"Año de la recuperación y consolidación de la economía peruana

UNIVERSIDAD PRIVADA SAN JUAN BAUTISTA

FACULTAD DE CIENCIAS DE LA SALUD

ESCUELA PROFESIONAL DE MEDICINA HUMANA

CURSO:

SISTEMATIZACIÓN Y MÉTODOS ESTADÍSTICOS

DOCENTE:

SEGUNDO VICENTE CASTRO LOPEZ

INTEGRANTES:

- **Marroquin Ticona Dayana**
- **Degregori Hinojosa Alexandra**
- **Loayza Quispe Leidy**
- **Quispe Valencia Evelyn**
- **Cayllahua Huarancca Mayrha**

```
---
title: "SEMANA 14 G1"
format: html
editor: visual
---
```

Integrantes

Dayana Lilian Marroquín Ticona

Alexandra Pamela Degregori Hinojosa

Cayllahua Huarancca Mayrha

Leidy Yojhana Loayza Quispe.

Evelyn Lizbeth Quispe Valencia

Instalar (si es necesario)

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

Cargar paquetes

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

1 Datos perdidos en investigación en salud

Es común encontrar datos faltantes en un conjunto de datos. Por ejemplo, al recolectar información a partir de historias clínicas de pacientes en un hospital, algunas variables pueden no estar disponibles porque no fueron medidas, anotadas o solicitadas por el personal de salud. En otro escenario, en estudios que utilizan encuestas, es posible que las personas encuestadas no respondan ciertas preguntas o que las respuestas sean ininteligibles.

Cuando se aplican métodos de regresión en investigaciones en ciencias de la salud, la práctica habitual consiste en eliminar las observaciones que contienen datos faltantes. Esta técnica se conoce como análisis de casos completos, y muchos paquetes estadísticos la implementan por defecto.

2 Imputación de datos

Siempre es preferible utilizar todas las observaciones en un análisis de regresión, ya que esto permite obtener estimaciones más precisas y cercanas a la realidad. En esta sesión, aplicaremos una técnica llamada imputación, que consiste en reemplazar los datos perdidos con una estimación de su valor verdadero.

Esta no es una técnica reciente. Enfoques anteriores de imputación —como, por ejemplo, reemplazar los valores perdidos con el promedio de la variable— han sido ampliamente utilizados, pero presentan limitaciones. Estas limitaciones han sido superadas por una técnica más moderna y actualmente muy popular: la imputación múltiple de datos.

3 El dataset para este ejercicio

Para ilustrar el proceso de imputación múltiple de datos, utilizaremos el conjunto de datos `data_almacsangre`

Este dataset incluye información de 316 pacientes con diagnóstico y tratamiento de cáncer de próstata. Las variables registradas comprenden aspectos clínicos y patológicos como el grupo etario (`Grupo_edad_GR`), la edad individual (`Edad`), la raza (afroamericana o no), antecedentes familiares, el volumen prostático (en cm^3), el volumen tumoral (clasificado como bajo, medio o alto), el estadio clínico (`Estadio_T`), la escala de Gleason al momento de la biopsia y tras la cirugía, el nivel de PSA preoperatorio (ng/mL), y variables relacionadas al tratamiento como terapia previa, radioterapia o terapia adyuvante.

Además, se incluyen desenlaces clínicos como la recurrencia bioquímica, el tiempo hasta dicha recurrencia y si el paciente fue censurado en el seguimiento.

Cargando los datos

```{r}
data_sm <- import(here("data", "almac_Sangre.csv"))
```

```
Registered S3 method overwritten by 'data.table':
  method           from
  print.data.table
```

Un vistazo a los datos

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

Description: df [6 × 20]

	Grupo_edad_GR <chr>	Edad_mediana... <int>	Edad <dbl>	Raza_afroame... <chr>
1	Mayor	25	72.1	No
2	Mayor	25	73.6	No
3	Mayor	25	67.5	No
4	Intermedio	15	65.8	No
5	Intermedio	15	63.2	No
6	Mayor	25	65.4	No

6 rows | 1-5 of 20 columns

4 Realizando la imputación de datos

4.1 ¿Donde estan los valores perdidos?

