SISTEMATIZACION Y MÉTODOS ESTADÍSTICOS

UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

ESTUDIANTES:

- MARÍA LUCIA JACOBO ATUNCAR
- GAMBOA CANALES MARIPAZ
- ARIANA ABIGIAL VIDAL ROMUCHO
- FERNANDA GIANELLA CASTILLA SALVADOR
- SEBASTIAN PALOMINO ROJAS
- KRISTY STEFANY ALVAREZ PEVES

CARGAR PAQUETES Y IMPORTAR LA DATA

```r
install.packages("mice")
install.packages("ggmice")
```

```r
library(mice)
library(tidyverse)
library(here)
library(rio)
library(ggmice)
library(gtsummary)
```

```r
cirrosis_4 <- import(here("cirrosi.CSV"))
```

Un vistazo a los datos

```r
head(cirrosis_4)
```

A tibble: 6 × 21

id <dbl>	dias_seguimie... <dbl>	estado <chr>	medicamento <chr>	edad <dbl>
1	400	Fallecido	D_penicilamina	21464
2	4500	Censurado	D_penicilamina	20617
3	1012	Fallecido	D_penicilamina	25594
4	1925	Fallecido	D_penicilamina	19994
5	1504	Censurado_tras...	Placebo	13918
6	2503	Fallecido	Placebo	24201

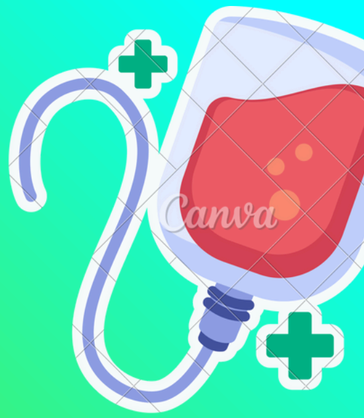
6 rows | 1-5 of 21 columns

pacientes total de : 418

El dataset para este ejercicio

Para ilustrar el proceso de imputación múltiple de datos, utilizaremos el conjunto de datos data_sm. Este dataset incluye información de **418** pacientes adultos.

Las variables registradas **: *ID, Dias_Seguimiento , Estado , Medicamento , Edad, Sexo , Ascitis , Hepatomegalia , Aracnoides ,Edema, Bilirrubina ,Colesterol , Albumina , Cobre, Fosfatasa_Alcalina , SGOT , Trigliceridos, Plaquetas , Tiempo_Protrombina, Etapa*** , entre otras. Algunos participantes presentan valores faltantes en al menos una de estas variables.



