UNIDAD DIDÁCTICA 1: Preprocesamiento de datos
Tema 2: Datos atípicos y datos ausentes

TÉCNICAS ESTADÍSTICAS PARA EL APRENDIZAJE I

Máster Universitario en Estadística Computacional
y Ciencia de Datos para la Toma de Decisiones

Índice

Datos faltantes

Introducción

Los datos ausentes son algo habitual en el análisis de datos.

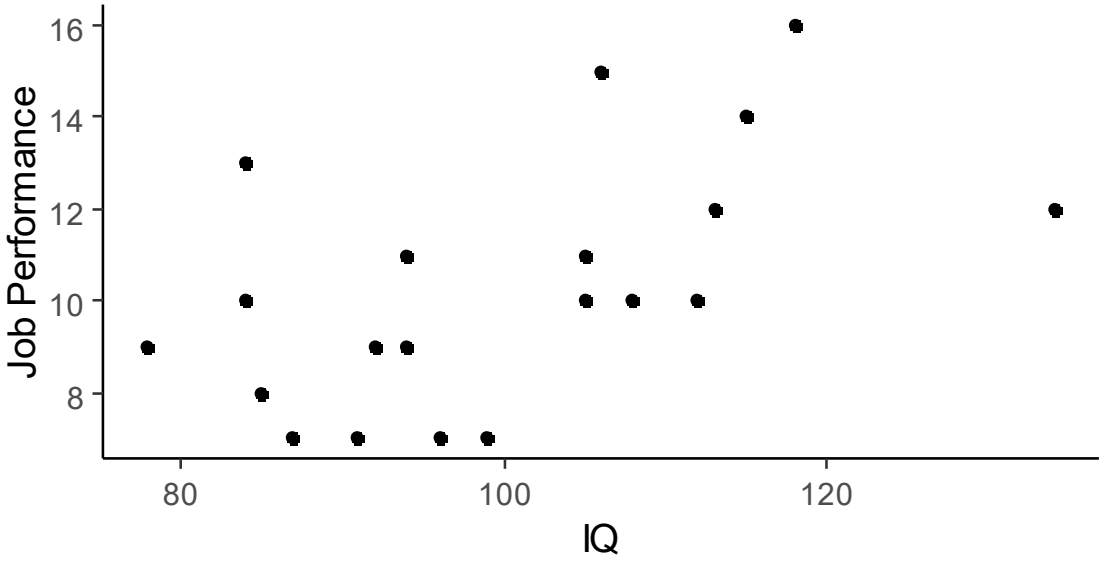
Nos planteamos cuestiones tales como:

- ¿Los datos faltantes se distribuyen aleatoriamente entre las observaciones o se pueden identificar distintas pautas?
- ¿En qué medida son relevantes los datos faltantes?

Introducción

Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Datos de Ejemplo.csv



Resumen de los datos faltantes: Resumen básico sobre datos faltantes

```
library(naniar)
```

- n_miss ()
- n_miss_row ()
- n_complete ()
- n_complete_row()
- prop_miss()
- prop_miss_case()
- prop_complete_case()

Resumen de los datos faltantes: Resumen del dataset sobre datos faltantes

Datos de Ejemplo.csv

library(naniar): miss_var_summary( ) y miss_var_table( )

```
# A tibble: 3 × 3
  variable                 n_miss      pct_miss
  <chr>                     <int>        <dbl>
1 job_perfMiss                10           50
2 IQ                           0            0
3 job_perf                     0            0
```

```
# A tibble: 2 × 3
  n_miss_in_var   n_vars      pct_vars
          <int>    <int>         <dbl>
1             0        2          66.7
2            10        1          33.3
```

Resumen de los datos faltantes: Resumen del dataset sobre datos faltantes

Datos de Ejemplo.csv

library(naniar):

miss_case_summary( ) y miss_case_table( )

```
# A tibble: 20 × 3
    case n_miss pct_miss
   <int>  <int>    <dbl>
 1     1      1     33.3
 2     2      1     33.3
 3     3      1     33.3
 4     4      1     33.3
 5     5      1     33.3
 6     6      1     33.3
 7     7      1     33.3
 8     8      1     33.3
 9     9      1     33.3
10    10      1     33.3
11    11      0      0
12    12      0      0
13    13      0      0
14    ……
```

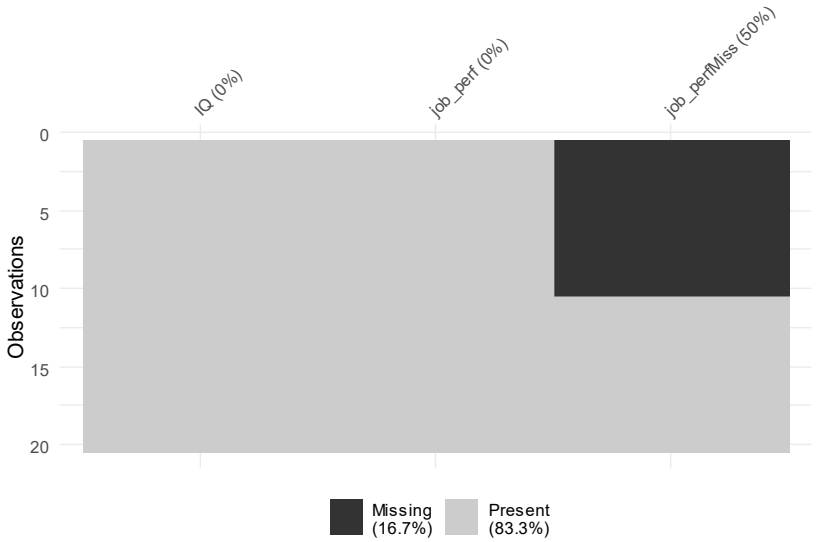
```
# A tibble: 2 × 3
  n_miss_in_case n_cases pct_cases
           <int>   <int>     <dbl>
1              0      10        50
2              1      10        50
```

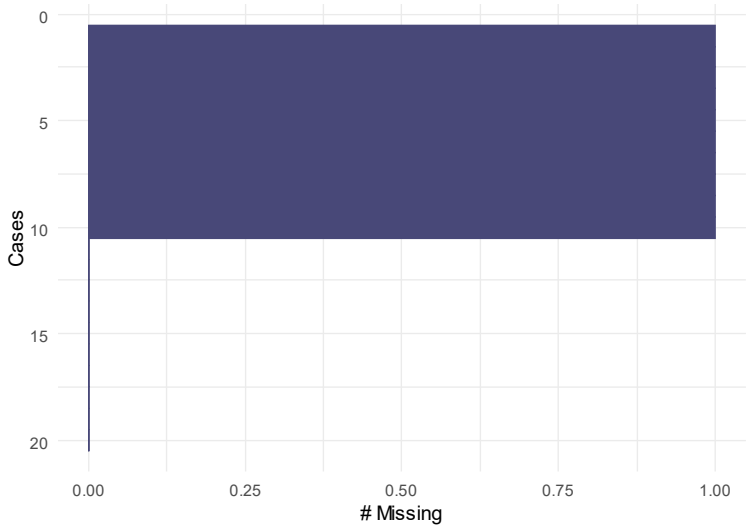
Resumen de los datos faltantes: Visualización de los valores perdidos

La visualización de los datos puede ayudar al investigador a hacerse una idea de como es el comportamiento de los valores perdidos.

library(naniar): vis_miss( ), gg_miss_case()

Datos de Ejemplo.csv

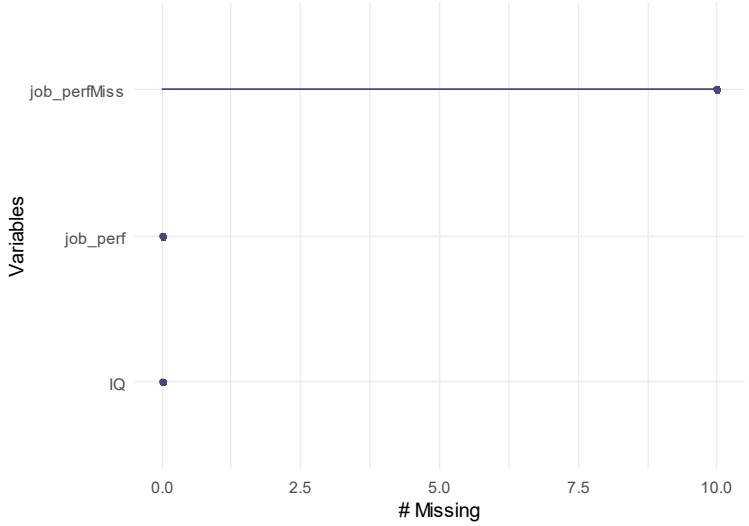




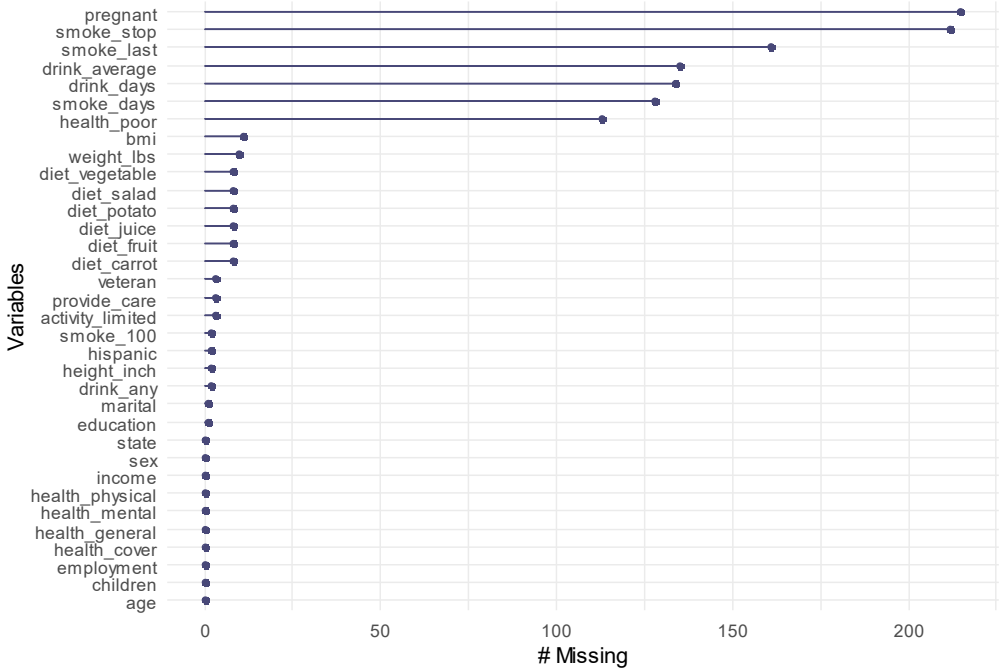
Resumen de los datos faltantes: Visualización de los valores perdidos

library(naniar): gg_miss_var (), gg_miss_var_cumsum()

