Comparación de los métodos de imputación con respecto al poder de separación del modelo de regresión logística

Por

Víctor López Vázquez

Tesis sometida en cumplimiento parcial de los requisitos para el grado de

MAESTRO EN CIENCIAS
en
MATEMÁTICAS
(Estadística)

UNIVERSIDAD DE PUERTO RICO
RECINTO UNIVERSITARIO DE MAYAGUEZ
Junio 2005

Aprobado por:

Edgardo Lorenzo, Ph.D. Miembro, Comité Graduado	Fecha
Tokuji Saito, Ph.D. Miembro, Comité Graduado	Fecha
Julio C. Quintana, Ph.D. Presidente, Comité Graduado	Fecha
Raúl E. Macchiavelli, Ph.D. Representante de Estudios Graduados	Fecha
Pedro Vásquez, D.Sc. Director de Departamento	Fecha

ABSTRACT

An **MCAR** (*Missing Completely at Random*) mechanism was used with different missing data proportions in order to generate iteratively missing values in some data sets obtained from the *Machine Learning Database Repository* at the University of California, Irvine, to compare the efficiency of single, hot deck, and multiple imputation techniques in a logistic regression model. The parameter of interest in these comparisons is the separation power of the logistic regression model obtained by the area under the Receiver Operating Characteristic (**ROC**) curve. We are implementing unconditional and conditional mean, median, and mode (IMEAN, ICMEAN, IMED, ICMED, IMOD, ICMOD) as the single imputation methods. And for the Hot-Deck imputation, we used the unconditional and conditional random sampling of the observed values (IRS, ICRS), and the k^{th} nearest neighbor imputation (KNN). The multiple one is the FRITZ (Federal Reserve Imputation Technique Zeta) algorithm implemented by [Kennickell, 1991] on the SCF (Survey of Consumer Finances). Several iterations for the separation power were obtained after a generation of missing data with a given proportions, and then fill-in these missing values by some imputation method. The average bias between the real separation power and the separation power for all the iterations was calculated for all the imputation methods and some missing data proportions. The testing of these estimated biases were made by using non-parametric comparison procedures. From these testing we have found that the ICRS technique generate the minor bias on the area under the ROC curve. Also, we found that under a MCAR mechanism there are imputation methods that have a good performance at proportions of missing data higher than 15 %.

RESUMEN

Un mecanismo MCAR (Datos faltantes por completa aleatoridad) se utilizó con diferenes proporciones de datos faltantes para generar recurrentemente valores faltantes en algunos conjuntos de datos obtenidos del *Machine Learning Database Repository de la Universidad de California en Irvine* con el propósito de comparar la eficiencia de técnicas de imputación sencilla, *hot deck* y múltiple en un modelo de regresión logística. El parámetro de interés en estas comparaciones es el poder de separación del modelo de regresión logística obtenido por el área bajo la curva *Receiver Operating Characteristic* (**ROC**). Los métodos de imputación simple que se implantaron fueron la media, mediana y moda incondicionales y condicionales (**IMEAN, ICMEAN, IMED, ICMED, IMOD, ICMOD**). Para la imputación *hot deck* se usó el muestreo aleatorio incondicional y condicional de los valores observados (**IRS, ICRS**) y el método por el k-ésimo vecino más cercano (**KNN**). El método múltiple usado fue el algoritmo **FRITZ** (**Federal Reserve Imputation Technique Zeta**) implantado por *Arthur B. Kennickell* en la encuesta **SCF** (*Survey of Consumer Finances*) [Kennickell, 1991, Kennickell, 1998].

Se obtuvieron recurrentemente estimados del poder de separación después de generarse datos faltantes con proporciones dadas y luego se sustituyeron por valores imputados por los distintos métodos. Se calculó el sesgo promedio entre el poder de separación real y el poder de separación estimado en todas las recurrencias, para todos los métodos de imputación y para algunas proporciones de datos faltantes. Las pruebas estadísticas de estos sesgos se hicieron usando procedimientos de comparación no paramétricos. De estas pruebas se encontró que la técnica **ICRS** genera el menor sesgo en el área bajo la curva **ROC**. También se encontró que bajo un mecanismo **MCAR** hay métodos de imputación que tienen una buena ejecución en proporciones de datos faltantes mayores del 15 %.

DEDICATORIA

A la persona más especial de mi vida ... mí esposa Yolanda

... por su amor, compañía, paciencia y apoyo incondicional ininterrumpido durante estos últimos cinco años y contando ...

AGRADECIMIENTOS

Al doctor Julio Quintana por su gran ayuda, paciencia y motivación y por haberme abierto muchas puertas en mi vida académica y profesional. Y por su inmensa ayuda en la redacción de este documento.

A mi maestro SiFu David González por estar siempre presente con su apoyo y por todas sus enseñanzas en pro de obtener una vida más integral y balanceada física, moral y espiritualmente. Gracias a usted he sabido tomar decisiones en mi vida de las cuales no me arrepiento.

A mi familia: mis padres Margarita y Alfredo y hermanos Javier y Alfredo (Dady) por su grandiosa aportación en mi formación moral y por su incondicional apoyo.

A mí comite graduado por su valiosa aportación a este documento.

A Caroline Rodríguez por su ayuda desde el bachillerato y sus valiosos consejos.

A mis amigos de siempre: Wilkins, Esbal, Billy, Kelvin, Elliot y Egui.

A mis nuevos y eternos compañeros y amigos de maestría: Santiago, Karen, José, Marggie, Alejo, Milena, Viviana, Edgardo Álvarez y Ángel por su ayuda y compañía incondicional y por también estar presente en todo momento.

A mis demás compañeros de maestría por todos esos agradables momentos.

Al Departamento de Matemáticas por su apoyo.

Tabla de Contenido

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

A través de los años el problema de los valores faltantes en un conjunto de datos ha cobrado gran interés en el análisis estadístico de varios estudios. Varios investigadores se han dado a la tarea de describir la causa de este problema y su efecto en diversos análisis estadísticos. En esta tesis se estudió el problema de datos faltantes en el modelo de regresión logística particularmente en el poder de separación del modelo obtenido mediante el cálculo del área bajo la curva **ROC** (del inglés *Receiver Operating Characteristic*). Esta área es utilizada como una medida de bondad de ajuste del modelo [Bradley, 1996, Le, 2003]. Para manejar este problema existen varias alternativas tales como; métodos ponderados, la estimación de parámetros basados en un modelo o simplemente la eliminación de aquellas unidades con valores faltantes. Esta última opción es conocida como análisis con datos completos (ADC) y es a su vez la opción estándar en la mayoría de los programados estadísticos existentes. En esta tesis la alternativa focal lo constituye los métodos de imputación. El interés por estas técnicas proviene de la idea de que al imputar se obtiene un conjunto de datos completos, permitiendo llevar a cabo un análisis con resultados de fácil interpretación y presentación aunque éstos pueden ser significativamente sesgados dependiendo del método de imputación utilizado y del parámetro que se esté estimando [Kalton and Kasprzyk, 1982]. Otro factor de interés para llevar a cabo esta tesis es la aplicabilidad del modelo de regresión logística en ramas importantes de la ciencia y de la medicina donde se generan con-

juntos de datos en donde usualmente se confronta el problema de datos faltantes. [Vach and Blettner, 1999, Perez et al., 2002].

En esta tesis se comparan empíricamente varios métodos de imputación y el ADC con respecto al poder de separación del modelo de regresión logística obtenido mediante la curva **ROC**. Se utilizaron cuatro conjuntos reales tomados del *Machine Learning Database Repository* de la Universidad de California, Irvine. Estos conjuntos originalmente no contienen datos faltantes por lo que se generaron los mismos con un mecanismo completamente aleatorizado **MCAR**. Luego, en el conjunto generado con datos faltantes, se aplicó cada método de imputación y se halló un poder de separación para cada uno de ellos. De estos estimados se calculó una diferencia o sesgo absoluto con el poder de separación de los datos originales. Este proceso de remover, imputar y estimar un poder de separación y luego calcular un sesgo se llevó a cabo varias veces generando varias simulaciones para cada método de imputación incrementando cada vez más la proporción de datos faltantes en cada conjunto de datos desde un 5 % hasta un 50 %. De ahí los métodos de imputación se compararon utilizando el método no paramétricos de Friedman con respecto a los sesgos entre los estimados del poder de separación de cada método y el poder de separación de los datos originales. Con esto se pretende alcanzar los siguientes objetivos:

- Mediante la prueba no paramétrica de Friedman, establecer si existe o no alguna diferencia significativa entre los métodos de imputación y el análisis con datos completos a través de todas las proporciones de datos faltantes en cada conjunto de datos con respecto al sesgo en el poder de separación.

- Mediante la comparación múltiple por las sumas de rangos de Friedman, establecer cuales métodos de imputación generan un sesgo menor con respecto a los estimados del poder de separación a través de todas las proporciones de datos faltantes en cada conjunto de datos.

- Mediante la prueba no paramétrica de Friedman, establecer si existe o no diferencia significativa entre los métodos de imputación y el análisis con

datos completos con respecto al sesgo en cada proporción dada de datos faltantes en cada conjunto de datos.

- Mediante la comparación múltiple por las sumas de rangos de Friedman, establecer qué métodos de imputación generan un sesgo menor con respecto a los estimados del poder de separación en cada proporción de datos faltantes en cada conjunto de datos.
- Examinar la variabilidad de los estimados del poder de separación de cada método de imputación en cada base de datos.
- Establecer una relación entre el sesgo y el poder de separación utilizando los estimados calculados a través de todas las bases de datos.

Además de las simulaciones, se llevó a cabo una parte de aplicación en donde se pusieron a prueba los métodos de imputación en un conjunto que contenía originalmente valores faltantes. Este conjunto de datos también proviene del *Machine Learning Database Repository* de la Universidad de California, Irvine.

En la literatura se han encontrado escritos que tratan de la comparación de métodos de imputación en diversos aspectos estadísticos tanto en estadísticos univariados simples [Herzog and Rubin, 1983] como bivariados [Santos, 1981, Kalton and Kasprzyk, 1982, Jinn, 2000]. También se han encontrado escritos donde se comparan métodos de imputación en la ejecución del modelo de regresión logística [Perez et al., 2002] y otros que tratan sobre el desempeño del modelo de regresión logística en conjuntos con datos faltantes [Iannacchione, 1999]. En ambos casos se utiliza el área bajo la curva ROC como medida del desempeño del modelo. En *Bradley (1996)* se utilizó el área bajo la curva ROC para comparar métodos de clasificación de forma empírica usando un experimento con los métodos de imputación análogos al de esta tesis con los métodos de imputación. La diferencia estriba en que Bradley utilizó análisis de varianza para comparar métodos de clasificación con respecto al área bajo la curva **ROC** directamente. Hasta el momento no se ha encontrado una comparación empírica de los métodos de imputación similar a la expuesta en esta tesis, donde se comparan ocho métodos de imputación

incluyendo el análisis con datos completos en el modelo de regresión logística utilizando métodos no paramétricos como Friedman.

Esta tesis está organizada de la siguiente forma. En el Capítulo 2 se describe el modelo de regresión logística y la teoría envuelta en la curva ROC y su relación con el modelo de regresión logística. En el Capítulo 3 se describe más a fondo el problema de datos faltantes en el desempeño de la regresión logística. Además, se describen los mecanismos que llevan o causan ausencia de datos y una descripción general de las soluciones existentes al problema de datos faltantes. En el Capítulo 4 se describen los métodos de imputación utilizados en esta tesis y sus ventajas y desventajas según la literatura encontrada. En el Capítulo 5 se describe en detalle la metodología a seguir para la comparación de los métodos de imputación a través de las simulaciones. En el Capítulo 6 se muestran los resultados del experimento y otros resultados para cumplir con los objetivos de la tesis y el análisis de los mismos respectivamente. En el Capítulo 7 se encuentra una parte de aplicación donde se comparan los métodos de imputación en un conjunto donde originalmente existen datos faltantes. Por último el Capítulo 8 muestra la discusión de los resultados, las conclusiones y algunas recomendaciones para investigaciones futuras.

1.1. Definición de términos

Esta sección presenta las definiciones de la abreviaturas utilizadas a través de la tesis. Aunque estos términos se definen a través de los capítulos, el lector puede referirse a esta sección en vez de buscar a través de todos los capítulos cuando quiera referirse hacia alguna abreviatura de algún término.

- **ROC**: del inglés *receiver operating characteristic*.
- **ABC**: área bajo la curva ROC calculada mediante la Regla Trapezoidal.
- **AMW**: área bajo la curva ROC calculada mediante el estadístico de Mann-Whitney, también conocido como el estimado de Wilcoxon.
- **Sesgo**: en esta tesis se refiere a la diferencia absoluta entre algún estimado del área bajo la curva ROC calculado luego de borrar e imputar datos en el conjunto y el valor del área bajo la curva ROC calculado de los datos originales.
- **MCAR**: del inglés *Missing Completely at Random* y se refiere al mecanismo de datos faltantes por completa aleatoreidad.
- **MAR**: del inglés *Missing at random* y se refiere al mecanismo de datos faltantes por aleatoreidad.
- **NMAR**: del inglés *Not missing at random* se refiere al mecanismo de datos faltantes que no se debe a aleatoridad. También se refiere cuando el mecanismo de datos faltantes no se puede ignorar.
- **ADC**: se refiere al análisis con datos completos, es decir cuando las unidades que contienen algún dato faltante para alguna de sus variables es eliminada. Sólo se toman en consideración las unidades completamente observadas.
- **IMEAN**: Imputación por el promedio de las observaciones en una variable.
- **ICMEAN**: Imputación por el promedio de las observaciones en una variable, condicional a las clases de la variable de respuesta.

- **IMED**: Imputación por la mediana de las observaciones en una variable.
- **ICMEAN**: Imputación por la mediana de las observaciones en una variable, condicional a las clases de la variable de respuesta.
- **IRS**: Imputación por muestreo aleatorio de las observaciones en una variable.
- **ICRS**: Imputación por muestreo aleatorio de las observaciones en una variable, pero condicional a las clases de la variable de respuesta.
- **KNN**: Imputación por el método de los k vecinos más cercanos (k^{th} *nearest neighbors*).
- **FRITZ**: Método de imputación múltiple utilizando el algoritmo establecido por Kennickell para la encuesta *Survey of Consumer Finances*. Las siglas del método provienen del inglés *Federal Reserve Imputation Technique Zeta*.

Capítulo 2

El modelo de regresión logística y la curva ROC

2.1. Introducción

En esta tesis se busca mostrar el efecto de los métodos de imputación en la regresión logística. Para es necesario tener clara cierta terminología sobre del modelo de regresión logística y el concepto de la curva **ROC** que consiste en una prueba no paramétrica de bondad de ajuste en la que se basan las comparaciones de los métodos de imputación que se describen más adelante. En este capítulo se describe el modelo de regresión logística, el cual tiene muchas aplicaciones, particularmente en la medicina y bioestadística. Un ejemplo de la aplicabilidad de la regresión logística es el modelo APACHE [Perez et al., 2002] que se utilizó para estimar la probabilidad de supervivencia en pacientes de cuidado intensivo en Colombia. A continuación presentaremos una descripción del modelo de regresión logística según la revisión literaria.

2.2. El modelo de regresión logística

Los métodos de regresión son una parte integral dentro del análisis estadístico de un conjunto de datos. Con ellos se busca describir una relación entre una variable de respuesta y una o varias variables explicativas. El modelo más conocido es el de regresión lineal donde la variable de respuesta es de tipo contínuo. Por otro lado, en

muchos casos la variable de respuesta puede ser de tipo discreto lo cual da lugar a un tipo de modelo distinto conocido como regresión logística. Las diferencias entre el modelo logístico y el lineal estriban en la selección del modelo paramétrico y de los supuestos de las variables, en especial la de respuesta. No obstante los principios generales utilizados en la regresión logística son similares a los de la regresión lineal [Hosmer and Lemeshow, 1989].

En esta tesis se analiza el caso en donde la variable de respuesta es una variable dicotómica o binaria, es decir que $\mathbf{Y}$ es una variable cuya distribución es de la forma

$$\boldsymbol{Pr(Y = y) = \pi^{y}(1 - \pi)^{1-y_i}} \tag{2.1}$$

donde $\boldsymbol{y = 0, 1}$.

Para un individuo i de una muestra ($\boldsymbol{i = 1, 2, \ldots, n}$), $\boldsymbol{Y_i}$ es una variable de *Bernoulli* con

$$\boldsymbol{Pr(Y_i = y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}} \tag{2.2}$$

Así, un modelo lineal no sería apropiado para ajustar la variable binaria $\mathbf{Y}$ con las variables explicativas $\mathbf{X}$. Para ajustar las variables se necesita un modelo de regresión logística que tiene la forma

$$\boldsymbol{\pi_i = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij})}}} \tag{2.3}$$

para $\boldsymbol{i = 1, 2, \ldots, n}$ donde

$$\boldsymbol{ln(\frac{\pi_i}{1 - \pi_i}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}} \tag{2.4}$$

La función logística básica está dada por

$$\boldsymbol{f(z_i) = \frac{1}{1 + e^{-z_i}}} \tag{2.5}$$

donde $\boldsymbol{z_i}$ se puede expresar como un modelo lineal de la forma

$$\boldsymbol{z_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j}} \tag{2.6}$$

El modelo de regresión logística se ha vuelto popular debido a dos razones principales [Le, 2003]:

1. El alcance de la función logística es el intervalo [0,1]; lo cual hace factible el uso de un modelo probabilístico para representar un riesgo individual en las unidades de observación.

2. La curva logística tiene una forma sigmoidal con un umbral que permite la aplicación a modelos biológicos, representando el riesgo de un individuo con respecto a ciertos factores a los cuales está expuesto.

Los parámetros del modelo se estiman mediante la siguiente función de máxima verosimilitud

$$\boldsymbol{L = \prod_{i=1}^{n} \frac{[e^{\beta_0 + \sum_{i=1,k} \beta_j x_{ji}}]^{y_i}}{1 + e^{\beta_0 + \sum_{i=1,k} \beta_j x_{ji}}}} \tag{2.7}$$

donde $\boldsymbol{y_i = 0, 1}$.

Los coeficientes del modelo no tienen una interpretación directa sino al evaluarse como $\boldsymbol{e^{\beta_i}}$, donde esta expresión representa un *odd ratio* para el cual se tienen dos interpretaciones dependiendo el tipo de variable al cual corresponda el coeficiente:

1. Si la variable asociada $\boldsymbol{X_i}$ es binaria entonces $\boldsymbol{e^{\beta_i}}$ es el *odd ratio* asociado con la exposición a $\boldsymbol{X_i}$ (expuesto si $\boldsymbol{X_i = 1}$ y no expuesto si $\boldsymbol{X_i = 0}$) cuando las demás variables permanecen constantes.