Cargando los datos

```{r}
data_sm <- import(here("cirrosis.csv"))
```

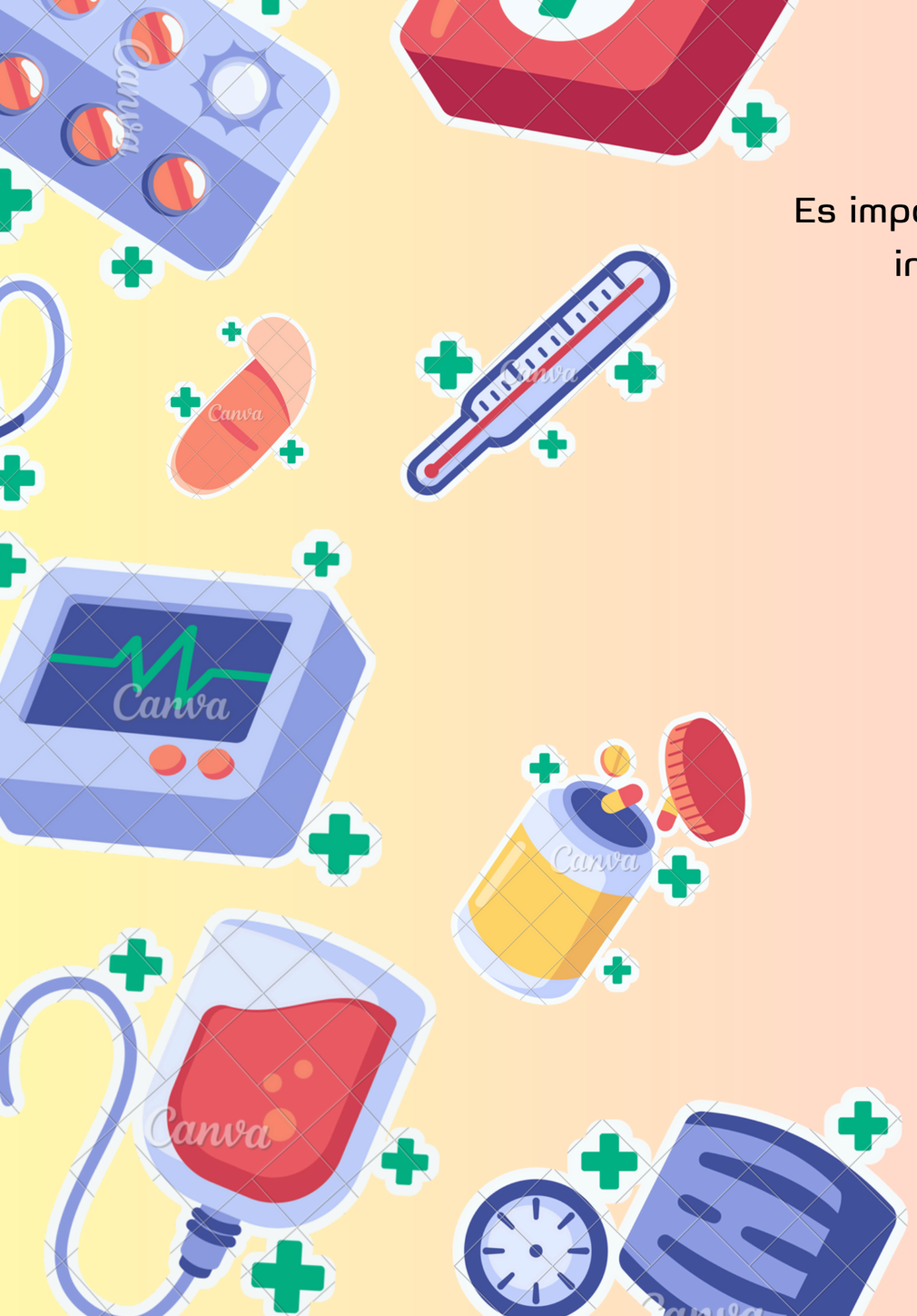
Un vistazo a los datos

```{r}
head(data_sm)
```

Description: df [6 × 20]

	ID <int>	Dias_Seguimiento <int>	Estado <chr>	Medicamento <chr>	Edad <int>	Sexo <chr>	Ascitis <chr>	Hepatomegalia <chr>	Aracnoides <chr>	Edema <chr>	Bilirrubina <dbl>	Colesterol <int>	Albumina <dbl>	Cobre <int>	Fosfatasa_Alcalina <dbl>	SGOT <dbl>
1	1	400	Fallecido	D_penicilamina	21464	Mujer	Sí	Sí	Sí	Severo	14.5	261	2.60	156	1718.0	137.95
2	2	4500	Censurado	D_penicilamina	20617	Mujer	No	Sí	Sí	Ausente	1.1	302	4.14	54	7394.8	113.52
3	3	1012	Fallecido	D_penicilamina	25594	Hombre	No	No	No	Leve	1.4	176	3.48	210	516.0	96.10
4	4	1925	Fallecido	D_penicilamina	19994	Mujer	No	Sí	Sí	Leve	1.8	244	2.54	64	6121.8	60.63
5	5	1504	Censurado_trasplante	Placebo	13918	Mujer	No	Sí	Sí	Ausente	3.4	279	3.53	143	671.0	113.15
6	6	2503	Fallecido	Placebo	24201	Mujer	No	Sí	No	Ausente	0.8	248	3.98	50	944.0	93.00

6 rows | 1-17 of 20 columns



Realizando la imputación de datos

¿Donde estan los valores perdidos?