Datos de Ejemplo.csv



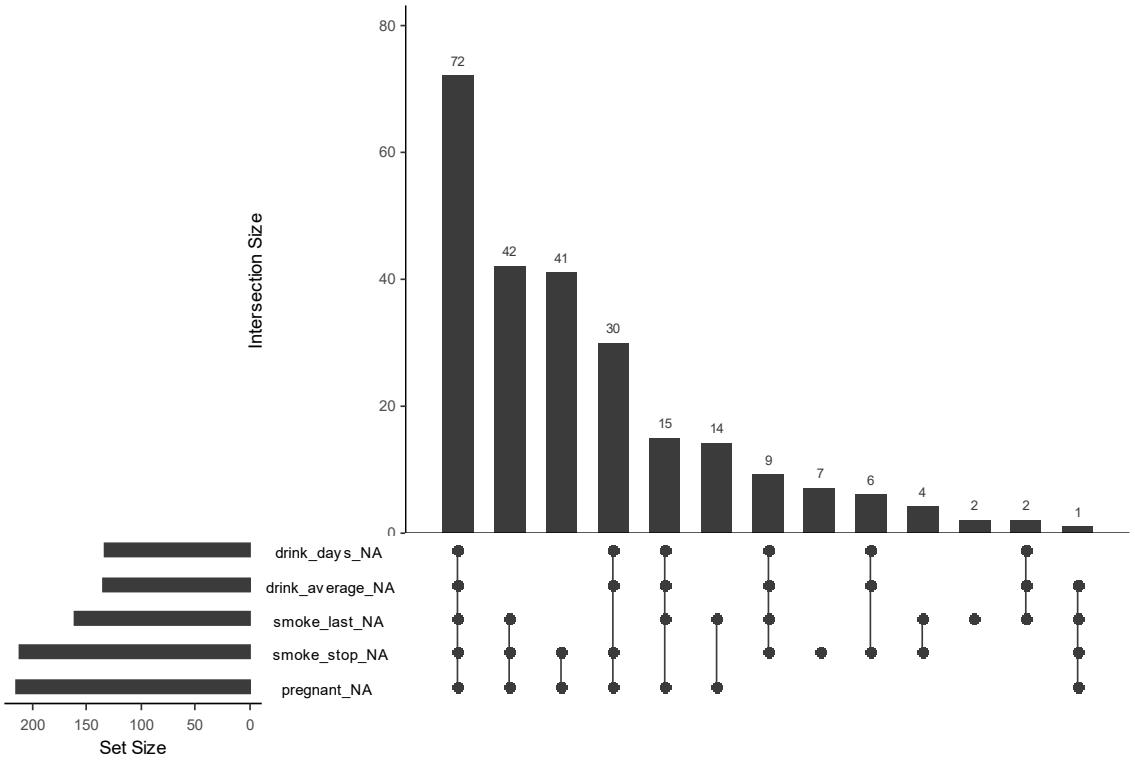
Datos: riskfactor



Resumen de los datos faltantes: Visualización de los valores perdidos

La visualización de patrones entre variables. Base de datos riskfactors

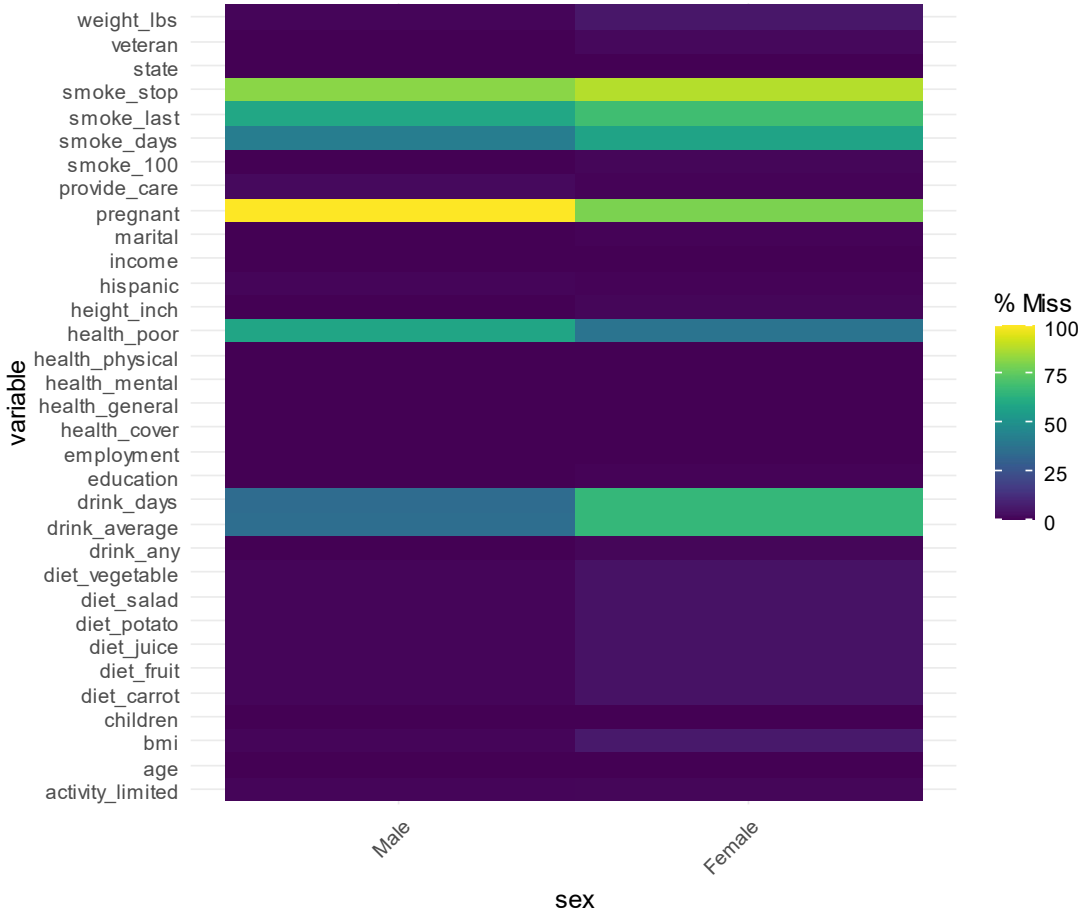
library(naniar): gg_miss_upset( ) .Para poder interpretar este gráfico se necesitan al menos dos variables. En este caso, hacemos uso de la base de datos “riskfactors”.



Resumen de los datos faltantes: Visualización de los valores perdidos

La visualización de patrones. Base de datos riskfactors

library(naniar): gg_miss_fct( )



Tipos de datos faltantes

Para decidir la solución de los datos faltantes, se debe averiguar el grado de aleatoriedad presente en los datos ausentes.

- **Datos ausentes prescindibles:**
  - Bajo el control del investigador, pueden ser identificados explícitamente.
  - No se necesitan soluciones específicas para la ausencia de datos.

Tipos de datos faltantes

- **Datos ausentes no prescindibles:**

No se encuentran bajo el control del investigador, no pueden ser identificados explícitamente:

- Errores en la entrada de datos
- Renuncia del encuestado a responder
- Respuestas inaplicables.

En estos casos se debe analizar si existen o no patrones sistemáticos que puedan sesgar los resultados obtenidos. Conviene analizar el grado de aleatoriedad presente en los mismos.

Según el grado de aleatoriedad, los datos ausentes se puede clasificar del siguiente modo:

- Missing Completely at Random (MCAR) (Faltante totalmente al azar)
- Missing not at Random (NMAR) (Faltante no al azar)
- Missing at Random (MAR) (Faltante al azar)

Tipos de datos faltantes

Missing Completely at Random (MCAR) (Faltante totalmente al azar)

La probabilidad de que falten datos en una variable no está relacionada con el valor de la misma o con los valores de cualquier otra variable del conjunto de datos.
Los puntos faltantes son un subconjunto aleatorio de los datos. No hay nada sistemático que haga que algunos datos tengan más probabilidades de faltar que otros.

$$P(Y_{missing}|Y,X) = P(Y_{missing})$$

Ejemplo:
En el conjunto de datos observamos que los datos faltantes aparecen tanto en la categoría A como en la B o en la C, y los valores faltantes pueden ser altos o bajos. Esto quiere decir que esos datos faltantes no dependen ni de la categoría ni del valor mismo de los datos, por lo que podemos decir que los datos faltantes de este ejemplo son MCAR.

Tipos de datos faltantes

Missing not at Random (NMAR) (Faltantes no al azar)

Datos ausentes no aleatorios, existen patrones sistemáticos en el proceso de datos ausentes y habría que evaluar la magnitud del problema calibrando, en particular, el tamaño de los sesgos.

Ejemplo:
Podemos ver que sistemáticamente los datos con valores más bajos faltan, tanto para las categorías A, B como C. Es decir que los valores faltantes dependen de la variable "V2", y por tanto la falta de datos en este caso NO es aleatoria

Tipos de datos faltantes

Missing at Random (MAR) (faltante al azar)

- Los datos ausentes obedecen a un proceso aleatorio (MAR) si los valores ausentes de Y dependen de X, pero no de Y.