2. Si la variable asociada $\boldsymbol{X_i}$ es continua entonces $\boldsymbol{e^{\beta_i}}$ es el *odd ratio* dado que $\boldsymbol{X_i}$ aumente en una unidad ($\boldsymbol{X_i = x + 1}$ **vs.** $\boldsymbol{X_i = x}$) cuando las demás variables permanecen constantes.

La idea del modelo es identificar aquellas variables explicativas o factores de riesgo importantes para describir un evento representado en la variable de respuesta como por ejemplo el padecer enfermedad o condición dependiendo en el contexto en que se esté hablando. La hipótesis en la que se enfocará esta tesis es aquella que

involucra a todas las variables explicativas, es decir, lo que se va a someter a prueba es si todas las $\boldsymbol{p}$ variables independientes explican en conjunto la variabilidad de la variable de respuesta. Las hipótesis a probar son la siguientes:

- $\boldsymbol{H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0}$
- $\boldsymbol{H_a}$**: Al menos alguna es diferente de cero.**

En este caso no se buscará someter a prueba esta hipótesis desde el punto de vista usual, es decir utilizando *la razón de máxima verosimilitud* ni la *prueba de scores*, [Le, 2003], sino que se utilizará una prueba no paramétrica de bondad de ajuste basada en el área bajo la curva **ROC** (del inglés *Receiver Operating Characteristic*). A esta área se le conoce como el *poder de separación* del modelo ajustado entre las variables explicativas y la variable de respuesta, que en este caso es binaria.

Este método no prueba directamente la hipótesis planteada anteriormente pero sí una variación de la misma. Cuando se busca determinar si un conjunto de variables explicativas explica la variabilidad de una variable binaria de dos eventos a través de un modelo, se busca en realidad determinar si las variables explicativas actúan como buenos separadores de esos eventos en la variable de respuesta. Es por esto que el poder de separación actúa como una prueba de bondad de ajuste del modelo de regresión logística y en este caso es apropiado para probar la hipótesis planteada.

En la siguiente sección se discutirá más a fondo en que consiste esta prueba no paramétrica de bondad de ajuste y la curva **ROC**.

2.3. La curva ROC

La teoría detrás de la curva **ROC** incluye una gama de conceptos de los que sólo se mencionarán aquellos que competen para cumplir con los objetivos de la tesis con respecto al modelo de regresión logística. Primero se describirá el proceso y la información requerida para construir esta curva y la razón por la que puede usarse como prueba de bondad de ajuste en el modelo de regresión logística.

La variable de respuesta $\mathbf{Y}$ induce dos clases $\boldsymbol{C_0}$ y $\boldsymbol{C_1}$ compuestas por n y m elementos respectivamente y donde la clase de interés es $\boldsymbol{C_1}$. Cuando se calcula el modelo de regresión logística cada valor ajustado tiene una probabilidad asociada, donde $\boldsymbol{\pi_i}$ es el valor ajustado para el *i-ésimo* individuo de la clase $\boldsymbol{C_1}$ y $\boldsymbol{\gamma_j}$ es el valor ajustado para *el j-ésimo* individuo de la clase $\boldsymbol{C_0}$. Es decir que si un valor de la variable de respuesta $\boldsymbol{y_i}$ es de la clase $\boldsymbol{C_1}$, entonces al ser ajustado por el modelo tendrá una probabilidad asociada $\boldsymbol{\pi_i}$. De manera similar un valor $\boldsymbol{y_j}$ de la clase $\boldsymbol{C_0}$ al ser ajustado por el modelo tendrá una probabilidad $\boldsymbol{\gamma_j}$.

La curva **ROC** está basada en las pruebas de ventanas que se derivan de los valores de las clases y de los valores ajustados por el modelo. Estas pruebas de ventana se conocen por *sensitividad* y la *especificidad* y ambas se definen como sigue.

Dado un número real $\boldsymbol{c}$ tenemos que:

$$\boldsymbol{sen_c} = \frac{\mathbf{1}}{\boldsymbol{m}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{I}(\boldsymbol{\pi_i} \geq \boldsymbol{c}) \tag{2.8}$$

$$\boldsymbol{esp_c} = \frac{\mathbf{1}}{\boldsymbol{n}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{I}(\boldsymbol{\gamma_j} < \boldsymbol{c}) \tag{2.9}$$

donde $\mathbf{I(S)}$ es una función indicadora que toma valor de 1 cuando el enunciado $\mathbf{S}$ es verdadero y valor de 0 cuando el enunciado $\mathbf{S}$ es falso, $\mathbf{m}$ y $\mathbf{n}$ representan la cantidad de individuos en la clase $\boldsymbol{C_1}$ y $\boldsymbol{C_0}$ respectivamente.

En otras palabras la sensitividad es la razón entre los elementos bien clasificados por el modelo para la clase de interés $\boldsymbol{C_1}$ y el número de individuos que en realidad pertenecen a la clase $\boldsymbol{C_1}$. La especificidad es análoga a la sensitividad, lo único que para la otra clase. También se puede interpretar la sensitividad como la habilidad del modelo para detectar una condición o enfermedad en personas que padecen realmente la condición. De manera similar, la especificidad es la habilidad del modelo para detectar ausencia de enfermedad en personas o pacientes que en realidad no padecen la condición.

Entonces para construir la curva **ROC** se grafican los pares ordenados $(\mathbf{1} - \boldsymbol{esp_c}, \boldsymbol{sen_c})$,

es decir los verdaderos positivos, ($\boldsymbol{FP_c}$), versus los falsos positivos, ($\boldsymbol{VP_c}$), del modelo. Estos se grafican tomando varios puntos o probabilidades de corte $\boldsymbol{c}$ para los valores ajustados del modelo de regresión. En este caso se escogen probabilidades de corte que van desde 0 hasta 1 variando por 0.01, por lo que los pares ordenados son de la forma ($\boldsymbol{FP_c, VP_c}$) para $\boldsymbol{c = 0, 0.01, 0.02, \ldots, 0.99, 1.}$

Para mostrar un ejemplo de la construcción de la curva, se tomó la base de datos *Estudiantes* donde la variable de respuesta (PoF) indica si un estudiante aprobó (P) o no aprobó (F) su primer curso de matemáticas de la Universidad de Puerto Rico - Mayaguez (U.P.R.M.). Las variables explicativas corresponden al historial de la escuela superior y las pruebas de admisión a la universidad del estudiante. Se ajustó un modelo de regresión logística con la ayuda del programado **R** utilizando la siguiente instrucción;

```
> estudiantes = read.table("D:/tesis/Investigacion/basesdatos/DatEst.txt",
+     header = T)
> model = glm(PoF ~ ., data = estudiantes, family = binomial)
```

La curva **ROC** para este modelo se presenta en la FIGURA 2.1. Para calcular los **VP's** y los **FP's** se diseñó en **R** la función; *ROC*.

La importancia de la curva **ROC** no estriba en la curva en sí sino en el área bajo la misma, la cual representa *el poder de separación del modelo* (**ABC**, Área Bajo la Curva). Un modelo donde la curva **ROC** se acerca a la diagonal o pasa por debajo de ella se considera inapropiado y llevaría a la conclusión de que las variables explicativas utilizadas en el mismo no son buenos separadores de las clases. Para estimar el área bajo la curva ROC existen varias formas. Una de ella es el proceso de estimación de áreas por la *Regla Trapezoidal* la cual se utilizó en las simulaciones y está dada por la siguiente ecuación:

$$\boldsymbol{ABC} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \frac{(\boldsymbol{FP_{k+1}} - \boldsymbol{FP_k})(\boldsymbol{VP_k} + \boldsymbol{VP_{k+1}})}{\boldsymbol{2}} \tag{2.10}$$

donde $\boldsymbol{FP_{k+1}}$ y $\boldsymbol{VP_{k+1}}$ son los cálculos de la proporciones de falsos positivos y verdaderos positivos respectivamente para el modelo para la $\boldsymbol{k+1}$ proba-

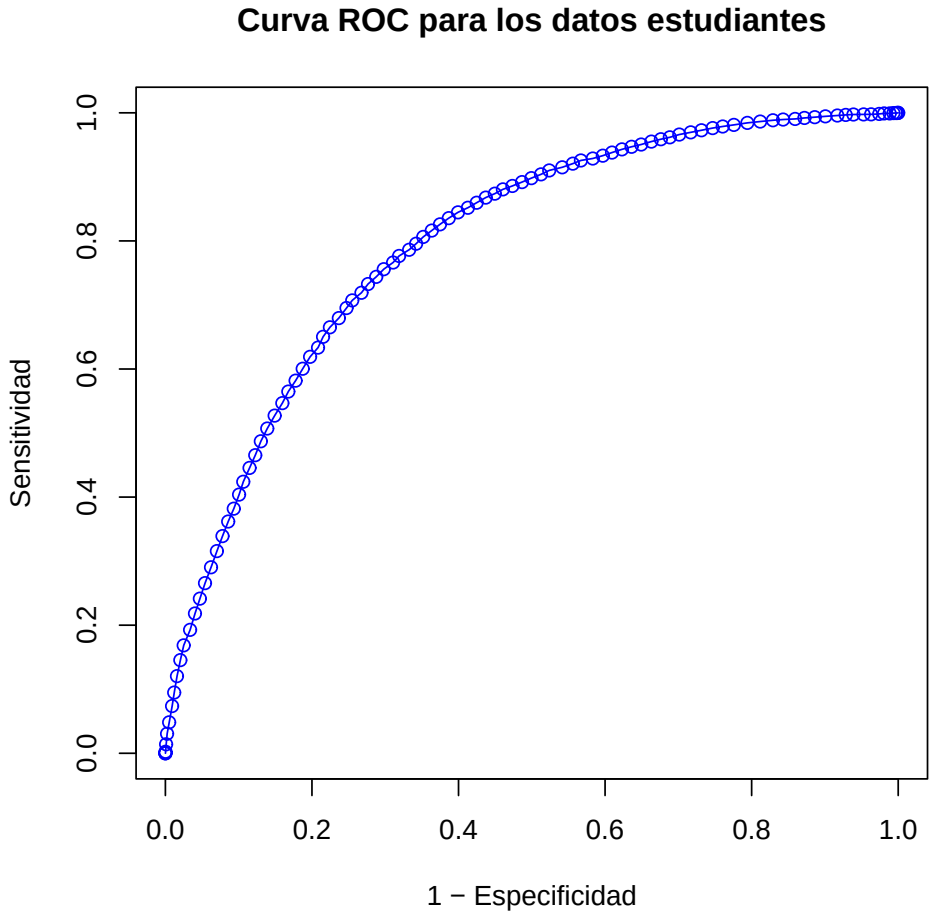


Figura 2.1: **Curva ROC para el modelo de regresión logística ajustado para los datos *estudiantes***

bilidad de corte. $\boldsymbol{FP_k}$ y $\boldsymbol{VP_k}$ son los valores para la *k-ésima* probabilidad de corte.

La estimación del *poder de separación* mediante la *Regla Trapezoidal* es equivalente al estimado de Mann-Whitney (**AMW**)[Delong et al., 1998] o también llamado el estimado de Wilcoxon el cual representa la probabilidad de que $\boldsymbol{\gamma_j}$ seleccionada aleatoriamente de la clase $\boldsymbol{C_0}$ sea menor o igual que una probabilidad $\boldsymbol{\pi_i}$ seleccionada aleatoriamente de la clase $\boldsymbol{C_1}$ [Hanley and McNeil, 1982]. Si se toma $\boldsymbol{\theta}$ como el área bajo la curva y a $\hat{\boldsymbol{\theta}} = \boldsymbol{AMW}$ como el estimado de la misma, el cálculo de $\hat{\boldsymbol{\theta}}$ es el siguiente:

$$\hat{\boldsymbol{\theta}} = \frac{\mathbf{1}}{\boldsymbol{mn}} \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{m}} \boldsymbol{\psi}(\boldsymbol{\pi_i}, \boldsymbol{\gamma_j}) \tag{2.11}$$

donde

$$\boldsymbol{\psi}(\boldsymbol{\pi}, \boldsymbol{\gamma}) = \begin{cases} \mathbf{1} & \text{si } \boldsymbol{\gamma} < \boldsymbol{\pi}, \\ \frac{\mathbf{1}}{\mathbf{2}} & \text{si } \boldsymbol{\gamma} = \boldsymbol{\pi}, \\ \mathbf{0} & \text{si } \boldsymbol{\gamma} > \boldsymbol{\pi}. \end{cases} \tag{2.12}$$

En términos de probabilidades $\boldsymbol{E}(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta} = \boldsymbol{Pr}(\boldsymbol{\gamma} < \boldsymbol{\pi}) + \frac{\mathbf{1}}{\mathbf{2}}\boldsymbol{Pr}(\boldsymbol{\gamma} = \boldsymbol{\pi})$ lo que indica que el estimado del *poder de separación* de *Mann-Whitney* es insesgado y por consiguiente su equivalente obtenido mediante la *Regla Trapezoidal* también lo es. Aunque el estimado por la Regla Trapezoidal tienda a subestimar un poco el área bajo la curva **ROC** para efectos prácticos las dos estimaciones son equivalentes. [Hanley and McNeil, 1982]

Continuando con el ejemplo de los datos *estudiantes*, el área bajo la curva estimada mediante la *Regla Trapezoidal* es la siguiente:

```
> ABC = Area(roc)
> ABC

[1] 0.7968355
```

En la FIGURA 2.2 se muestra la forma de la curva y el área bajo la misma. Examinando estos resultados se puede concluir que el modelo ajusta bien los

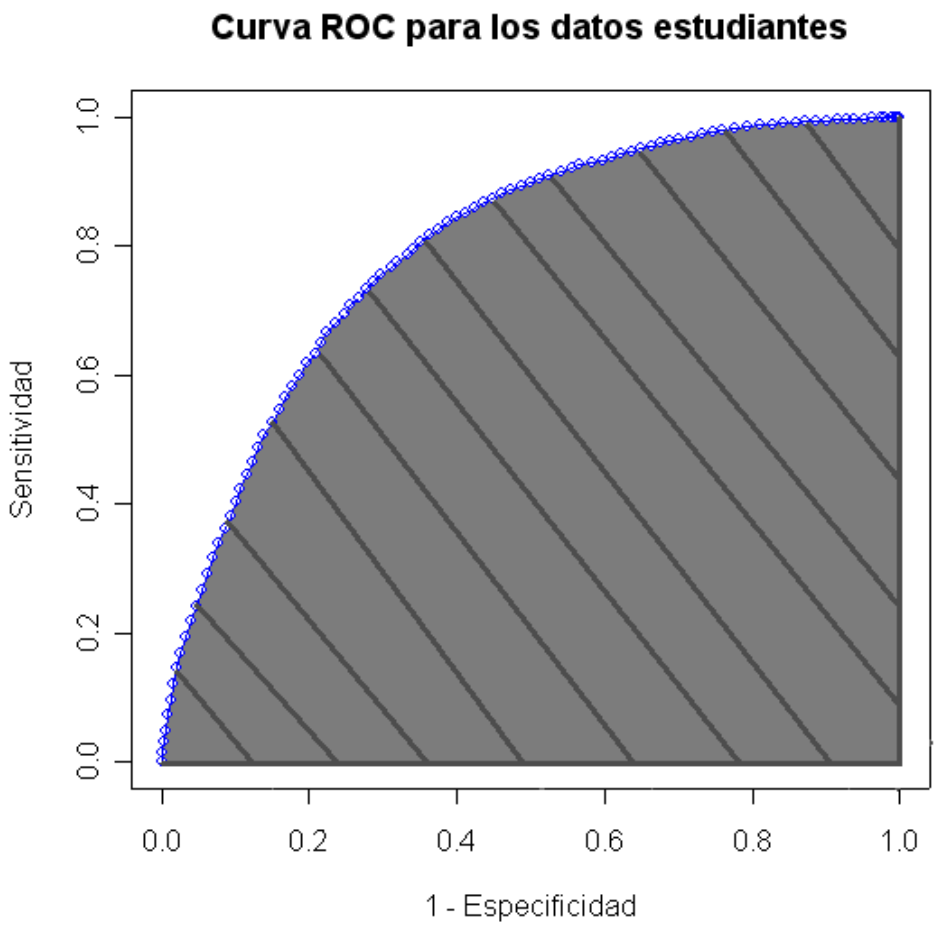


Figura 2.2: **Curva ROC con el área bajo la curva sombreada**

datos y que las variables explicativas separan de forma adecuada el que un estudiante apruebe o no su primer curso de matemáticas en la Universidad de Puerto Rico en Mayaguez. Aunque en la literatura no se ha encontrado o desarrollado algún intervalo de confianzapara el área bajo la curva ROC sí se ha conseguido acotar ésta medida. Las cotas del área bajo la curva ROC se presentan a continuación.

Para obtener las cotas del **ABC** se utilizarón las ideas presentadas en [Shapiro, 1998] donde se describe cómo obtener las cotas inferior y superior de dicho intervalo.

Para poder definir el las cotas es necesario encontrar el punto de corte óptimo ***O*** en la curva **ROC**. El punto ***O*** es aquel cuya distancia euclideana con el punto (0,1) sea la más pequeña, es decir, el punto o probabilidad de corte donde se alcanza un nivel de sensitividad y especificidad más cercano al cien por ciento de ambos. La gráfica en la FIGURA 2.3 presenta el punto óptimo de corte en la curva **ROC**.

El punto óptimo ***O*** se utiliza para calcular ambas cotas. La cota superior toma el valor del área bajo la recta tangente al punto de corte óptimo ***O***. El límite inferior se calcula con el área bajo los segmentos de recta que van desde el punto (0,0) al punto ***O*** y de ahí al punto (1,1). La recta tangente de la cota superior y los segmentos de la cota inferior se presentan en la gráfica de la FIGURA 2.4.

