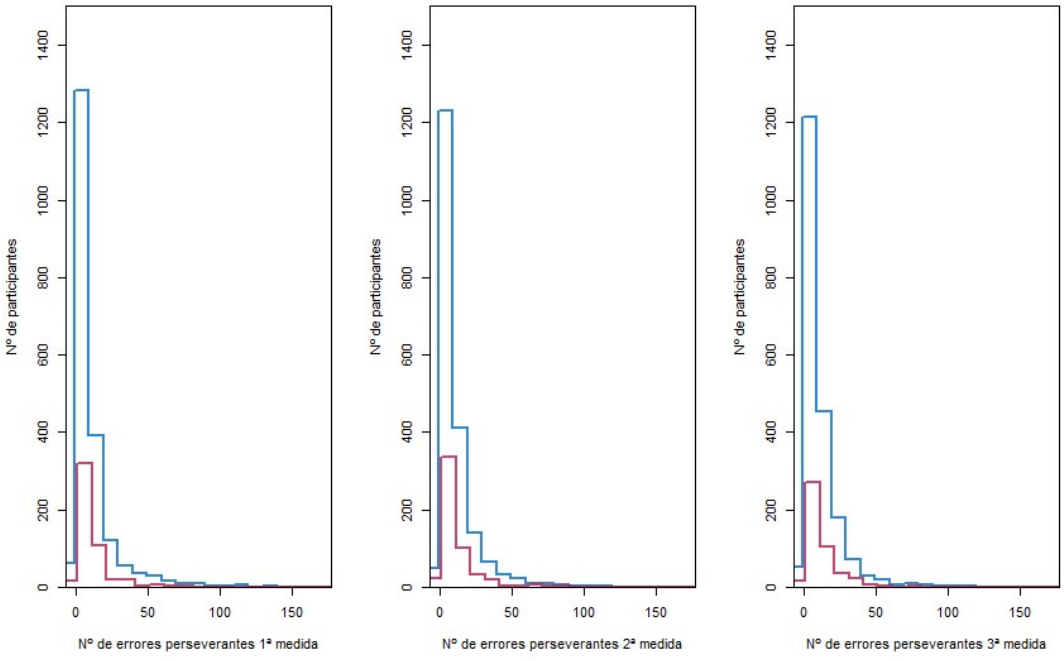


De modo similar se imputan los datos faltantes para el nº de errores perseverantes en cada una de las 3 mediciones, utilizando el método de imputación múltiple (IM).

Figura 58. Imputación datos faltantes por IM (MICE-PMM): Nº de errores perseverantes (20%)

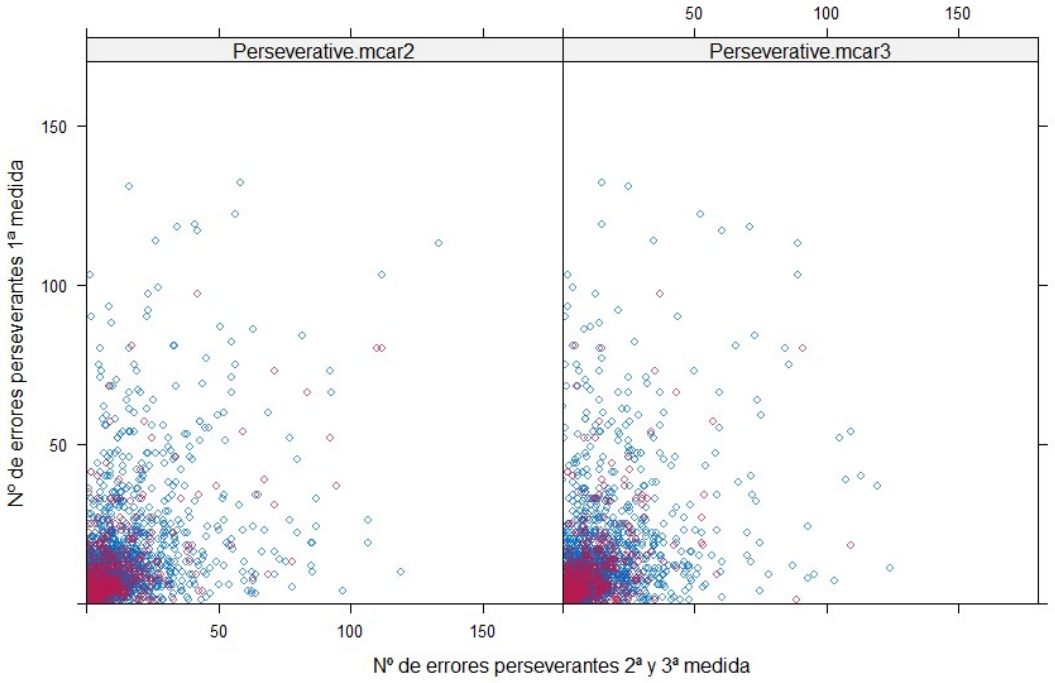
```
## Multiply imputed data set
## Call:
## mice(data = RFFT_wide_perseverative_20[, 2:7], printFlag = FALSE,
##     seed = 123456)
## Number of multiple imputations:  5
## Missing cells per column:
##                  Age               Gender            Education
##                    0                    0                    0
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##                  495                  533                  468
## Imputation methods:
##                  Age               Gender            Education
```

```
##                   ""                   ""                   ""
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##               "pmm"                "pmm"                "pmm"
## VisitSequence:
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##                   4                    5                    6
## PredictorMatrix:
##                     Age Gender Education Perseverative.mcar1
## Age                   0      0         0                    0
## Gender                0      0         0                    0
## Education             0      0         0                    0
## Perseverative.mcar1   1      1         1                    0
## Perseverative.mcar2   1      1         1                    1
## Perseverative.mcar3   1      1         1                    1
##                     Perseverative.mcar2 Perseverative.mcar3
## Age                                   0                    0
## Gender                                0                    0
## Education                             0                    0
## Perseverative.mcar1                   1                    1
## Perseverative.mcar2                   0                    1
## Perseverative.mcar3                   1                    0
## Random generator seed value:  123456
```



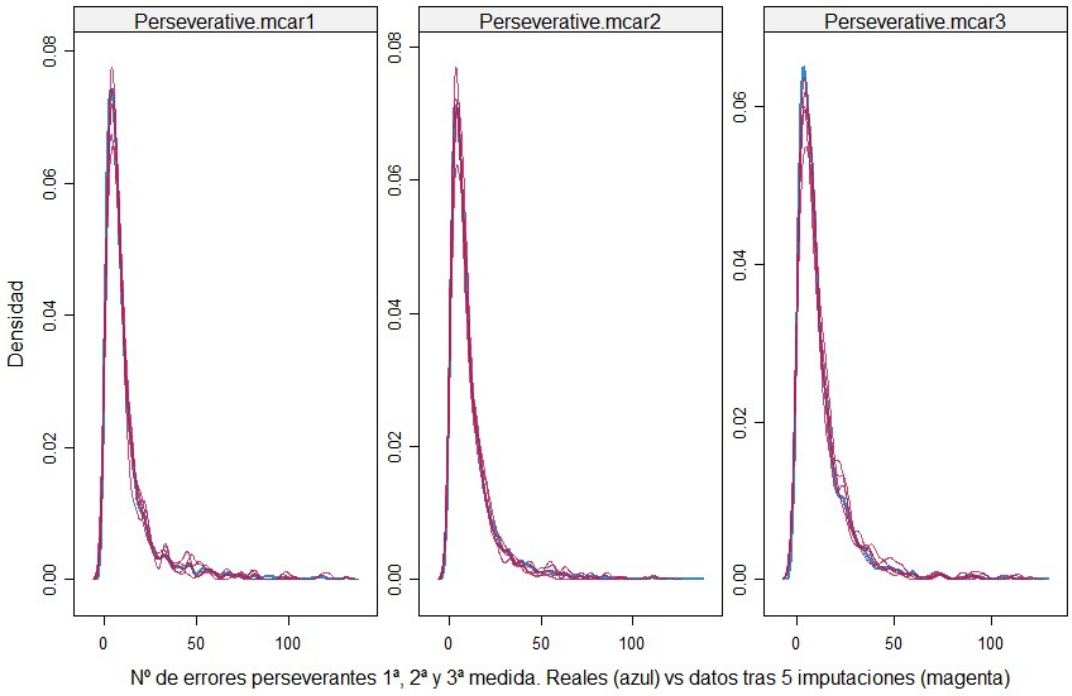
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las variables de medición del nº de errores perseverantes (dos a dos).

Figura 59. Valores reales vs imputados IM (MICE-PMM): Nº de errores perseverantes (20%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de las imputaciones de los datos faltantes.

Figura 60. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de errores perseverantes (20%)



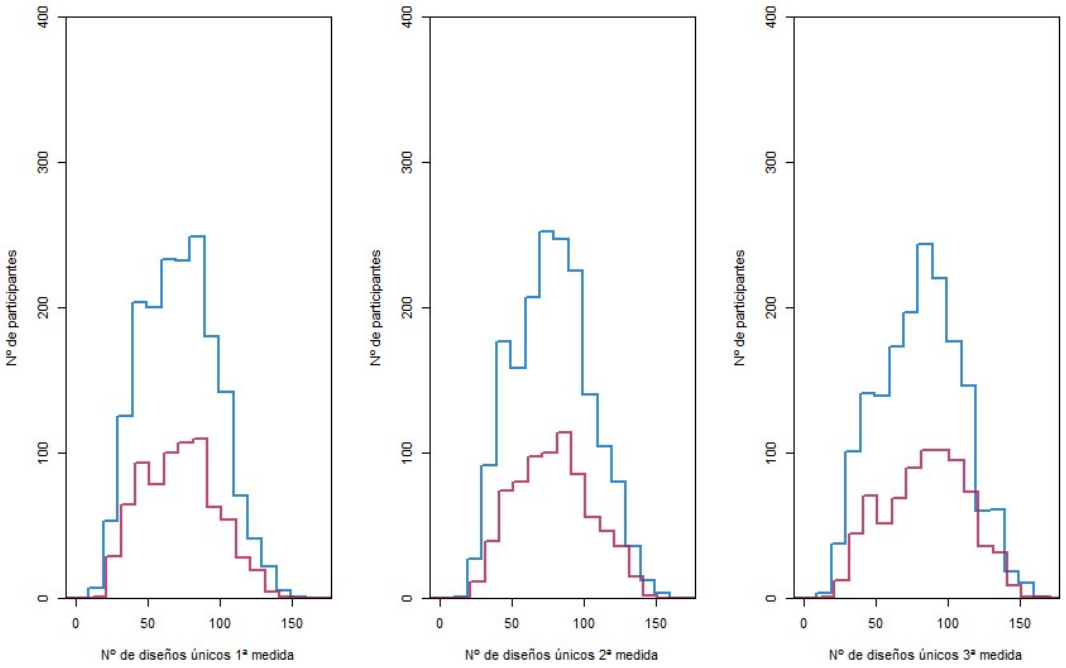
6.12 Imputación múltiple (método PMM) (30%)

Se generan dos dataframes (por separado para los *Nº de diseños únicos* y *Errores perseverantes*) utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*, en el caso con 30% de datos faltantes.. Este método obtiene buenos resultados tanto para variables continuas como categóricas (binarias o más categorías) sin necesidad de calcular los errores (residuals) ni ajustar por máxima verosimilitud.

PMM es por defecto el algoritmo que utiliza la función *"mice"* para imputar datos faltantes, por lo que en la opción *meth* no sería necesario especificar que el método a utilizar es *pmm*. Se recomienda que la imputación múltiple (IM) se efectue con al menos 5 imputaciones (también es el nº que *mice* emplea por defecto) y el nº de iteraciones también por defecto es de 5. Una vez imputados, mediante la función *pool* se obtiene el valor medio para todas las imputaciones Y se representa gráficamente la distribución de los datos imputados en cada una de las 3 variables de la medición del nº de diseños únicos.