- Las observaciones faltantes están condicionadas por otras variables explicativas en el conjunto de datos, aunque no con la variable respuesta

$$P(Y_{missing}|Y,X) = P(Y_{missing}|X)$$

Ejemplo:
Los datos faltantes corresponden únicamente a datos en la categoría B, y que estos datos faltantes van desde los más pequeños a los más grandes. Esto quiere decir que los valores faltantes dependen sólo de la variable "V1" (la categoría) y no de la propia variable "V2".

	V2	
V1	Valor real	MAR
A	85	85
A	94	94
A	111	111
A	130	130
B	80	?
B	97	?
B	117	?
B	125	?
C	88	88
C	91	91
C	123	123
C	132	132

Datos faltantes: Diagnóstico de aleatoriedad

Test de Little:

Realizar un test conjunto de aleatoriedad que determine si los datos ausentes pueden ser clasificados como MCAR.

Se suele utilizar **la prueba de Little** recogido en el trabajo: Little, Roderick J. A. 1988. "A Test of Missing Completely at Random for Multivariate Data with Missing Values." Journal of the American Statistical Association 83 (404). https://www.jstor.org/stable/2290157.

Si el test es **no significativo**, los datos ausentes pueden ser clasificados como MCAR.

Tipos de datos faltantes

Ejemplo: Con los datos de ejemplo.csv

bind_shadow(): Vinculación de columnas sombra con la indicación de dato faltante

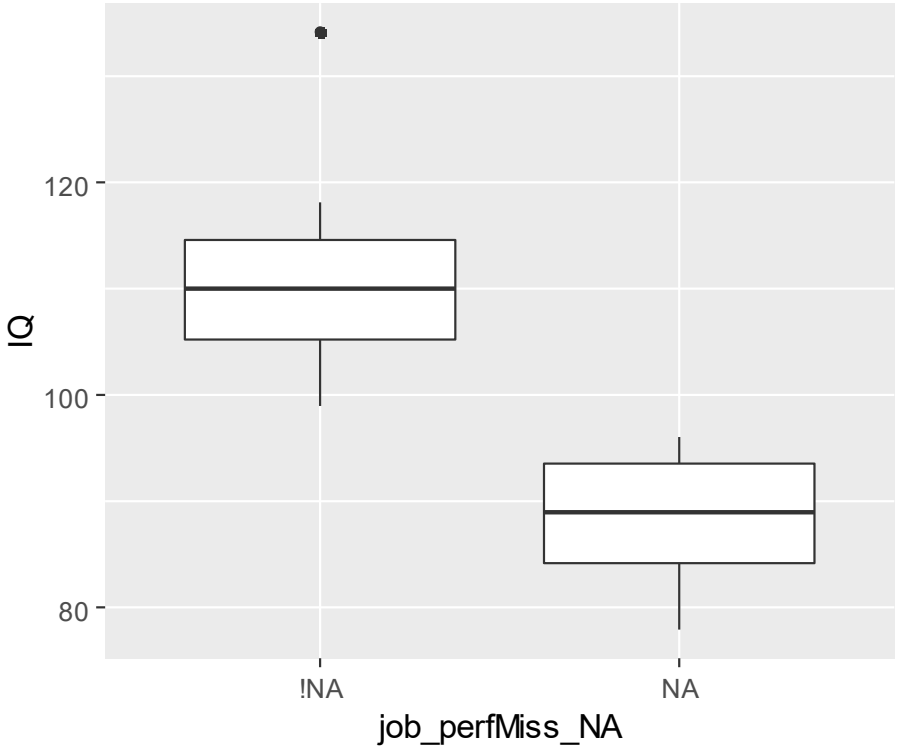
```
# A tibble: 20 x 4
      IQ job_perf job_perfMiss job_perfMiss_NA
   <int>    <int>        <int> <fct>
 1    78        9           NA NA
 2    84       13           NA NA
 3    84       10           NA NA
 4    85        8           NA NA
 5    87        7           NA NA
 6    91        7           NA NA
 7    92        9           NA NA
 8    94        9           NA NA
 9    94       11           NA NA
10    96        7           NA NA
11    99        7            7 !NA
12   105       10           10 !NA
13   105       11           11 !NA
14   106       15           15 !NA
15   108       10           10 !NA
```

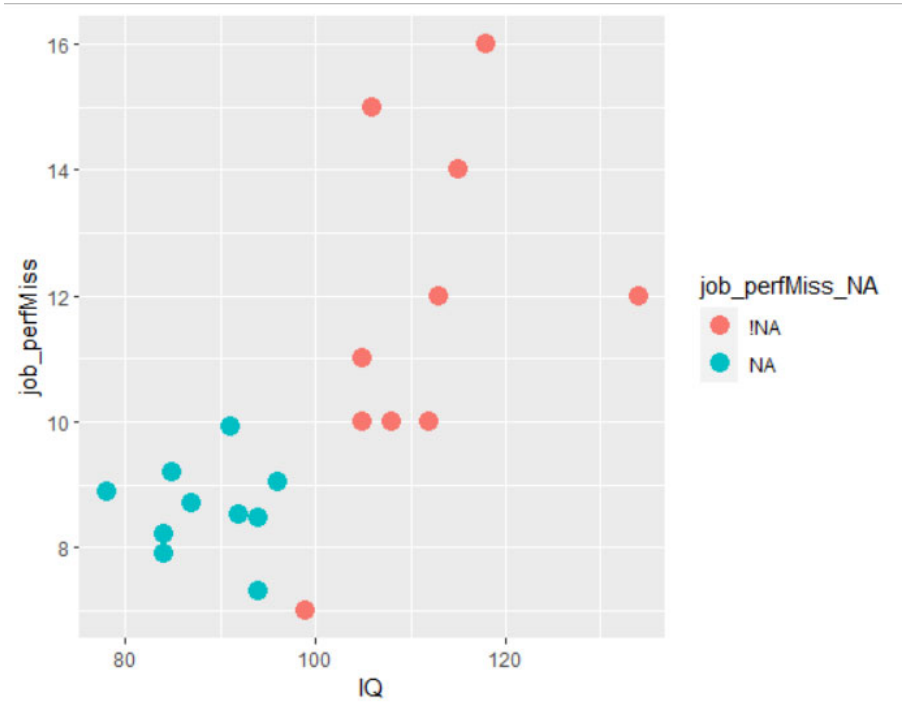
```
# A tibble: 2 x 2
  job_perfMiss_NA  mean
  <fct>           <dbl>
1 !NA              112.
2 NA                88.5
```

Media de IQ, de los valores no faltantes y de los faltantes

Ejemplo: Con los datos de ejemplo.csv

bind_shadow(): Vinculación de columnas sombra con la indicación de dato faltante





Tipos de datos faltantes

Realizamos el test de aleatoriedad a los datos de ejemplo.csv:

El test implementado se puede consultar en:

https://journals.sagepub.com/doi/pdf/10.1177/1536867X1301300407

mcar_test( )

```
# A tibble: 1 × 4
  statistic    df      p.value   missing.patterns
      <dbl> <dbl>      <dbl>              <int>
1      14.9     2      0.000592               2
```

En el ejemplo, **el test es significativo**, por tanto, los datos ausentes no pueden ser clasificados como MCAR

Soluciones a los datos faltantes

Soluciones a los datos faltantes

1.- Utilizar sólo aquellas observaciones con datos completos (Listwise Deletion)

Suele ser el método por defecto.

Esta aproximación debería usarse sólo si los datos ausentes son **MCAR,** porque los datos ausentes que no lo son tienen elementos no aleatorios que sesgarían los resultados.

No utiliza toda la información.

Reduce la potencia estadística (porque disminuye la n).

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

Soluciones a los datos faltantes

Ejemplo: con los datos de ejemplo.csv
na.omit()

Complete data		Missing data
IQ	Job performance	Job Performance
78	9	—
84	13	—
84	10	—
85	8	—
87	7	—
91	7	—
92	9	—
94	9	—
94	11	—
96	7	—
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

	IQ	job_perf	job_perfMiss
11	99	7	7
12	105	10	10
13	105	11	11
14	106	15	15
15	108	10	10
16	112	10	10
17	113	12	12
18	115	14	14
19	118	16	16
20	134	12	12

Soluciones a los datos faltantes

2.- Supresión de caso(s) y/o variable(s) (Pairwise Deletion)

Suprimir el caso(s) y/o variable(s) que peor se comporta(n) respecto a los datos ausentes.

Ventaja:
Utiliza toda la información posible con cada análisis. Mantiene el mayor número posible de casos para cada análisis.

Desventaja:
No se pueden comparar los análisis porque la muestra es diferente cada vez. El error estándar calculado por la mayoría de los softwares utiliza el tamaño medio de la muestra en todos los análisis. Esto tiende a producir errores estándar subestimados o sobreestimados.

Gender	8th grade math test score	12th grade math score
F	45	.
M	.	99
F	55	86
F	85	88
F	80	75
.	81	82
F	75	80
M	95	.
M	86	90
F	70	75
F	85	.

3.- Métodos de imputación

- Tratar con datos ausentes mediante uno de los muchos métodos de imputación.
- La imputación es el proceso de estimación de valores ausentes basado en valores válidos de otras variables y/o casos de la muestra.
- El objetivo es emplear relaciones conocidas que puedan identificarse en los valores válidos de la muestra para ayudar en la estimación de valores ausentes.

La imputación **se debe realizar con precaución**:

- La imputación sólo puede aplicarse correctamente a una pequeña gama de problemas.
- Si hay datos que faltan en y (variable dependiente), es probable que no se pueda realizar ninguna imputación de forma adecuada.
- Si tiene cierto tipo de datos perdidos (por ejemplo, datos perdidos no aleatorios) en las variables independientes, entonces se podría corregir con la imputación.

Soluciones a los datos faltantes

3.1.- Sustitución por la media:

Sustituir los valores ausentes por el valor medio que se calcula sobre todas las respuestas válidas.

Ventajas:
Proceso sencillo que proporciona una información completa para todos los casos.

Desventajas:
Puede producir estimaciones sesgadas de los parámetros (por ejemplo, las varianzas).
La distribución real de los valores se encuentra distorsionada por la sustitución por la media.
Este método modifica la correlación observada porque todos los datos ausentes tendrán un valor único constante.