El punto de corte óptimo para el modelo en el caso de la base de datos *estudiantes* es el siguiente:

```
> O = Poptimo(roc)
> O

[1] 0.6
```

Las cotas para el **ABC** en este caso es;

```
> B = Bounds(roc)
> B
```

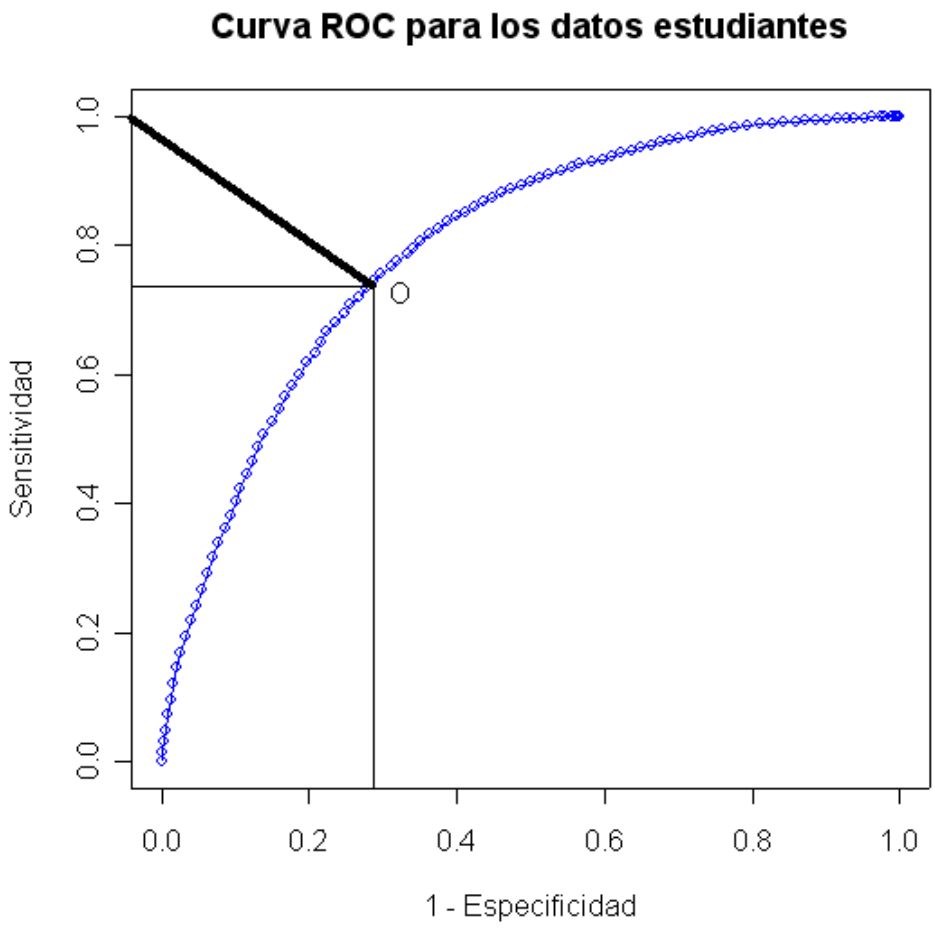


Figura 2.3: **Curva ROC y el punto de corte identificado el cual es el punto donde se alcanza el máximo nivel de sesitividad y especificidad**

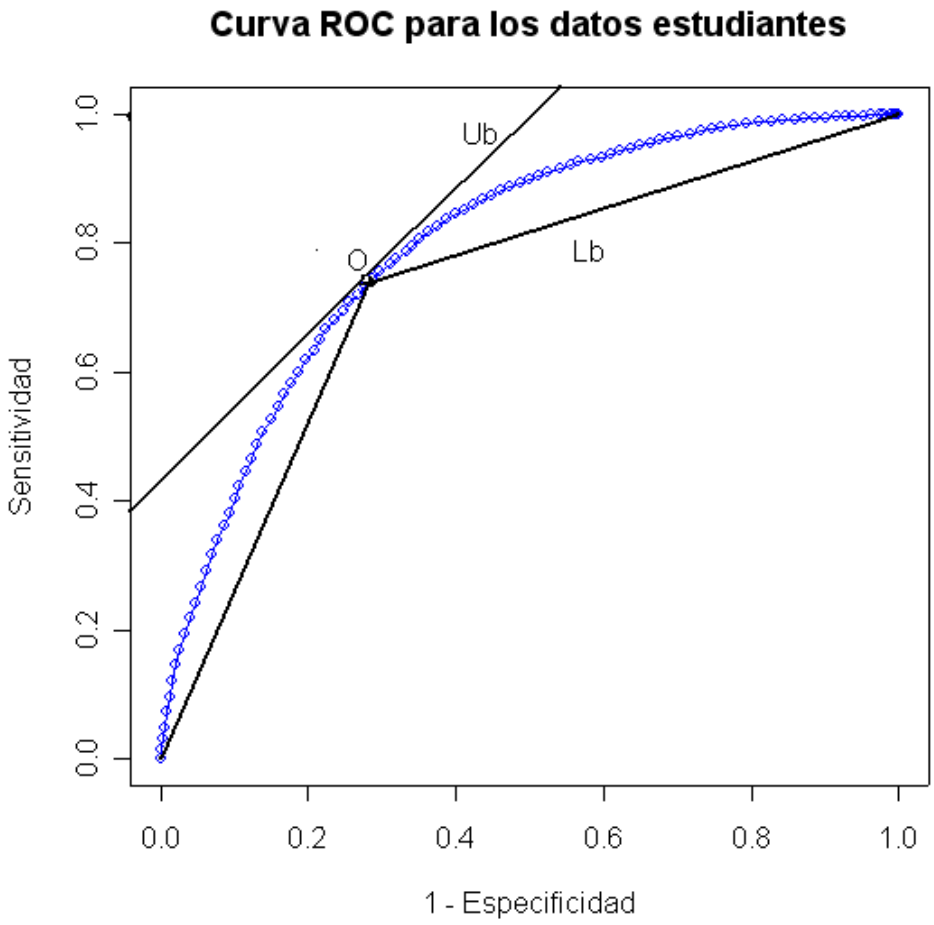


Figura 2.4: **Curva ROC y las cotas del poder de separación**

```
           Lb        Ub
[1,] 0.7280932 0.8506508
```

Otro aspecto importante del poder de separación es su error estándar el cual se calculó para el área bajo la curva **ROC** obtenida mediante el **AMW**. No obstante como el $\hat{\boldsymbol{\theta}} = \boldsymbol{AMW} \approx \boldsymbol{ABC}$ [Hanley and McNeil, 1982] el error estándar **SE** se calculará para el **ABC** de la siguiente forma:

$$\boldsymbol{SE}(\hat{\boldsymbol{\theta}}) = \sqrt{\frac{\hat{\boldsymbol{\theta}}(\mathbf{1}-\hat{\boldsymbol{\theta}}) + (\boldsymbol{n_1}-\mathbf{1})(\boldsymbol{Q_1}-\hat{\boldsymbol{\theta}}^{\mathbf{2}}) + (\boldsymbol{n_0}-\mathbf{1})(\boldsymbol{Q_2}-\hat{\boldsymbol{\theta}}^{\mathbf{2}})}{\boldsymbol{n_1 n_0}}} \tag{2.13}$$

donde:

- $\boldsymbol{Q_1} = \hat{\boldsymbol{\theta}} \div (\mathbf{2} - \hat{\boldsymbol{\theta}})$, representa en este caso la probabilidad de escoger aleatoriamente dos individuos de la clase $\boldsymbol{C_1}$ cuyas probabilidades de pertenecer a esa clase sean mayores que la de un individuo de la clase $\boldsymbol{C_0}$ escogido aleatoriamente.
- $\boldsymbol{Q_2} = \mathbf{2}\hat{\boldsymbol{\theta}}^{\mathbf{2}} \div (\mathbf{1} + \hat{\boldsymbol{\theta}})$, representa la probabilidad de escoger aleatoriamente un individuo de la clase $\boldsymbol{C_1}$ cuya probabilidad de pertenecer a esa clase sea mayor que la de dos individuos escogidos aleatoriamente de la clase $\boldsymbol{C_0}$.
- $\boldsymbol{n_0}$ es el número de individuos en la clase $\boldsymbol{C_0}$.
- $\boldsymbol{n_1}$ es el número de individuos en la clase $\boldsymbol{C_1}$.

Este error estándar se utilizará para comparar las áreas bajo la curva estimadas con los diversos métodos de imputación en los Capítulos 6 y 7. Para calcular del error estándar **SE** se diseñó la función *StdError* en el programado **R**. Continuando con los datos *estudiantes* el error estándar para el estimado del **ABC** es:

```
> SE = StdError(estudiantes$PoF, ABC)
> SE
```

```
[1] 0.003821757
```

Como observamos en este caso el **ABC $\approx$ 0.80** y el error estándar es de **SE** = 0.0038 lo cual resulta aceptable en este caso. Además la relación entre el **ABC** y el **SE** es inversa, es decir que a mayor **ABC** menor **SE** [Hanley and McNeil, 1983].

Ya explicada la teoría del modelo de regresión logística y el *poder de separación* del modelo se procederá a describir el problema que compete a esta tesis. En el próximo capítulo se presentará cómo afecta el problema de datos faltantes al *poder de separación* (**ABC**) del modelo de regresión logística.

Capítulo 3

El problema de los datos faltantes

3.1. Introducción

Los métodos estadísticos fueron desarrollados para analizar conjuntos rectangulares de datos que formen matrices donde las filas representan unidades, o también llamadas casos, observaciones o sujetos dependiendo del contexto, y las columnas representan variables o atributos asociados a esas unidades. El problema que confrontan muchos estudios surge cuando algunas de las entradas de esa matriz bajo investigación no son observables, ya sea por problemas en la recolección de la información, problemas con el equipo de investigación o, en los casos de encuestas o estudios con personas, por la acción de éstas al no responder a las preguntas ya sea por falta de información, desconocimiento, vergüenza o temor a contestar, entre otras. En los casos de encuestas es menos natural considerar los valores no observados como datos faltantes. En vez de esto se prefiere agrandar el espacio muestral y añadirle eventos que identifiquen los estratos de la población que no tienen preferencia o tengan desconocimiento del tema programado, esta opción está sujeta a condiciones de presupuesto y tiempo. Otro origen para el problema de datos faltantes es la edición de datos. En este proceso se busca de alguna forma identificar valores en las unidades para las cuales no existe coherencia a través

de los atributos o variables o a través de los grupos a los cuales pertenecen dichas unidades. Por lo tanto al encontrarse tal valor éste usualmente tiende a eliminarse dejando así un valor faltante en el conjunto de datos.

En este capítulo se mostrará cómo el problema de datos faltantes afecta al modelo de regresión logística en la investigación y al poder de separación del mismo. También se discutirá cuáles son las causas para tal problema y bajo qué mecanismos probabilísticos se rigen. Además se mostrarán algunas soluciones para lidiar con la falta de datos.

3.2. El problema de los datos faltantes y la regresión logística

Muchos programas estadísticos están diseñados para llevar a cabo un análisis con los datos completos **ADC**, excluyendo aquellas unidades donde para alguna variable asociada no se observa o no se obtiene valor alguno. Esto puede resultar inapropiado pues el investigador usualmente está interesado en desarrollar inferencias acerca de la población de interés completa en vez de una porción de esa población. Con respecto al poder de separación de la regresión logística este problema se incrementa dependiendo de la cantidad de valores ausentes en el conjunto de datos. A continuación se mostrará cómo, mediante el uso del método **ADC**, se va incrementando el sesgo en el parámetro que en este caso es el poder de separación del modelo. Para esto se utilizará nuevamente la base de datos *estudiantes*.

Si se elimina gradualmente desde el 5 por ciento hasta el 50 por ciento de los valores del conjunto de datos y se calculan las áreas bajo la curva **ROC** para cada conjunto restante, en cada proporción se observará que rápidamente ocurre una disminución en el poder de separación del modelo. La FIGURA 3.1 muestra las áreas bajo las curvas **ROC** para cada porcentaje de datos faltantes. Como se puede observar, las áreas estimadas van decreciendo drásticamente, lo cual hace ver de antemano la inefectividad del **ADC**.

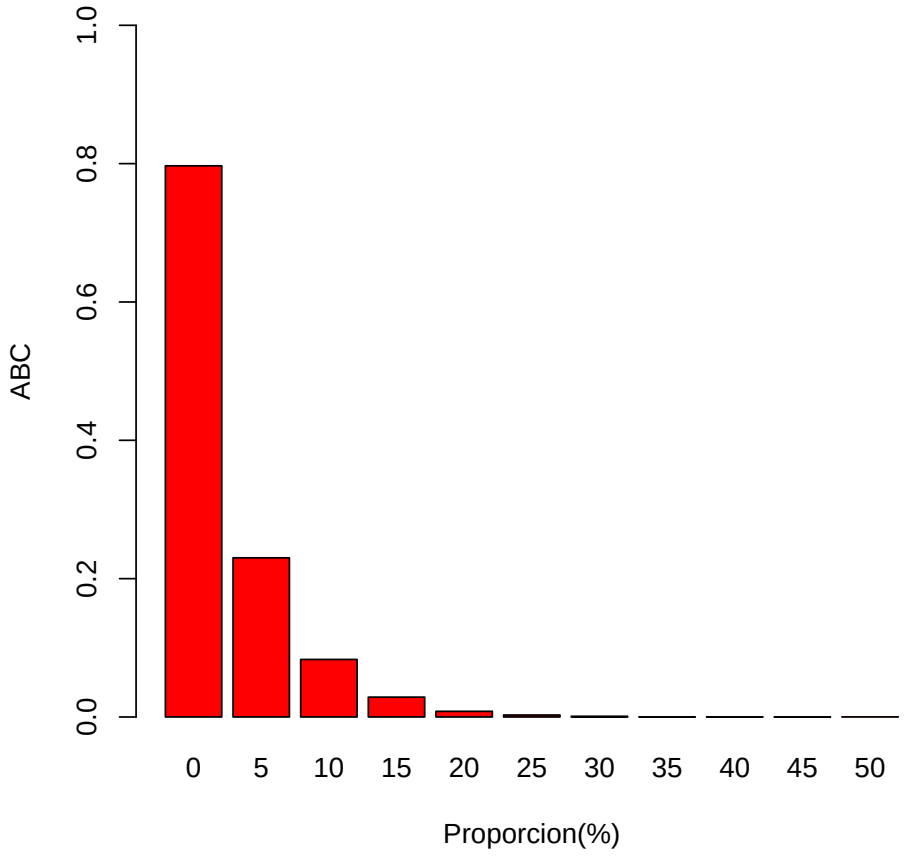


Figura 3.1: **Valores de ABC por proporción de datos eliminados para los datos *estudiantes***

3.3. Mecanismos que llevan a datos faltantes

Antes de considerar algún método para manejar el problema de datos faltantes es necesario saber cuáles son las causas que llevan a la ausencia de estos datos y cuál es la naturaleza de los valores faltantes en las variables y su relación con los valores de otras variables dentro del conjunto de datos. Definamos a $\mathbf{Y} = (\mathbf{y_{ij}})$ como la matriz de datos a analizarse con *i filas* y *j columnas* y sea $\mathbf{M} = (\mathbf{m_{ij}})$ la matriz indicadora de datos faltantes donde la misma toma valor de 1 cuando el dato es faltante y 0 cuando el dato está disponible. Los mecanismos o causas que conllevan a la ausencia de valores en un conjunto de datos están sujetos a una probabilidad condicional entre los datos $\mathbf{Y}$ y la matriz indicadora $\mathbf{M}$ dada por $\boldsymbol{f(M|Y,\phi)}$ donde $\boldsymbol{\phi}$ denota los parámetros a estimarse. Se describen tres tipos de mecanismos posibles para explicar la ausencia de valores en un conjunto de datos. [Little and Rubin, 2002].

- Si la falta de datos no depende en sí de la matriz de datos $\mathbf{Y}$, es decir, que si tenemos;

$$\boldsymbol{f(M|Y,\phi) = f(M|\phi)} \tag{3.1}$$

  para toda $\mathbf{Y}$,$\boldsymbol{\phi}$ los datos se conocen como faltantes por completa aleatoridad llamados por sus siglas en inglés por **MCAR** (*Missing Completely at Random*).

- Por otro lado se define la siguiente partición; $(\mathbf{Y_{obs}}, \mathbf{Y_{miss}})$ en los datos de la matriz $\mathbf{Y}$ donde $\mathbf{Y_{obs}}$ representan los datos observados y $\mathbf{Y_{miss}}$ representan los datos faltantes. Entonces un mecanismo menos restrictivo que el anterior es aquel donde la falta de datos depende sólo de $\mathbf{Y_{obs}}$, es decir que:

$$\boldsymbol{f(M|Y,\phi) = f(M|Y_{obs},\phi)} \tag{3.2}$$

para toda $\boldsymbol{Y_{miss}}, \boldsymbol{\phi}$ los datos se conocen como faltantes por aleatoridad y se denotan como **MAR** *(missing at random)*, por sus siglas en inglés.

- En última instancia si la distribución de **M** depende de $\mathbf{Y_{miss}}$, y también puede que de $\mathbf{Y_{obs}}$, entonces decimos que los datos no son faltantes por aleatoridad y se denotan por sus siglas en inglés **NMAR** (*not missing at random*). Este mecanismo obedece la siguiente distribución:

$$\boldsymbol{f(M|Y,\phi) = f(M|Y_{miss}, Y_{obs}, \phi)} \tag{3.3}$$

Los mecanismos de datos faltantes se definen de acuerdo a cuán restrictivos son. El más restrictivo es el mecanismo **MCAR**, lo cual lo hace poco ocurrente en la vida real pues la falta de datos es más usual que se deba a causas asociadas a un sector específico de la población. Por ejemplo, en una encuesta donde se pregunte si una persona es o no alcohólica, por lo general, si esta persona lo es en realidad, contestará que no, o como es lo más común, rehusará contestar. En este ejemplo los datos faltantes se rigen bajo un mecanismo **MAR** pues contestar o no contestar la pregunta depende de si la persona es o no alcohólica. No obstante el mecanismo **MAR** no es el más común de todos pues todavía tiene la restricción de que la falta de datos depende sólo de los valores del conjunto de datos que son observados. Por lo que el mecanismo más frecuente en la realidad es el **NMAR**. Tomando el mismo ejemplo de alcoholismo, si se le pregunta a una persona alcohólica que conteste cuántas veces consume alcohol al día, lo más probable es que tampoco conteste esta pregunta pues las personas por lo general tienden a ser muy cuidadosas al contestar para no contradecirse. Por lo tanto, la ausencia de respuesta a la pregunta de cuántas veces consume alcohol al día puede tener estrecha relación con la ausencia a la contestación en cuanto a la aceptación de ser alcohólico o no, que usualmente corresponde a la subpoblación de personas alcohólicas.

Por su naturaleza, los mecanismos **MAR** y **NMAR** son muy difíciles de simular pues al ser los más comunes y menos restrictivos el crear simulaciones para todos los casos sería practicamente imposible. No obstante el mecanismo **MCAR**, al ser el más restrictivo de todos, nos permite escoger una simulación dentro de un marco limitado de simulaciones.

Para efectos de esta tesis los hallazgos se basarán en el mecanismo por completa aleatoridad **MCAR**. Esto se debe a que se están poniendo a prueba varios métodos de imputación a través de ciertas porporciones de datos faltantes. Todos los datos analizados en las simulaciones están completamente observados y basándonos en este mecanismo es que se logrará remover algunos de ellos, creando artificialmente datos faltantes para cada conjunto de datos.