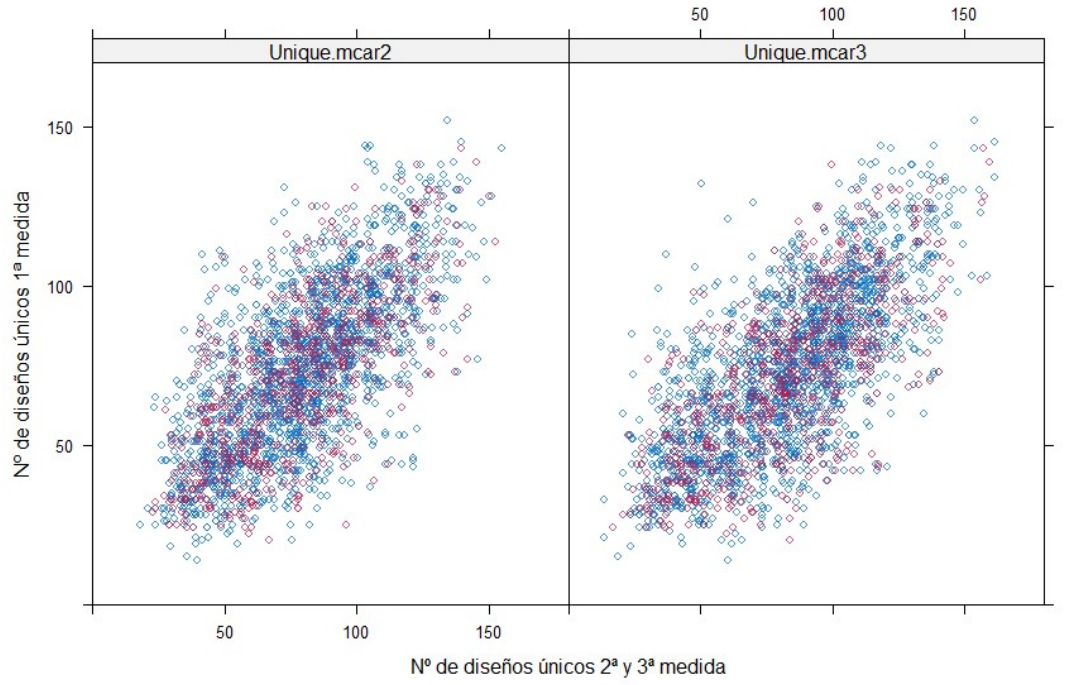
Figura 61. Imputación datos faltantes por IM (MICE-PMM): Nº de diseños únicos (30%)

```
## Multiply imputed data set
## Call:
## mice(data = RFFT_wide_unique_30[, 2:7], printFlag = FALSE, seed = 1234
56)
## Number of multiple imputations:  5
## Missing cells per column:
##          Age       Gender    Education Unique.mcar1 Unique.mcar2
##            0            0            0          751          755
## Unique.mcar3
##          787
## Imputation methods:
##          Age       Gender    Education Unique.mcar1 Unique.mcar2
##           ""           ""           ""        "pmm"        "pmm"
## Unique.mcar3
##        "pmm"
## VisitSequence:
## Unique.mcar1 Unique.mcar2 Unique.mcar3
##            4            5            6
## PredictorMatrix:
##              Age Gender Education Unique.mcar1 Unique.mcar2 Unique.mca
r3
## Age            0      0         0            0            0
0
## Gender         0      0         0            0            0
0
## Education      0      0         0            0            0
0
## Unique.mcar1   1      1         1            0            1
1
## Unique.mcar2   1      1         1            1            0
1
## Unique.mcar3   1      1         1            1            1
0
## Random generator seed value:  123456
```



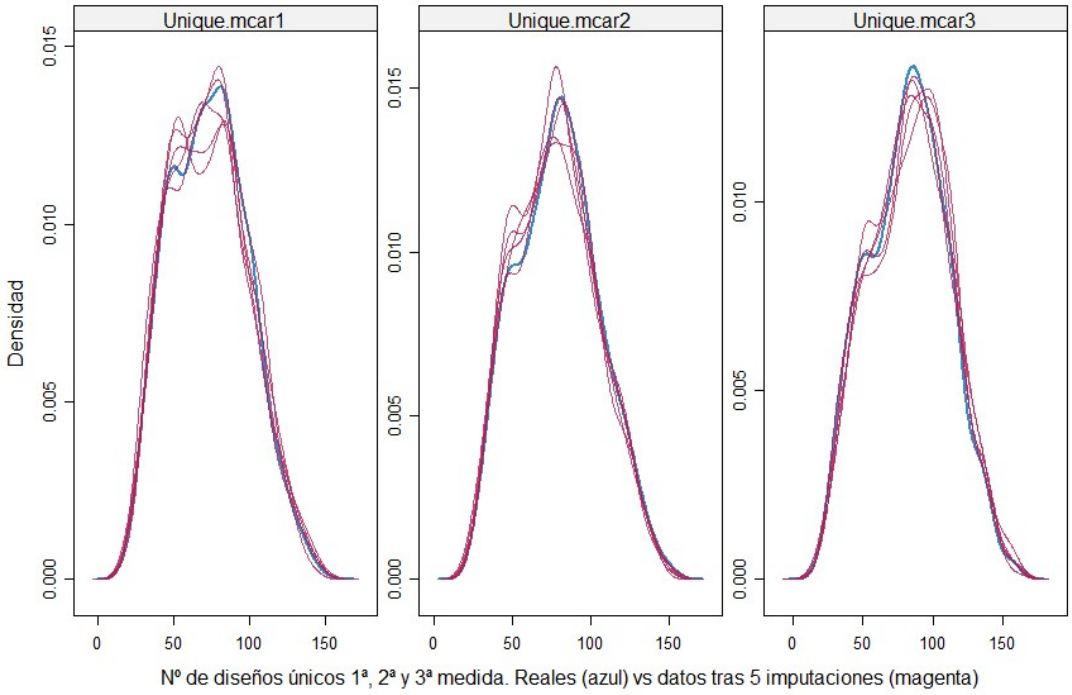
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las variables de medición del nº de diseños únicos (dos a dos).

Figura 62. Valores reales vs imputados IM (MICE-PMM): Nº de diseños únicos (30%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de las imputaciones de los datos faltantes.

Figura 63. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de diseños únicos (30%)

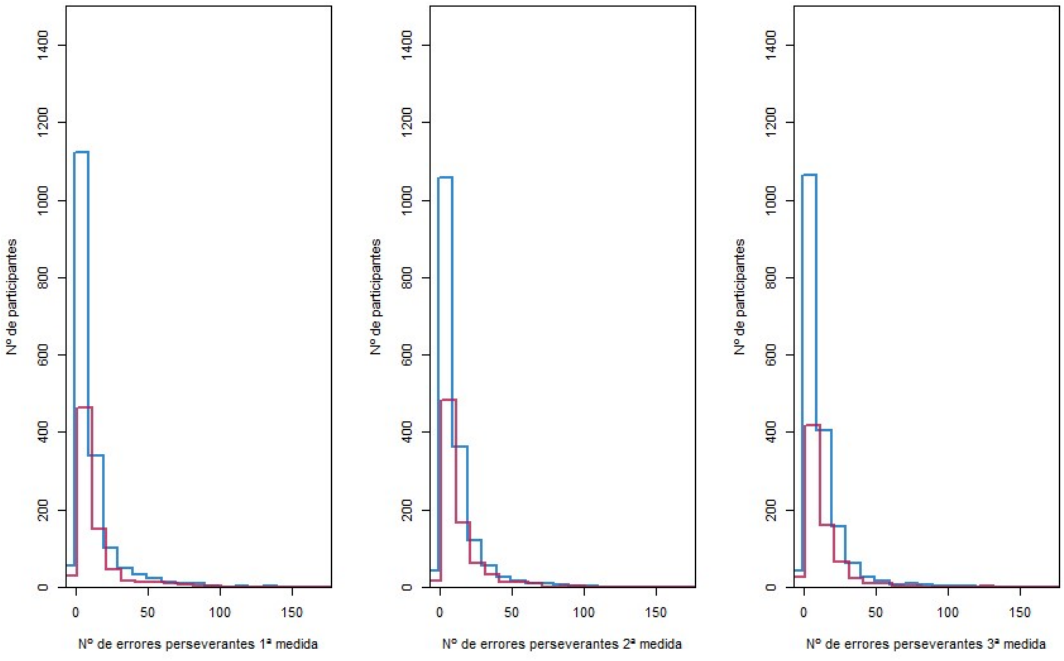


De modo similar se imputan los datos faltantes para el nº de errores perseverantes en cada una de las 3 mediciones, utilizando el método de imputación múltiple (IM).

Figura 64. Imputación datos faltantes por IM (MICE-PMM): Nº de errores perseverantes (30%)

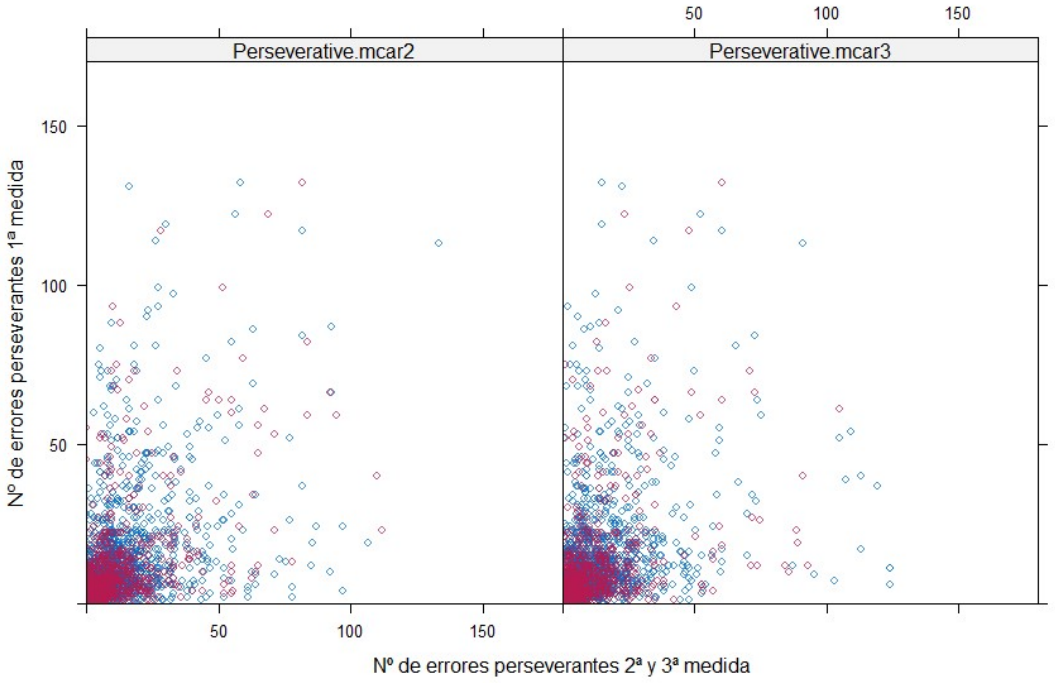
```
## Multiply imputed data set
## Call:
## mice(data = RFFT_wide_perseverative_30[, 2:7], printFlag = FALSE,
##     seed = 123456)
## Number of multiple imputations:  5
## Missing cells per column:
##                  Age               Gender            Education
##                    0                    0                    0
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##                  752                  799                  715
## Imputation methods:
##                  Age               Gender            Education
```

```
##                   ""                   ""                   ""
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##               "pmm"                "pmm"                "pmm"
## VisitSequence:
## Perseverative.mcar1 Perseverative.mcar2 Perseverative.mcar3
##                   4                    5                    6
## PredictorMatrix:
##                     Age Gender Education Perseverative.mcar1
## Age                   0      0         0                    0
## Gender                0      0         0                    0
## Education             0      0         0                    0
## Perseverative.mcar1   1      1         1                    0
## Perseverative.mcar2   1      1         1                    1
## Perseverative.mcar3   1      1         1                    1
##                     Perseverative.mcar2 Perseverative.mcar3
## Age                                   0                    0
## Gender                                0                    0
## Education                             0                    0
## Perseverative.mcar1                   1                    1
## Perseverative.mcar2                   0                    1
## Perseverative.mcar3                   1                    0
## Random generator seed value:  123456
```