Soluciones a los datos faltantes

3.1.- Sustitución por la media:

library(naniar):

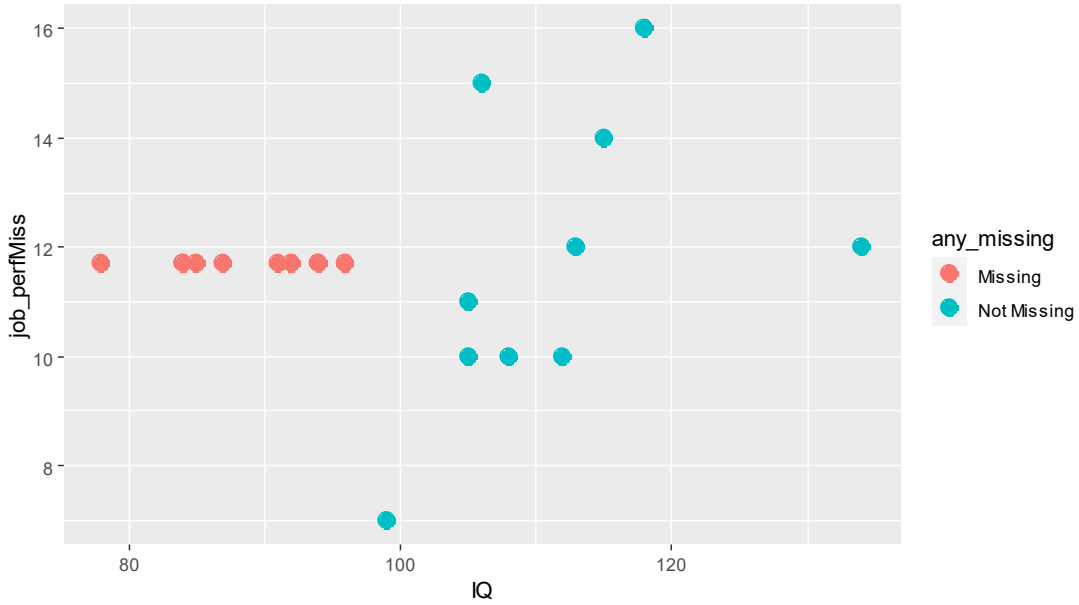
- bind_shadow( ): creación de una matriz sombra para identificar los valores faltantes
- impute_mean_all(): imputar los valores faltantes con la media
- add_label_shadow(): añadir etiquetas a los valores que se han imputado

Ejemplo: Con los datos de ejemplo.csv realiza la imputación por la media en los datos faltantes

```
# A tibble: 6 x 5
     IQ job_perf job_perfMiss job_perfMiss_NA any_missing
  <dbl>    <dbl>        <dbl> <fct>           <chr>
1    78        9         11.7 NA              Missing
2    84       13         11.7 NA              Missing
3    84       10         11.7 NA              Missing
4    85        8         11.7 NA              Missing
5    87        7         11.7 NA              Missing
6    91        7         11.7 NA              Missing
```

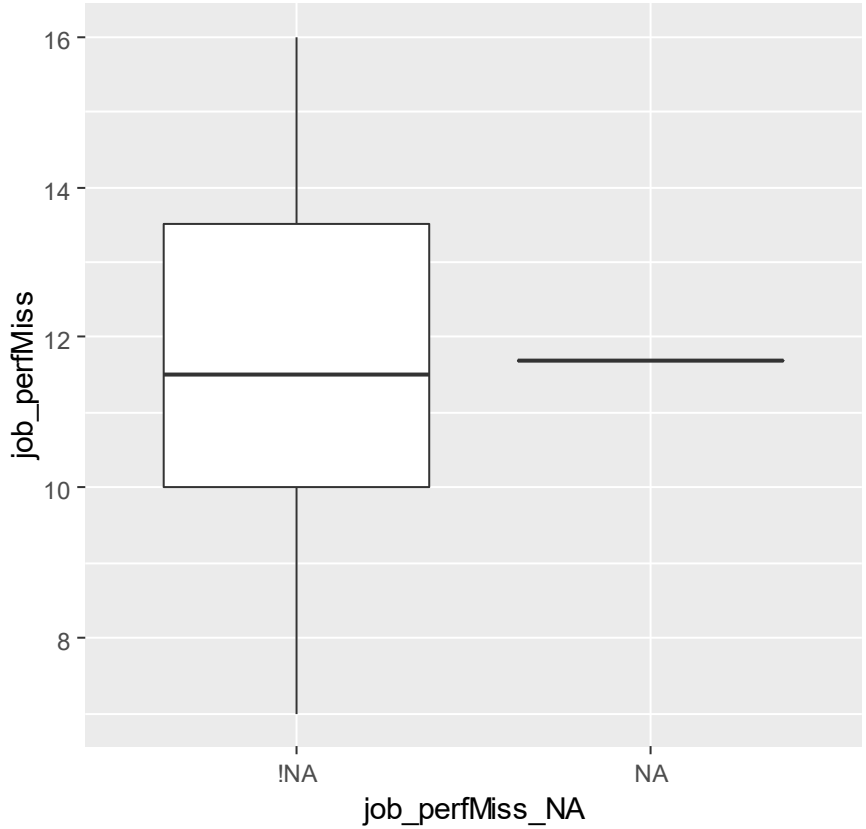
Soluciones a los datos faltantes

Exploramos la imputación:



Soluciones a los datos faltantes

Exploramos la imputación:



3.2.- Imputación por regresión:

También conocida como imputación de la media condicional. Se usa el análisis de regresión para predecir los valores ausentes de una variable.

Ventajas:
Método prometedor en aquellos casos donde las relaciones entre las variables están lo suficientemente establecidas.

Desventajas:
Refuerza las relaciones ya existentes en los datos.
Se subestima la varianza de la distribución.
Los valores predichos pueden no corresponder a rangos válidos (por ejemplo, predecir un valor de 11 para una escala de 10 puntos).

Soluciones a los datos faltantes

Ejemplo: library(simputation)

Con los datos ejemplo.csv, utilizamos los 10 casos completos para estimar la regresión de las puntuaciones de rendimiento laboral sobre el IQ.

lm(data = ejemplo, job_perfMiss ~ IQ)

```
Call:
lm(formula = job_perfMiss ~ IQ, data = ejemplo)

Coefficients:
(Intercept)           IQ
    -2.0646       0.1234
```

$$JP_i^* = \hat{\beta}_0 + \hat{\beta}_1(IQ_i) = -2.065 + 0.123(IQ_i)$$

Soluciones a los datos faltantes

library(simputation)

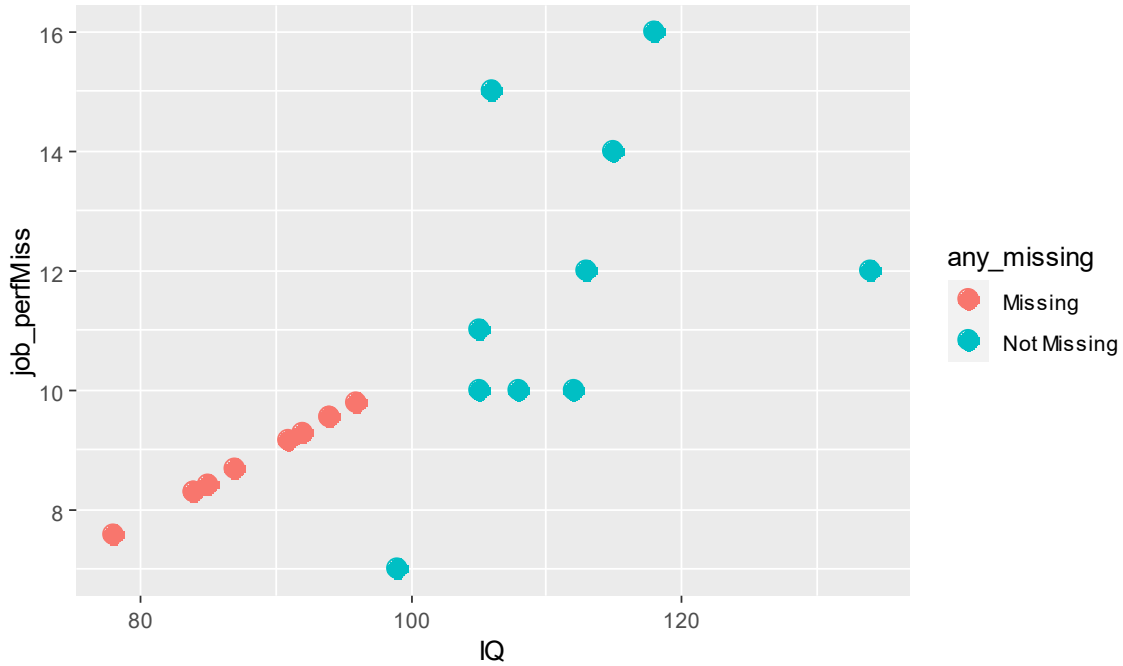
impute_lm (): imputar valores usando un modelo lineal a través de la función

```
# A tibble: 20 × 7
      IQ job_perf job_perfMiss IQ_NA job_perf_NA job_perfMiss_NA any_missing
  * <dbl>    <dbl>        <dbl> <fct> <fct>      <fct>           <chr>
 1    78        9         7.56 !NA   !NA        NA              Missing
 2    84       13         8.31 !NA   !NA        NA              Missing
 3    84       10         8.31 !NA   !NA        NA              Missing
 4    85        8         8.43 !NA   !NA        NA              Missing
 5    87        7         8.68 !NA   !NA        NA              Missing
 6    91        7         9.17 !NA   !NA        NA              Missing
 7    92        9         9.29 !NA   !NA        NA              Missing
 8    94        9         9.54 !NA   !NA        NA              Missing
 9    94       11         9.54 !NA   !NA        NA              Missing
10    96        7         9.79 !NA   !NA        NA              Missing
11    99        7         7    !NA   !NA        !NA             Not Missing
12   105       10        10    !NA   !NA        !NA             Not Missing
13   105       11        11    !NA   !NA        !NA             Not Missing
14   106       15        15    !NA   !NA        !NA             Not Missing
15   108       10        10    !NA   !NA        !NA             Not Missing
16   112       10        10    !NA   !NA        !NA             Not Missing
17   113       12        12    !NA   !NA        !NA             Not Missing
18   115       14        14    !NA   !NA        !NA             Not Missing
19   118       16        16    !NA   !NA        !NA             Not Missing
20   134       12        12    !NA   !NA        !NA             Not Missing
```

datos de ejemplo.csv

Soluciones a los datos faltantes

Estas puntuaciones predichas completan las calificaciones de rendimiento laboral que faltan y sirven como datos para todos los análisis posteriores.