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

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

```
           Grupo_edad_GR          Edad_mediana_GR                     Edad
 Raza_afroamericana         Historia_familiar          Volumen_prostata
                         0                         0                     0
 0                       0                         9
         Volumen_tumoral                 Estadio_T          Gleason_biopsia
 Confinamiento_organo          PSA_preoperatorio          Terapia_previa
                         6                        13                     2
 0                       3                         0
   Unidades_transfundidas       Gleason_quirurgico        Terapia_adyuvante
 Radioterapia_adyuvante    Recurrencia_bioquimica                   Censor
                         0                         0                     0
 0                       0                         0
 Tiempo_hasta_recurrencia              BN_positivo
                         1                         0
```

Interpretacion:

Los valores perdidos se encontraron en las siguientes variables:

Volumen_prostata: 9 valores faltantes

Volumen_tumoral: 6 valores faltantes

Estadio_T: 13 valores faltantes

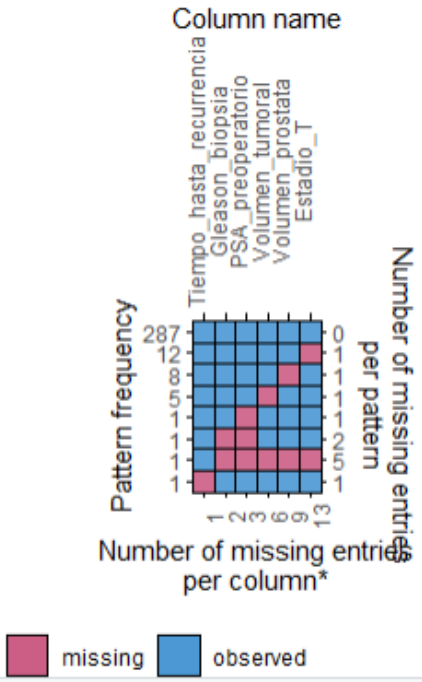
Gleason_biopsia: 2 valores faltantes

PSA_preoperatorio: 3 valores faltantes

Tiempo_hasta_recurrencia: 1 valor faltante

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

```{r}
data_sm |>
  select(
    Volumen_prostata,
    Volumen_tumoral,
    Estadio_T,
    Gleason_biopsia,
    PSA_preoperatorio,
    Tiempo_hasta_recurrencia
  ) |>
  ggmice::plot_pattern(
    square = TRUE,
    rotate = TRUE
  )
```



Interpretacion:

Primero que nada, el total de valores faltantes es bajo: 34 en total. Es decir, la gran mayoría de los datos está completa. La variable con más valores ausentes es Estadio_T con 13 casos faltantes, seguida por Volumen_prostata, Volumen_tumoral y Gleason_biopsia que tienen entre 2 y 9 valores faltantes. PSA_preoperatorio y Tiempo_hasta_recurrencia tienen solo 1 dato faltante cada una, lo cual es mínimo.

Lo interesante de este gráfico es que no solo muestra qué columnas tienen datos faltantes en color rosado sino también cuántas observaciones comparten el mismo patrón de ausencia. Por ejemplo, hay un pequeño grupo de observaciones que tienen 3 o más variables vacías al mismo tiempo (lo que se ve en las filas más bajas del gráfico), pero en general, la mayoría de las filas están completas en azul

4.2 Comparación de participantes con y sin valores perdidos

Una buena práctica antes de iniciar la imputación de datos es también evaluar cómo difieren los valores de las otras variables entre el grupo de participantes con valores perdidos y el grupo sin valores perdidos. Esto es importante porque puede darnos pistas sobre si realmente es necesaria la imputación o, dicho de otra forma, si es seguro usar el análisis de casos completos.
¿Cómo? Si la distribución de las otras variables no difiere entre el grupo con valores perdidos y el grupo sin valores perdidos, entonces podría no ser necesaria la imputación.