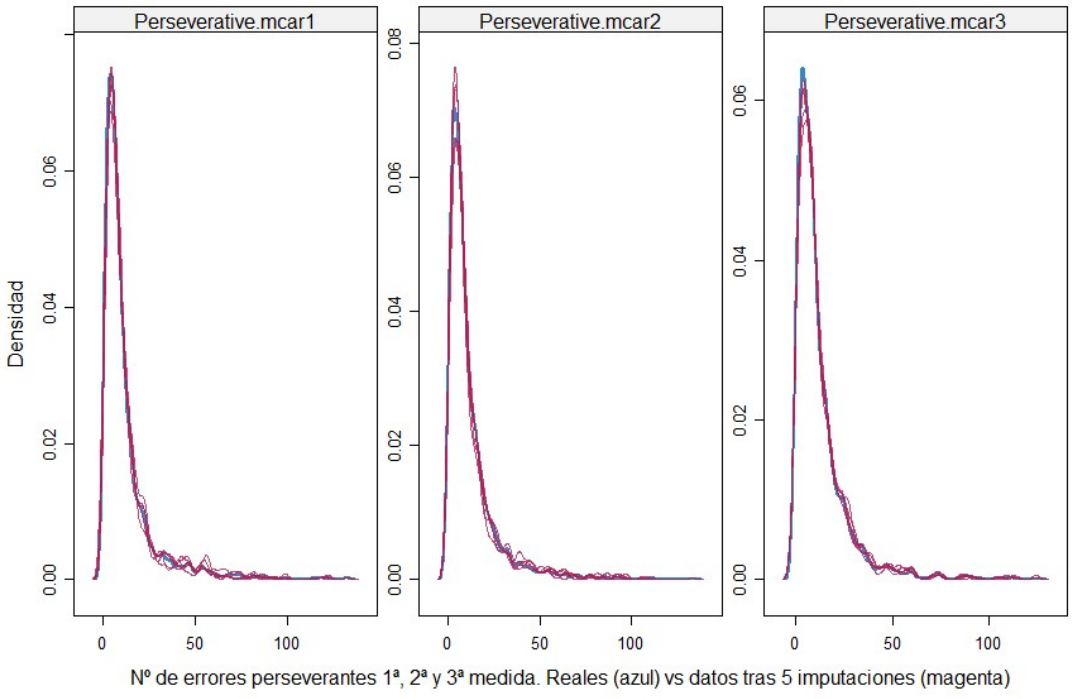
Se puede utilizar la función *xyplot* para representar los valores reales frente a los imputados en las variables de medición del nº de errores perseverantes (dos a dos).

Figura 65. Valores reales vs imputados IM (MICE-PMM): Nº de errores perseverantes (30%)



Para evaluar visualmente la eficacia de la imputación mediante el algoritmo PMM, utilizando la función ***densityplot*** existe la opción de obtener una figura comparativa de los gráficos de densidad de los datos contenidos en una variable determinada antes y en cada una de las imputaciones de los datos faltantes.

Figura 66. Densidad. Imputaciones IM (MICE-PMM) frente a datos reales: Nº de errores perseverantes (30%)



7. REPRODUCCIÓN DE LOS ANÁLISIS ORIGINALES

7.1 Reproducción de los análisis con datos originales

Los autores reportaron un aumento significativo ($p_{tendencia} < 0.001$) en el número medio (SD) de diseños únicos en el RFFT incrementándose de 73 (26) en la primera medición, a 79 (27) en la segunda medición y a 83 (26) en la tercera. Además dicho aumento se asoció negativamente con la edad (disminuyó con 0,23 por incremento de un año) y no se halló relación con el nivel educativo. Resultados similares se obtuvieron para los errores perseverantes: la mediana (IQR) en la primera medición fue 7 (3-13), en la segunda medición 7 (4-14) y aumentando en la tercera medición 8 (4-15) ($p_{tendencia} = 0.002$)(43).

A partir de la base de datos original se reproducen los resultados publicados:

Perfiles de los valores medios
Para esta sección se utiliza la BBDD en formato long original. Se analizan los valores medios y medianos de *diseños únicos* y de *errores perseverantes* en cada visita y se

presentan junto con el nº de participantes analizables en cada visita. Para reproducir los análisis del artículo original (*Marlise E. A. van Eersel et al.*), se utilizan el análisis de varianza (ANOVA) y el test de Kruskal Wallis, para analizar el cambio a lo largo de las 3 mediciones en el *nº de diseños únicos* y el *nº de errores perseverantes*.

```
##                          mean       sd IQR   0%  25% 50%   75%  100%
## 1ª medida (2003-06) 72.93917 25.59792  38 14.0 53.0  72  91.0 157.0
## 2ª medida (2006-08) 78.69245 26.59011  38 13.0 58.5  79  96.5 155.0
## 3ª medida (2008-12) 82.70795 28.43950  41 13.5 62.0  84 103.0 161.5
##                     Unique:n
## 1ª medida (2003-06)     2515
## 2ª medida (2006-08)     2515
## 3ª medida (2008-12)     2515

##                 Df  Sum Sq Mean Sq F value Pr(>F)
## Measurement_r    2  121268   60634   83.78 <2e-16 ***
## Residuals     7542 5458127     724
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##  Kruskal-Wallis rank sum test
##
## data:  Perseverative by Measurement
## Kruskal-Wallis chi-squared = 12.766, df = 2, p-value = 0.00169
```

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas.

Tabla 1. Resumen resultados medidas consecutivas (datos originales)

Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Todos los participantes	73 (26)	79 (27)	83 (28)	p<0.001
Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Todos los participantes	7 (3-13)	7 (4-14)	8 (4-15)	0.002

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis

Para reproducir los resultados obtenidos por los autores se utiliza la función *lme* (paquete *nlme*) que permite ajustar los efectos de modelos lineales mixtos y analizar los datos longitudinales. Los resultados obtenidos son muy similares a los que se presentan en el artículo (diferencias mínimas en algunos estimadores).

7.1.1 Reproducción del análisis original: Diseños únicos

Se ajustan los 3 modelos especificados en el artículo: En el primer modelo se incluyen *Edad*, *Género*, *Nivel de educación*, en el 2º se añade *Medida (nº consecutivo)* y en el 3º la *interacción entre edad y medida (nº consecutivo)*.

Con los resultados obtenidos se genera una tabla resumen de las medidas consecutivas del *Nº de diseños únicos*.

Tabla 2. Modelos lineales mixtos de los resultados RFFT: nº de diseños únicos

Variables	**Modelo 1**	**Modelo 2**	**Modelo 3**
———————-	***B*, IC 95% (p-valor)**	***B*, IC 95% (p-valor)**	***B*, IC 95% (p-valor)**
Edad	-1.12, -1.2 to -1.05 (<0.001)	-1.12, -1.2 to -1.05 (<0.001)	-0.66, -0.76 to -0.55 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.
Sexo: Mujer	1.36, -0.06 to 2.79 (0.061)	1.38, -0.05 to 2.8 (0.058)	1.38, -0.04 to 2.8 (0.058)
NE: Escuela primaria	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.69, 3.6 to 9.77 (<0.001)	6.64, 3.56 to 9.72 (<0.001)	6.63, 3.55 to 9.7 (<0.001)
NE: Secundaria superior	13.11, 9.97 to 16.25 (<0.001)	13.06, 9.93 to 16.19 (<0.001)	13.04, 9.91 to 16.17 (<0.001)
NE: Universitaria	25.41, 22.34 to 28.48 (<0.001)	25.36, 22.3 to 28.42 (<0.001)	22.28, 25.35 to 28.41 (<0.001)
Medida (nº consecutivo)		4.88, 4.48 to 5.29 (<0.001)	17.11, 15.09 to 19.14 (<0.001)
Edad x Medida (nº consecutivo)			-0.23, -0.27 to -0.19 (<0.001)

B: Coeficiente no estandarizado; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se comparan los 3 modelos utilizando el criterio de Akaike (AIC) que se puede obtener mediante la función *anova*:

```
##               Model df      AIC      BIC    logLik   Test  L.Ratio p-v
alue
## model1_unique     1  9 65749.14 65811.49 -32865.57
```

```
## model2_unique     2 10 65237.97 65307.24 -32608.98 1 vs 2 513.1752  <.
0001

## [1] 65749.14

## [1] 65237.97

## [1] 1.292358e-113

##               Model df      AIC      BIC    logLik   Test  L.Ratio p-v
alue
## model2_unique     1 10 65237.97 65307.24 -32608.98
## model3_unique     2 11 65103.72 65179.92 -32540.86 1 vs 2 136.2465  <.
0001

## [1] 65237.97

## [1] 65103.72

## [1] 1.762372e-31
```

Se inspeccionan la información numérica que aporta *summary* acerca de estos modelos:

- El AIC del modelo 1 es 65749.1.
- El AIC del modelo 2 es 65238.
- El AIC del modelo 3 es 65103.7.
- la varianza residual del modelo 1 es 16.646 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*), para el modelo 2 toma el valor 14.719 y para el modelo 3: 14.273.

Al comparar el modelo 1 con el 2, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de 511.18, que resulta estadísticamente significativa en beneficio del modelo 2, (<0.001). Al comparar el modelo 2 con el 3, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de 134.25, que resulta estadísticamente significativa en beneficio del modelo 3, (<0.001).

En este caso el modelo que se selecciona para comparar con las bases de datos generadas con datos faltantes es el modelo 3.

7.1.2 Reproducción de los análisis: Errores perseverantes

Se ajustan los 3 modelos especificados en el artículo: En el primer modelo se incluyen *Edad*, *Género*, *Nivel de educación*, en el 2º se añade *Medida (nº consecutivo)* y en el 3º la *interacción entre edad y medida (nº consecutivo)*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de errores perseverantes*.