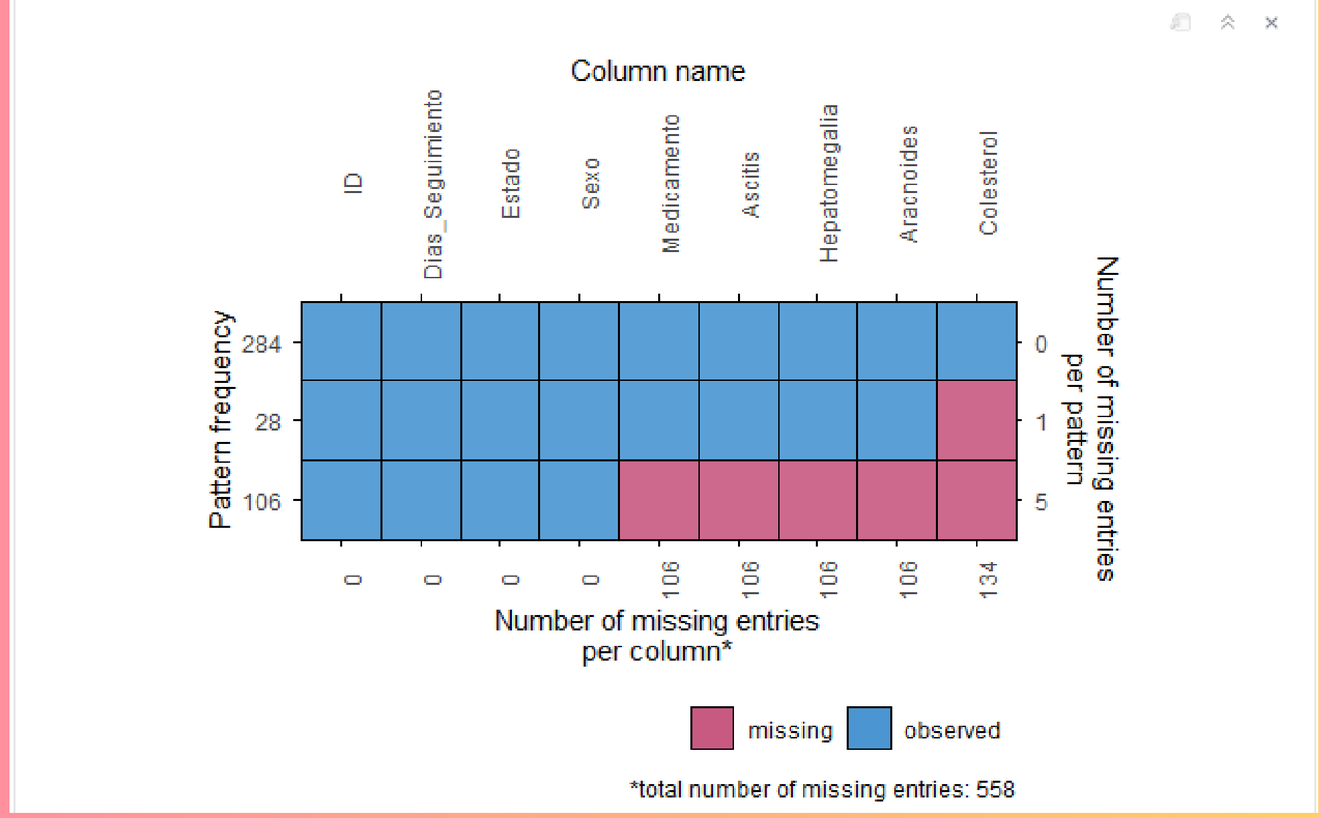
Es importante saber en qué variables se encuentran los datos antes de iniciar la inputación. Una forma rápida es usando la función colSums() es is.na().

```{r}
colSums(is.na(data_sm))
```

ID	Dias_Seguimiento
0	0
Estado	Medicamento
0	106
Edad	Sexo
0	0
Ascitis	Hepatomegalia
106	106
Aracnoides	Edema
106	0
Bilirrubina	Colesterol
0	134
Albumina	Cobre
0	108
Fosfatasa_Alcalina	SGOT
106	106
Trigliceridos	Plaquetas
136	11
Tiempo_Protrombina	Etapa
2	6

Incluso mejor, podemos visualizar los datos perdidos en un mapa de calor usando la función plot_pattern() de ggmice.

```{r}
data_sm |>
  select(ID , Dias_Seguimiento , Estado ,Medicamento , Sexo ,Ascitis , Hepatomegalia
,Aracnoides ,Colesterol) |>
  ggmice::plot_pattern(
    square = TRUE,
    rotate = TRUE)
```