Evaluamos esto en nuestro dataset para las variables PSA_preoperatorio y Volumen_prostata, que presentan algunos datos faltantes.

```r
# Comparación según valores perdidos en PSA_preoperatorio
tabla_psa <- data_sm |>
  dplyr::select(
    Edad,
    Raza_afroamericana,
    Historia_familiar,
    Estadio_T,
    Gleason_biopsia,
    Volumen_prostata,
    PSA_preoperatorio,
    Terapia_previa,
    Gleason_quirurgico,
    Recurrencia_bioquimica
  ) |>
  mutate(missing = factor(
    is.na(PSA_preoperatorio),
    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 según valor perdido en
**PSA_preoperatorio**") |>
  bold_labels()
```

```r
# Comparación según valores perdidos en Volumen_prostata
tabla_volumen <- data_sm |>
  dplyr::select(
    Edad,
    Raza_afroamericana,
    Historia_familiar,
    Estadio_T,
    Gleason_biopsia,
    Volumen_prostata,
    PSA_preoperatorio,
    Terapia_previa,
    Gleason_quirurgico,
    Recurrencia_bioquimica
  ) |>
  mutate(missing = factor(
    is.na(Volumen_prostata),
    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 según valor perdido en
**Volumen_prostata**") |>
  bold_labels()
```

```r
# Unir ambas tablas
tabla <- tbl_merge(
  tbls = list(tabla_psa, tabla_volumen),
  tab_spanner = c("**PSA preoperatorio**", "**Volumen prostático**")
)
```

```r
{r}
tabla
```

Características de los participantes según valor perdido en **PSA_preoperatorio**

Variable	PSA preoperatorio		Volumen prostático	
	Sin valores perdidos N = 313 (99.1%)[1]	**Con valores perdidos** N = 3 (0.95%)[1]	**Sin valores perdidos** N = 307 (97.2%)[1]	**Con valores perdidos** N = 9 (2.85%)[1]
Edad	61 (7)	67 (8)	61 (7)	66 (7)
Raza_afroamericana				
No	259 (83%)	2 (67%)	252 (82%)	9 (100%)
Sí	54 (17%)	1 (33%)	55 (18%)	0 (0%)
Historia_familiar				
No	245 (78%)	3 (100%)	241 (79%)	7 (78%)
Sí	68 (22%)	0 (0%)	66 (21%)	2 (22%)
Estadio_T				
T1-T2a	267 (89%)	2 (100%)	262 (89%)	7 (88%)
T2b-T3	34 (11%)	0 (0%)	33 (11%)	1 (13%)
Unknown	12	1	12	1
Gleason_biopsia				
Gleason 0-6	188 (60%)	1 (100%)	183 (60%)	6 (75%)
Gleason 7	93 (30%)	0 (0%)	91 (30%)	2 (25%)
Gleason 8-10	32 (10%)	0 (0%)	32 (10%)	0 (0%)
Unknown	0	2	1	1
Volumen_prostata	57 (30)	48 (13)	56 (30)	NA (NA)

Volumen_prostata	57 (30)	48 (13)	56 (30)	NA (NA)
Unknown	8	1	0	9
PSA_preoperatorio	8.2 (6.0)	NA (NA)	8.3 (6.1)	4.8 (1.1)
Unknown	0	3	2	1
Terapia_previa				
No	276 (88%)	2 (67%)	269 (88%)	9 (100%)
Sí	37 (12%)	1 (33%)	38 (12%)	0 (0%)
Gleason_quirurgico				
Gleason 0-6	84 (27%)	1 (33%)	82 (27%)	3 (33%)
Gleason 7	171 (55%)	1 (33%)	166 (54%)	6 (67%)
Gleason 8-10	21 (6.7%)	0 (0%)	21 (6.8%)	0 (0%)
No asignado	37 (12%)	1 (33%)	38 (12%)	0 (0%)
Recurrencia_bioquimica				
No	261 (83%)	1 (33%)	255 (83%)	7 (78%)
Sí	52 (17%)	2 (67%)	52 (17%)	2 (22%)

[1] Mean (SD); n (%)

Interpretacion:

Comparación según valores perdidos en PSA preoperatorio

De los 316 pacientes totales, solo 3 tienen valores faltantes en `PSA_preoperatorio`, lo que representa menos del 1%. En promedio, estos pacientes tienen una edad de 67 años, mientras que los que tienen el dato presente tienen 61 años. Es decir, los pacientes con datos faltantes son en promedio 6 años mayores.

En cuanto a la raza, el 33% de los pacientes con dato faltante son afroamericanos, mientras que entre los que sí tienen el valor registrado, solo el 17% lo son.

También se nota que todos los pacientes con dato faltante en PSA no tienen antecedentes familiares de cáncer, en cambio en el grupo sin datos perdidos, el 22% sí tenía historia familiar.

Respecto al estadio clínico, el 100% de los casos con dato perdido están en estadio T1-T2a, y ninguno en estadio T2b-T3. En la escala de Gleason (biopsia), los tres pacientes con dato perdido están en la categoría más baja (0-6).