Tabla 3. Modelos lineales mixtos de los resultados RFFT: nº de errores perseverantes

Variables	Modelo 1	Modelo 2	Modelo 3
__	***B*, IC 95% (p-valor)**	***B*, IC 95% (p-valor)**	***B*, IC 95% (p-valor)**
Edad	-0.02, -0.06 to 0.03 (0.491)	-0.02, -0.06 to 0.03 (0.491)	0.08, 0 to 0.15 (0.046)
Sexo: Hombre	Ref.	Ref.	Ref.
Sexo: Mujer	2.64, 1.76 to 3.53 (<0.001)	2.64, 1.76 to 3.53 (<0.001)	2.64, 1.76 to 3.53 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33, -0.58 to 3.24 (0.171)	1.33, -0.58 to 3.24 (0.171)	1.33, -0.57 to 3.24 (0.171)
NE: Secundaria superior	-0.26, -2.21 to 1.68 (0.792)	-0.26, -2.21 to 1.68 (0.792)	-0.26, -2.21 to 1.68 (0.792)
NE: Universitaria	-1.25, -3.15 to 0.65 (0.196)	-1.25, -3.15 to 0.65 (0.196)	-3.15, -1.25 to 0.65 (0.196)
Medida (nº consecutivo)		0.09, -0.23 to 0.41 (0.562)	2.54, 0.89 to 4.19 (0.002)
Edad x Medida (nº consecutivo)			-0.05, -0.08 to -0.02 (0.003)

B: Coeficiente no estandarizado; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se comparan los 3 modelos utilizando el criterio de Akaike (AIC) que se puede obtener mediante la función *anova*:

```
##                       Model df      AIC      BIC    logLik   Test  L.Ra
tio
## model1_perseverative      1  9 60778.18 60840.53 -30380.09
## model2_perseverative      2 10 60781.63 60850.91 -30380.81 1 vs 2 1.452
029
##                       p-value
## model1_perseverative
## model2_perseverative  0.2282

## [1] 60778.18

## [1] 60781.63

## [1] 0.2282026
```

```
##                       Model df      AIC      BIC    logLik   Test  L.Ra
tio
## model2_perseverative     1 10 60781.63 60850.91 -30380.81
## model3_perseverative     2 11 60781.30 60857.50 -30379.65 1 vs 2 2.334
206
##                       p-value
## model2_perseverative
## model3_perseverative  0.1266

## [1] 60781.63

## [1] 60781.3

## [1] 0.1265595

##                       Model df      AIC      BIC    logLik   Test   L.R
atio
## model1_perseverative     1  9 60778.18 60840.53 -30380.09
## model3_perseverative     2 11 60781.30 60857.50 -30379.65 1 vs 2 0.882
1769
##                       p-value
## model1_perseverative
## model3_perseverative  0.6433

## [1] 60778.18

## [1] 60781.3

## [1] 0.6433358
```

Se inspeccionan la información numérica que aporta *summary* acerca de estos modelos:

- El AIC del modelo 1 es 60778.2.
- El AIC del modelo 2 es 60781.6.
- El AIC del modelo 3 es 60781.3.
- la varianza residual del modelo 1 es 11.589 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*), para el modelo 2 toma el valor 11.591 y para el modelo 3: 11.573.

Al comparar el modelo 1 con el 2, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de -3.45, que no resulta estadísticamente significativa en beneficio del modelo 2, 0.228.
Al comparar el modelo 2 con el 3, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de 0.33, que no resulta estadísticamente significativa en beneficio del modelo 3, 0.127.
Al comparar el modelo 1 con el 3, mediante los valores del criterio Akaike (AIC) se obtiene una diferencia de -3.12, que no resulta estadísticamente significativa en beneficio del modelo 3, 0.643.

En este caso el modelo que se selecciona para comparar con las bases de datos generadas con datos faltantes es el modelo 1.

---

8. REPRODUCCIÓN DE LOS ANÁLISIS CON 10% DE DATOS FALTANTES

8.2 Reproducción de los análisis: Eliminación de los casos (listwise) (10%)

Se utilizan los dataframes generados anteriormente mediante el método de *eliminación de los casos con datos faltantes (listwise)* para reprodcucir el análisis original, en el caso de 10% de datos faltantes.

Perfiles de los valores medios
Para esta sección se utiliza la BBDD en formato long original con datos faltantes. Se analizan los valores medios y medianos de *diseños únicos* y de *errores perseverantes* en cada visita y se presentan junto con el nº de participantes analizables en cada visita. Para reproducir los análisis del artículo original (*Marlise E. A. van Eersel et al.*), se utilizan el análisis de varianza (ANOVA) y el test de Kruskal Wallis, para analizar el cambio a lo largo de las 3 mediciones en el *nº de diseños únicos* y el *nº de errores perseverantes*.

```
##                   mean       sd  IQR   0%   25% 50%    75%  100% Uniqu
e:n
## Measurement 1 73.63526 25.70821 38.0 14.0 53.00  73  91.00 152.0     1
815
## Measurement 2 79.00386 26.47429 37.5 13.0 59.25  79  96.75 155.0     1
815
## Measurement 3 82.94325 28.28829 41.0 13.5 62.00  84 103.00 161.5     1
815

##                Df  Sum Sq Mean Sq F value Pr(>F)
## Measurement2    1   78625   78625   109.1 <2e-16 ***
## Residuals    5443 3922536     721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##  Kruskal-Wallis rank sum test
##
## data:  Perseverative by Measurement2
## Kruskal-Wallis chi-squared = 12.577, df = 2, p-value = 0.001858
```

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas.

Tabla 4. Resumen resultados medidas consecutivas (Eliminación de los casos (listwise))

Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Todos los participantes	74 (26)	79 (26)	83 (28)	p<0.001
Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Todos los participantes	7 (3-13)	7 (4-14)	8 (4-15)	0.002

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis

Para reproducir los resultados obtenidos por los autores se utiliza la función *lme* (paquete *nlme*) que permite ajustar los efectos de modelos lineales mixtos y analizar los datos longitudinales. Los resultados obtenidos son muy similares a los que se presentan en el artículo (diferencias mínimas en algunos estimadores).

8.2.1 Reproducción de los análisis. Modelo nº diseños únicos: Eliminación de los casos (listwise) (10%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de diseños únicos*.

Tabla 5. Modelos lineales mixtos de los resultados RFFT: nº de diseños únicos (Eliminación de los casos (listwise))

Variables	**Modelo 3**
_	***B* (DE), IC 95% (p-valor)**
Edad	-0.69 (0.062), -0.81 to -0.57 (<0.001)
Sexo: Hombre	Ref.
Sexo: Mujer	1.14 (0.853), -0.53 to 2.82 (0.18)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	6.66 (1.866), 3 to 10.32 (<0.001)
NE: Secundaria superior	12.24 (1.896), 8.52 to 15.96 (<0.001)
NE: Universitaria	20.94 (1.846), 24.56 to 28.18 (<0.001)
Medida (nº consecutivo)	16.34 (1.207), 13.97 to 18.71 (<0.001)
Edad x Medida (nº consecutivo)	-0.22 (0.023), -0.27 to -0.18 (<0.001)

B: Coeficiente no estandarizado; DE: Desviación estándard; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 47020.5.

- la varianza residual es 14.224 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

# #

8.2.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Eliminación de los casos (listwise) (10%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de errores perseverantes*.

Tabla 6. Modelos lineales mixtos de los resultados RFFT: nº de errores perseverantes (Eliminación de los casos (listwise))

Variables	**Modelo 1**
__	***B* (DE), IC 95% (p-valor)**
Edad	-0.02 (0.026), -0.07 to 0.68 (0.398)
Sexo: Hombre	Ref.
Sexo: Mujer	2.33 (0.526), 1.3 to 3.36 (<0.001)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	0.75 (1.133), -1.47 to 2.98 (0.506)
NE: Secundaria superior	-0.37(1.154), -2.63 to 1.9 (0.751)
NE: Universitaria	-1.54 (1.129), -3.75 to 0.68 (0.174)

B: Coeficiente no estandarizado; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 44393.4.

- la varianza residual es 11.281 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

8.3 Reproducción de los análisis: Media (10%)

Se utilizan los dataframes generados anteriormente mediante el método de *Imputación por media* , que reemplaza los datos faltantes por el valor medio del resto de datos conocidos en de la variable. Y con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

Perfiles de los valores medios

Para esta sección se utiliza la BBDD en formato long con datos imputados. Se analizan los valores medios y medianos de *diseños únicos* y de *errores perseverantes* en cada visita y se presentan junto con el nº de participantes analizables en cada visita. Para reproducir los análisis del artículo original (*Marlise E. A. van Eersel et al.*), se utilizan el análisis de varianza (ANOVA) y el test de Kruskal Wallis, para analizar el cambio a lo largo de las 3 mediciones en el *nº de diseños únicos* y el *nº de errores persevarentes*.

```
##                   mean       sd  IQR   0%  25%      50% 75%  100% Uniq
ue:n
## Measurement 1 73.34336 24.28495 33.0 14.0 56.0 73.34336  89 157.0
2515
## Measurement 2 78.81443 25.06033 33.5 13.0 61.5 78.81443  95 155.0
2515
## Measurement 3 82.62411 26.85676 35.0 13.5 65.0 82.62411 100 161.5
2515

##                   mean       sd IQR 0% 25% 50% 75%  100% Perseverative
:n
## Measurement 1 11.75809 14.65574   8  0   4 8.0  12 132.0            25
15
## Measurement 2 11.73730 13.53611   9  0   4 8.5  13 133.5            25
15
## Measurement 3 11.98561 13.03623  10  0   4 9.0  14 124.0            25
15

##                Df  Sum Sq Mean Sq F value Pr(>F)
## Measurement2    1  108311  108311   167.6 <2e-16 ***
## Residuals    7543 4875965     646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##  Kruskal-Wallis rank sum test
##
## data:  Perseverative by Measurement2
## Kruskal-Wallis chi-squared = 11.766, df = 2, p-value = 0.002786
```

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas.

Tabla 7. Resumen resultados medidas consecutivas (reemplazamiento por la media)

Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Todos los participantes	73 (24)	79 (25)	83 (27)	p<0.001
Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Todos los participantes	8 (4-12)	8 (4-13)	9 (4-14)	0.003

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis

Para reproducir los resultados obtenidos por los autores se utiliza la función *lme* (paquete *nlme*) que permite ajustar los efectos de modelos lineales mixtos y analizar los datos longitudinales. Los resultados obtenidos son muy similares a los que se presentan en el artículo (diferencias mínimas en algunos estimadores).

8.3.1 Reproducción de los análisis. Modelo nº diseños únicos: Media (10%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de diseños únicos*.

Tabla 8. Modelos lineales mixtos de los resultados RFFT: nº de diseños únicos (Media)

Variables	**Modelo 3**
_	***B* (DE), IC 95% (p-valor)**
Edad	-0.61 (0.053), -0.72 to -0.51 (<0.001)
Sexo: Hombre	Ref.
Sexo: Mujer	1.01 (0.68), -0.33 to 2.34 (0.139)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	5.63 (1.469), 2.75 to 8.51 (<0.001)
NE: Secundaria superior	10.94 (1.496), 8.01 to 13.87 (<0.001)
NE: Universitaria	19.1 (1.462), 21.97 to 24.84 (<0.001)
Medida (nº consecutivo)	15.2 (1.098), 13.04 to 17.35 (<0.001)
Edad x Medida (nº consecutivo)	-0.2 (0.02), -0.24 to -0.16 (<0.001)

B: Coeficiente no estandarizado; DE: Desviación estándard; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 65528.5.

- la varianza residual es 15.147 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

# #

8.3.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Media (10%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de errores perseverantes*.

Tabla 9. Modelos lineales mixtos de los resultados RFFT: nº de errores perseverantes (Media)

Variables	**Modelo 1**
_	***B* (DE), (p-valor)**
Edad	-0.015 (0.021), (0.489)
Sexo: Hombre	Ref.
Sexo: Mujer	2.261 (0.417), (<0.001)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	1.272 (0.902), (0.158)
NE: Secundaria superior	-0.156(0.918), (0.865)
NE: Universitaria	-1.076 (0.898), (0.231)

B: Coeficiente no estandarizado; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación
Para este modelo no se pudieron calcular los IC al 95% por "Non-positive definite approximate variance-covariance"

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 60020.2.