Comparación de participantes con y sin valores perdidos

```r
{r}
tabla_ID = data_sm |>
  dplyr::select(ID , Dias_Seguimiento , Estado ,Medicamento , Sexo ,Ascitis ,
Hepatomegalia ,Aracnoides ,Colesterol) |>
  mutate(missing = factor(
    is.na(ID),
    levels = c(FALSE, TRUE),
    labels = c("Sin valores perdidos", "Con valores perdidos")
  )) |>
  tbl_summary(
    by = missing,
    statistic = list(
      all_continuous()  ~ "{mean} ({sd})",
      all_categorical() ~ "{n}    ({p}%)")
    ) |>
  modify_header(label = "**Variable**",
                all_stat_cols() ~ "**{level}**<br>N = {n} ({style_percent(p, digits
=1)}%)") |>
  modify_caption("Características de los participantes segun valor perdido") |>
  bold_labels()

tabla_Colesterol = data_sm |>
  dplyr::select(ID , Dias_Seguimiento , Estado ,Medicamento , Sexo ,Ascitis ,
Hepatomegalia ,Aracnoides ,Colesterol) |>
  mutate(missing = factor(
    is.na(Colesterol),
    levels = c(FALSE, TRUE),
    labels = c("Sin valores perdidos", "Con valores perdidos")
  )) |>
  tbl_summary(
    by = missing,
    statistic = list(
      all_continuous()  ~ "{mean} ({sd})",
      all_categorical() ~ "{n}    ({p}%)")
    ) |>
  modify_header(label = "**Variable**",
                all_stat_cols() ~ "**{level}**<br>N = {n} ({style_percent(p, digits
=1)}%)") |>
  modify_caption("Características de los participantes segun valor perdido") |>
  bold_labels()

tabla <- tbl_merge(
  tbls = list(tabla_ID, tabla_Colesterol),
  tab_spanner = c("**numero_caso**", "**leucocitos_sangre**")
)
```

Características de los participantes segun valor perdido

Variable	numero_caso		leucocitos_sangre	
	Sin valores perdidos N = 418 (100.0%)†	Con valores perdidos N = 0 (0%)†	Sin valores perdidos N = 284 (67.9%)†	Con valores perdidos N = 134 (32.1%)†
ID	210 (121)	NA (NA)	158 (91)	318 (103)
Dias_Seguimiento	1,918 (1,105)	NA (NA)	1,992 (1,112)	1,761 (1,075)
Estado				
Censurado	232 (56%)	0 (NA%)	152 (54%)	80 (60%)
Censurado_trasplante	25 (6.0%)	0 (NA%)	18 (6.3%)	7 (5.2%)
Fallecido	161 (39%)	0 (NA%)	114 (40%)	47 (35%)
Medicamento				
D_penicilamina	158 (51%)	0 (NA%)	140 (49%)	18 (64%)
Placebo	154 (49%)	0 (NA%)	144 (51%)	10 (36%)
Unknown	106	0	0	106
Sexo				
Hombre	44 (11%)	0 (NA%)	35 (12%)	9 (6.7%)
Mujer	374 (89%)	0 (NA%)	249 (88%)	125 (93%)
Ascitis				
No	288 (92%)	0 (NA%)	263 (93%)	25 (89%)
Sí	24 (7.7%)	0 (NA%)	21 (7.4%)	3 (11%)
Unknown	106	0	0	106

¿Qué variables debo incluir en el proceso de imputación?

```{r}
input_data =
  data_sm |>
    dplyr::select(ID , Dias_Seguimiento , Estado ,Medicamento , Sexo ,Ascitis ,
Hepatomegalia ,Aracnoides ,Colesterol) |>
  mutate(ID = as.factor(Colesterol))
```

La función mice() para imputar datos

```r
names(input_data)
```

```
[1] "ID"              "Dias_Seguimiento" "Estado"          "Medicamento"
[5] "Sexo"            "Ascitis"          "Hepatomegalia"   "Aracnoides"
[9] "Colesterol"
```

```r
data_imputada =
  mice(
    input_data,
    m = 20,
    method = c(
      "",
      "",
      "",
      "",
      "pmm",
      "",
      "pmm",
      "",
      "",
      "logreg"),
    maxit = 20,
    seed = 3,
    print = F
    )
```