En términos generales, los pacientes con valores faltantes en `PSA_preoperatorio` parecen ser levemente mayores, sin historia familiar, y con tumores más localizados y menos agresivos según Gleason. Sin embargo, el número es muy pequeño (n = 3), por lo que cualquier diferencia debe tomarse con precaución.

Comparación según valores perdidos en Volumen prostático

Hay 9 pacientes con datos faltantes en `volumen_prostata`, lo cual representa casi un 3%. Estos pacientes tienen una edad promedio de 66 años, mientras que quienes tienen el valor presente tienen un promedio de 61 años.

En cuanto a la raza, el 100% de los pacientes con dato faltante no son afroamericanos, frente al 82% del grupo sin valores perdidos.

Los antecedentes familiares están distribuidos de forma muy parecida en ambos grupos: 78% sin antecedentes en ambos.

En la escala de Gleason por biopsia, 75% de los pacientes con datos faltantes están en la categoría 0-6, lo que sugiere tumores menos agresivos. No hay pacientes en la categoría 8-10 entre los que tienen valores perdidos.

El valor promedio de PSA preoperatorio en este grupo también fue más bajo: 4.8 ng/mL frente a 8.3 ng/mL del grupo con datos presentes. Finalmente, todos los pacientes con valores perdidos en esta variable no recibieron terapia previa.

4.3 ¿Qué variables debo incluir en el proceso de imputación?

Debemos incluir todas las variables que se utilizarán en los análisis posteriores, incluso aquellas que no presentan valores perdidos. La razón es que el modelo de imputación debe ser tan completo como el análisis que se realizará después. De lo contrario, se podría perder información relevante. Además, aunque algunas variables no tengan valores faltantes, su inclusión puede mejorar la estimación de los valores imputados.

También es importante recordar que las variables categóricas deben estar en formato `factor`. A continuación, se seleccionan las variables que serán incluidas y se transforman aquellas que lo requieren.

```{r}
input_data <- data_sm |>
  dplyr::select(
    Edad,
    Raza_afroamericana,
    Historia_familiar,
    Estadio_T,
    Gleason_biopsia,
    Gleason_quirurgico,
    PSA_preoperatorio,
    Volumen_prostata,
    Volumen_tumoral,
    Recurrencia_bioquimica,
    Terapia_previa,
    Radioterapia_adyuvante,
    Censor,
    Tiempo_hasta_recurrencia,
    BN_positivo
  ) |>
  mutate(
    Raza_afroamericana = as.factor(Raza_afroamericana),
    Historia_familiar = as.factor(Historia_familiar),
    Estadio_T = as.factor(Estadio_T),
    Gleason_biopsia = as.factor(Gleason_biopsia),
    Gleason_quirurgico = as.factor(Gleason_quirurgico),
    Recurrencia_bioquimica = as.factor(Recurrencia_bioquimica),
    Terapia_previa = as.factor(Terapia_previa),
    Radioterapia_adyuvante = as.factor(Radioterapia_adyuvante),
    Censor = as.factor(Censor),
    Volumen_tumoral = as.factor(Volumen_tumoral),
    BN_positivo = as.factor(BN_positivo)
  )
```

4.4 La función `mice()` para imputar datos

Para imputar datos utilizaremos la función mice() del paquete del mismo nombre. Entre sus argumentos, debemos especificar:

- el número de imputaciones con m,
- una semilla (seed) para que los resultados sean reproducibles, y
- el método de imputación con method.

Para los métodos de imputación:

Usaremos "pmm" (predictive mean matching) para variables continuas con valores faltantes como PSA_preoperatorio, Volumen_prostata y Tiempo_hasta_recurrencia.

Usaremos "logreg" para variables binarias con valores faltantes como Estadio_T o Gleason_biopsia (si tienen solo dos niveles).

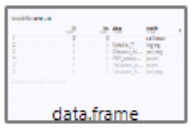
Para las variables sin valores faltantes, colocamos simplemente "".

```{r}
names(input_data)
```

```
 [1] "Edad"                    "Raza_afroamericana"      "Historia_familiar"
"Estadio_T"               "Gleason_biopsia"         "Gleason_quirurgico"
 [7] "PSA_preoperatorio"       "Volumen_prostata"        "Volumen_tumoral"
"Recurrencia_bioquimica"  "Terapia_previa"           "Radioterapia_adyuvante"
[13] "Censor"                  "Tiempo_hasta_recurrencia" "BN_positivo"
```

```{r}
data_imputada <- mice(
  input_data,
  m = 20,
  method = c(
    "",
    "",
    "",
    "",
    "logreg",
    "polyreg",
    "",
    "pmm",
    "pmm",
    "polyreg",
    "",
    "",
    "",
    "",
    "",
    "pmm",
    ""
  ),
  maxit = 20,
  seed = 3,
  print = FALSE
)
```

```r
data_imputada
```