Para evitar variabilidad dentro de los resultados del experimento se utilizó un mecanismo **MCAR** extremadamente restricivo en donde cada valor dentro del conjunto de datos tiene exactamente la misma probabilidad $\boldsymbol{p}$ de estar ausente. Por lo tanto, se define el mecanismo como sigue;

$\mathbf{Pr}(\mathbf{m_{ij}} = \mathbf{1}) = \mathbf{p}$ tal que $\boldsymbol{p} \in$ 0.05, 0.10, 0.15, ..., 0.50

En esta tesis estaremos analizando los métodos de imputación bajo el mecanismo **MCAR** en diez proporciones distintas. No obstante en la literatura, [Pyle, 1999], ya se han dividido varias clases o categorías de proporciones de datos faltantes y los efectos de las mismas en el análisis estadístico que se lleve a cabo. Las clases son las siguientes:

- Si la proporción de datos faltantes es menor del 1 por ciento, el efecto en las estimaciones de los parámetros es trivial.
- Entre el 1 y el 5 por ciento global, el problema es manejable por algún método poco sofisticado.
- Entre el 5 y el 15 por ciento global, el problema se puede lidiar por un método mucho más sofisticado de manejo de datos faltantes.

- Si la proporción es mayor del 15 por ciento el problema de datos faltantes tendrá un efecto impactante en el análisis estadístico y en la estimación de parámetros en el estudio.

A continuación se presenta un ejemplo simulado de cómo se generan datos faltantes bajo este mecanismo. Se presenta el código para el programa que genera datos faltantes en **R**. Los datos utilizados corresponden a la base *estudiantes*. En la siguiente gráfica se representan los datos faltantes en la variables del conjunto utilizando la función ***imagmiss*** de la librería ***dprep*** [Rodríguez, 2004].
Programa para generar datos faltantes por el mecanismo **MCAR** utilizado en esta tesis.

```
> GenPat <- function(X, prob) {
+     n = dim(X)[1]
+     p = dim(X)[2]
+     gmv = matrix(0, n, p)
+     for (j in 1:p) {
+         gmv[, j] = rbinom((n), 1, prob)
+         w = which(gmv[, j] == 1)
+         X[w, j] = NA
+     }
+     X
+ }
```

En la FIGURA 3.2 se muestran los datos faltantes utilizando el mecanismo con una proporción del 30 por ciento por lo que la cantidad de datos faltantes en el conjunto es sustancial. En este caso se muestra un patrón general de datos faltantes, siendo éste el más común en los datos bajo el mecanismo *MCAR*, pues por su aleatoridad no existe un patrón definido. Los otros dos mecanismos, **MAR** y **NMAR**, generan otros patrones de datos faltantes. En Little y Rubin (2002) se presenta más información acerca de estos otros patrones.

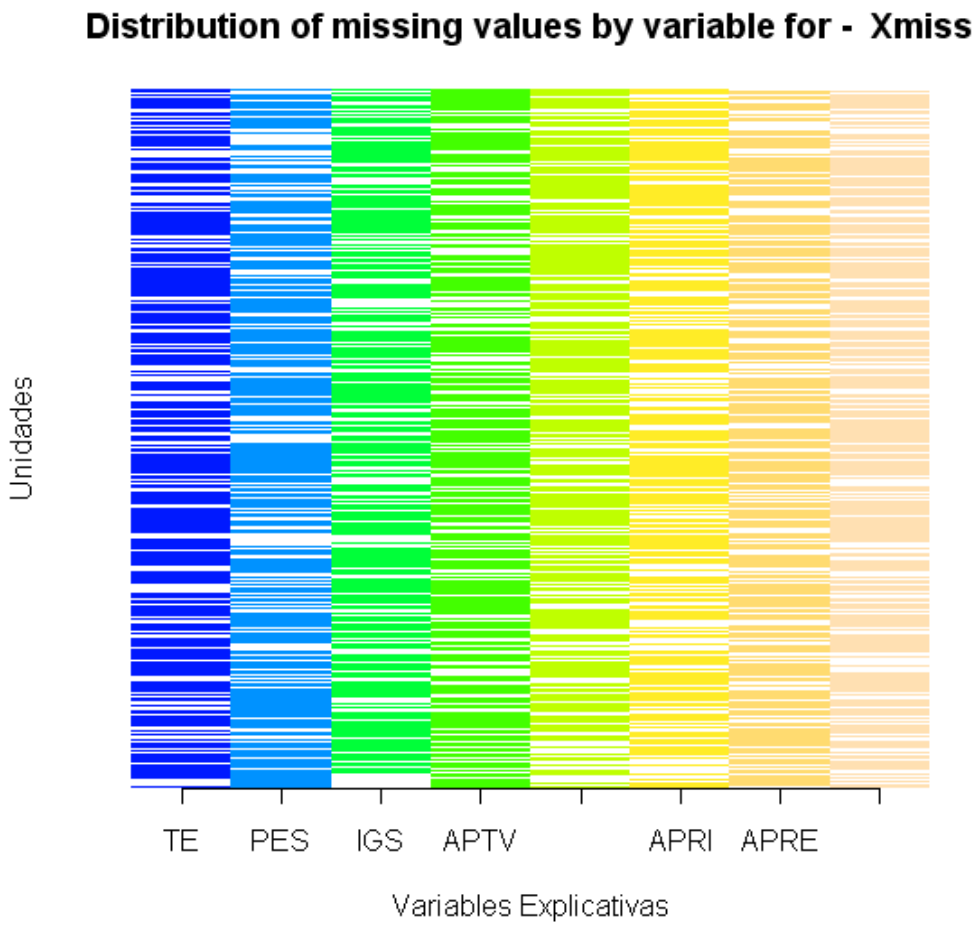


Figura 3.2: **Patrón general de datos faltantes generados con la función *imagmiss* con una proporción de datos faltantes del 30 %**

3.4. Soluciones al problema de datos faltantes

Aunque nuestro objetivo es analizar la efectividad de diversos métodos de imputación en la problemática de datos faltantes, en la literatura se presentan varios métodos para resolver este problema. Aquí se mencionarán varios de ellos y se describirá brevemente la metodología que emplean. Los métodos de imputación se describen en el próximo capítulo.

3.4.1. Método de omisión de observaciones

Este método tambien conocido como *Análisis con datos completos* ya se mencionó en la segunda sección de este capítulo. Es un método muy fácil de implantar y consiste en eliminar las observaciones, es decir las filas, en la matriz de datos en donde aparezca al menos un valor faltante para una variable o atributo. Este método como ya se pudo observar crea sesgos considerables dependiendo de la cantidad de datos faltantes en la matriz. Además incrementa la variabilidad de las estimaciones. Si la proporción de valores faltantes es mínima [Pyle, 1999] la eliminación de datos crea resultados satisfactorios, aunque se debe tomar en consideración el mecanismo que genera tales datos faltantes.

3.4.2. Método de omisión de observaciones con ponderación

En este caso el análisis con datos completos es acompañado por unas ponderaciones provenientes del diseño de muestreo y de la tasa de respuesta. Es decir a cada observación completa $\boldsymbol{i}$ se le asigna una ponderación $\mathbf{w_i}$ calculada a base del diseño de muestreo y a base de la tasa de respuesta en el experimento. Las ponderaciones se pueden calcular de la siguiente forma.

Sea $\mathbf{D_S}$ el conjunto de la observaciones muestreadas con una probabilidad dada. Sea $\mathbf{D_C}$ el conjunto de las observaciones completas para todas las variables en el conjunto de datos. Entonces se define la ponderación $\boldsymbol{w_i}$ como:

$$\boldsymbol{w_i} = \frac{\mathbf{1}}{\boldsymbol{Pr(i\epsilon(D_S \bigcap D_C))}}$$
$$\boldsymbol{w_i} = \frac{\mathbf{1}}{\boldsymbol{Pr(i\epsilon D_S)}} \times \frac{\mathbf{1}}{\boldsymbol{Pr(i\epsilon D_C | i\epsilon D_S)}}$$
$$\boldsymbol{w_i} = \boldsymbol{w_{i,D_S}} \times \boldsymbol{w_{i,D_C}}$$

Donde $\mathbf{w_{i,D_S}}$ es la ponderación por muestreo, que usualmente es conocida y $\mathbf{w_{i,D_C}}$ es la ponderación por respuesta u observación. Esta última usualmente es desconocida pero se puede estimar experimentalmente.

Los métodos ponderados aunque son relativamente simples, tienden a controlar el sesgo, pero a su vez también introducen mucha variabilidad. Además el análisis y la inferencia sobre las ponderaciones es algo controversial pues usualmente los programados estadísticos calculan errores estándar para estudios complejos pero no toman en cuenta las ponderaciones. Además este método no es muy útil cuando se tiene una cantidad de variables o atributos considerable y cuando el patrón de datos faltantes no es monótono [Little and Raghunathan, 2004].

3.4.3. Estimación de parámetros

En este caso se estima el parámetro deseado utilizando una mezcla entre un modelo estadístico y la información que provea el conjunto de datos con valores faltantes. El modelo estadístico requiere de la información completa del conjunto de datos y del mecanismo de datos faltantes. No obstante, si por ejemplo se estiman lo parámetros mediante la función de máxima verosimilitud y el mecanismo de datos faltantes es **MCAR** o **MAR**, entonces el mecanismo puede obviarse del modelo, no así si el mecanismo es **NMAR**. Además el método por máxima verosimilitud requiere de fuertes supuestos en la distribución de los datos lo cual hace complicada su aplicación [Little and Rubin, 2002, Little and Raghunathan, 2004].

Otra forma de estimación de parámetros es el algoritmo **EM** (del inglés *Expectation Maximization*). Este método es un puente entre el método de máxima verosimilitud y la imputación de datos. Se compone de dos pasos;

el paso E y el paso M. Primero se comienza con unos estimados iniciales provenientes de los datos completos y en el paso E se busca imputar los datos utilizando los datos y los parámetros iniciales. Finalmente en el paso M se busca maximizar los estimados de los parámetros utilizando los datos obtenidos en el paso E. Este método es uno iterativo entre el paso E y el paso M hasta hallar convergencia en las estimación del parámetro deseado. El algoritmo **EM** incrementa la verosimilitud del estimado conforme se incrementan las iteraciones bajo ciertas condiciones especiales [Shaefer, 1997].

La estimación de parámetros es un método efectivo cuando el mecamismo de datos faltantes es fuertemente restrictivo [Little and Raghunathan, 2004], es decir **MCAR**. Se abundará más al respecto cuando se discuta el método **FRITZ** que se utiliza como método de imputación múltiple.

Capítulo 4

Métodos de imputación

4.1. Introducción

Los métodos de imputación se pueden definir simplemente como promedios o selecciones provenientes de una distribución de predicción de los valores faltantes que se basa en los valores observados. Existen dos formas genéricas para obtener tal distribución [Little and Rubin, 2002]; los modelos explícitos y los modelos implícitos. En los modelos explícitos la distribución está basada en un modelo estadístico formal lo cual lleva a que los supuestos sean explícitos. Por otro lado los modelos implícitos están enfocados en el uso de algoritmos, los cuales implican un modelo detrás de éstos y por lo tanto los supuestos son implícitos. A continuación se describen varios métodos de imputación simple de ambos modelos los cuales se pondrán a prueba en esta tesis. Además describiremos el método de imputación múltiple utilizado por *Kennickell* en el SCF (*Survey of Consumer Finances*) [Kennickell, 1998]. Este método, como ya se indicó en el capítulo anterior, utiliza los principios del algoritmo **EM** (*Expectation Maximization*).

4.2. Métodos de imputación sencilla

4.2.1. Métodos por modelos explícitos

Entre las ventajas de estos métodos es que funcionan para cualquier tipo de variable, estabilizan los valores imputados, no requieren de algún supuesto de distribución y se pueden utilizar análisis estadísticos tradicionales en el conjunto completado mediante tales imputaciones. No obstante los métodos bajo estos modelos tienen la desventaja de que dependen de la distribución de los valores observados. Además subestiman la varianza de los estimados y distorsionan la matriz de correlación. Otra desventaja es que requieren el mecanismo de datos faltantes por completa aleatoridad **MCAR** para obtener buenos estimados [Little and Rubin, 2002].

Los métodos explícitos a analizarse son: imputación por la media muestral, imputación por la mediana e imputación por la moda. Estos tres métodos se pueden clasificar a su vez en condicionales y no condicionales. Los métodos condicionales son aquéllos donde la imputación está sujeta a ser seleccionada según la clase a la cual pertenezca la unidad. Por el contrario, los métodos no condicionales utilizan los valores observados sin tomar el consideración las clases.

Métodos no condicionales

1. **Imputación por la media muestral (IMEAN)**

   La imputación por la media muestral es uno de los métodos más antiguos de rellenar los datos faltantes en una matriz. El método consiste en rellenar los datos que faltan de la matriz $\boldsymbol{X} = (\boldsymbol{x_{ij}})$ variable por variable, promediando los valores observados en cada variable y tomando ese promedio como la imputación o relleno para los datos faltantes. Es decir que para cada variable $\boldsymbol{x_j}$ se tienen $\boldsymbol{r}$ valores observados y $\boldsymbol{n\text{-}r}$ datos faltantes. Por lo tanto, para los $\boldsymbol{n\text{-}r}$ datos faltantes se tiene que los valores imputados en la variable $\boldsymbol{x_j}$ se calculan por la fórmula:

$$\boldsymbol{x_{imp(j)} = \frac{\sum_{i=1}^{r} x_{i,obs(j)}}{r}} \tag{4.1}$$

Este método de imputación es útil para variables numéricas continuas. Además es uno de los métodos más fáciles de aplicar y casi todos los programados estadísticos lo tienen. Sin embargo, por ser la media muestral una medida de tendencia central tiende a disminuir la variabilidad de los datos en la variable [Little and Rubin, 2002].

La variable (*PES*) que representa el promedio de escuela superior en los datos *estudiantes* es una variable numérica contínua. Se calculó la varianza de esta variable en los datos originales y se generaron datos faltantes en ella bajo el mecanismo **MCAR**. Luego se imputaron utilizando la media muestral de los datos observados en la variable y se calculó nuevamente su varianza. Se ejecutaron 100 iteraciones utilizando el programa **R**. La FIGURA 4.1 muestra que la imputación por la media subestima la varianza real de la variable *PES*, la cual es representada por la línea recta.

2. **Imputación por la mediana (IMED)**

   Además de la imputación por la media muestral, otra medida de tendencia central utilizada es la mediana. Este método de imputación se utiliza para variables continuas y ordinales. Al igual que la media muestral, la mediana es un método de fácil aplicación y aparece en la mayoría de los programados estadísticos. Cada dato faltante es sustituido por la mediana de una misma variable creando así un estimado más robusto para ese valor, es decir, con resistencia a valores influenciales ya que es un estadístico de orden. La imputacion por la mediana se define como sigue:

$$\boldsymbol{x_{imp(j)} = Mediana(x_{obs(j)})} \tag{4.2}$$

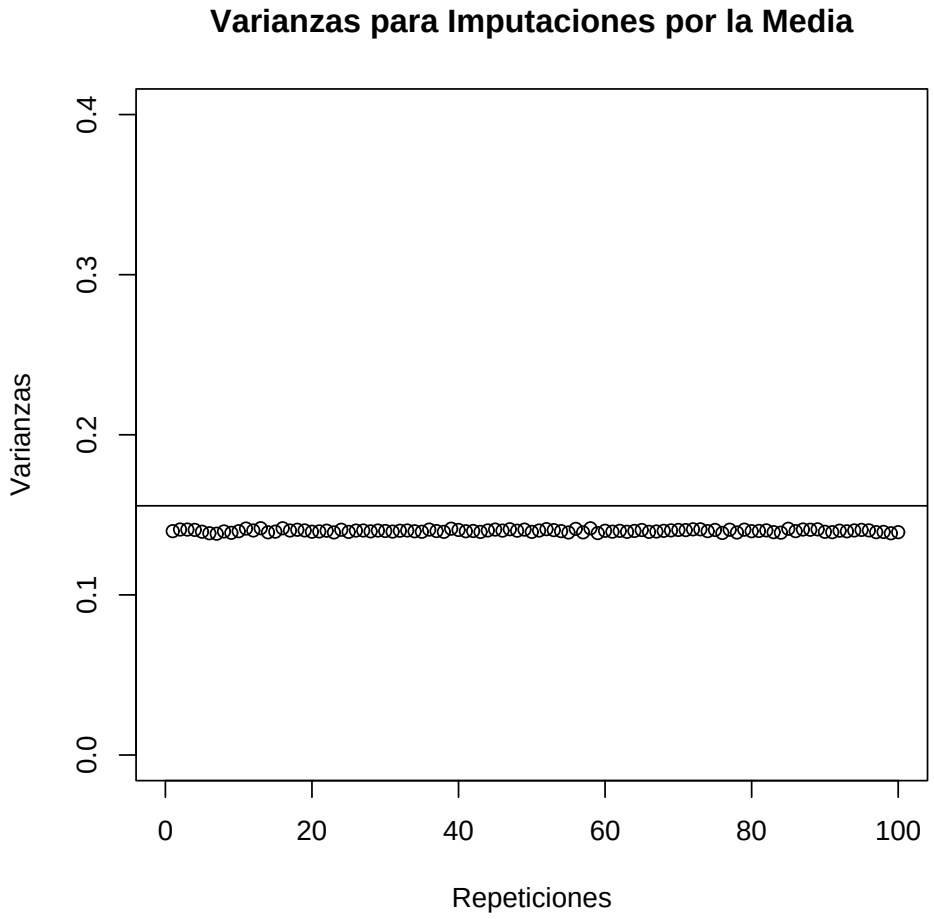


Figura 4.1: **Gráfica de las varianzas para la imputación por la media muestral al borrar mediante un mecanismo MCAR el 10 % de las observaciones en la variable *PES* del conjunto de datos *estudiantes***

No obstante la imputación por la mediana también afecta la variabilidad de los datos en una variable, tanto o más que la imputación por la media. Esto se observará con la varianza de la variable *PES* del conjunto de datos *estudiantes*. En la FIGURA 4.2 se muestran los estimados de las varianzas similar al caso de imputación por la media muestral.