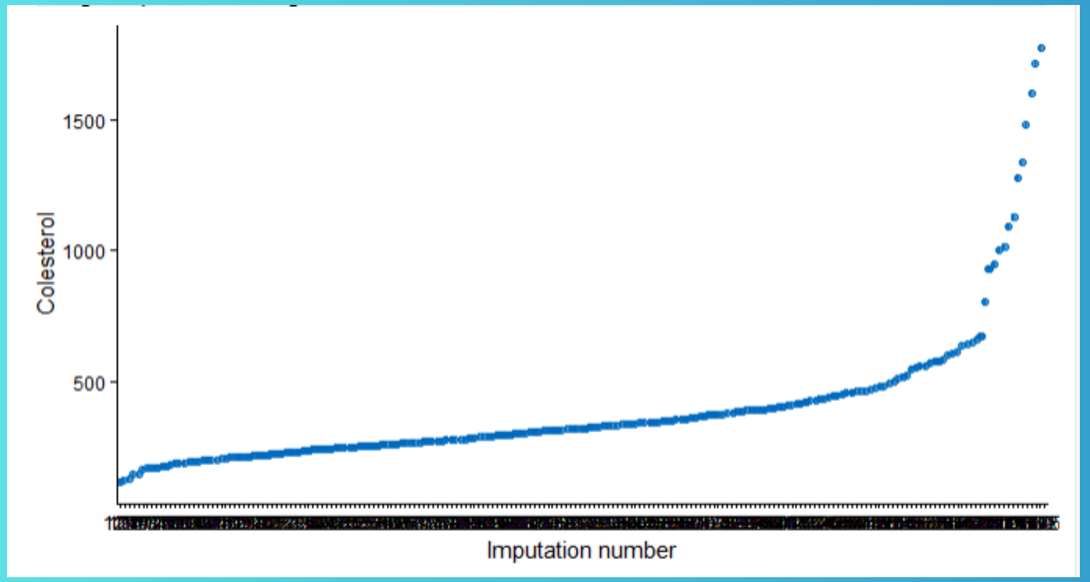
```r
data_imputada
```

Description: df [6 × 5]

	it <dbl>	im <dbl>	dep <chr>	meth <chr>	out <chr>
1	0	0		constant	Estado
2	0	0		constant	Medicamento
3	0	0		constant	Sexo
4	0	0		constant	Ascitis
5	0	0		constant	Hepatomega...
6	0	0		constant	Aracnoides

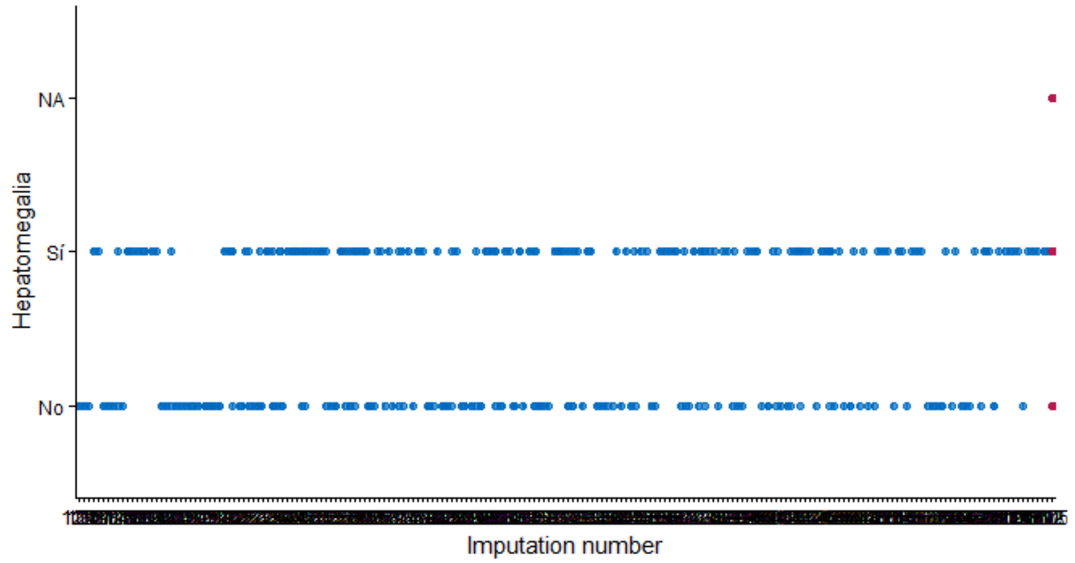
Analizando los datos imputados

```r
ggmice(data_imputada, aes(x = ID, y = Colesterol)) +
  geom_jitter(height = 0, width = 0.25) +
  geom_boxplot(width = 0.5, size = 1, alpha = 0.55, outlier.shape = NA) +
  labs(x = "Imputation number")
```



Para la variables Hepatomegalia

```r
ggmice(data_imputada, aes(x = ID, y = Hepatomegalia)) +
  geom_jitter(height = 0, width = 0.25) +
  geom_boxplot(width = 0.5, size = 1, alpha = 0.55, outlier.shape = NA) +
  labs(x = "Imputation number")
```