Description: df [6 × 5]

	it <dbl>	im <dbl>	dep <chr>	meth <chr>	out <chr>
1	0	0		collinear	Censor
2	1	1	Estadio_T	logreg	Gleason_qui...
3	1	1	Gleason_bio...	polyreg	Gleason_qui...
4	1	1	PSA_preoper...	pmm	Gleason_qui...
5	1	1	Volumen_pr...	pmm	Terapia_pre...
6	1	1	Volumen_tu...	polyreg	Gleason_qui...

6 rows

Interpretacion:

Se imputaron seis variables con datos faltantes. PSA preoperatorio y Volumen prostático se completaron usando un método para variables numéricas. Estadio T, Gleason en la biopsia y Volumen tumoral se completaron usando modelos para variables categóricas. Todo el proceso se realizó correctamente y se puede continuar con el análisis.

5 Analizando los datos imputados

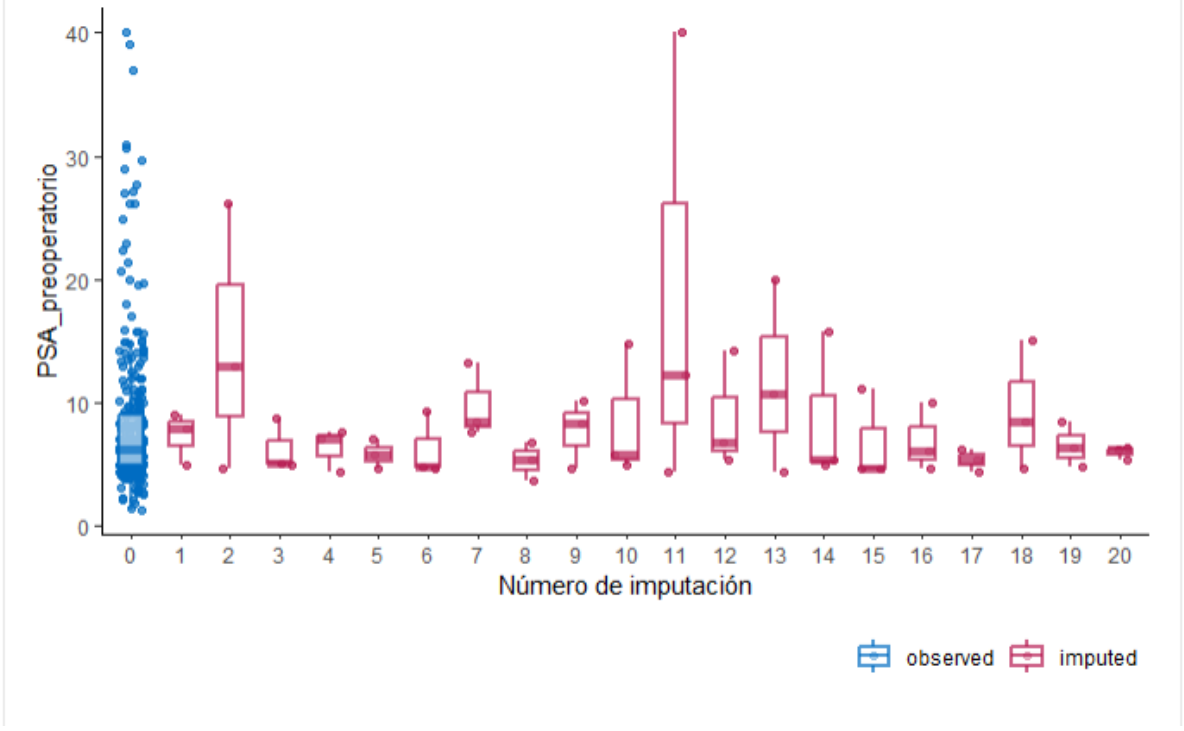
Antes de hacer cualquier análisis adicional con los datos imputados, es importante revisar si los valores imputados son razonables. Lo ideal es que no se vean muy distintos a los valores originales que sí estaban en el dataset.