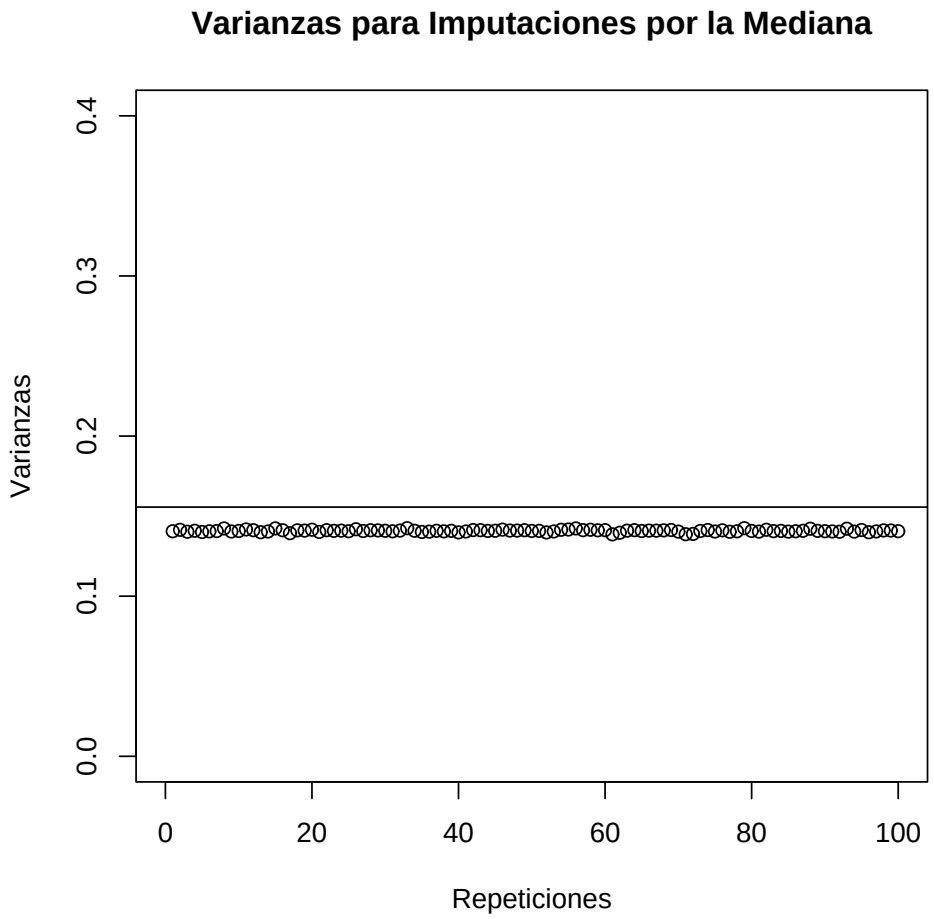


Figura 4.2: **Gráfica de las varianzas para la imputación por la mediana al borrar mediante un mecanismo MCAR el 10 % de las observaciones en la variable *PES* del conjunto de datos *estudiantes***

Como se observa en ambas gráficas las medidas de tendencia central reemplazan los valores faltantes y hacen subestimar la varianza de una

variable. No obstante más interesante aún es la subestimación de la covarianza que crean estos dos métodos entre dos variables $\boldsymbol{x_j}$ y $\boldsymbol{x_k}$. A continuación se mostrará un ejemplo simulado parecido al ejemplo anterior. En este caso se muestra la covarianza entre las variables *IGS* y *PES* del mismo conjunto de datos y las simulaciones con 100 iteraciones donde se remueven el 10 % de las observaciones en cada variable y luego se imputan las dos variables con un método, ya sea la media muestral o la mediana. La FIGURA 4.3 muestra en línea entera las covarianzas luego de imputar por la media y en línea con puntos luego de imputar por la mediana. En este caso, como se tienen dos variables distintas las probabilidades del mecanismo **MCAR** son diferentes y escogidas al azar en un rango del 10 al 20 por ciento de datos faltantes. El código en **R** que muestra la covarianza real entre IGS y PES.

```
> IGS = estudiantes$IGS
> PES = estudiantes$PES
> cov(cbind(IGS, PES))[1, 2]

[1] 9.855576
```

La importancia que tiene esta reducción en la covarianza es el efecto que esto pueda tener sobre el modelo de regresión logística en especial sobre los estimadores de los coeficientes del modelo. No obstante ese efecto lo veremos más adelante en la parte de los resultados.

3. **Imputación por la moda (IMOD)**

   La imputación por la moda se utiliza para variables categóricas. En caso de haber más de una moda se escoge aleatoriamente entre las modas y se imputa el valor faltante por ese valor. Se utiliza la imputación por la moda para este tipo de variable en vez de la media muestral o la mediana para evitar que el número de clases en la variable no se

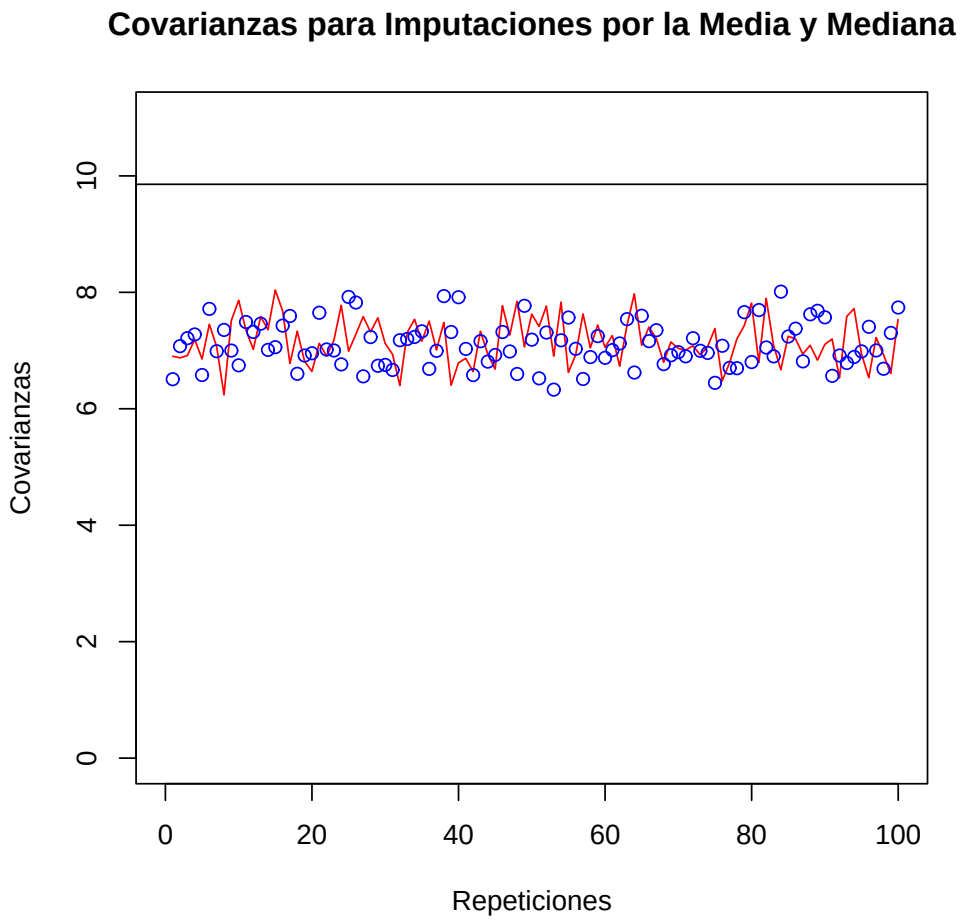


Figura 4.3: **Covarianzas para IMEAN e IMED al eliminar el 10 % de los datos en las variables *IGS* y *PES* del conjunto de datos estudiantes**

afecte. La función *moda* utilizada en la función **IMOD** pertenece al programado en **R** llamado **dprep** [Rodríguez, 2004].

Estos tres métodos de imputación simple se utilizan por separado en los conjuntos de datos con variables contínuas. En los conjuntos mixtos, es decir los conjuntos de datos con diferentes tipos de variables, se comparan sólo **IMEAN** e **IMED**. La imputación por la moda, **IMOD**, se utiliza junto con ambos métodos de imputación para imputar sólo las variables categóricas. Esta última también tiende a subestimar la varianza. Como ejemplo se utiliza la imputación por la moda para la variable tipo de escuela (TE) del conjunto de datos *estudiantes* cuya gráfica se puede observar en la FIGURA 4.4.

Métodos condicionales

La imputación condicional está sujeta a las clases de interés de la variable de respuesta $\mathbf{Y}$. En el caso de regresión logística, las clases que se inducen son $\boldsymbol{C_0}$ y $\boldsymbol{C_1}$. Para cada variable $\boldsymbol{x_j}$ en la clase $\boldsymbol{C_0}$ de tamaño $\boldsymbol{n_0}$ se tiene $\boldsymbol{r_0}$ valores observados y $\boldsymbol{n_0 - r_0}$ valores faltantes. Similarmente en la clase $\boldsymbol{C_1}$ de tamaño $\boldsymbol{n_1}$ se tiene $\boldsymbol{r_1}$ valores observados y $\boldsymbol{n_1 - r_1}$ valores faltantes para la misma variable. Entonces se aplican los mismos métodos no condicionales pero por clases.

1. **Imputación por la media muestral por clases (ICMEAN)**

   Este tipo de imputación es similar a **IMEAN** pero aplicado a ambas clases. La imputación condicional por la media muestral obedece a la siguiente ecuación.

$$\boldsymbol{x_{imp(j),C_k}} = \frac{\sum_{i=1,r_k} \boldsymbol{x_{i,obs(j),C_k}}}{\boldsymbol{r_k}} \tag{4.3}$$

   para $\boldsymbol{k = 0, 1}$.

2. **Imputación por la mediana condicional (ICMED)**

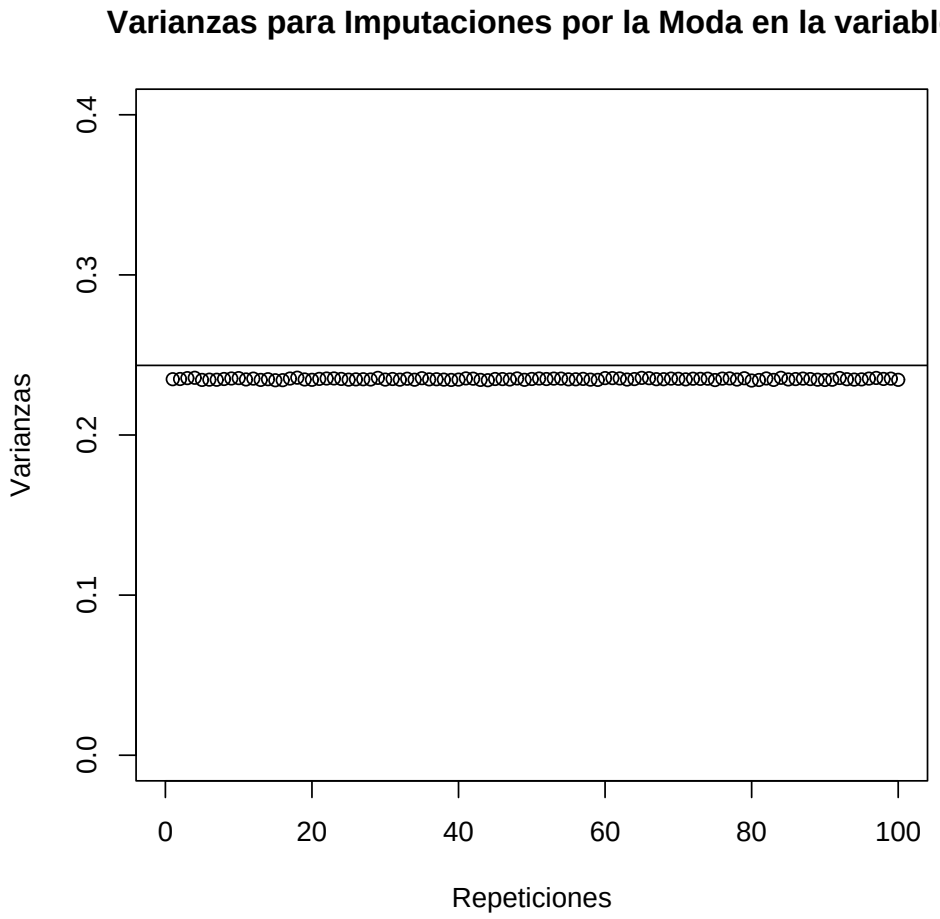


Figura 4.4: **Varianzas para la imputación por la moda en la variable *TE* del conjunto de datos *estudiantes* luego de haber eliminado el 10 % de las observaciones**

Este tipo de imputación es tambien similar a la imputación por la mediana pero en este caso está sujeto a clases. La ecuación para **ICMED** es la siguiente;

$$\boldsymbol{x_{imp(j),C_k} = Mediana(x_{obs(j),C_k})} \tag{4.4}$$

para $\boldsymbol{k = 0, 1}$.

La imputación por la moda condicional nuevamente es para variables categóricas y es similar a las anteriores; lo único que se aplica a las clases $\boldsymbol{C_0}$ y $\boldsymbol{C_1}$.

4.3. Métodos por modelos implícitos

Los métodos implícitos como ya se había indicado corresponden al diseño de algoritmos y no a modelos estadísticos formales. En cada algoritmo para que un valor faltante pueda ser imputado usualmente depende de otras variables auxiliares dentro del conjunto de datos. Los métodos de este modelo se pueden utilizar en cualquier tipo de variable. Se genera mayor variabilidad que en los métodos del modelo explícito pues provee una selección aleatoria de los valores a utilizarse en la imputación, característica que hace en cierta forma superior este modelo implícito sobre el explícito [Little and Rubin, 2002]. Además los valores imputados por los métodos bajo este modelo preservan aproximadamente la distribución de los datos observados [Rubin, 1987]. Por lo tanto, para generar los aproximados no se requieren fuertes supuestos distribucionales, por consiguiente el conjunto que incluye los datos imputados puede analizarse por métodos estadísticos tradicionales con resultados aceptables [Little and Rubin, 2002].

No obstante, se requiere alguna programación para poder implantar estos métodos. Además se requiere información completa en las variables auxiliares por lo que como se mencionó antes, los estimados dependen sobremanera de los valores de las variables auxiliares. Esto implica que no existe

un algoritmo fijo para imputar mediante este modelo. Además se requiere el supuesto de un mecanismo restrictivo como **MCAR** o **MAR**. A continuación se presenten los métodos de imputación implícitos que se utilizarán en esta tesis: *imputación por muestreo aleatorio no condicional*(**IRS**), *imputación por muestreo aleatorio condicional* (**ICRS**), e *imputación por vecinos más cercanos* (**KNN**).

1. **Imputación por muestreo aleatorio (IRS)**

   Este tipo de imputación es muy sencillo y funciona para cualquier tipo de variable. Consiste en tomar un muestreo aleatorio con reemplazo de los valores observados en una variable y usarlos como reemplazos o sustituciones para los valores faltantes dentro de la misma variable.

   Sea $\boldsymbol{X_j}$ la variable de interés que deseamos imputar con r valores observados y $\boldsymbol{n-r}$ valores faltantes. Sea $\boldsymbol{X_j^r}$ el conjunto de valores observados para la variable $\boldsymbol{X_j}$. Entonces las función de imputación para este método está dada por:

   $$\boldsymbol{x_{imp(j)} = x_s} \tag{4.5}$$

   donde $\boldsymbol{x_s \ \epsilon \ X_j^r}$ para $\boldsymbol{s = 1, \ldots, r}$. Como el muestreo es con reemplazo $\boldsymbol{x_s}$ puede utilizarse más de una vez como valor de imputación con probabilidad de $\frac{1}{r}$ de ser escogido. Una ventaja de este método es que no distorsiona la distribución de $\boldsymbol{X_j}$ en comparación con los métodos anteriores. No obstante incrementa la varianza de una forma considerable [Little and Rubin, 2002] por lo que los resultados finales de este experimento bajo este método no son adecuados (*Ver Capítulo 6*). Por lo tanto se requiere de métodos más elaborados como *Hot Deck* para aliviar el problema de la estimación sobre la varianza.

2. **Imputación por muestreo aleatorio condicional (ICRS)**

Para controlar la variabilidad introducida por el **IRS** se lleva a cabo una variación al método. En vez del muestreo aleatorio por todas las observaciones de la variable $\boldsymbol{X_j}$, se utiliza por las clases inducidas por la variable de respuesta, similar a los métodos condicionales antes mencionados. Sea $\boldsymbol{X_{j,C_k^{r_k}}}$ el conjunto de $\boldsymbol{r_k}$ valores observados para la variable $\boldsymbol{X_j}$ en la clase $\boldsymbol{C_k}$, entonces

$$\boldsymbol{x_{imp(j),C_k} = x_{s,C_k}} \tag{4.6}$$

donde $\boldsymbol{x_{s,C_k}}$ $\boldsymbol{\epsilon}$ $\boldsymbol{X_{j,C_k^{r_k}}}$ para $\boldsymbol{s = 1, \ldots, r_k}$ y $k = 1, 2$. En este caso se obtuvo resultados aceptables con este método de imputación (*Ver Capítulo 6*).

3. **Imputación por vecinos más cercanos (KNN)**

   En un intento por buscar un método general que fuera más certero en sus estimaciones se han creado métodos que hacen uso de métricas para medir distancias entre unidades basadas en los valores de las variables asociadas dentro del mismo conjunto de datos. Luego se calculan tales distancias y se procede a imputar los valores faltantes utilizando las unidades del conjunto de unidades completas más cercanas según la métrica [Little and Rubin, 2002]. El algoritmo **KNN** imputa valores de esta forma utilizando como métrica la distancia euclideana entre las unidades. Los pasos del algoritmo se explican a continuación [Rodríguez, 2004].

   - Particionar el conjunto de datos **D** en dos partes: Las unidades completamente observadas $\boldsymbol{D_c}$ y las unidades con valores faltantes $\boldsymbol{D_m}$.
   - Para cada unidad $\boldsymbol{x_i}$, en $\boldsymbol{D_m}$ calcular las distancias entre $\boldsymbol{x_i}$, y las unidades completas $\boldsymbol{x_c}$, y escoger las $\boldsymbol{k}$ unidades más cercanas según la distancia euclideana. Este conjunto escogido para $\boldsymbol{x_i}$, se

llama el conjunto $\boldsymbol{D_k}$ de los $\boldsymbol{k}$ vecinos más cercanos a $\boldsymbol{x_i}$, el cual contiene valores faltantes para una o varias variables o atributos.

- Con los valores en $\boldsymbol{D_k}$ se imputarán los valores faltantes en $\boldsymbol{x_i}$, en cada variable $\boldsymbol{j}$ dependiendo del tipo de ésta. Si la variable $\boldsymbol{j}$ es del tipo contínuo entonces se hace un promedio de los $\boldsymbol{k}$ vecinos en esa variabe $\boldsymbol{j}$ y ese promedio pasa a ser el valor de imputación para $\boldsymbol{x_{i,j}}$. Por otro lado si la variable es de tipo binaria u ordinal se buscará el valor que más se repita dentro de los k vecinos y ese pasará a ser el valor de imputación.
- El proceso termina cuando las unidades en $\boldsymbol{D_m}$ se han imputado completamente.