- la varianza residual es 11.035 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

8.4 Reproducción de los análisis: Regresión (10%)

Se utilizan los dataframes generados anteriormente mediante el método de *Imputación por regresión simple (media condicional)*, que reemplaza los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto. Con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

Perfiles de los valores medios

Para esta sección se utiliza la BBDD en formato long con datos imputados. Se analizan los valores medios y medianos de *diseños únicos* y de *errores perseverantes* en cada visita y se presentan junto con el nº de participantes analizables en cada visita. Para reproducir los análisis del artículo original (*Marlise E. A. van Eersel et al.*), se utilizan el análisis de varianza (ANOVA) y el test de Kruskal Wallis, para analizar el cambio a lo largo de las 3 mediciones en el *nº de diseños únicos* y el *nº de errores persevarentes*.

```
##                   mean       sd      IQR   0% 25%      50%     75%  100
%
## Measurement 1 73.2052 25.14272 37.00000 14.0  53 73.00000  90.000 157.
0
## Measurement 2 78.6060 26.02469 37.00000 13.0  59 78.94425  96.000 155.
0
## Measurement 3 82.5522 27.90424 40.41202 13.5  62 83.50000 102.412 161.
5
##               Unique:n
## Measurement 1     2515
## Measurement 2     2515
## Measurement 3     2515

##                   mean       sd       IQR 0% 25% 50%      75%  100%
## Measurement 1 11.72458 14.84073  9.228735  0   4 7.0 13.22874 132.0
## Measurement 2 11.77212 13.76294 10.000000  0   4 8.0 14.00000 133.5
## Measurement 3 12.05268 13.21597 11.000000  0   4 8.5 15.00000 124.0
##               Perseverative:n
## Measurement 1            2515
## Measurement 2            2515
## Measurement 3            2515

##                Df  Sum Sq Mean Sq F value Pr(>F)
## Measurement2    1  109863  109863   157.8 <2e-16 ***
## Residuals    7543 5250339     696
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Perseverative by Measurement2
## Kruskal-Wallis chi-squared = 14.07, df = 2, p-value = 0.0008807
```

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas.

Tabla 10. Resumen resultados medidas consecutivas (Regresión)

Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Todos los participantes	73 (25)	79 (26)	83 (28)	p<0.001
Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Todos los participantes	7 (4-13)	8 (4-14)	8 (4-15)	0.001

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis

Para reproducir los resultados obtenidos por los autores se utiliza la función *lme* (paquete *nlme*) que permite ajustar los efectos de modelos lineales mixtos y analizar los datos longitudinales. Los resultados obtenidos son muy similares a los que se presentan en el artículo (diferencias mínimas en algunos estimadores).

8.4.1 Reproducción de los análisis. Modelo nº diseños únicos: Regresión (10%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de diseños únicos*.

Tabla 11. Modelos lineales mixtos de los resultados RFFT: nº de diseños únicos (Regresión)

Variables	**Modelo 3**
_	***B* (DE), IC 95% (p-valor)**
Edad	-0.67 (0.051), -0.76 to -0.57 (<0.001)
Sexo: Hombre	Ref.
Sexo: Mujer	1.27 (0.719), -0.14 to 2.68 (0.077)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	6.57 (1.555), 3.52 to 9.62 (<0.001)
NE: Secundaria superior	12.76 (1.583), 9.66 to 15.86 (<0.001)
NE: Universitaria	22 (1.547), 25.04 to 28.07 (<0.001)

Medida (nº consecutivo)	16.64 (0.952), 14.77 to 18.51 (<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.018), -0.26 to -0.19 (<0.001)

B: Coeficiente no estandarizado; DE: Desviación estándard; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 64291.2.

- la varianza residual es 13.163 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

# #

8.4.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Regresión (10%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de errores perseverantes*.

Tabla 12. Modelos lineales mixtos de los resultados RFFT: nº de errores perseverantes (Regresión)

Variables	**Modelo 1**
_	***B* (DE), IC 95% (p-valor)**
Edad	-0.02 (0.022), -0.06 to 0.7 (0.47)
Sexo: Hombre	Ref.
Sexo: Mujer	2.52 (0.44), 1.65 to 3.38 (<0.001)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	1.54 (0.952), -0.33 to 3.41 (0.106)
NE: Secundaria superior	-0.14(0.969), -2.04 to 1.76 (0.883)
NE: Universitaria	-1.16 (0.948), -3.01 to 0.7 (0.223)

B: Coeficiente no estandarizado; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 59790.

- la varianza residual es 10.753 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

8.5 Reproducción de los análisis: MI (PMM) (10%)

Se utilizan los dataframes generados anteriormente utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*. Con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

Perfiles de los valores medios

Para esta sección se utiliza la BBDD en formato long con datos imputados. Se analizan los valores medios y medianos de *diseños únicos* y de *errores perseverantes* en cada visita y se presentan junto con el nº de participantes analizables en cada visita. Para reproducir los análisis del artículo original (*Marlise E. A. van Eersel et al.*), se utilizan el análisis de varianza (ANOVA) y el test de Kruskal Wallis, para analizar el cambio a lo largo de las 3 mediciones en el *nº de diseños únicos* y el *nº de errores persevarentes*.

```
##                   mean       sd   IQR   0%  25%  50%    75%  100% Uniq
ue:n
## Measurement 1 73.17535 25.67941 38.00 14.0 53.0 73.0  91.00 157.0
2515
## Measurement 2 78.61093 26.33853 37.25 13.0 59.0 78.5  96.25 155.0
2515
## Measurement 3 82.56501 28.46342 41.50 13.5 61.5 84.0 103.00 161.5
2515

##                   mean       sd IQR 0% 25% 50%  75%  100% Perseverativ
e:n
## Measurement 1 11.71531 15.41559  10  0 3.0 7.0 13.0 132.0            2
515
## Measurement 2 12.01511 14.72812  11  0 3.5 7.5 14.5 133.5            2
515
## Measurement 3 12.01431 13.69773  11  0 4.0 8.0 15.0 124.0            2
515

##                Df  Sum Sq Mean Sq F value Pr(>F)
## Measurement2    1  110868  110868   153.7 <2e-16 ***
## Residuals    7543 5439497     721
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##
##  Kruskal-Wallis rank sum test
##
```

```
## data:  Perseverative by Measurement2
## Kruskal-Wallis chi-squared = 15.988, df = 2, p-value = 0.0003374
```

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas.

Tabla 13. Resumen resultados medidas consecutivas (Imputación múltiple (PMM))

Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Todos los participantes	73 (26)	79 (26)	83 (28)	p<0.001
Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Todos los participantes	7 (3-13)	8 (4-14)	8 (4-15)	0

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis

Para reproducir los resultados obtenidos por los autores se utiliza la función *lme* (paquete *nlme*) que permite ajustar los efectos de modelos lineales mixtos y analizar los datos longitudinales. Los resultados obtenidos son muy similares a los que se presentan en el artículo (diferencias mínimas en algunos estimadores).

8.5.1 Reproducción de los análisis. Modelo nº diseños únicos: MI (PMM) (10%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de diseños únicos*.

Tabla 14. Modelos lineales mixtos de los resultados RFFT: nº de diseños únicos (MI (PMM))

Variables	**Modelo 3**
_	***B* (DE), IC 95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.56 (<0.001)
Sexo: Hombre	Ref.
Sexo: Mujer	1.16 (0.723), -0.25 to 2.58 (0.108)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	6.7 (1.562), 3.64 to 9.77 (<0.001)
NE: Secundaria superior	12.84 (1.59), 9.72 to 15.95 (<0.001)
NE: Universitaria	22.09 (1.555), 25.14 to 28.19 (<0.001)

Medida (nº consecutivo)	16.9 (1.021), 14.9 to 18.9 (<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)

B: Coeficiente no estandarizado; DE: Desviación estándard; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 65187.4.

- la varianza residual es 14.057 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

# #

8.5.2 Reproducción de los análisis. Modelos nº de errores perseverantes: MI (PMM) (10%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

Se utilizan los resultados obtenidos para generar una tabla resumen de los resultados en las medidas consecutivas del *Nº de errores perseverantes*.

Tabla 15. Modelos lineales mixtos de los resultados RFFT: nº de errores perseverantes (MI (PMM))

Variables	**Modelo 1**
_	***B* (DE), IC 95% (p-valor)**
Edad	-0.01 (0.023), -0.06 to 0.88 (0.58)
Sexo: Hombre	Ref.
Sexo: Mujer	2.56 (0.452), 1.68 to 3.45 (<0.001)
NE: Escuela primaria	Ref.
NE: Secundaria inicial	1.63 (0.976), -0.29 to 3.54 (0.095)
NE: Secundaria superior	-0.04(0.994), -1.99 to 1.91 (0.968)
NE: Universitaria	-1.02 (0.972), -2.93 to 0.88 (0.293)

B: Coeficiente no estandarizado; IC 95: Intervalo de confianza al 95% del coeficiente B; Ref.: Categoría de referencia; NE: Nivel de educación

Se inspecciona la información numérica que aporta *summary* acerca de este modelo:

- El AIC del modelo es 60753.2.

- la varianza residual es 11.559 (la salida da 'sigma'), que es la estimación de la varianza de los residuos (como std.dev), por lo tanto reporta la variabilidad entre observaciones (*"within-participante variability"*).

---

9. REPRODUCCIÓN DE LOS ANÁLISIS CON 20% DE DATOS FALTANTES

Se reproducen los modelos multivariados mediante los 4 métodos de tratamiento de datos faltantes con los dataframes que contiene 20% de datos faltantes.

9.1 Reproducción de los análisis: Eliminación de los casos (listwise) (20%)

Se utilizan los dataframes generados anteriormente mediante el método de *eliminación de los casos con datos faltantes (listwise)* para reprodcucir el análisis original, en el caso de 20% de datos faltantes.

9.1.1 Reproducción de los análisis. Modelo nº diseños únicos: Eliminación de los casos (listwise) (20%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

9.1.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Eliminación de los casos (listwise) (20%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

9.2 Reproducción de los análisis: Media (20%)

Se utilizan los dataframes generados anteriormente mediante el método de *Imputación por media* , que reemplaza los datos faltantes por el valor medio del resto de datos conocidos en de la variable. Y con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

9.2.1 Reproducción de los análisis. Modelo nº diseños únicos: Media (20%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

9.2.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Media (20%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

9.3 Reproducción de los análisis: Regresión (20%)

Se utilizan los dataframes generados anteriormente mediante el método de *Imputación por regresión simple (media condicional)*, que reemplaza los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto. Con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

9.3.1 Reproducción de los análisis. Modelo nº diseños únicos: Regresión (20%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

9.3.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Regresión (20%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

9.4 Reproducción de los análisis: MI (PMM) (20%)

Se utilizan los dataframes generados anteriormente utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*. Con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

9.4.1 Reproducción de los análisis. Modelo nº diseños únicos: MI (PMM) (20%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

9.4.2 Reproducción de los análisis. Modelos nº de errores perseverantes: MI (PMM) (20%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

---

10. REPRODUCCIÓN DE LOS ANÁLISIS CON 30% DE DATOS FALTANTES

Se reproducen los modelos multivariados mediante los 4 métodos de tratamiento de datos faltantes con los dataframes que contiene 30% de datos faltantes.

10.1 Reproducción de los análisis: Eliminación de los casos (listwise) (30%)

Se utilizan los dataframes generados anteriormente mediante el método de *eliminación de los casos con datos faltantes (listwise)* para reprodcucir el análisis original, en el caso de 30% de datos faltantes.