Analizando los datos imputados

```r
{r}
data_imputada_l <- complete(data_imputada, "long", include = TRUE)
```

```r
{r}
data_imputada_l <- data_imputada_l %>%
  mutate(imputed = .imp > 0,
         imputed = factor(imputed,
                          levels = c(F,T),
                          labels = c("Observado", "Imputado")))

prop.table(table(data_imputada_l$Colesterol,
                 data_imputada_l$Hepatomegalia),
           margin = 2)
```

	No	Sí
120	0.007299270	0.000000000
127	0.007299270	0.000000000
132	0.007299270	0.000000000
149	0.000000000	0.006802721
151	0.000000000	0.006802721
168	0.007299270	0.000000000
172	0.007299270	0.000000000
174	0.007299270	0.000000000
175	0.007299270	0.006802721
176	0.007299270	0.000000000
178	0.000000000	0.013605442
187	0.000000000	0.006802721
188	0.000000000	0.006802721
191	0.000000000	0.006802721
193	0.000000000	0.006802721
194	0.000000000	0.006802721
196	0.000000000	0.006802721
198	0.007299270	0.000000000
200	0.007299270	0.000000000
201	0.007299270	0.006802721
204	0.007299270	0.000000000
205	0.007299270	0.000000000
206	0.007299270	0.000000000
210	0.007299270	0.000000000
212	0.007299270	0.000000000
213	0.007299270	0.000000000
215	0.014598540	0.000000000
216	0.007299270	0.000000000
217	0.014598540	0.000000000
219	0.014598540	0.000000000
220	0.000000000	0.006802721
222	0.000000000	0.013605442
223	0.007299270	0.006802721
225	0.007299270	0.000000000
226	0.007299270	0.006802721
227	0.007299270	0.006802721
231	0.007299270	0.000000000
232	0.014598540	0.006802721
233	0.007299270	0.000000000
235	0.000000000	0.013605442
236	0.014598540	0.006802721
239	0.007299270	0.006802721
242	0.014598540	0.006802721
243	0.000000000	0.006802721
244	0.000000000	0.013605442
246	0.000000000	0.013605442
247	0.007299270	0.006802721
248	0.007299270	0.013605442
250	0.000000000	0.006802721
251	0.000000000	0.006802721
252	0.000000000	0.013605442
253	0.007299270	0.006802721
255	0.007299270	0.000000000
256	0.014598540	0.000000000
257	0.000000000	0.020408163
258	0.007299270	0.006802721
259	0.014598540	0.006802721
260	0.007299270	0.020408163
261	0.000000000	0.006802721
262	0.000000000	0.013605442
263	0.021897810	0.000000000
266	0.007299270	0.000000000
267	0.000000000	0.013605442
268	0.007299270	0.000000000
269	0.000000000	0.006802721

Procedimientos adicionales luego de la imputación

```r
data_Colesterol <- factor(data_imputada$Colesterol, levels = c(0, 1))
```

```r
str(data_imputada$Colesterol)
table(data_imputada$Hepatomegalia)
```

```
 NULL
< table of extent 0 >
```

```r
tabla_multi <-
  data_imputada |>
  with(glm(ID  ~ Dias_Seguimiento  + Estado  +Sexo + Ascitis  + Hepatomegalia  +
Aracnoides + Colesterol ,
           family = binomial(link = "logit"))) |>
  tbl_regression(exponentiate = TRUE,
                 label = list(
                   Dias_Seguimiento  ~ "DIAS",
                   Estado ~ "FallecidO/Censurado_trasplante/Censurado",
                   Sexo ~ "Mujer/Hombre",
                   Ascitis ~ "SI/NO",
                   Hepatomegalia ~ "SI/NO",
                   Aracnoides  ~ "SI/NO",
                   Colesterol  ~ "mg/dL")) |>
  bold_p(t = 0.05) |>
  modify_header(estimate = "**OR ajustado**", p.value = "**p valor** ")
```

```r
tabla_multi
```

Characteristic	OR ajustado
DIAS	1.02
FallecidO/Censurado_trasplante/Censurado	
Censurado	—
Censurado_trasplante	0.00
Fallecido	140,866,799,727,199
Mujer/Hombre	
Hombre	—
Mujer	0.00
SI/NO	







UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA SJB

MUCHAS GRACIAS