Como se puede observar el **KNN** posee las ventajas mencionadas anteriormente para los métodos de modelos implícitos, además de que toma en consideración la estructura de la correlación del conjunto de datos. Sin embargo, para este método no se provee un criterio de selección de la métrica. En este caso se utilizó la distancia euclideana pero pudieron usarse otras como por ejemplo la Manhattan, la Mahalanobis, y la Pearson, entre otras [Little and Rubin, 2002]. Además el **KNN** sólo imputa valores escogiendo $\boldsymbol{k}$ vecinos dentro de una misma clase, lo cual lo convierte en un método condicional. Otra desventaja de este método es que tampoco provee un criterio específico para la selección de un número $\boldsymbol{k}$ de vecinos. En la literatura se han encontrado buenos resultados con un valor de $\boldsymbol{k = 10}$. No obstante un valor muy pequeño de $\boldsymbol{k}$ deteriora la ejecución de un clasificador pues brinda mayor énfasis a una pocas unidades en la estimación de la imputaciones. Por otro lado un número muy grande de $\boldsymbol{k}$ distorsionaría las imputaciones en el sentido de que se escogerían unidades muy diferentes para estimar la imputación. Por lo tanto muchos investigadores optan por utilizar un número de $\boldsymbol{k}$ menor de 10 para conjuntos pequeños [Rodríguez, 2004].

En esta tesis, debido a lo restrictivo del mecanismo y a la gran cantidad

de unidades con valores faltantes se utilizó un sólo vecino ($\boldsymbol{k = 1}$), lo cual pueda parecer poco apropiado pero de esa forma nos aseguramos de encontrar siempre algún vecino y por ende algún valor para poder imputar. El efecto de esta decisión en el modelo de regresión logística se discutirá en el Capítulo 8.

Los métodos de imputación sencilla pueden resultar muy atractivos en el sentido de que como resultado final se obtiene un conjunto de datos rectangular completo al cual se le puede aplicar análisis estadísticos tradicionales. No obstante el problema con las imputaciones obviamente es que los valores faltantes siguen desconocidos y la aplicación de cualquier análisis en el conjunto de datos imputados se ejecuta como si los valores faltantes fueran conocidos. Por lo tanto, aún conociendo el mecanismo que genera datos faltantes, las inferencias basadas en los valores imputados pueden representar algún riesgo debido a que la variabilidad añadida por el desconocimiento de los valores faltantes no se toma en consideración. Por consiguiente los parámetros que dependen de la varianza, como por ejemplo la correlación, pueden resultar altamente sesgadas. Más aún, el peligro puede ser mayor si no se conoce el mecanismo más apropiado para el método de imputación empleado [Rubin, 1987].

Existen métodos de imputación simple donde se toma en consideración una distribución de predicción para los valores faltantes. Por ejemplo la imputación por regresión ha brindado buenos resultados en la estimación de parámetros en el modelo de regresión lineal simple [Jinn, 2000]. No obstante este tipo de imputación no se tomó en consideración como método de imputación sencilla pues requiere que las observaciones en las variables explicativas sean observadas al menos para la unidad donde se encuentra el valor faltante. Sin embargo, la regresión lineal se utilizó como parte del método de imputación múltiple que se discutirá en la proxima sección.

4.4. Métodos de imputación múltiple

La imputación múltiple retiene las virtudes de la imputación sencilla y corrige la mayoría de sus fallas. La idea general del método consiste en que para cada espacio faltante se generan varios valores para su imputación, sea $\boldsymbol{m}$ el número de valores imputados para una observación faltante. Esos $\boldsymbol{m}$ valores están ordenados en el sentido de que el primer conjunto de imputaciones para los valores faltantes son utilizados para formar el primer conjunto completo de datos, luego se tiene un segundo conjunto de imputaciones que serán utilizados para formar un segundo conjunto completo de datos, y así sucesivamente. Por lo que las $\boldsymbol{m}$ imputaciones de los valores faltantes crean $\boldsymbol{m}$ conjuntos completos de datos. La cantidad de $\boldsymbol{m}$ valores imputados utilizada depende de la cantidad de valores faltantes. Para una cantidad moderada de datos faltantes, un cantidad moderada $\boldsymbol{m}$ de imputaciones puede variar de 2 a 10 [Rubin, 1987].

4.4.1. Ventajas y desventajas de la imputación múltiple

La imputación múltiple comparte dos ventajas básicas con la imputación sencilla; (1) la habilidad de generar un conjunto completo de datos para llevar a cabo cualquier tipo de análisis y (2) la habilidad para incorporar el conocimiento del encuestador [Rubin, 1987]. Este método también corrige la desventaja de la imputación sencilla con respecto a la variabilidad de los datos faltantes pues, como las $\boldsymbol{m}$ imputaciones son repeticiones bajo un modelo de predicción de los valores faltantes, el análisis en los datos completados puede combinarse fácilmente para crear inferencias que tomen en consideración la variabilidad del muestreo y por ende la variabilidad de los valores faltantes. Además, como las $\boldsymbol{m}$ imputaciones provienen de más de un modelo, la incertidumbre de hallar el modelo correcto queda diluida por la variabilidad en las inferencias a través de los modelos utilizados para las $\boldsymbol{m}$ imputaciones

[Little and Rubin, 2002].

Pero el método de imputación múltiple presenta varios inconvenientes. La necesidad de hallar un modelo probabilístico en el conjunto de datos completos puede ser una ardua tarea pues en la mayoría de los casos resulta muy complicado ajustar un modelo que tome en consideración la información de los valores faltantes y que al mismo tiempo consiga convergencia en los estimados. Además contrario a la imputaciónn sencilla, la imputación múltiple requiere de un gran esfuerzo computacional y de una cantidad considerable de tiempo, además, la disponibilidad de *software* es limitada.

Sin embargo, existen varios métodos de este tipo que se han probado con buenos resultados. En esta tesis se pondrá a prueba un método de imputación múltiple conocido por **FRITZ** (*Federal Reserve Imputation Technique Zeta*) diseñado e implantado por el economista Arthur B. Kennickell para solucionar el problema de datos faltantes en el **SCF** (*Survey of Consumer Finances*) [Kennickell, 1991, Kennickell, 1998].

4.4.2. El algoritmo FRITZ

A continuación se describe el algoritmo **FRITZ** de imputación múltiple. La estructura del mismo es influenciada por ideas del muestreo de GIBBS. La estructura del algoritmo es de alta complejidad y su construcción es motivada por obtener un marco lo suficientemente honesto y uniforme, basándose en la información de los datos disponibles. Es un método secuencial en el sentido de que busca imputar los datos faltantes variable por variable y valor por valor. Además **FRITZ** es un modelo iterativo que busca que las $\boldsymbol{m}$ imputaciones realizadas se tomen en consideración en todo momento. Las ideas fundamentales del algoritmo son tomadas de *Kennickell (1991)* y la descripción del modelo es tomada de las mismas fuentes. El modelo que se utiliza en esta tesis no es exactamente el propuesto por Kennickell, esto pues para variables de tipo multinomial no se imputaron bajo un modelo de predicción. Para estas variable se utilizó el método **KNN**. Sin embargo para las variables

de tipo contínuo y binario se pudo utilizar un modelo de predicción.

Sea $\boldsymbol{X^m}$ la matriz de datos con valores faltantes de $\boldsymbol{n}$ unidades por p variables, la cual se desea imputar. Entonces la idea principal es tomar variable por variable de la matriz $\boldsymbol{X^m}$ e imputar los valores observación por observación donde haya valores faltantes. Sea $\boldsymbol{X_j^m}$ una variable de la matriz $\boldsymbol{X^m}$ donde $\boldsymbol{X_j^m}$ cuenta con $\boldsymbol{r}$ unidades observadas y $\boldsymbol{n}$ **-** $\boldsymbol{r}$ unidades faltantes. Entonces cada variable $\boldsymbol{X_j^m}$ se imputará de la siguiente forma:

Primera iteración

- Para cada *i-ésimo* valor faltante $\boldsymbol{x_{ij}^m}$ de la variable $\boldsymbol{X_j^m}$ se identifican aquellas variables $\boldsymbol{X_k^{r_i}}$ donde hay valores observados en la *i-ésima* unidad correspondiente a ese *i-ésimo* valor faltante.

- Luego se ajusta un modelo donde $\boldsymbol{X_j^m}$ es la variable de respuesta y $\boldsymbol{X_k^{r_i}}$ es el conjunto de variables explicativas. El tipo de modelo depende del tipo de variable que se esté imputando, si $\boldsymbol{X_j^m}$ es contínua, el modelo ajustado es el modelo de regresión lineal gaussiano. Si $\boldsymbol{X_j^m}$ es discreta binaria, el modelo ajustado es el de regresión logística y si $\boldsymbol{X_j^m}$ es ordinal se imputará por el algoritmo de vecinos más cercanos **KNN**.

- Luego, para las variables contínuas se utilizan los valores de $\boldsymbol{X_j^{r_i}}$ para la **i - ésima** observación faltante denotados por $\boldsymbol{x_{ij}^{r_i}}$ y se obtiene una predicción $\boldsymbol{\hat{x}_{ij}^m}$ para el valor faltante $\boldsymbol{x_{ij}^m}$. Es decir que para cada valor faltante en una variable contínua la imputación tiene la siguiente forma:

$$\boldsymbol{\hat{x}_{ij}^m = \beta_0 + \sum_{k=1}^{r} \beta_k X_k^{r_i} + e_{ij}} \tag{4.7}$$

  donde $\boldsymbol{e_{ij}}$ es un residual del modelo ajustado escogido aleatoriamente.

  Para las variables discretas binarias, se ajusta un modelo de regresión logística y hallando el punto óptimo de corte y estimando una predicción $\boldsymbol{\hat{x}_{ij}^m}$ para el valor faltante $\boldsymbol{x_{ij}^m}$.

En el caso de las variables multinomiales, éstas no se imputan una a una, sino que se imputan todas al mismo tiempo para las unidades donde haya valores faltantes. Sea $\boldsymbol{X_o^m}$ el conjunto de variables ordinales con valores faltantes, entonces las imputaciones para esas variables son de la forma:

$$\hat{\boldsymbol{X}}_{\boldsymbol{o}}^{\boldsymbol{m}} = \boldsymbol{KNN}(\boldsymbol{X}_{\boldsymbol{o}}^{\boldsymbol{m}}, \boldsymbol{Y}) \tag{4.8}$$

donde $\mathbf{Y}$ es la variable de respuesta del conjunto de datos en general. Cabe señalar que lo ideal sería hacer imputación por **KNN** en el conjunto $\boldsymbol{X_o^m}$ utilizando todas las variables $\boldsymbol{X^m}$. Por lo que las imputaciones para las variables ordinales permanecerán fijas a través de todas las iteraciones.

Al final de la primera iteración se obtiene un conjunto de datos completos denotados por $\boldsymbol{X^{fill,1}}$.

Iteraciones subsiguientes

Para las siguientes iteraciones se imputará de la misma forma que en la primera iteración. A continuación se describe la iteración $\boldsymbol{t}$ donde al final se obtiene un conjunto de datos completos $\boldsymbol{X^{fill,t}}$. Para lograr este cometido se utilizan los datos completos obtenidos en la iteración anterior $\boldsymbol{X^{fill,t-1}}$. El procedimiento es como sigue:

- El conjunto de datos $\boldsymbol{X^m}$ se imputa de la misma forma variable por variable y observación por observación al igual que en la primera iteración. La diferencia en este caso es que las variables explicativas $\boldsymbol{X_k^{r_i}}$ no son selecionadas del conjunto de datos $\boldsymbol{X^m}$ sino del conjunto $\boldsymbol{X^{fill,t-1}}$.

- Es decir, que las imputaciones $\hat{\boldsymbol{x}}_{\boldsymbol{ij}}^{\boldsymbol{m,t}}$ para la iteración $\boldsymbol{t}$ utiliza las observaciones $\boldsymbol{x_{ij,fill}^{k,t-1}}$. En una variable contínua las imputaciones se rep-

resentan como sigue:

$$\hat{\boldsymbol{x}}_{ij}^{m,t} = \boldsymbol{\beta}_{0,t-1} + \sum_{k=1}^{r} \boldsymbol{\beta}_{k,t-1} \boldsymbol{X}_k^{fill,t-1} + \boldsymbol{e}_{ij}^{t} \tag{4.9}$$

Las variables discretas se imputan de forma similar pero con un modelo de regresión logística y como ya se indicó, las variables ordinales se imputan con **KNN** y permanecen constantes.

Como se observa, en el algoritmo **FRITZ** se utilizan las imputaciones anteriores para obtener nuevas imputaciones.

$$\boldsymbol{X}^{fill,1} \rightarrow \boldsymbol{X}^{fill,2} \rightarrow \cdots \rightarrow \boldsymbol{X}^{fill,t-1} \rightarrow \boldsymbol{X}^{fill,t} \tag{4.10}$$

Con el conjunto imputado $\boldsymbol{X}^{fill,t}$ se ajusta el modelo de regresión logística con la variable de respuesta $\mathbf{Y}$ y se halla el poder de separación del modelo. En ese sentido **FRITZ** simula el algoritmo **EM**, pues en cada iteración se obtienen estimados de forma iterativa de tal forma que se halle convergencia en la estimación del parámetro de interés. Como la iteración $\boldsymbol{t}$ utiliza sólo la iteración anterior ***t-1*** el algoritmo **FRITZ** se convierte en un método ***markoviano*** de imputación.

En esta tesis, debido al tiempo requerido para imputar mediante **FRITZ**, sólo se utilizaron cinco iteraciones. Es decir que el conjunto de datos completos que se utiliza al final es $\boldsymbol{X}^{fill,5}$. Como se puede observar los métodos de imputacion múltiple, como en este caso el algoritmo **FRITZ** son relativamente complicados y requieren un enorme esfuerzo computacional. El código de **FRITZ** se modifica dependiendo de la base de datos con que se esté trabajando. Los resultados se muestran en el Capítulo 6.

Capítulo 5

Metodología para la obtención de resultados

5.1. Introducción

La metodología que se muestra a continuación es distintiva de esta tesis. La idea general es comparar los métodos de imputación en un modelo de regresión logística a través del área bajo la curva **ROC**. Para esto se eliminarán valores de un conjunto de datos completos en el conjunto de variables explicativas **X**. En la variable de respuesta **Y** se supondrá que hay completa observación en las unidades. Luego de la eliminación de subconjuntos de datos en **X** se imputará en esa misma base de datos mediante todos los métodos de interés y luego se ajustará un modelo de regresión logística y se estimará el área (**ABC**)bajo la curva **ROC** para cada método de imputación. De ese estimado se caculará un sesgo con respecto al parámetro (*área real bajo la curva ROC*). Esos sesgos se compararán utilizando métodos no paramétricos. El mismo experimento se llevará a cabo con 4 bases de datos distintas.

5.2. Cálculo de los estimados

Para comenzar se calculará el parámetro de interés que se define como el área bajo la curva **ROC** en el conjunto de datos completos, es decir del conjunto sin haberle removido ningún valor. Luego se calcularán los estimados

obtenidos por cada método y para cada proporción de datos faltantes. De ahí se calcularán los sesgos entre los estimados y el parámetro.

5.2.1. Cálculo del parámetro

Para generar resultados mediante las simulaciones primero se debe calcular el parámetro $\boldsymbol{\theta}$ que representa el área bajo la curva **ROC** del modelo de regresión logística. Éste se obtiene de la siguiente forma:

1. Para el conjunto de datos; se definen cuáles serán las variables explicativas (**X**) y cuál será la variable de respuesta (**Y**). La variable de respuesta debe estar codificada con 0's y 1's, los cuales conforman las clases $\boldsymbol{C_0}$ y $\boldsymbol{C_1}$ respectivamente.

2. Se ajustan las variables a través de un modelo de regresión logística (**Y vs. X**) utilizando la función *glm* con la instrucción *family = binomial* del programado estadístico **R** y se hallan los valores ajustados $\hat{\boldsymbol{Y}}$.

3. Se calculan la sensitividad (*sen*) y la especificidad (*esp*) con probabilidades de corte $\boldsymbol{p = 0.01,0.02,\ldots,0.99}$. Luego se hallan los pares ordenados **(*sen, 1 - esp*)** que componen la curva **ROC**. Para llevar a cabo este cálculo se programó la función *ROC* en el programado **R**

4. Se calcula el área **ABC** bajo la curva **ROC**. Esta área es el parámetro $\boldsymbol{\theta}$ con el cual se calcularán todos los sesgos de las simulaciones en el conjunto de datos. Para calcular el **ABC** se utilizó la *Regla Trapazoidal* la cual se programó bajo el nombre de la función *Area* en el programado **R**.

5.2.2. Los estimados en un proceso recurrente

Ya que se tiene el parámetro, se procederá a describir un proceso recurrente o simulación para hallar un estimado del área bajo la curva **ROC** con

una proporción de datos faltantes dada. Se enumeran los pasos a seguir para conseguir un estimado.

1. Se generan valores faltantes en las variables explicativas **X** con la función *GenPat* descrita en el Capítulo 3, donde cada entrada en la matriz tiene la misma probabilidad de ser eliminada. La función *GenPat* recibe como entrada la matriz de datos **X** y una proporción **p** de datos faltantes. El conjunto con datos faltantes se denota por $\boldsymbol{X_{miss}}$.

2. Se imputan los datos faltantes en $\boldsymbol{X_{miss}}$ mediante todos los métodos de imputación, uno por uno, y se obtiene el conjunto completo $\boldsymbol{X_{imp,j}}$ donde j representa uno de los siguientes métodos de imputación: **IMEAN, ICMEAN, IMED, ICMED, IRS, ICRS, KNN, ADC y FRITZ**. Es decir, se obtienen nueve conjuntos de datos distintos $\boldsymbol{(X_{imp,1}, \ldots, X_{imp,9})}$, uno por cada método de imputación. **ADC** no es un método de imputación sino un método para manejar datos faltantes que aparece en la mayoría de los programados estadísticos como opción estándar y funciona tal y como se describió en el Capítulo 3.

3. Se ajusta el modelo de regresión logística, $\mathbf{Y}$ **vs** $\boldsymbol{X_{imp}}$ para cada una de las nueve bases de datos imputadas y se calcula la sensitividad y la especificidad y se halla la curva **ROC** para cada método de imputación.

4. Por último se calcula el **ABC** para cada método y la proporción de datos faltantes dada.

Debido a la particularidad de cada base de datos, el número de procesos recurrentes que se ejecutarán será entre 25 y 50 dependiendo del conjunto. Por lo tanto se calcularán de 20 a 50 estimados de las áreas bajo la curva para cada método y cada proporción dada con el conjunto de datos. Entonces se define $\boldsymbol{\hat{\theta}_{ijk}}$ como el estimado del **ABC** para la *i-ésima* proporción de datos faltantes ($\boldsymbol{p_i}$) el *j-ésimo* método de imputación ($\boldsymbol{M_j}$) y la *k-ésima* simulación. Por lo

tanto se obtiene la TABLA 5.1 donde $\boldsymbol{k = 1, 2, \cdots, 50}$ y varía dependiendo del conjunto de datos.