10.1.1 Reproducción de los análisis. Modelo nº diseños únicos: Eliminación de los casos (listwise) (30%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

10.1.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Eliminación de los casos (listwise) (30%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

10.2 Reproducción de los análisis: Media (30%)

Se utilizan los dataframes generados anteriormente mediante el método de *Imputación por media* , que reemplaza los datos faltantes por el valor medio del resto de datos conocidos en de la variable. Y con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

10.2.1 Reproducción de los análisis. Modelo nº diseños únicos: Media (30%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

10.2.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Media (30%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

10.3 Reproducción de los análisis: Regresión (30%)

Se utilizan los dataframes generados anteriormente mediante el método de *Imputación por regresión simple (media condicional)*, que reemplaza los datos faltantes por el valor predicho a partir de la aplicación de la regresión simple sobre el resto de datos conocidos en el sujeto. Con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

10.3.1 Reproducción de los análisis. Modelo nº diseños únicos: Regresión (30%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

10.3.2 Reproducción de los análisis. Modelos nº de errores perseverantes: Regresión (30%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

10.4 Reproducción de los análisis: MI (PMM) (30%)

Se utilizan los dataframes generados anteriormente utilizando el algoritmo de *Múltiple Imputación utilizando el método PMM (Predictive mean matching)*. Con los nuevos dataframes se reproducen los *mejores* modelos del análisis original. En primer lugar se transponen los datos de formato *wide* a formato *long*.

10.4.1 Reproducción de los análisis. Modelo nº diseños únicos: MI (PMM) (30%)

Se ajusta el modelo 3 especificado en el artículo: se incluyen *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*.

10.4.2 Reproducción de los análisis. Modelos nº de errores perseverantes: MI (PMM) (30%)

Se ajusta el 1º modelo especificado en el artículo, se incluyen: *Edad*, *Género* y *Nivel de educación*.

---

11. RESUMEN DE RESULTADOS

A continuación se resumen los resultados obtenidos originalmente por los autores con las 4 simulaciones que se han propuesto anteriormente para los dataframes con datos faltantes en el 10% de los casos. Los resultados son similares tanto para el nº de diseños únicos como para el nº de errores perseverantes.

Tabla 16. Resumen resultados medidas consecutivas del nº de diseños únicos (10%)

Análisis: Nº de diseños únicos, media(DE)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (anova)**
Datos originales	72.9 (25.6)	78.7 (26.6)	82.7 (28.4)	p<0.001
Eliminación de los casos (listwise)	73.6 (25.7)	79 (26.5)	82.9 (28.3)	p<0.001
Reemplazamiento por la media	73.3 (24.3)	78.8 (25.1)	82.6 (26.9)	p<0.001
Regresión	73.2 (25.1)	78.6 (26)	82.6 (27.9)	p<0.001
Imputación múltiple (*PMM*)	73.2 (25.7)	78.6 (26.3)	82.6 (28.5)	p<0.001

Tabla 17. Resumen resultados medidas consecutivas del nº de errores perseverantes

Análisis: Nº errores perseverantes, mediana(RIQ)	**1ª (2003-06)**	**2ª (2006-08)**	**3ª (2008-12)**	**p-valor (K-W)**
Datos originales	7 (3-13)	7 (3.5-14)	8 (4-15)	0.002
Eliminación de los casos (listwise)	7 (3-13)	7 (3.5-14)	8 (4-15)	0.002
Reemplazamiento por la media	8 (4-12)	8.5 (4-13)	9 (4-14)	0.003
Regresión	7 (4-13.2)	8 (4-14)	8.5 (4-15)	0.001
Imputación múltiple (*PMM*)	7 (3-13)	7.5 (3.5-14.5)	8 (4-15)	0

DE: Desviación estándard; RIQ: rango inter-cuartílico; K-W: Kruskal-Wallis; PMM: Predictive Mean Matching o Equiparación de media predictiva

11.1 Comparación de los modelos: Original vs 10%

A modo de análisis de sensibilidad se comparan los resultados de los modelos con las 4 técnicas de tratamiento de los datos faltantes, con los resultados originales. En el caso de los diseños únicos el 3er modelo, que incluye *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*, fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 18. Resumen comparativo resultados de los modelos: nº de diseños únicos (10%)

Tipo de análisis	**Análisis original**	**Listwise**	**Media**	**Regresión**	**IM (5, PMM)**
Modelo	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.69 (0.062) , -0.81 to -0.57 (<0.001)	-0.61 (0.053), -0.72 to -0.51 (<0.001)	-0.67 (0.051), -0.76 to -0.57 (<0.001)	-0.66 (0.053), -0.76 to -0.56 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	1.14 (0.853) , -0.53 to 2.82 (0.18)	1.01 (0.68), -0.33 to 2.34 (0.139)	1.27 (0.719), -0.14 to 2.68 (0.077)	1.16 (0.723), -0.25 to 2.58 (0.108)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	6.66 (1.866), 3 to 10.32 (<0.001)	5.63 (1.469), 2.75 to 8.51 (<0.001)	6.57 (1.555), 3.52 to 9.62 (<0.001)	6.7 (1.562), 3.64 to 9.77 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	12.24 (1.896), 8.52 to 15.96 (<0.001)	10.94 (1.496), 8.01 to 13.87 (<0.001)	12.76 (1.583), 9.66 to 15.86 (<0.001)	12.84 (1.59), 9.72 to 15.95 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	24.56 (1.846), 20.94 to 28.18 (<0.001)	21.97 (1.462), 19.1 to 24.84 (<0.001)	25.04 (1.547), 22 to 28.07 (<0.001)	25.14 (1.555), 22.09 to 28.19 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	24.56 (1.207), 13.97 to 18.71(0.18)	15.2 (1.098), 13.04 to 17.35(<0.001)	16.64 (0.952), 14.77 to 18.51(<0.001)	16.9 (1.021), 14.9 to 18.9(<0.001)

Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.22 (0.023), -0.27 to -0.18 (<0.001)	-0.2 (0.02), -0.24 to -0.16 (<0.001)	-0.23 (0.018), -0.26 to -0.19 (<0.001)	-0.23 (0.019), -0.27 to -0.19 (<0.001)
AIC	65103.7	47020.5*	65528.5	64291.2	65187.4
Varianza residual	14.3	14.2*	15.1	13.2	14.1

*B: Coeficiente no estandarizado; DE: Desviación estándard; IC: Intervalo de confianza; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike * : Se ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

Tabla 19. Resumen comparativo resultados de los modelos: nº de errores perseverantes (10%)

Tipo de análisis	**Análisis original**	**Listwise**	**Media**	**Regresión**	**IM (5, PMM)**
Modelo	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.02 (0.023), -0.06 to 0.03 (0.491)	-0.02 (0.026), -0.07 to 0.03(0.398)	-0.015 (0.021), – to – (0.489)	-0.02 (0.022), -0.06 to 0.03 (0.47)	-0.01 (0.023), -0.06 to 0.03 (0.58)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	2.64 (0.451), 1.76 to 3.53 (<0.001)	2.33 (0.526), 1.3 to 3.36 (<0.001)	2.261 (0.417), – to – (<0.001)	2.52 (0.44), 1.65 to 3.38 (<0.001)	2.56 (0.452), 1.68 to 3.45 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33 (0.974), -0.58 to 3.24 (0.171)	0.75 (1.133), -1.47 to 2.98 (0.506)	1.272 (0.902), – to – (0.158)	1.54 (0.952), -0.33 to 3.41 (0.106)	1.63 (0.976), -0.29 to 3.54 (0.095)
NE: Secundaria superior	-0.26 (0.992), -2.21 to 1.68 (0.792)	-0.37 (1.154), -2.63 to 1.9 (0.751)	-0.156 (0.918), – to – (0.865)	-0.14 (0.969), -2.04 to 1.76 (0.883)	-0.04 (0.994), -1.99 to 1.91 (0.968)
NE: Universitaria	-1.25 (0.97), -3.15 to 0.65 (0.196)	-1.54 (1.129), -3.75 to 0.68 (0.174)	0.898 (0.898), – to – (0.231)	-1.16 (0.948), -3.01 to 0.7 (0.223)	-1.02 (0.972), -2.93 to 0.88 (0.293)
AIC	60778.2	44393.4*	60020.2	59790	60753.2
Varianza residual	11.6	11.3*	11	10.8	11.6

*B: Coeficiente no estandarizado; DE: Desviación estándard; IC: Intervalo de confianza; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike * : Se ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

11.2 Comparación de los modelos: Original vs 20%

A modo de análisis de sensibilidad se comparan los resultados de los modelos con las 4 técnicas de tratamiento de los datos faltantes, con los resultados originales. En el caso de los diseños únicos el 3er modelo, que incluye *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*, fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias

estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 20. Resumen comparativo resultados de los modelos: nº de diseños únicos (20%)

Tipo de análisis	**Análisis original**	**Listwise**	**Media**	**Regresión**	**IM (5, PMM)**
Modelo	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.69 (0.075) , -0.83 to -0.54 (<0.001)	-0.54 (0.053), -0.65 to -0.44 (<0.001)	-0.64 (0.049), -0.74 to -0.55 (<0.001)	-0.68 (0.053), -0.78 to -0.57 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	0.68 (1.023) , -1.33 to 2.69 (0.506)	0.97 (0.629), -0.26 to 2.2 (0.123)	1.39 (0.712), -0.01 to 2.79 (0.051)	1.25 (0.733), -0.19 to 2.69 (0.088)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	6.87 (2.302), 2.35 to 11.38 (0.003)	4.95 (1.36), 2.28 to 7.62 (<0.001)	6.48 (1.54), 3.46 to 9.5 (<0.001)	6.74 (1.585), 3.64 to 9.85 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	11.43 (2.344), 6.84 to 16.03 (<0.001)	9.6 (1.385), 6.88 to 12.31 (<0.001)	12.81 (1.568), 9.73 to 15.88 (<0.001)	12.62 (1.614), 9.46 to 15.79 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	23.44 (2.278), 18.97 to 27.91 (<0.001)	19.38 (1.354), 16.72 to 22.03 (<0.001)	25.11 (1.533), 22.11 to 28.12 (<0.001)	25.6 (1.578), 22.51 to 28.69 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	23.44 (1.465), 14.6 to 20.35(0.506)	13.92 (1.149), 11.67 to 16.18(<0.001)	16.8 (0.88), 15.07 to 18.52(<0.001)	15.84 (1.02), 13.85 to 17.84(<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.24 (0.027), -0.3 to -0.19 (<0.001)	-0.18 (0.021), -0.22 to -0.13 (<0.001)	-0.23 (0.016), -0.26 to -0.2 (<0.001)	-0.21 (0.019), -0.25 to -0.18 (<0.001)
AIC	65103.7	32348.2*	65665	63483.5	65193.4
Varianza residual	14.3	14.3*	15.8	12.1	14.1

B: Coeficiente no estandarizado; DE: Desviación estándard; IC: Intervalo de confianza; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike
** : Se ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

Tabla 21. Resumen comparativo resultados de los modelos: nº de errores perseverantes (20%)

Tipo de análisis	**Análisis original**	**Listwise**	**Media**	**Regresión**	**IM (5, PMM)**
Modelo	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.02 (0.023), -0.06 to 0.03 (0.491)	-0.02 (0.031), -0.08 to 0.04(0.513)	-0.01 (0.019), – to – (0.617)	-0.01 (0.022), -0.06 to 0.03 (0.595)	-0.01 (0.022), -0.06 to 0.03 (0.572)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	2.64 (0.451), 1.76 to 3.53 (<0.001)	2.44 (0.615), 1.23 to 3.64 (<0.001)	2.089 (0.385), – to – (<0.001)	2.66 (0.445), 1.78 to 3.53 (<0.001)	2.61 (0.446), 1.74 to 3.49 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.