3.4.- Stochastic Regression Imputation:

Utiliza ecuaciones de regresión para predecir los valores faltantes pero se introduce un término residual normalmente distribuido.

La adición de los residuos a los valores imputados restablece la variabilidad perdida de los datos y elimina eficazmente los sesgos asociados a los esquemas de imputación de regresión estándar.

Para ilustrar el proceso de imputación, reconsideremos el conjunto de datos del ejemplo. La ecuación de regresión de imputación es la siguiente:

$$JP_i^* = \hat{\beta}_0 + \hat{\beta}_1(IQ_i) = -2.065 + 0.123(IQ_i) + z_i$$

Soluciones a los datos faltantes

$$JP_i^* = \hat{\beta}_0 + \hat{\beta}_1(IQ_i) = -2.065 + 0.123(IQ_i) + z_i$$

Los coeficientes de regresión de la ecuación son idénticos a los vistos en el método anterior.

Sin embargo, esta ecuación tiene un término z_i adicional. Este término residual es un valor aleatorio de una distribución normal con una media de cero y una varianza igual a la varianza residual de la regresión del rendimiento laboral sobre la variable IQ.

El análisis de regresión del caso completo produjo una estimación de la varianza residual de $\hat{\sigma}^2_{JP|IQ} = 6{,}650$.

- Generamos 10 puntuaciones a partir de una distribución normal con una media de cero y una varianza de 6,650, mediante técnicas de simulación (Random residual).
- Sumando los residuos a las puntuaciones predichas se obtienen los valores de la **Imputación estocástica**.
- Estas puntuaciones completan las calificaciones de rendimiento laboral que faltan y dan lugar a un conjunto de datos completo.

Soluciones a los datos faltantes

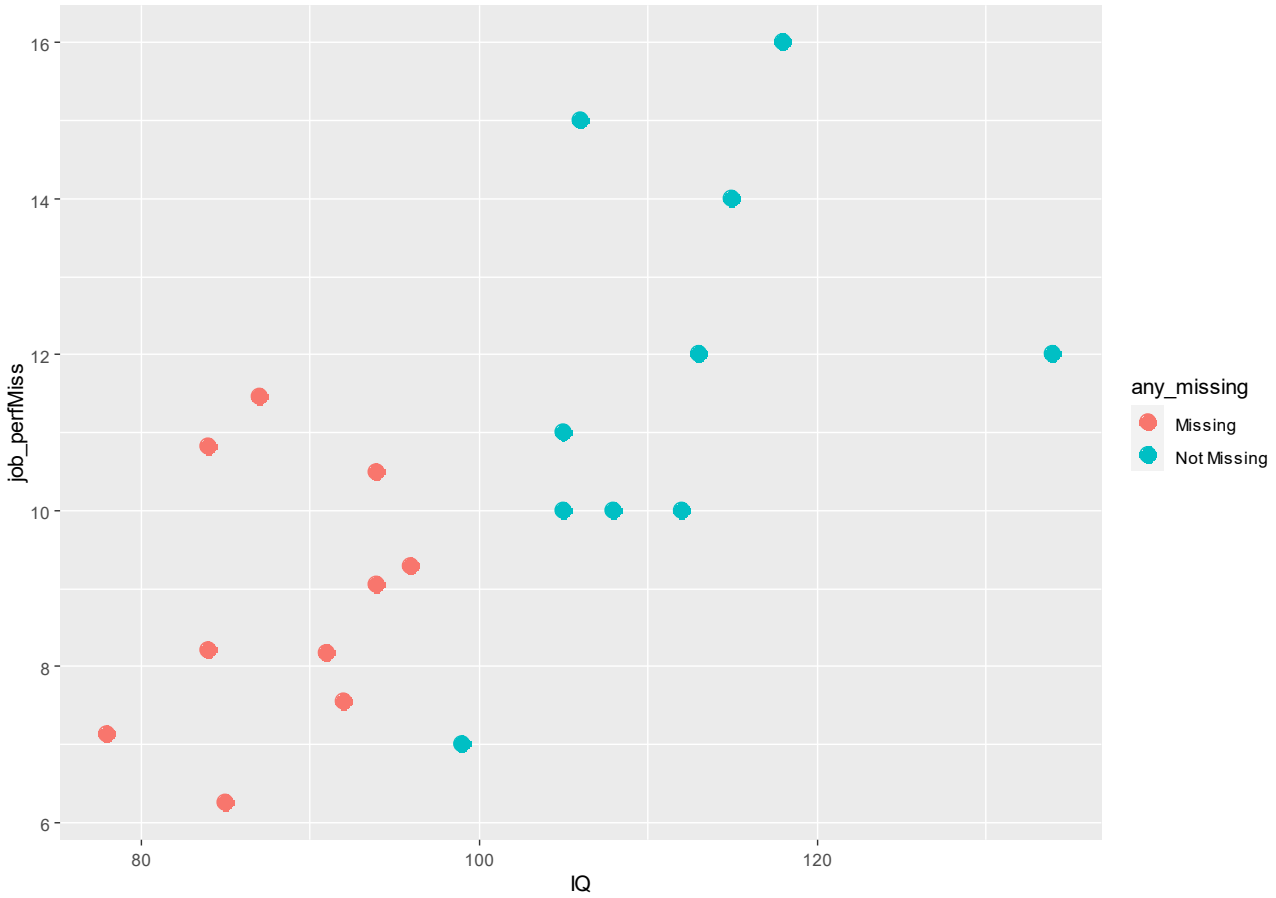
Ejemplo: Con los datos de ejemplo.csv realiza la imputación por regresión estocástica en los datos faltantes

impute_lm(job_perfMiss ~ IQ, add_residual = "normal")

```
# A tibble: 20 x 7
      IQ job_perf job_perfMiss IQ_NA job_perf_NA job_perfMiss_NA any_missing
   * <dbl>    <dbl>        <dbl> <fct> <fct>      <fct>           <chr>
 1    78        9         7.13  !NA   !NA        NA              Missing
 2    84       13         8.21  !NA   !NA        NA              Missing
 3    84       10        10.8   !NA   !NA        NA              Missing
 4    85        8         6.26  !NA   !NA        NA              Missing
 5    87        7        11.5   !NA   !NA        NA              Missing
 6    91        7         8.16  !NA   !NA        NA              Missing
 7    92        9         7.56  !NA   !NA        NA              Missing
 8    94        9         9.04  !NA   !NA        NA              Missing
 9    94       11        10.5   !NA   !NA        NA              Missing
10    96        7         9.28  !NA   !NA        NA              Missing
11    99        7         7     !NA   !NA        !NA             Not Missing
12   105       10        10     !NA   !NA        !NA             Not Missing
13   105       11        11     !NA   !NA        !NA             Not Missing
14   106       15        15     !NA   !NA        !NA             Not Missing
15   108       10        10     !NA   !NA        !NA             Not Missing
16   112       10        10     !NA   !NA        !NA             Not Missing
17   113       12        12     !NA   !NA        !NA             Not Missing
18   115       14        14     !NA   !NA        !NA             Not Missing
19   118       16        16     !NA   !NA        !NA             Not Missing
20   134       12        12     !NA   !NA        !NA             Not Missing
```

Soluciones a los datos faltantes

$$JP_i^* = \hat{\beta}_0 + \hat{\beta}_1(IQ_i) = -2.065 + 0.123(IQ_i) + z_i$$

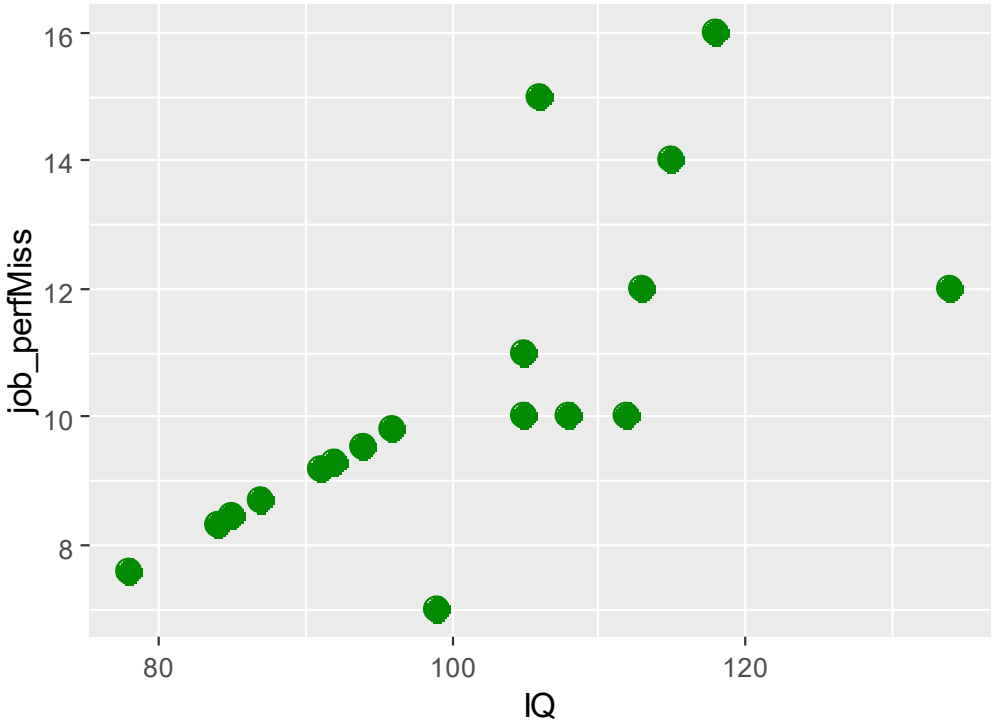


Ejemplo: Imputación por regresión lineal con librería mice

Con los datos de ejemplo.csv realiza la imputación por regresión en los datos faltantes

mice(method = "norm.predict", m = 1, maxit = 1)