TABLA 5.1: **Estimados de las áreas bajo la curva**

	M_1	M_2	$\cdots$	M_j	$\cdots$	M_9
p_1	$\hat{\theta}_{111}$	$\hat{\theta}_{121}$	$\cdots$	$\hat{\theta}_{1j1}$	$\cdots$	$\hat{\theta}_{191}$
	$\hat{\theta}_{112}$	$\hat{\theta}_{122}$	$\cdots$	$\hat{\theta}_{1j2}$	$\cdots$	$\hat{\theta}_{192}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	$\hat{\theta}_{11k}$	$\hat{\theta}_{12k}$	$\cdots$	$\hat{\theta}_{1jk}$	$\cdots$	$\hat{\theta}_{19k}$
p_2	$\hat{\theta}_{211}$	$\hat{\theta}_{221}$	$\cdots$	$\hat{\theta}_{2j1}$	$\cdots$	$\hat{\theta}_{291}$
	$\hat{\theta}_{212}$	$\hat{\theta}_{222}$	$\cdots$	$\hat{\theta}_{2j2}$	$\cdots$	$\hat{\theta}_{292}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ldots$	$\vdots$
	$\hat{\theta}_{21k}$	$\hat{\theta}_{22k}$	$\cdots$	$\hat{\theta}_{2jk}$	$\cdots$	$\hat{\theta}_{29k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_i	$\hat{\theta}_{i11}$	$\hat{\theta}_{i21}$	$\cdots$	$\hat{\theta}_{ij1}$	$\cdots$	$\hat{\theta}_{i91}$
	$\hat{\theta}_{i12}$	$\hat{\theta}_{i22}$	$\cdots$	$\hat{\theta}_{ij2}$	$\cdots$	$\hat{\theta}_{i92}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	$\hat{\theta}_{i1k}$	$\hat{\theta}_{i2k}$	$\cdots$	$\hat{\theta}_{ijk}$	$\cdots$	$\hat{\theta}_{i9k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_{10}	$\hat{\theta}_{1011}$	$\hat{\theta}_{1021}$	$\cdots$	$\hat{\theta}_{10j1}$	$\cdots$	$\hat{\theta}_{1091}$
	$\hat{\theta}_{1012}$	$\hat{\theta}_{1022}$	$\cdots$	$\hat{\theta}_{10j2}$	$\cdots$	$\hat{\theta}_{1092}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	$\hat{\theta}_{101k}$	$\hat{\theta}_{102k}$	$\cdots$	$\hat{\theta}_{10jk}$	$\cdots$	$\hat{\theta}_{109k}$

5.2.3. Los sesgos absolutos

El cálculo de los sesgos absolutos corresponde a la siguiente ecuación.

$$\boldsymbol{S_{ijk} = |\hat{\theta}_{ijk} - \theta|} \tag{5.1}$$

Por lo tanto se obtiene la tabla que aparece en la TABLA 5.2.

El sesgo absoluto es la variable de interés que se pondrá a prueba a través de los métodos de imputación y las proporciones de datos faltantes. No obs-

TABLA 5.2: **Sesgos absolutos**

	M_1	M_2	$\cdots$	M_j	$\cdots$	M_9
p_1	S_{111}	S_{121}	$\cdots$	S_{1j1}	$\cdots$	S_{191}
	S_{112}	S_{122}	$\cdots$	S_{1j2}	$\cdots$	S_{192}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	S_{11k}	S_{12k}	$\cdots$	S_{1jk}	$\cdots$	S_{19k}
p_2	S_{211}	S_{221}	$\cdots$	S_{2j1}	$\cdots$	S_{291}
	S_{212}	S_{222}	$\cdots$	S_{2j2}	$\cdots$	S_{292}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	S_{21k}	S_{22k}	$\cdots$	S_{2jk}	$\cdots$	S_{29k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_i	S_{i11}	S_{i21}	$\cdots$	S_{ij1}	$\cdots$	S_{i91}
	S_{i12}	S_{i22}	$\cdots$	S_{ij2}	$\cdots$	S_{i92}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	S_{i1k}	S_{i2k}	$\cdots$	S_{ijk}	$\cdots$	S_{i9k}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
p_{10}	$S_{10,11}$	$S_{10,21}$	$\cdots$	$S_{10,j1}$	$\cdots$	$S_{10,91}$
	$S_{10,12}$	$S_{10,22}$	$\cdots$	$S_{10,j2}$	$\cdots$	$S_{10,92}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	$S_{10,1k}$	$S_{10,2k}$	$\cdots$	$S_{10,jk}$	$\cdots$	$S_{10,9k}$

tante para las pruebas se utilizará el sesgo promedio por método y por proporción. El arreglo de la TABLA 5.3 presenta los sesgos promedios;

TABLA 5.3: **Sesgos absolutos promedios**

	M_1	M_2	$\cdots$	M_j	$\cdots$	M_9
p_1	$\bar{S}_{11}$	$\bar{S}_{12}$	$\cdots$	$\bar{S}_{1j}$	$\cdots$	$\bar{S}_{19}$
p_2	$\bar{S}_{21}$	$\bar{S}_{22}$	$\cdots$	$\bar{S}_{2j}$	$\cdots$	$\bar{S}_{29}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
p_i	$\bar{S}_{i1}$	$\bar{S}_{i2}$	$\cdots$	$\bar{S}_{ij}$	$\cdots$	$\bar{S}_{i9}$
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
p_{10}	$\bar{S}_{10,1}$	$\bar{S}_{10,2}$	$\cdots$	$\bar{S}_{10,j}$	$\cdots$	$\bar{S}_{10,9}$

5.3. Comparación de los métodos de imputación

Los métodos de imputación se comparan de dos maneras distintas utilizando métodos no paramétricos. Primero se comparan de forma global utilizando la prueba de *Friedman* que es análoga a la prueba de ANOVA en dos direcciones. En *Friedman* los bloques corresponden a las proporciones $\boldsymbol{p_i}$ de datos faltantes y los tratamientos a los métodos de imputación $\boldsymbol{M_j}$. Luego se comparan los métodos de imputación en cada proporción utilizando la prueba de *Friedman* pero en este caso los bloques serán los proesos recurrentes los cuales varían en cantidad dependiendo el conjunto de datos trabajado. La estimación del **ABC** en un proceso recurrente no es independiente pues se está estimando sobre el mismo conjunto de datos faltantes para cada método de imputación, esto fuerza a tratar cada proceso recurrente como un bloque separado aunque pertenezca a la misma proporción de datos faltates. Cabe además señalar que la prueba de Friedman asume que no existe interacción entre los tratamientos (métodos de imputación) y los bloques (proporción de datos faltantes ó procesos recurrentes).

5.3.1. Comparación global mediante la prueba no paramétrica de Friedman

Esta comparación es la primera que se lleva a cabo pues va a indicar cómo se comporta cada método de imputación a través de todas las proporciones de datos faltantes en general. Los supuestos de esta comparación con la prueba de *Friedman* son los siguientes:

1. El modelo para la prueba está dado por;

$$\boldsymbol{y_{ij}} = \boldsymbol{\mu} + \boldsymbol{\beta_i} + \boldsymbol{\tau_j} + \boldsymbol{e_{ij}} \tag{5.2}$$

   donde $\boldsymbol{y_{ij}}$ corresponde al sesgo promedio para la *i-ésima* proporción de datos faltantes y el *j-ésimo* método de imputación. $\boldsymbol{\mu}$ es la media global desconocida, $\boldsymbol{\beta_i}$ es el efecto del la *i-ésima* proporción y $\boldsymbol{\tau_j}$ es el efecto del *j-ésimo* método de imputación.

2. Los errores $\boldsymbol{e_{ij}}$ son mutuamente excluyentes. No son independientes pues en la prueba de Friedman no es necesario el supuesto de distribución alguna [Hollander and Wolfe, 1973].

3. Cada error proviene de la misma población contínua.

Entonces, las hipótesis que se someten a prueba son las siguientes:
$\boldsymbol{H_0 : \tau_1 = \tau_2 = ... = \tau_9}$
$\boldsymbol{H_a}$: Al menos existe un $\boldsymbol{\tau_j}$ que difiere de los demás.

La prueba de *Friedman* se ejecuta con la instrucción en **R** *friedman.test*. Luego utilizando esta misma prueba se lleva a cabo una comparación múltiple entre los métodos para saber cuál de ellos es el más adecuado. Se definirá como el método más adecuado aquél que produzca el menor sesgo promedio.

5.3.2. Comparación global múltiple mediante la prueba de Friedman

Para saber cual método en general provoca el menor sesgo promedio es necesario llevar a cabo comparaciones múltiples entre todos los métodos. Esta comparación se ejecuta siempre y cuando se rechace la hipótesis nula en la prueba de *Friedman*. Para esto se necesita una función que calcule los rangos para cada método de imputación. Se programó la función *Score* para el cálculo de los rangos en la prueba de *Friedman* y la salida de ésta se ilustra en la TABLA 5.4.

TABLA 5.4: **Rangos para comparación múltiple utilizando la prueba no paramétrica de Friedman**

	M_1	M_2	$\cdots$	M_j	$\cdots$	M_9
p_1	R_{11}	R_{12}	$\cdots$	R_{1j}	$\cdots$	R_{19}
p_2	R_{21}	R_{22}	$\cdots$	R_{2j}	$\cdots$	R_{29}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
p_i	R_{i1}	R_{i2}	$\cdots$	R_{ij}	$\cdots$	R_{i9}
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
p_{10}	$R_{10,1}$	$R_{10,2}$	$\cdots$	$R_{10,j}$	$\cdots$	$R_{10,9}$

donde $\mathbf{1} \leq \boldsymbol{R_{ij}} \leq \mathbf{9}$ y $\boldsymbol{R_{il}} \neq \boldsymbol{R_{im}}$ para $\boldsymbol{l} \neq \boldsymbol{m}$. Estos rangos se calculan individualemente para cada proporción de datos faltantes asignándole de forma ascendente un rango al sesgo promedio de cada método en esa proporción. Es decir que en la proporción dada el menor rango (1) corresponde al método con el menor sesgo promedio y el método con el mayor sesgo promedio recibe el rango mayor (9). Si hay empates entre los rangos de los métodos de imputación se le asigna el rango promedio a cada uno. Luego del cálculo de los rangos se procede a sumar los mismos y para esto se programó la función *Rsums*. Como resultado se obtiene el siguiente vector:

$$\langle \boldsymbol{\Sigma R_{\cdot 1}}, \boldsymbol{\Sigma R_{\cdot 2}}, \ldots, \boldsymbol{\Sigma R_{\cdot 9}} \rangle \tag{5.3}$$

De ahí se procede a comparar las sumas de los rangos de cada método a través de todas las proporciones de datos faltantes. Los métodos: $\boldsymbol{M_l}$ y el $\boldsymbol{M_m}$ se comparan de la siguiente forma:

$$\boldsymbol{D_{l,m} = |\Sigma R_{\cdot l} - \Sigma R_{\cdot m}| > r(\alpha, w, z)} \tag{5.4}$$

donde $\boldsymbol{r(\alpha, w, z)}$ corresponde al estadístico para la comparación múltiple de sumas de rangos de *Friedman*, [Hollander and Wolfe, 1973]. Los valores de $\boldsymbol{w}$ y $\boldsymbol{z}$ corresponden al número de métodos de imputación y al número de proporciones de datos faltantes respectivamente. En este caso $\boldsymbol{w = 9}$ y $\boldsymbol{z = 10}$.

Se programaron las funciones *Differences* y *Significant* para obtener las diferencias entre todas las sumas de rangos de todos los métodos y obtener cuáles de ellas son significativas respectivamente. Las diferencias forman una matriz triangular simétrica. Sólo se considerarán los elementos superiores de esta matriz. La forma de la matriz de diferencias se ilustra en la TABLA 5.5.

TABLA 5.5: **Matriz de diferencias**

	M_2	$\cdots$	M_9
M_1	$D_{1,2}$	$\cdots$	$D_{1,9}$
M_2		$\cdots$	$D_{1,9}$
		$\ddots$	$\vdots$
M_8			$D_{8,9}$

La matriz de significancia tiene la misma forma y los valores de la misma son 0's y 1's donde los 0's implican que no hay diferencias significativas y los 1's indican diferencias significativas. Antes de llevar a cabo las diferencias y las significancias es necesario ordenar las sumas de rangos ascendentemente.

5.3.3. Comparación por proporción p_i de datos faltantes

Para esta comparación se ejecuta nuevamente la prueba de *Friedman* para probar la siguiente hipótesis en cada proporción de datos faltantes.

$\boldsymbol{H_{0_i}}$ **:** $\boldsymbol{\gamma_{i1} = \gamma_{i2} = \ldots = \gamma_{i9}}$; para $\boldsymbol{i = 1, ..., 10}$

$\boldsymbol{H_{a_i}}$ **: Al menos existe algún** $\boldsymbol{\gamma_{ij}}$

En este caso la variable de interés en la prueba es el sesgo absoluto y no el sesgo absoluto promedio como en la prueba global. Aquí los bloques corresponden a los procesos recurrentes en cada proporción de datos faltantes, por lo que la asignación de rangos es para cada sesgo asoluto y no para el sesgo promedio por método de imputación y por proporción de datos faltantes. Por lo tanto se llevaran a cabo 10 nuevas pruebas de *Friedman*, una por cada proporción de datos faltantes en cada conjunto de datos. El arreglo de los rangos para la *i-ésima* proporción se muestra en la TABLA 5.6.

TABLA 5.6: **Rangos de la prueba de Friedman en cada proporción**

	M_1	M_2	$\ldots$	M_j	$\ldots$	M_9
$Proc_1$	r_{i11}	r_{i21}	$\ldots$	r_{ij1}	$\ldots$	r_{i91}
$Proc_2$	r_{i12}	r_{i22}	$\ldots$	r_{ij2}	$\ldots$	r_{i92}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$Proc_k$	r_{i1k}	r_{i2k}	$\ldots$	r_{ijk}	$\ldots$	r_{i9k}

donde $\boldsymbol{1 \leq r_{ijk} \leq 9}$ y $\boldsymbol{r_{ijk}}$ es el rango de la ***i-ésima*** porporción de datos faltantes $\boldsymbol{p_i}$, el ***j-ésimo método*** de imputación $\boldsymbol{M_j}$ y el ***k -ésimo*** proceso recurrente. En cada porporción $\boldsymbol{p_i}$ se aplica la prueba de Friedman de la misma manera queen la comparación global.

5.3.4. Comparación múltiple por probabilidad p_i para los métodos de imputación

Aquí se muestra como se comparó cada método de imputación en cada proporción de datos faltantes. De rechazarse la hipótesis nula anterior, se

lleva a cabo la siguiente comparación entre cada par de métodos utilizando la suma de rango de Friedman representada por un vector de la forma:

$$\langle \boldsymbol{\Sigma r_{i1}}, \boldsymbol{\Sigma r_{i2}}, \ldots, \boldsymbol{\Sigma r_{i9}} \rangle \tag{5.5}$$

Este vector se ordena ascendentemente y se le aplican las funciones *Differences* y *Significant* y así se ordenan los métodos de imputación de acuerdo a su efectividad para la proporción $\boldsymbol{p_i}$. Entonces, cada par de métodos se comparan con respecto a sus sumas de rangos de la siguiente forma;

$$\boldsymbol{d_{l,m} = |\Sigma r_l - \Sigma r_m| > q_{\alpha,J,\infty} \sqrt{\frac{K(J)(J+1)}{12}}} \tag{5.6}$$

donde $\boldsymbol{q_{\alpha,J,\infty}}$ es el estadísico de Friedman para diferencias entre sumas de rangos cuando el número de bloques $\boldsymbol{K}$ es grande para un nivel de significancia $\boldsymbol{\alpha}$ y el número de tratamientos $\boldsymbol{J}$ [Hollander and Wolfe, 1973]. En este caso $\boldsymbol{\alpha = 0.05}$, $\boldsymbol{J = 9}$ en todas las pruebas y $\boldsymbol{K = 25, 50}$ dependiendo del conjunto de datos. Si la desigualdad se cumple, los métodos $\boldsymbol{M_l}$ y $\boldsymbol{M_m}$ difieren significativamente con respecto a las sumas de rangos las cuales representan los sesgos absolutos en la proporción de datos faltantes $\boldsymbol{p_i}$.

En este capítulo se discutió la metodología a seguir en las simulaciones y en el próximo capítulo se discutirán los resultados de estos procedimientos en las cuatro bases de datos.

Capítulo 6

Resultados del Experimento

6.1. Introducción

A continuación se presentan las descripciones de los conjuntos de datos y los resultados de las pruebas de *Friedman* global y porporción de datos faltantes y de las comparaciones múltiples. Además se muestran otros resultados interesantes que nos ayudarán en la descripción de los resultados principales como medidas de variabilidad delos estimados. Tambíen se observó la relación entre el área bajo la curva ROC y el sesgo generado.

6.2. Descripción de los datos

Los datos para este proyecto provienen de varias fuentes, una de ellas es el *Machine Learning Database Repository* de la Universidad de Califonia, Irvine. Las bases de datos de esta universidad fueron facilitadas por el Dr. Edgar Acuña Fernández (Universidad de Puerto Rico en Mayagüez). La TABLA 6.1 muestra la definición de las bases de datos para las simulaciones y sus respectivas descripciones. Ninguna de las siguientes bases de datos contienen originalmente datos faltantes. En el número de variables no está incluida la variable de respuesta.

Las variables de respuesta de los conjuntos de datos están codificadas con 0's y 1's y representan las dos clases de interés. Las variables predictoras en

TABLA 6.1: **Descripción de los datos**

Datos	**Número de filas**	**Número de columnas**
Bupa	345	6
Diabetes	768	9
Bajopeso	189	10
German	1000	20

los conjuntos son de distintos tipos. En *Bupa* y *Diabetes* todas las variables predictoras son de tipo continuo. Las otras dos bases de datos son de tipo mixto, es decir, tienen variables de varios tipos, continuas y discretas. Las discretas las subdividiremos en binarias y ordinales. Las variables explicativas de los datos *Bajopeso* se componen de 2 variables continuas, 2 binarias y 5 variables ordinales. Por último la base de datos *German* se compone de 14 variables ordinales, 3 variables continuas y 3 variables binarias.
El número de repeticiones en cada conjunto es el siguiente:

1. Datos Bupa: 50
2. Datos Diabetes: 50
3. Datos Bajopeso: 25
4. Datos German: 25

El número de simulaciones se escogió debido al tiempo disponible para llevar a cabo el experimento, pues para una iteración se debía imputar por todos los métodos antes de pasar a la siguiente. El mayor tiempo invertido por un proceso recurrente fue consumido por el algoritmo FRITZ de imputación múltiple. Otro factor para escoger el número de iteraciones fue el mecanismo de datos faltantes MCAR. En ciertos métodos de manejo e imputación de datos como el ADC, KNN y FRITZ, el número de unidades completamente observadas es de suma importancia y la ejecución de los mismos se puede ver afectada seriamente. Bajo el mecanismo **MCAR** utilizado en esta tesis cada

entrada de la matriz de datos tiene una probabilidad de ser eliminada,por lo que la probabilidad de que cada unidad sea completamente eliminada es directamente proporcional al número de variables del conjunto y claro está, a la proporción de datos faltantes utilizada. Por lo que puede darse el caso de que para una proporción grande de datos faltantes KNN no encuentre donantes en un conjunto con muchas variables, y no se pueda ejecutar. En el caso de los datos German no había datos disponibles que sirvieran de donantes en el algoritmo cuando la proporción de datos era de 35 % o más por lo que en el caso de de esta base de datos se consideraron 25 procesos recurrentes por cada proporción de datos faltantes (las cuales sólo llegaron hasta el 30 %). Además de eso, los datos German tienen una cantidad considerable de filas, por lo que FRITZ resultaba demasiado lento en proporciones grandes de datos faltantes. En los datos Bajopeso aunque sólo había 10 variables no se pudieron realizar con facilidad más de 25 procesos recurrentes, aunque sí se logró llevar a cabo hasta el 50 % de datos faltantes.