Una forma de ver esto es usando gráficos de caja para comparar las distribuciones de los valores observados con los valores imputados en cada una de las 20 imputaciones.

A continuación, mostramos los gráficos para dos variables: `PSA_preoperatorio` y `Volumen_prostata`.

Para la variable PSA_preoperatorio

```{r}
ggmice(data_imputada, aes(x = .imp, y = PSA_preoperatorio)) +
  geom_jitter(height = 0, width = 0.25) +
  geom_boxplot(width = 0.5, size = 1, alpha = 0.55, outlier.shape = NA) +
  labs(x = "Número de imputación")
```



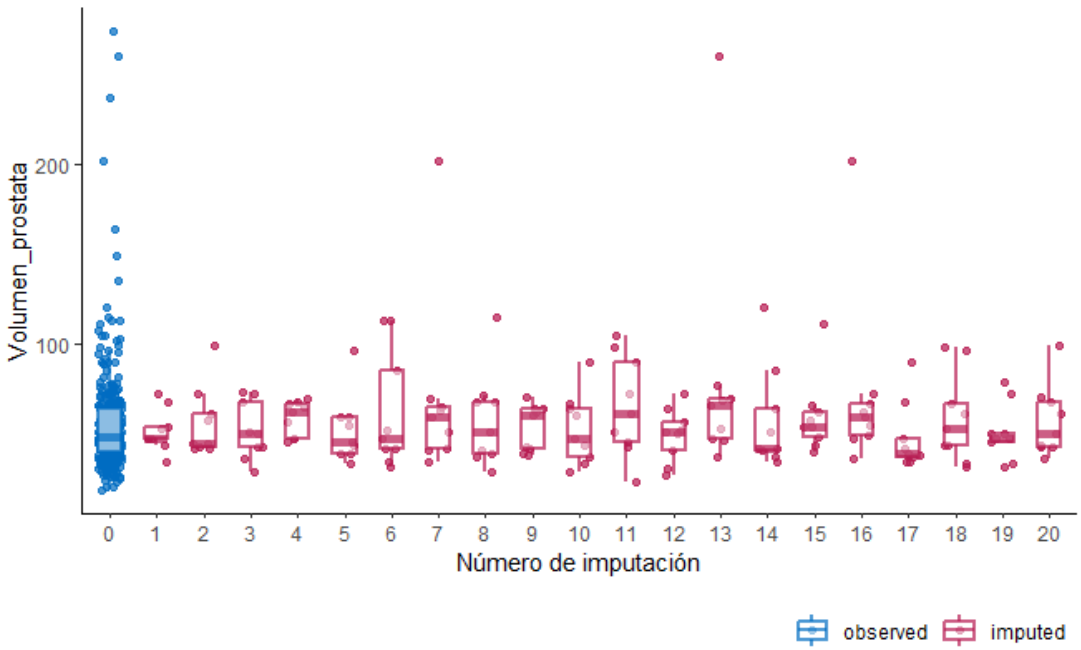
Interpretacion:

El gráfico muestra la distribución del PSA preoperatorio antes y después de la imputación. Los puntos azules a la izquierda (en la posición 0) representan los valores originales observados, mientras que los boxplots en color rosado desde el 1 al 20 muestran las 20 imputaciones generadas por el modelo.

Podemos ver que la mayoría de los valores imputados están dentro del rango de los valores reales. Aunque hay algo de variación entre imputaciones, en general los valores imputados se distribuyen de forma razonable y siguen un patrón similar al de los datos observados.

Para la variable Volumen_prostata

```r
{r}
ggmice(data_imputada, aes(x = .imp, y = Volumen_prostata)) +
  geom_jitter(height = 0, width = 0.25) +
  geom_boxplot(width = 0.5, size = 1, alpha = 0.55, outlier.shape = NA) +
  labs(x = "Número de imputación")
```



Interpretacion:

El gráfico muestra cómo se distribuyen los valores del volumen prostático antes y después de la imputación. A la izquierda, en la posición 0, están los valores reales observados (en azul), y del 1 al 20 se muestran las imputaciones realizadas por el modelo (en rosado).

La mayoría de los valores imputados se mantienen dentro del rango general de los datos originales. Aunque algunos boxplots muestran mayor dispersión y unos pocos valores extremos (especialmente entre las imputaciones 10 y 14), en general la forma y el centro de los datos imputados son parecidos a los observados.