	IQ	job_perfMiss
1	78	7.564442
2	84	8.305139
3	84	8.305139
4	85	8.428588
5	87	8.675487
6	91	9.169285
7	92	9.292735
8	94	9.539634
9	94	9.539634
10	96	9.786533
11	99	7.000000
12	105	10.000000
13	105	11.000000
14	106	15.000000
15	108	10.000000
16	112	10.000000
17	113	12.000000
18	115	14.000000
19	118	16.000000
20	134	12.000000

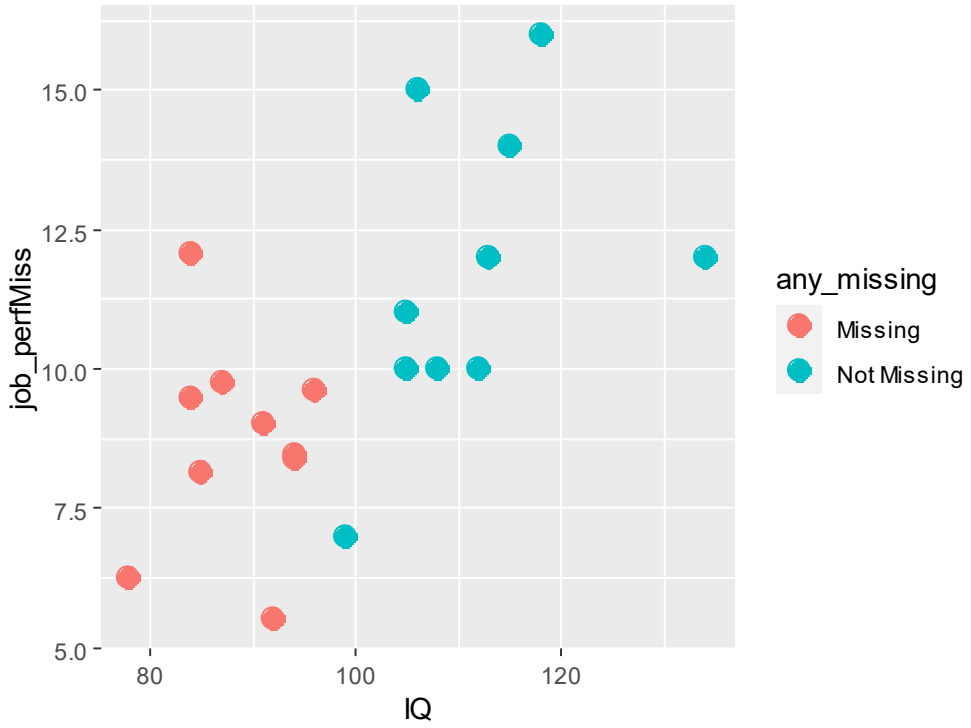


Ejemplo: Imputación por regresión lineal estocástica con librería mice

Con los datos de ejemplo.csv realiza la imputación por regresión estocástica en los datos faltantes

mice("norm.nob", m = 1, maxit = 1)

```
 iter imp variable
  1   1  job_perfMiss
     IQ job_perfMiss
1    78     6.246279
2    84     9.457319
3    84    12.050743
4    85     8.145223
5    87     9.744219
6    91     9.020956
7    92     5.496460
8    94     8.395578
9    94     8.452656
10   96     9.623018
11   99     7.000000
12  105    10.000000
13  105    11.000000
14  106    15.000000
15  108    10.000000
16  112    10.000000
17  113    12.000000
18  115    14.000000
19  118    16.000000
20  134    12.000000
```



3.5.- Hot-deck Imputation

- Estadísticos de la oficina Census Bureau desarrollaron originalmente la imputación **Hot-deck**, el procedimiento tiene una larga historia en encuestas (Scheuren, 2005).
- **Imputación con puntuaciones "similares".**
- Este método es un **procedimiento de duplicación**. Cuando falta información en un registro se duplica un valor ya existente en la muestra para reemplazarlo. Las unidades muestrales se clasifican en grupos de forma que sean lo más homogéneas posible dentro de los grupos. A cada valor que falte, se le asigna un valor del mismo grupo. Se está suponiendo que dentro de cada grupo la no respuesta sigue la misma distribución que los que responden.
- Este método suele **preservar las distribuciones univariantes** de los datos y **no disminuye la variabilidad**.
- Método **no adecuado para estimar medidas de asociación** y pueden producir estimaciones sustancialmente sesgadas de correlaciones y coeficientes de regresión (Brown, 1994; Schafer y Graham, 2002).

Soluciones a los datos faltantes

Ejemplo: Utilizamos el dataset "retailers" de la librería validate

```
library(simputacion) #librería con funciones de imputación
library(validate)#contiene la base de datos "retailers"
data(retailers)
```

```
> head(retailers, 10)
   size incl.prob staff turnover other.rev total.rev staff.costs total.costs profit  vat
1   sc0      0.02    75       NA        NA      1130          NA       18915  20045   NA
2   sc3      0.14     9     1607        NA      1607         131        1544     63   NA
3   sc3      0.14    NA     6886       -33      6919         324        6493    426   NA
4   sc3      0.14    NA     3861        13      3874         290        3600    274   NA
5   sc3      0.14    NA       NA        37      5602         314        5530     72   NA
6   sc0      0.02     1       25        NA        25          NA          22      3   NA
7   sc3      0.14     5       NA        NA      1335         135         136      1 1346
8   sc1      0.02     3      404        13       417          NA         342     75   NA
9   sc3      0.14     6     2596        NA      2596         147        2486    110   NA
10  sc2      0.05     5       NA        NA        NA          NA          NA     NA   NA
```

Soluciones a los datos faltantes

Ejemplo: Utilizamos el dataset “retailers” de la libería validate

```
ret1_hd<- impute_rhd(retailers, turnover + other.rev + total.rev ~ size )
```

Con el paquete **simputacion**, las celdas de imputación se determinan por el lado derecho de la fórmula que especifica el modelo

```
> head(ret1_hd, 10)
   size incl.prob staff turnover other.rev total.rev staff.costs total.costs profit  vat
1   sc0      0.02    75      359         9      1130          NA       18915  20045   NA
2   sc3      0.14     9     1607     98350      1607         131        1544     63   NA
3   sc3      0.14    NA     6886       -33      6919         324        6493    426   NA
4   sc3      0.14    NA     3861        13      3874         290        3600    274   NA
5   sc3      0.14    NA     2649        37      5602         314        5530     72   NA
6   sc0      0.02     1       25       622        25          NA          22      3   NA
7   sc3      0.14     5     4445        20      1335         135         136      1 1346
8   sc1      0.02     3      404        13       417          NA         342     75   NA
9   sc3      0.14     6     2596        32      2596         147        2486    110   NA
10  sc2      0.05     5     1175         4       206          NA          NA     NA   NA
```

Soluciones a los datos faltantes

Ejemplo:

En el hot deck secuencial, se ordena el conjunto de datos utilizando una o más variables, y los valores perdidos en un registro se toman del primer registro anterior o posterior que tenga un valor.

```
ret1_shd <- impute_shd(retailers, turnover ~ size + profit)
```

```
> head(ret1_shd)
  size incl.prob staff turnover other.rev total.rev staff.costs total.costs profit vat
1  sc0      0.02    75      839        NA      1130          NA       18915  20045  NA
2  sc3      0.14     9     1607        NA      1607         131        1544     63  NA
3  sc3      0.14    NA     6886       -33      6919         324        6493    426  NA
4  sc3      0.14    NA     3861        13      3874         290        3600    274  NA
5  sc3      0.14    NA     1607        37      5602         314        5530     72  NA
6  sc0      0.02     1       25        NA        25          NA          22      3  NA
```

Datos atípicos (outliers)

Datos atípicos (outliers)

- Observaciones con una combinación única de características identificables que les diferencia claramente de las otras observaciones.
- Cuando son beneficiosos, aunque diferentes, pueden ser indicativos de las características segmento de la población que se llegarían a descubrir en el curso normal del análisis.
- Los casos atípicos problemáticos no son representativos de la población y están en contra de los objetivos del análisis. Pueden distorsionar seriamente los test estadísticos.
- Error de procedimiento, error en la entrada de datos o un error de codificación, deberían eliminarse o recodificarse como datos ausentes.

Detección de casos atípicos

Los casos atípicos pueden identificarse desde una perspectiva univariante, bivariante o multivariante.