NE: Secundaria inicial	1.33 (0.974), -0.58 to 3.24 (0.171)	-0.44 (1.33), -3.05 to 2.17 (0.739)	1.146 (0.832), – to – (0.169)	1.79 (0.963), -0.1 to 3.68 (0.063)	1.88 (0.963), -0.01 to 3.76 (0.052)
NE: Secundaria superior	-0.26 (0.992), -2.21 to 1.68 (0.792)	-1.5 (1.347), -4.14 to 1.14 (0.265)	-0.078 (0.847), – to – (0.927)	0.13 (0.98), -1.79 to 2.06 (0.892)	0.15 (0.981), -1.78 to 2.07 (0.882)
NE: Universitaria	-1.25 (0.97), -3.15 to 0.65 (0.196)	-2.87 (1.319), -5.46 to -0.28 (0.03)	0.829 (0.829), – to – (0.262)	-0.88 (0.959), -2.76 to 1 (0.358)	-0.72 (0.959), -2.6 to 1.16 (0.454)
AIC	60778.2	31082.9*	59336.4	58776	60347.5
Varianza residual	11.6	11.6*	10.8	10.3	11.4

B: Coeficiente no estandarizado; DE: Desviación estándard; IC: Intervalo de confianza; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike

* *: Se ha de tener en cuenta que este caso presenta una* ***n*** *distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

11.3 Comparación de los modelos: Original vs 30%

A modo de análisis de sensibilidad se comparan los resultados de los modelos con las 4 técnicas de tratamiento de los datos faltantes, con los resultados originales. En el caso de los diseños únicos el 3er modelo, que incluye *Edad*, *Género*, *Nivel de educación*, *Medida (nº consecutivo)* y la *interacción entre edad y medida (nº consecutivo)*, fue el que presentó un menor valor del criterio Akaike-AIC (hallándose diferencias estadísticamente significativas respecto a los modelos más parsimoniosos), y es el utilizado para hacer la comparación con las 4 simulaciones.

Tabla 22. Resumen comparativo resultados de los modelos: nº de diseños únicos (30%)

Tipo de análisis	**Análisis original**	**Listwise**	**Media**	**Regresión**	**IM (5, PMM)**
Modelo	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**	**Modelo 3**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.66 (0.053), -0.76 to -0.55 (<0.001)	-0.58 (0.093) , -0.76 to -0.4 (<0.001)	-0.5 (0.053), -0.6 to -0.4 (<0.001)	-0.66 (0.047), -0.75 to -0.57 (<0.001)	-0.64 (0.053), -0.75 to -0.54 (<0.001)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	1.38 (0.725), -0.04 to 2.8 (0.058)	1.24 (1.245) , -1.2 to 3.69 (0.318)	1 (0.575), -0.12 to 2.13 (0.081)	1.6 (0.708), 0.21 to 2.99 (0.024)	1.76 (0.726), 0.34 to 3.18 (0.015)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	6.63 (1.568), 3.55 to 9.7 (<0.001)	8.09 (2.975), 2.26 to 13.93 (0.007)	4.21 (1.242), 1.77 to 6.64 (0.001)	6.23 (1.531), 3.23 to 9.23 (<0.001)	6.05 (1.569), 2.98 to 9.13 (<0.001)
NE: Secundaria superior	13.04 (1.597), 9.91 to 16.17 (<0.001)	11.43 (3.021), 5.5 to 17.36 (<0.001)	7.91 (1.265), 5.43 to 10.39 (<0.001)	12.61 (1.559), 9.55 to 15.67 (<0.001)	12.15 (1.597), 9.02 to 15.28 (<0.001)
NE: Universitaria	25.35 (1.561), 22.28 to 28.41 (<0.001)	25.03 (2.951), 19.23 to 30.82 (<0.001)	16.88 (1.237), 14.45 to 19.3 (<0.001)	25.06 (1.524), 22.07 to 28.04 (<0.001)	24.48 (1.562), 21.42 to 27.54 (<0.001)
Medida (nº consecutivo)	17.11 (1.032), 15.09 to 19.14 (<0.001)	25.03 (1.819), 14.99 to 22.13(0.318)	12.03 (1.183), 9.71 to 14.34(<0.001)	16.47 (0.805), 14.9 to 18.05(<0.001)	17.26 (1.038), 15.22 to 19.29(<0.001)
Edad x Medida (nº consecutivo)	-0.23 (0.019), -0.27 to -0.19 (<0.001)	-0.27 (0.034), -0.33 to -0.2 (<0.001)	-0.15 (0.022), -0.19 to -0.1 (<0.001)	-0.23 (0.015), -0.26 to -0.2 (<0.001)	-0.23 (0.019), -0.27 to -0.2 (<0.001)
AIC	65103.7	21661.3*	65482.6	62548.6	65369.7
Varianza residual	14.3	14.5*	16.3	11.1	14.3

B: Coeficiente no estandarizado; DE: Desviación estándard; IC: Intervalo de confianza; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike

** : Se ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

Tabla 23. Resumen comparativo resultados de los modelos: nº de errores perseverantes (30%)

Tipo de análisis	Análisis original	Listwise	Media	Regresión	IM (5, PMM)
Modelo	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**	**Modelo 1**
Variables	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**	***B* (DE), IC95% (p-valor)**
Edad	-0.02 (0.023), -0.06 to 0.03 (0.491)	-0.06 (0.039), -0.13 to 0.02(0.153)	-0.013 (0.018), – to – (0.449)	-0.02 (0.022), -0.06 to 0.03 (0.421)	-0.01 (0.023), -0.05 to 0.04 (0.774)
Sexo: Hombre	Ref.	Ref.	Ref.	Ref.	Ref.
Sexo: Mujer	2.64 (0.451), 1.76 to 3.53 (<0.001)	2.13 (0.773), 0.61 to 3.65 (0.006)	1.755 (0.348), – to – (<0.001)	2.65 (0.44), 1.79 to 3.51 (<0.001)	2.82 (0.454), 1.93 to 3.71 (<0.001)
NE: Escuela primaria	Ref.	Ref.	Ref.	Ref.	Ref.
NE: Secundaria inicial	1.33 (0.974), -0.58 to 3.24 (0.171)	0.07 (1.649), -3.17 to 3.3 (0.969)	0.956 (0.752), – to – (0.204)	1.38 (0.952), -0.49 to 3.25 (0.147)	1.28 (0.982), -0.65 to 3.2 (0.193)
NE: Secundaria superior	-0.26 (0.992), -2.21 to 1.68 (0.792)	-1.8 (1.669), -5.08 to 1.47 (0.28)	-0.152 (0.766), – to – (0.843)	-0.33 (0.97), -2.23 to 1.57 (0.731)	-0.42 (1), -2.38 to 1.54 (0.675)
NE: Universitaria	-1.25 (0.97), -3.15 to 0.65 (0.196)	-3.06 (1.631), -6.26 to 0.14 (0.061)	0.749 (0.749), – to – (0.212)	-1.32 (0.948), -3.18 to 0.54 (0.164)	-1.42 (0.977), -3.33 to 0.5 (0.148)
AIC	60778.2	19935.6*	58464.5	57633.1	60576.1
Varianza residual	11.6	11.7*	10.5	10.3	11.9

B: Coeficiente no estandarizado; DE: Desviación estándard; IC: Intervalo de confianza; IM: Imputación Múltiple; PMM: Predictive Mean Matching o Equiparación de media predictiva; AIC: Criterio Akaike

** : Se ha de tener en cuenta que este caso presenta una **n** distinta al resto por lo que el valor del AIC no es comparable al resto de modelos.*

Figura 67. Correlaciones valores ajustados diseños únicos (Modelo 3) (10%)

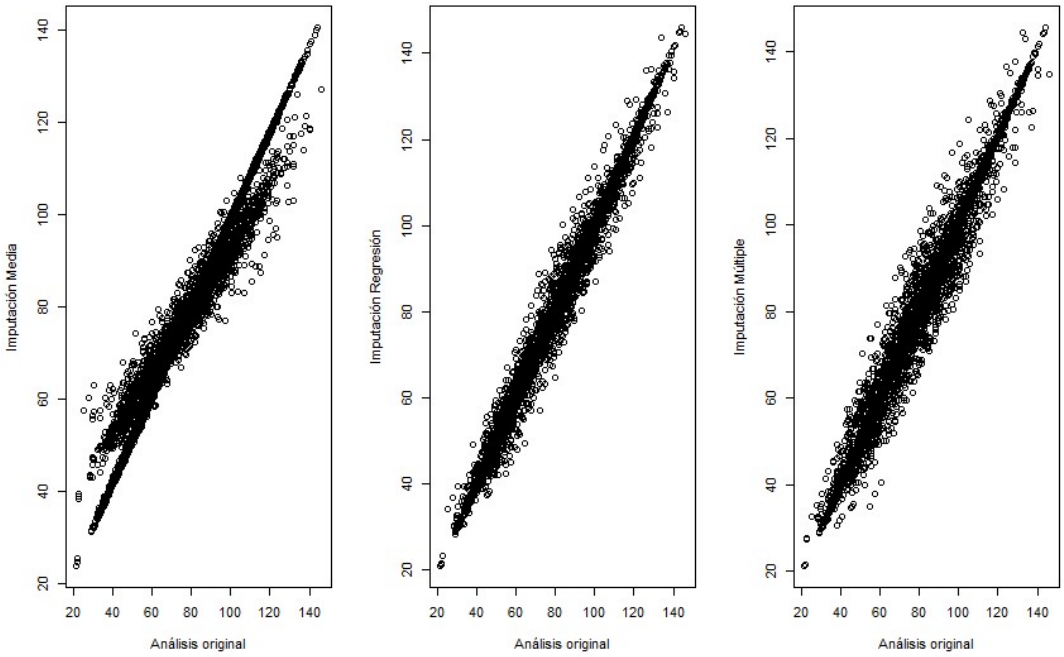


Figura 68. Correlaciones valores ajustados errores perseverantes (Modelo 1) (10%)

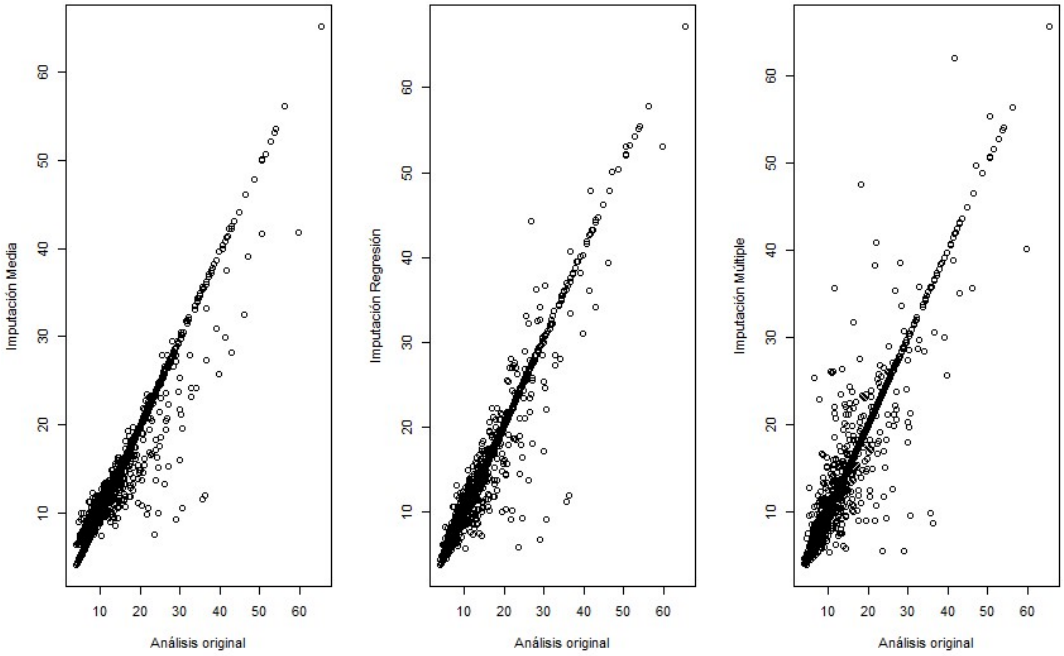


Figura 69. Correlaciones valores ajustados diseños únicos (Modelo 3) (20%)