Para datos categóricos, podemos crear una tabla de dos entradas comparando la distribución de la variable con datos completos e incompletos. Esto requiere primero crear la versión "long" de la data imputada.

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

Ahora la tabla.

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

prop.table(
  table(data_imputada_l$Gleason_biopsia, data_imputada_l$imputed),
  margin = 2
)
```

```
              Observado  Imputado
  Gleason 0-6  0.6019108 0.5988924
  Gleason 7    0.2961783 0.2968354
  Gleason 8-10 0.1019108 0.1042722
```

Interpretacion:

La tabla muestra la proporción de cada categoría de la variable Gleason_biopsia en los datos observados y en los datos imputados. Los resultados son muy parecidos entre ambos grupos.

Por ejemplo, el 60% de los datos observados eran Gleason 0-6, y en los imputados también fue casi 60%. Lo mismo pasa con Gleason 7, que aparece en el 29.6% de los casos tanto en observados como imputados. Para Gleason 8-10, la diferencia es mínima: 10.2% en imputados y 10.1% en observados.

Esto indica que la imputación fue adecuada, ya que mantuvo la distribución original de la variable.

5.1 Procedimientos adicionales luego de la imputación

Una vez realizada la imputación, se puede hacer una regresión logística multivariada usando los datos imputados. Como estás usando el paquete `gtsummary`, no necesitas usar la función `pool()` por separado, ya que `gtsummary` ya combina los resultados de las múltiples imputaciones de forma automática.

```{r}
tabla_multi <-
  data_imputada |>
  with(glm(Recurrencia_bioquimica ~ Edad + Raza_afroamericana + Historia_familiar +
             Estadio_T + Gleason_biopsia + PSA_preoperatorio + Volumen_prostata +
Terapia_previa,
           family = binomial(link = "logit"))) |>
  tbl_regression(
    exponentiate = TRUE,
    label = list(
      Raza_afroamericana ~ "Raza afroamericana",
      Historia_familiar ~ "Antecedente familiar",
      Estadio_T ~ "Estadio clínico T",
      Gleason_biopsia ~ "Gleason biopsia",
      PSA_preoperatorio ~ "PSA preoperatorio (ng/mL)",
      Volumen_prostata ~ "Volumen prostático (cm³)",
      Terapia_previa ~ "Terapia previa",
      Edad ~ "Edad"
    )
  ) |>
  bold_p(t = 0.05) |>
  modify_header(estimate = "**OR ajustado**", p.value = "**p valor**")
```

```{r}
tabla_multi
```

```{r}
tabla_multi
```

Characteristic	OR ajustado	95% CI	p valor
Edad	1.03	0.98, 1.08	0.3
Raza afroamericana			
No	—	—	
Sí	2.05	0.93, 4.52	0.074
Antecedente familiar			
No	—	—	
Sí	0.48	0.18, 1.28	0.14
Estadio clínico T			
T1-T2a	—	—	

```{r}
tabla_multi
```

T2b-T3	1.52	0.59, 3.91	0.4
Gleason biopsia			
Gleason 0-6	—	—	
Gleason 7	2.64	1.20, 5.81	**0.016**
Gleason 8-10	6.44	2.34, 17.7	**<0.001**
PSA preoperatorio (ng/mL)	1.05	1.00, 1.10	0.067
Volumen prostático (cm^3)	1.00	0.98, 1.01	0.5
Terapia previa			
No	—	—	
Sí	1.83	0.74, 4.49	0.2

Abbreviations: CI = Confidence Interval, OR = Odds Ratio

Interpretacion:

En general, la mayoría de las variables no mostraron diferencias estadísticamente significativas, ya que sus valores p son mayores a 0.05. Sin embargo, hay dos excepciones importantes en la variable Gleason biopsia.

Los pacientes con Gleason 7 tienen un riesgo más de dos veces mayor de recurrencia en comparación con los que tienen Gleason 0–6. El valor p es 0.016, por lo tanto es significativo.

Para los pacientes con Gleason 8–10, el riesgo es todavía más alto, más de seis veces mayor, y el resultado es muy significativo con un valor p menor a 0.001.

El resto de las variables como edad, raza, antecedentes familiares, estadio clínico, PSA, volumen prostático y terapia previa no mostraron asociaciones significativas. Esto quiere decir que en este análisis, el valor de Gleason en la biopsia fue el factor que más se relacionó con el riesgo de recurrencia bioquímica.