Medidas robustas univariantes

Una regla simple y automática es considerar sospechosas aquellas observaciones tales que:

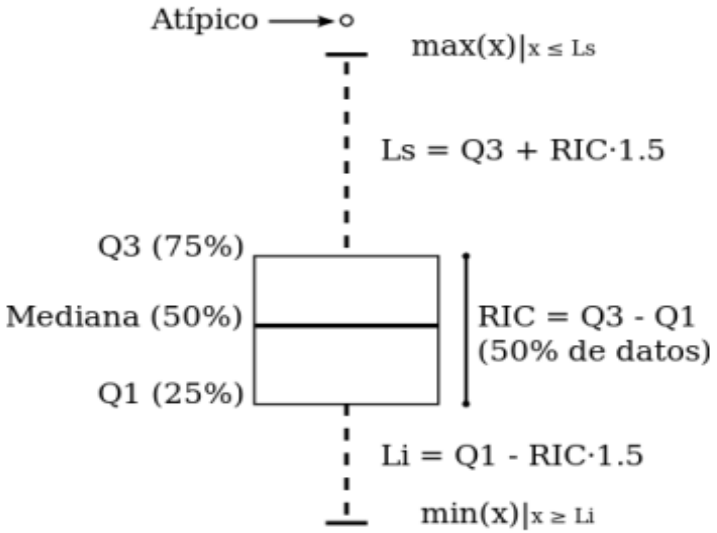
$$\frac{|x_i - med(x)|}{MAD(x)} > 4.5$$

Donde $med(x)$ es la mediana de las observaciones
MAD (Desviación Mediana Absoluta): para un conjunto de datos univariados $x_1, x_2, \ldots, x_n$, el MAD se define como la mediana de las desviaciones absolutas con respecto a la mediana de los datos.

$$MAD = mediana_i(|X_i - mediana_j(X_j)|)$$

Detección de casos atípicos

Diagrama de caja (Boxplot)



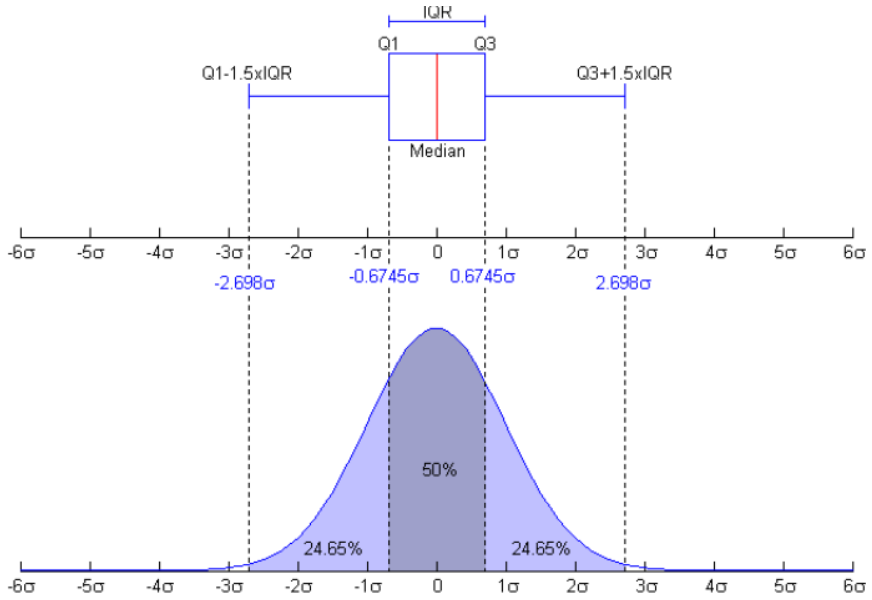
$$L_s = q3 + RIC * 1{,}5$$

$$L_i = q1 - RIC * 1{,}5$$

Detección de casos atípicos

Una observación es declarada como **outlier extremo** si esta cae fuera del intervalo (q1- RIC*3, q3+ RIC*3). Método llamado "**Hampel filter**"

Una observación es declarada como **outlier leve** si esta cae fuera del intervalo (q1- RIC*1.5, q3+ RIC*1.5).



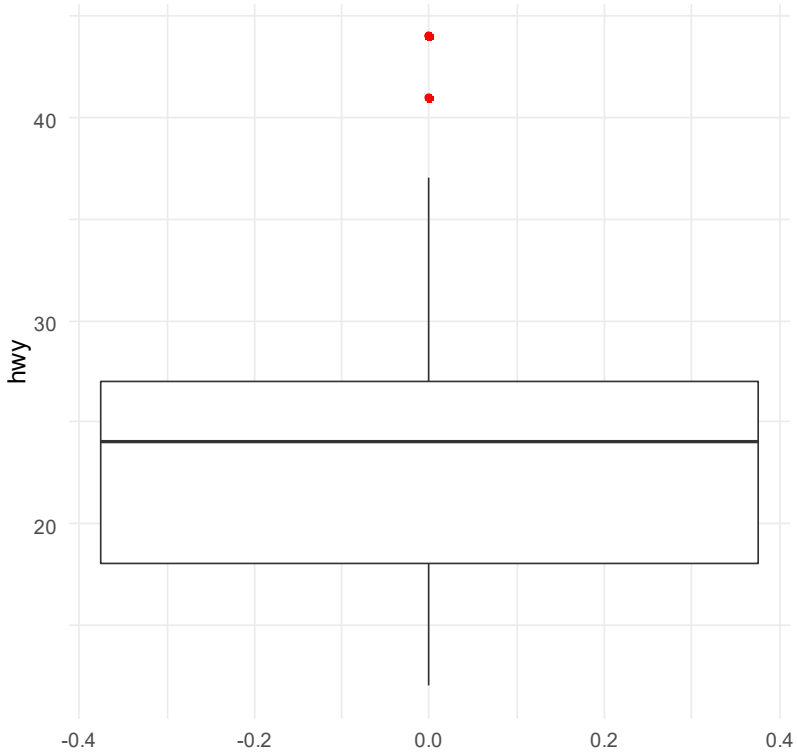
Detección de casos atípicos

Ejemplo:

Datos del dataset "mpg"

library(ggplot2), geom_boxplot( )

boxplot.stats( )

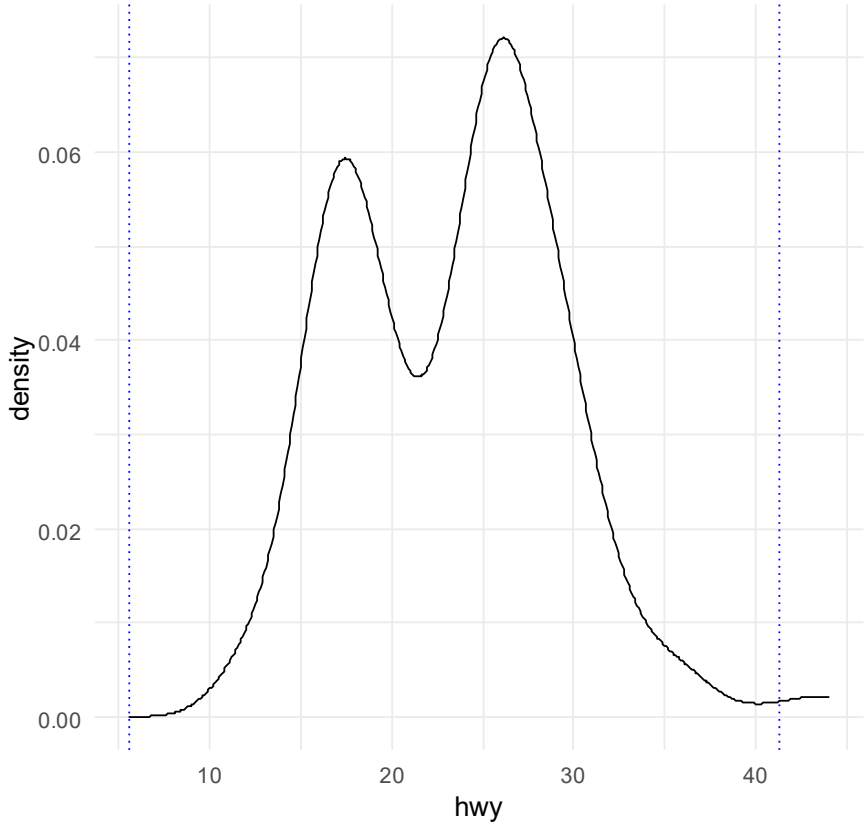


Detección de casos atípicos

Ejemplo:

Datos del dataset "mpg"

library(ggplot2), geom_density( )



Detección de casos atípicos

Prueba de Grubbs

La prueba de Grubbs permite detectar si el valor más alto o más bajo en un conjunto de datos es un valor atípico.

H_0: no hay datos atípicos en la muestra
H_1: hay al menos un dato atípico

$$G = \frac{\max_{i=1,\ldots,N} |Y_i - \bar{Y}|}{s}$$

Se rechazará la hipótesis nula H_0 de no existencia de dato atípico si G excede de cierto valor crítico:

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

Ejemplo: Datos del dataset mpg

```
library(outliers)
grubbs.test( )
```

```
        Grubbs test for one outlier

data:  mpg$hwy
G = 3.45274, U = 0.94862, p-value = 0.05555
alternative hypothesis: highest value 44 is an outlier
```

Detección de casos atípicos

Prueba de Q de Dixon

La prueba de Dixon determina si el valor más extremo de una muestra es un valor atípico. El Q-test se basa en la distribución estadística de datos ordenados, extraídos de la misma población normal. Test para pocos datos, **n ≤ 30**.
Se ordenan los datos en orden creciente para seleccionar el valor discordante, $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n)}$ (supuesto discordante, el valor más grande)

H_0: *no hay datos atípicos en la muestra*
H_1: *hay al menos un dato atípico*

$$Q = \frac{|x_{(n)} - x_{(n-1)}|}{(x_{(n)} - x_{(1)})}$$

$Q > Q_{crit}$ se rechaza la H_0

Ejemplo:

Seleccionamos los primeros 20 registros de la base de datos mpg y comprobamos si hay valores atípicos en la variable hwy

```
library(outliers)
dixon.test( )
```

```
        Dixon test for outliers

data:  submpg$hwy
Q = 0.57143, p-value = 0.006508
alternative hypothesis: lowest value 15 is an outlier
```

Detección de casos atípicos

Generalized ESD Test for Outliers (Prueba de Rosner)

Dado el límite superior, r, la prueba realiza r pruebas separadas: una prueba para un solo valor atípico, una prueba para dos valores atípicos, y así sucesivamente hasta r valores atípicos.

H_0: No hay ningún valor atípico en el conjunto de datos
H_1: Hay hasta r valores atípicos en el conjunto de datos

$$R_i = \frac{\max_i |x_i - \bar{x}|}{s}$$

Se van eliminando observaciones y calculando de forma secuencial los valores de Ri. Repetir y continuar el proceso hasta que se hayan eliminado r observaciones. Entonces los resultados en r pruebas, R_1, R_2, ..., R_r.