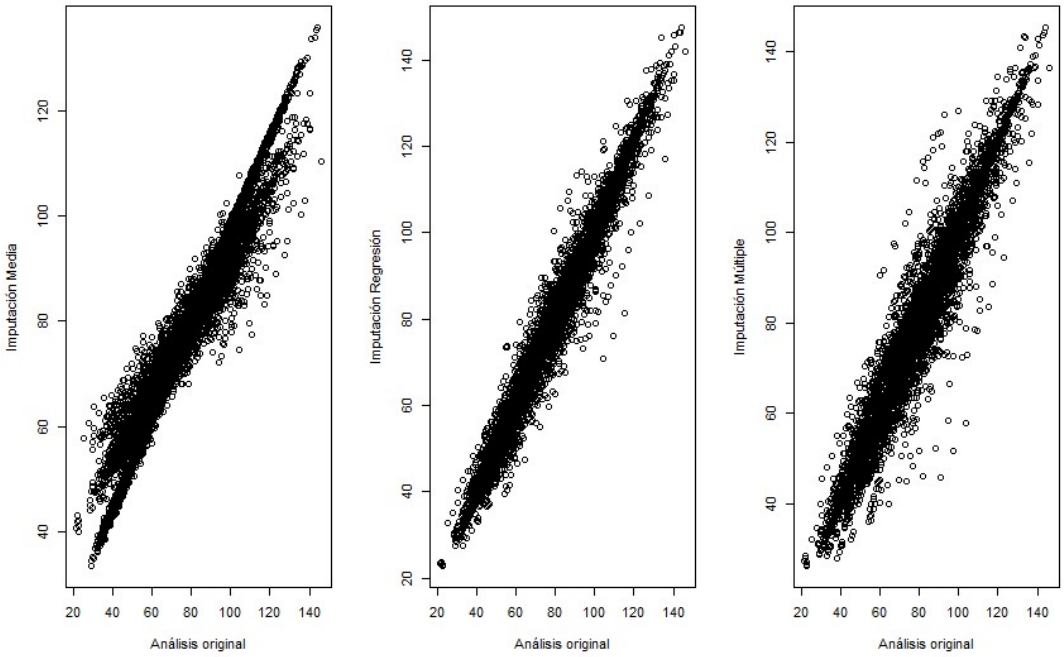


Figura 70. Correlaciones valores ajustados errores perseverantes (Modelo 1) (20%)

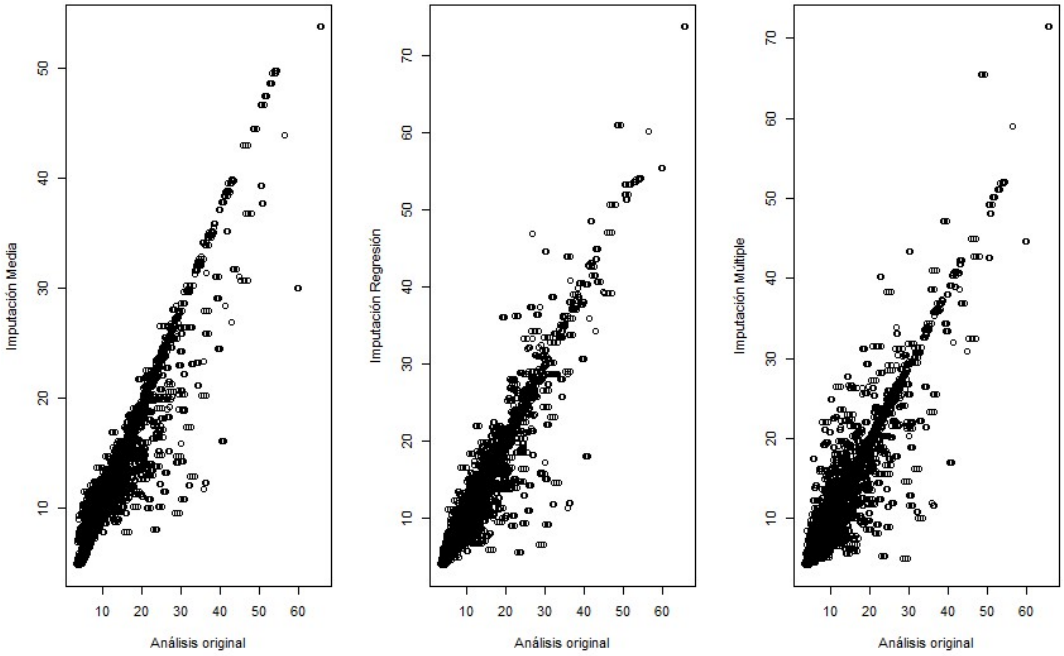


Figura 71. Correlaciones valores ajustados diseños únicos (Modelo 3) (30%)

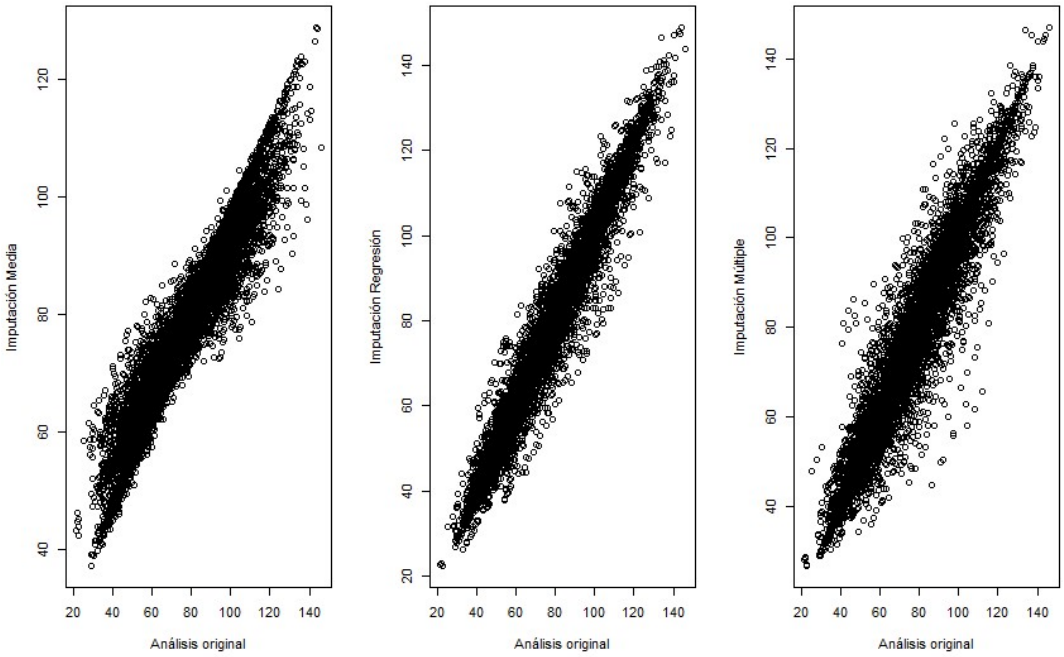


Figura 71. Correlaciones valores ajustados errores perseverantes (Modelo 1) (30%)

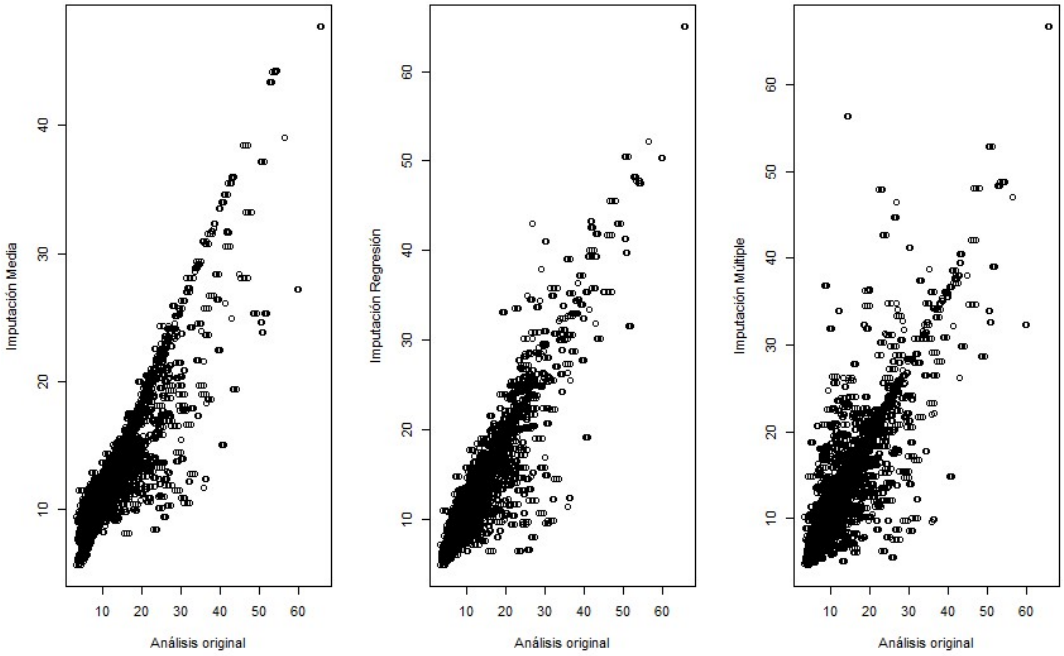


Tabla 24. Resumen correlaciones entre valores ajustados: Modelo Original vs Modelos con datos faltantes en Diseños únicos

Correlaciones	**Original vs Impt Media**	**Original vs Regresión**	**Original vs IM (5)**
Diseño	**Coef. Pearson, IC95%, (p-valor)**	**Coef. Pearson, IC95%, (p-valor)**	**Coef. Pearson, IC95%, (p-valor)**
Modelo 3: 10% datos faltantes	0.9838, 0.9831 to 0.9845, (<0.001)	0.9939, 0.9936 to 0.9941, (<0.001)	0.9881, 0.9876 to 0.9886, (<0.001)
Modelo 3: 20% datos faltantes	0.9681, 0.9666 to 0.9695, (<0.001)	0.9863, 0.9856 to 0.9869, (<0.001)	0.9718, 0.9706 to 0.9731, (<0.001)
Modelo 3: 30% datos faltantes	0.9501, 0.9479 to 0.9523, (<0.001)	0.9752, 0.974 to 0.9762, (<0.001)	0.9522, 0.9501 to 0.9543, (<0.001)

Coef Pearson: Coeficiente de correlación de Pearson; IC: Intervalo de confianza

Tabla 25. Resumen correlaciones entre valores ajustados: Modelo Original vs Modelos con datos faltantes en Errores Perseverantes

Correlaciones	**Original vs Impt Media**	**Original vs Regresión**	**Original vs IM (5)**
Diseño	**Coef. Pearson, IC95%, (p-valor)**	**Coef. Pearson, IC95%, (p-valor)**	**Coef. Pearson, IC95%, (p-valor)**
Modelo 3: 10% datos faltantes	0.9683, 0.9669 to 0.9697, (<0.001)	0.9721, 0.9708 to 0.9733, (<0.001)	0.9445, 0.9421 to 0.9469, (<0.001)
Modelo 3: 20% datos faltantes	0.9376, 0.9349 to 0.9403, (<0.001)	0.9432, 0.9406 to 0.9456, (<0.001)	0.9136, 0.9098 to 0.9173, (<0.001)
Modelo 3: 30% datos faltantes	0.9035, 0.8993 to 0.9076, (<0.001)	0.9179, 0.9143 to 0.9214, (<0.001)	0.8434, 0.8368 to 0.8498, (<0.001)

Coef Pearson: Coeficiente de correlación de Pearson; IC: Intervalo de confianza