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i-1+t^2_{p,n-i-1})(n-i+1)}}$$

donde i = 1, 2, ... , r.
$t_{p,v}$ es el punto porcentual de la distribución t con v grados de libertad y $p = 1 - \frac{\alpha}{2(n-i+1)}$

El número de valores atípicos se determina encontrando el mayor valor de i tal que $R_i > \lambda_i$

Detección de casos atípicos

Ejemplo:

Realizamos el test de Rosner para comprobar si hay valores atípicos en la variable hwy de base de datos mpg

```
library(EnvStats)
rosnerTest( )
```

```
> test$all.stats
  i   Mean.i     SD.i Value Obs.Num    R.i+1 lambda.i+1 Outlier
1 0 23.44017 5.954643    44     213 3.452739   3.652091   FALSE
2 1 23.35193 5.812124    44     222 3.552586   3.650836   FALSE
3 2 23.26293 5.663340    41     223 3.131909   3.649575   FALSE
```

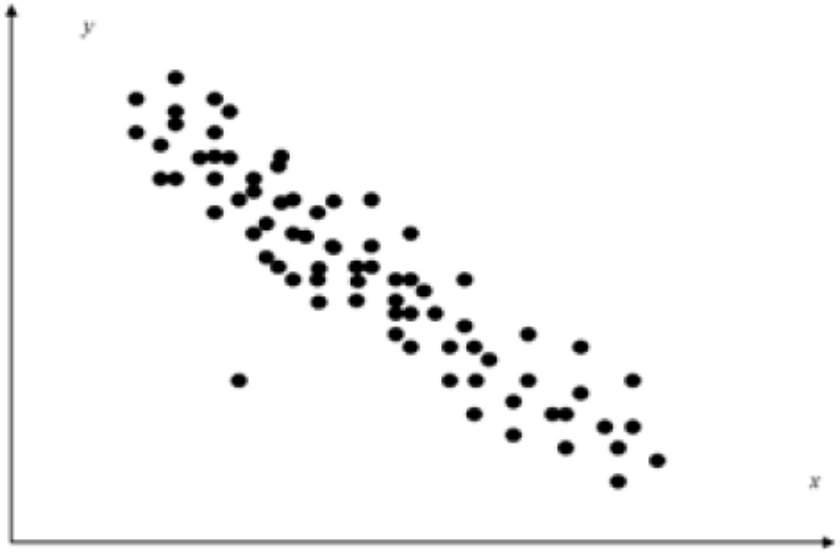
Detección de casos atípicos

Detección multivariante:

Pueden evaluarse conjuntamente pares de variables mediante un gráfico de dispersión.

Casos que caigan manifiestamente fuera del rango del resto de las observaciones pueden identificarse como puntos aislados en el gráfico de dispersión.

Para ayudar a determinar el rango esperado de las observaciones, se puede superponer sobre el gráfico de dispersión una elipse que represente un intervalo de confianza especificado para una distribución normal bivariante.



Detección de casos atípicos

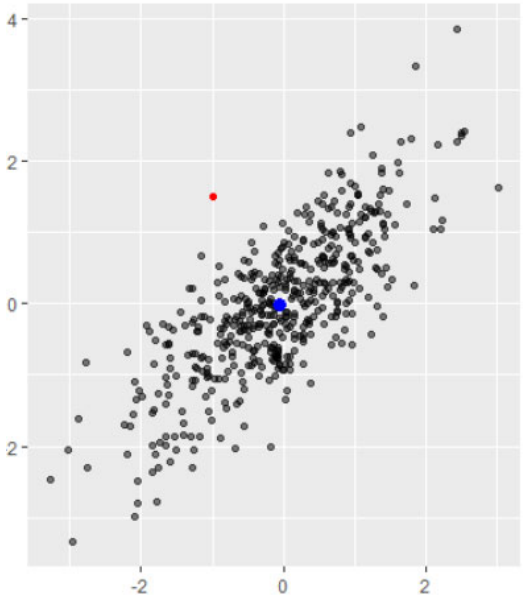
Detección de outliers basado en estadística robusta

La distancia de Mahalanobis es una medida de la distancia de cada observación en un espacio multidimensional respecto del centro medio de las observaciones.

Si x es un vector aleatorio (de dimensión p) con matriz de covarianza muestral S, la distancia de Mahalanobis se define como:

$$D^2 = (x - \bar{x})' S^{-1} (x - \bar{x}) \sim \chi^2_p.$$

Si $D^2 > \chi^2_{p,\alpha}$ se considera potencialmente atípico

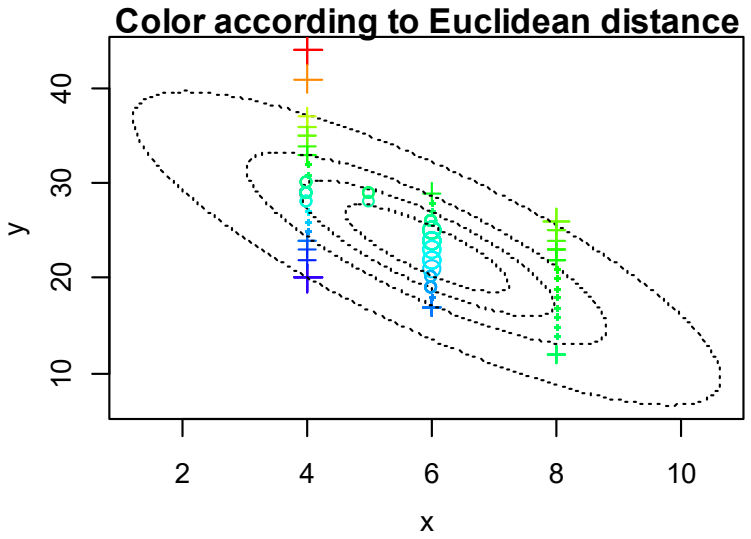


Detección de casos atípicos

Ejemplo: Datos del dataset mpg

```
library(mvoutlier)

Z <- cbind(mpg$cyl, mpg$hwy)
color.plot(Z, quan=0.75)
```



La función color.plot traza los datos (bidimensionales) utilizando diferentes símbolos según la distancia mahalanobis robusta con ajuste y utilizando diferentes colores según las distancias euclidianas de las observaciones.

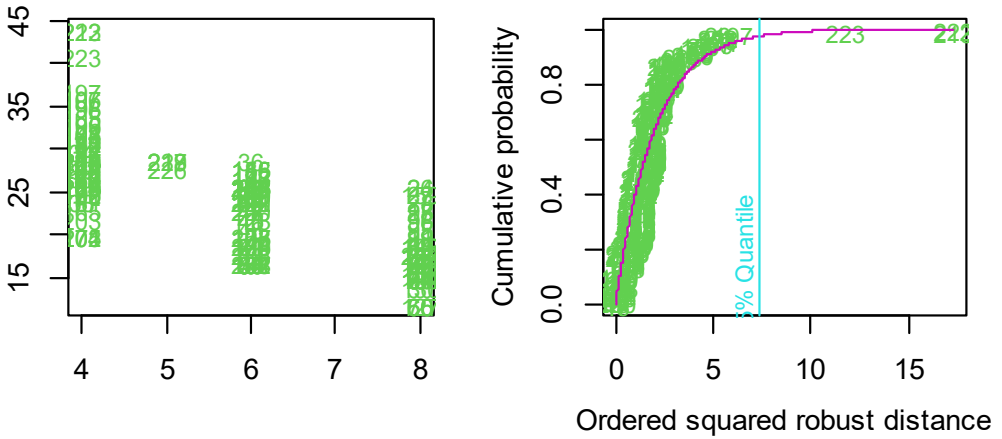
Detección de casos atípicos

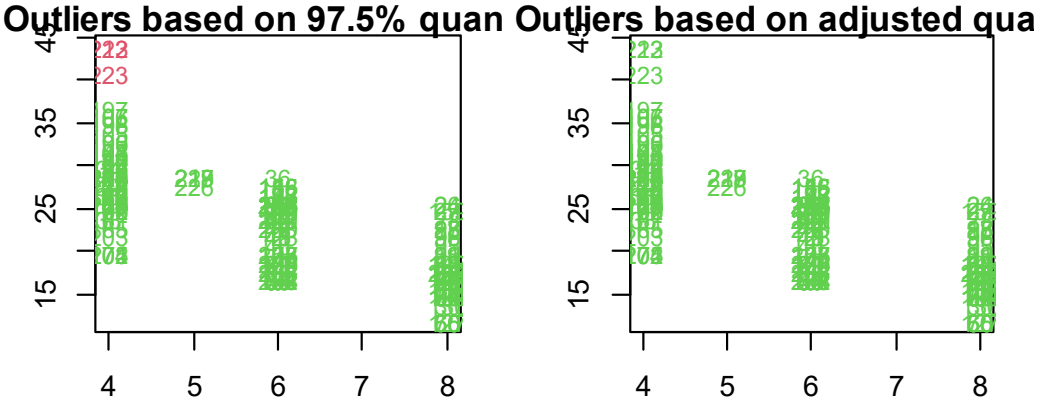
Ejemplo: Datos del dataset mpg

```
library(mvoutlier)

Y <- as.matrix(mpg[, c("cyl", "hwy")])
res <- aq.plot(Y)
```

La función aq.plot traza las distancias de Mahalanobis robustas al cuadrado ordenadas contra la función de distribución empírica. Además, se traza la función de distribución chi^2 y líneas verticales correspondientes al cuantil (por defecto es 0,975). Se crean tres gráficos adicionales (el primero muestra los datos, el segundo muestra los valores atípicos detectados por el cuantil especificado de la distribución y el tercero muestra estos valores atípicos detectados por el cuantil ajustado).





Detección de casos atípicos

- Peña, D. (2002). Análisis de datos multivariantes. McGraw-Hill.
- Hair, Anderson, Tatham, Black. (2001). Análisis Multivariante.
- Enders, C. K. (2010). Applied missing data analysis. Guilford press.
- Lai, M. (2019). Course Handouts for Bayesian Data Analysis Class. https://bookdown.org/marklhc/notes_bookdown/missing-data.html
- Nguyen, M. (2021). A Guide on Data Analysis. https://bookdown.org/mike/data_analysis/
- Hawkins, D. M. (1980). Identification of outliers (Vol. 11). London: Chapman and Hall.
- Aldás Manzano, J., & Uriel Jimenez, E. (2017). Análisis multivariante aplicado con R. Ediciones Paraninfo, SA