6.3. Resultados de las pruebas de Friedman

La TABLA 6.2 muestra los resultados de la prueba no paramétrica de Friedman para cada conjunto de datos. Se muestra el estadístico Chi-cuadrado de Friedman y el p-valor de la prueba. Nótese que en todas las pruebas los grados de libertad (**df**) son 8 pues, se están sometiendo a prueba 9 métodos. Nótese además que el p-valor es menor de 0.05 para cada conjunto de datos, por lo que se rechaza la hipótesis nula de la prueba y se concluye que existe suficiente evidencia para decir que al menos un método de imputación difiere significativamente de los demás con respecto al sesgo promedio.
Los gráficos de los métodos de imputación, **Sesgos promedios vs. Proporción de datos faltantes**, se muestran por conjunto de datos en las FIGURAS 6.1 - 6.4. Similarmente, los gráficos de **Áreas promedios vs. Proporción de datos faltantes** por conjunto de datos se muestran en las

TABLA 6.2: **Resumen de la prueba de Friedman para todas las bases de datos**

Datos	χ^2 de Friedman	P-valor
Bupa	78.24	1.104469e-13
Diabetes	79.44	6.337137e-14
Bajopeso	69.3867	6.509196e-12
German	40	3.20372e-06

FIGURAS 6.5 - 6.8. Nótese que al 0 % de datos faltantes el ABC corresponde al obtenido en los datos originales.

6.4. Comparaciones globales múltiples

En esta sección se muestran los vectores de las sumas de rangos de Friedman de los métodos de imputación para cada conjunto de datos. Los métodos se muestran de izquierda a derecha de menor a mayor suma de rangos, por lo que están acomodados de menor a mayor sesgo promedio a través de las proporciones de datos faltantes. Junto con cada uno de los vectores se muestran los resultados de las pruebas de comparación múltiple que se obtuvieron de las matrices de significancia (*Ver Apéndice*) las cuales provinieron de los resultados de las matrices de diferencias significativas (*Ver Apéndice*) entre las sumas de rangos. Para los datos *bupa*, *diabetes* y *bajopeso*, dos sumas de rangos son diferentes significativamente si la diferencia es mayor de 38. En los datos *german* el valor pico para diferencia significativa es 29. Nótese que las rayas aparecen bajo los métodos que según la prueba no son diferentes significativamente. Por ejemplo, en el primer caso de los datos *bupa*, la primera fila de rayas muestra que no existe diferencia significativa entre los métodos: ICMED, ICMEAN, ICRS, IMED e IMEAN, los cuales presentan la menor suma de rangos y por ende son los métodos que menor sesgo generan con respecto al área bajo la curva **ROC**. La segunda fila de rayas representa que no hay diferencia significativa entre las sumas de rangos de los métodos:

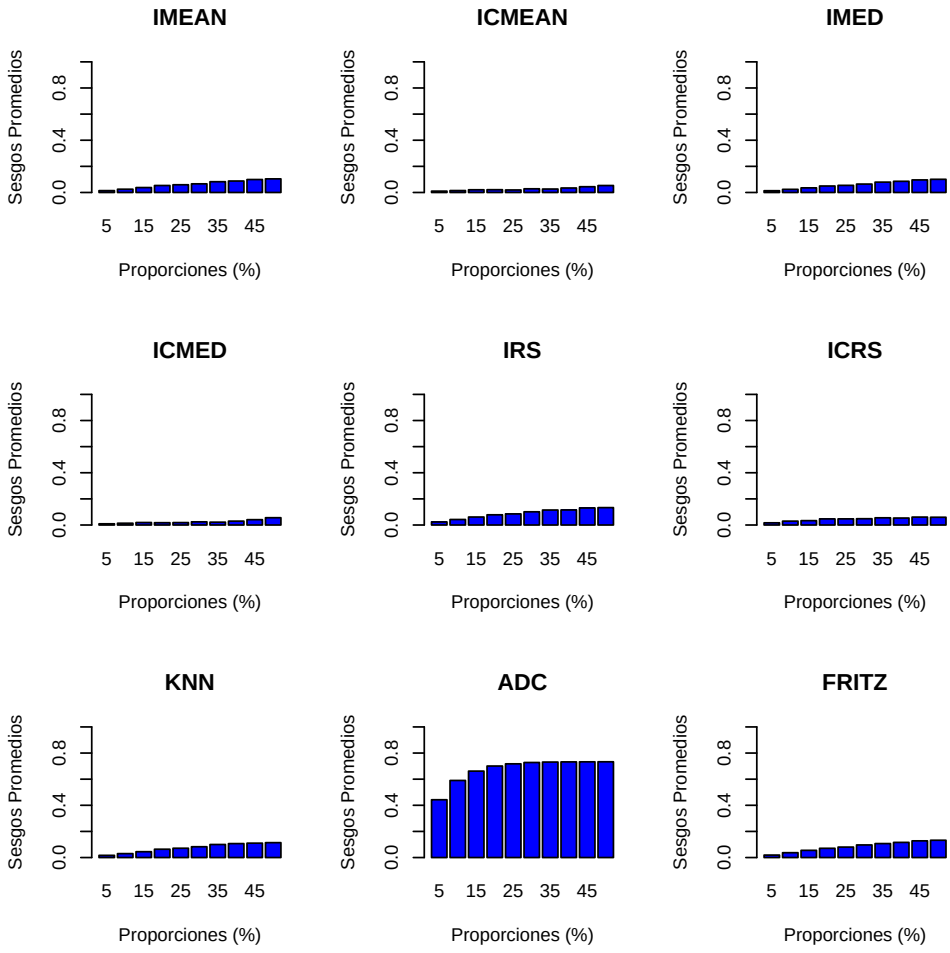


FIGURA 6.1: **Sesgos Promedio vs. Proporción de datos faltantes para los datos *bupa***

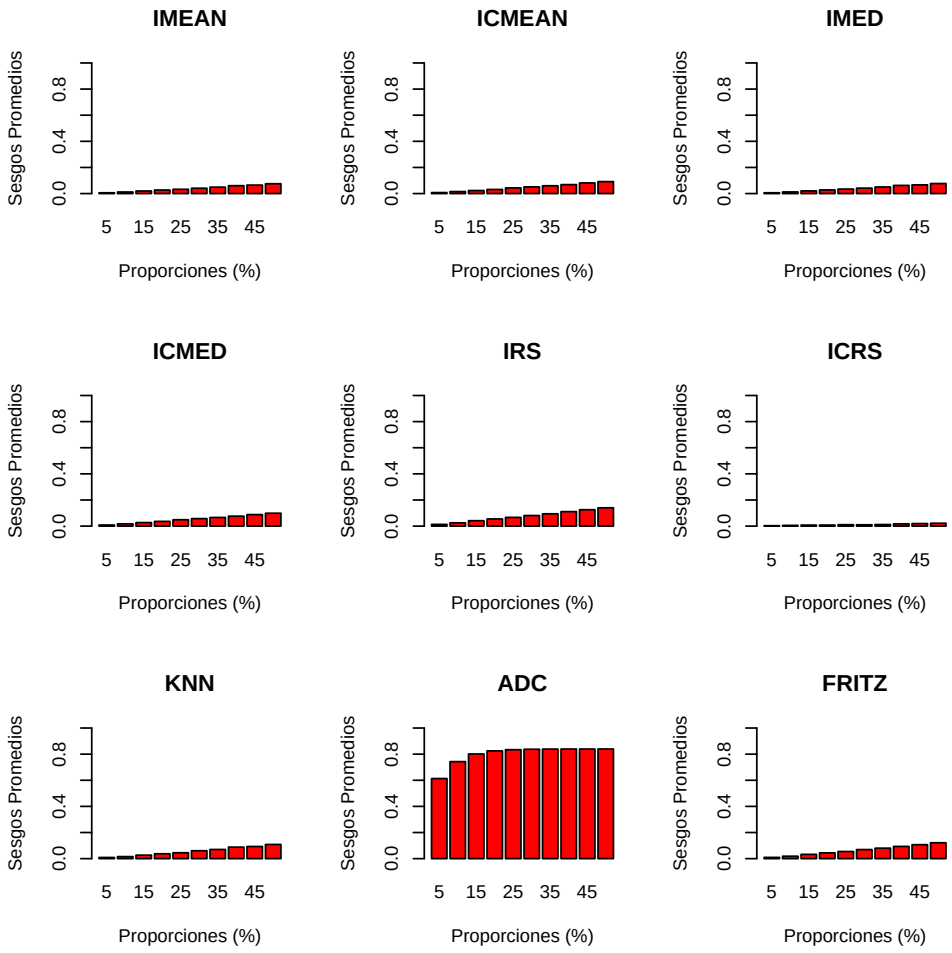


FIGURA 6.2: **Sesgos Promedio vs. Proporción de datos faltantes para los datos *diabetes***

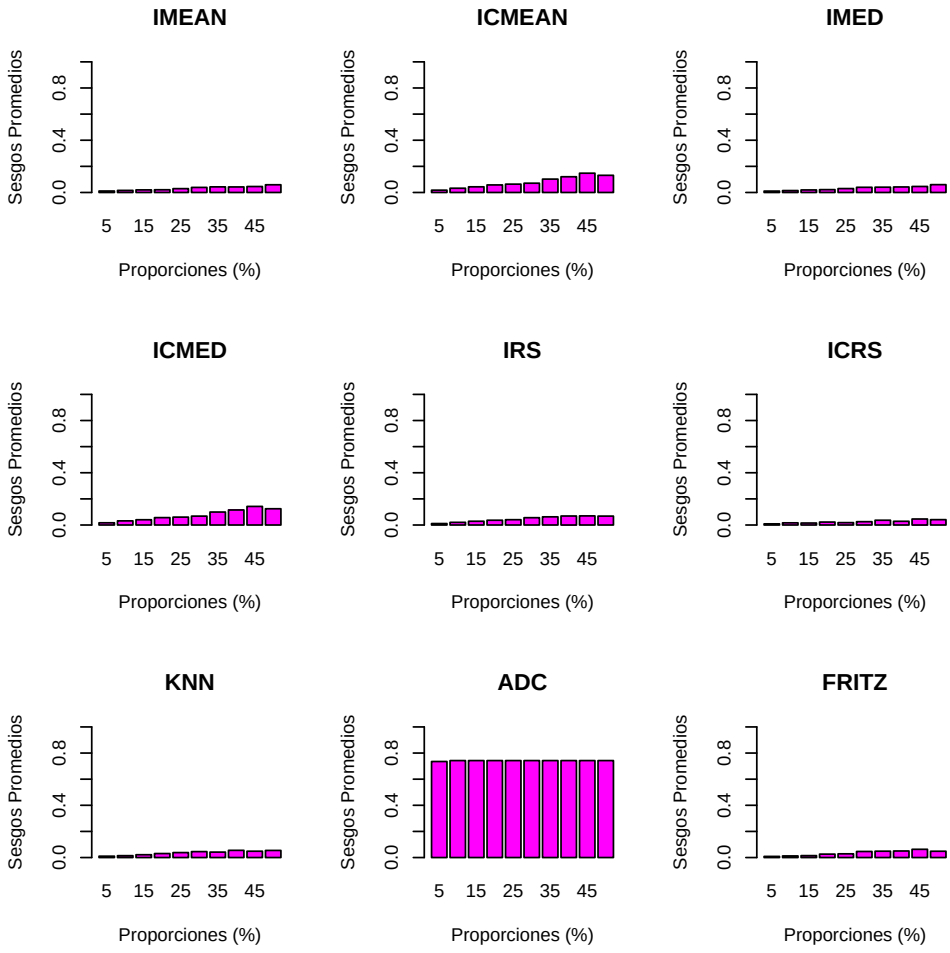


FIGURA 6.3: **Sesgos Promedio vs. Proporción de datos faltantes para los datos *bajopeso***

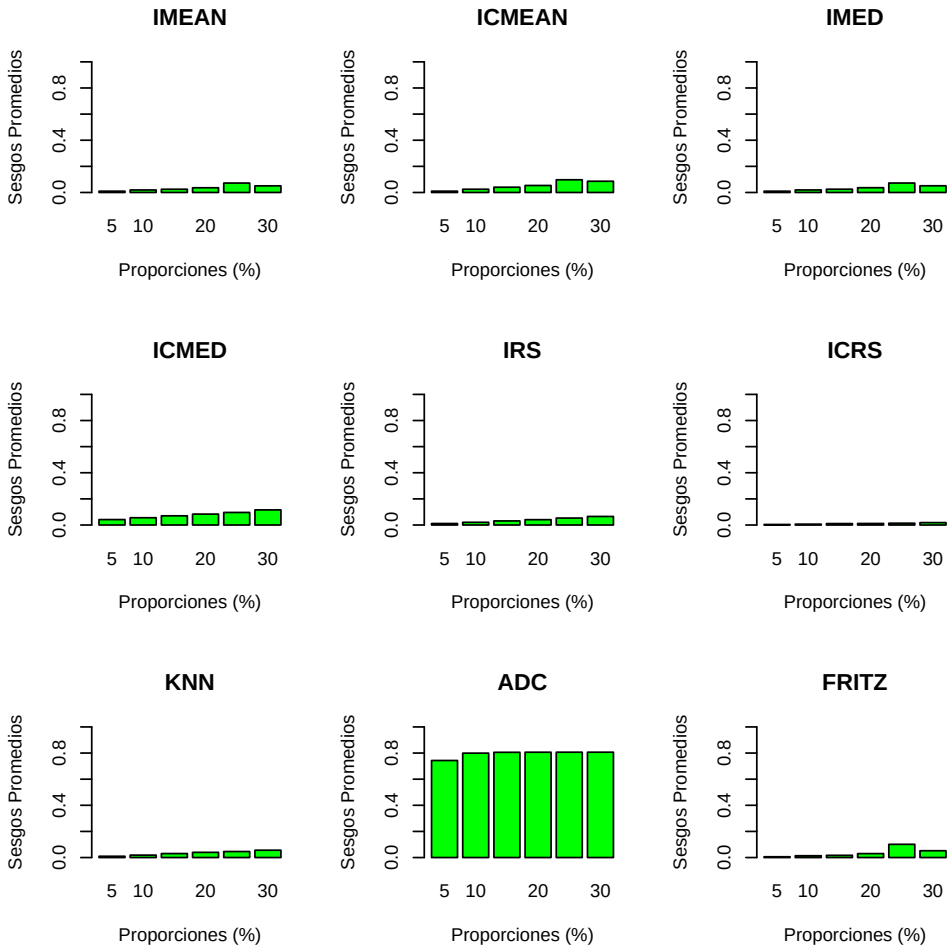


FIGURA 6.4: **Sesgos Promedio vs. Proporcion de datos faltantes para los datos *german***

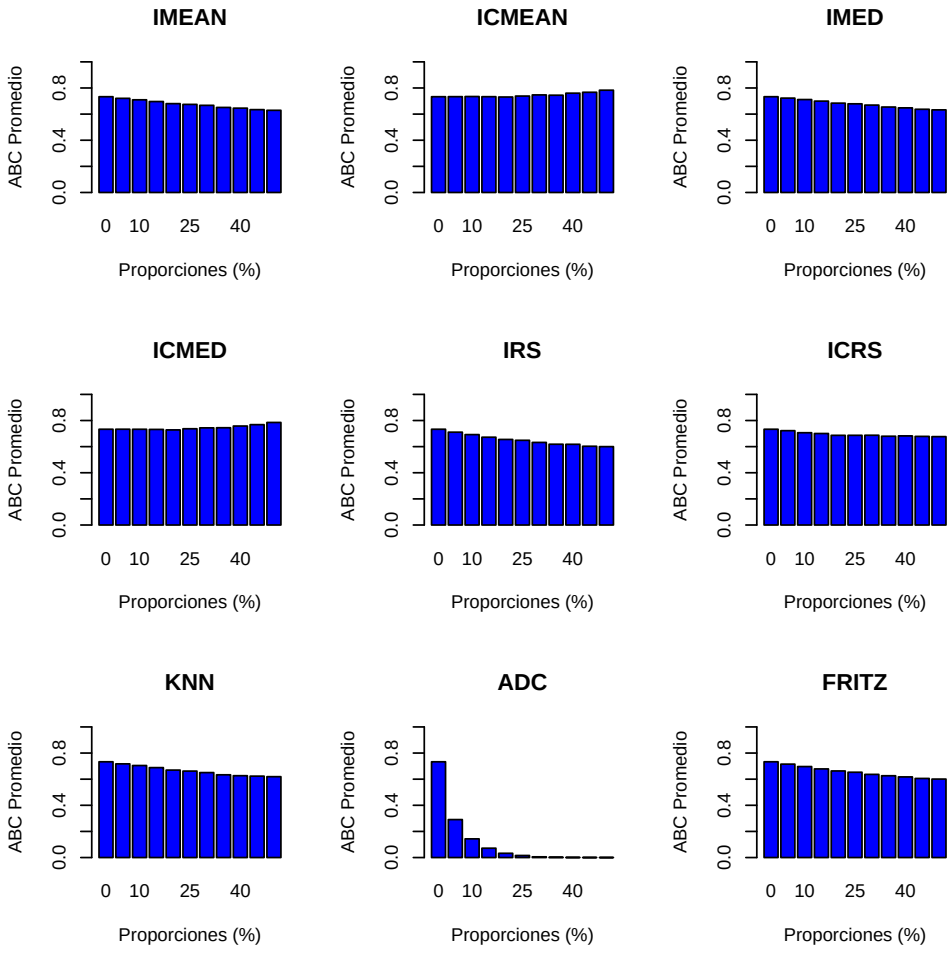


FIGURA 6.5: **Áreas Promedio vs. Proporción de datos faltantes para los datos *bupa***

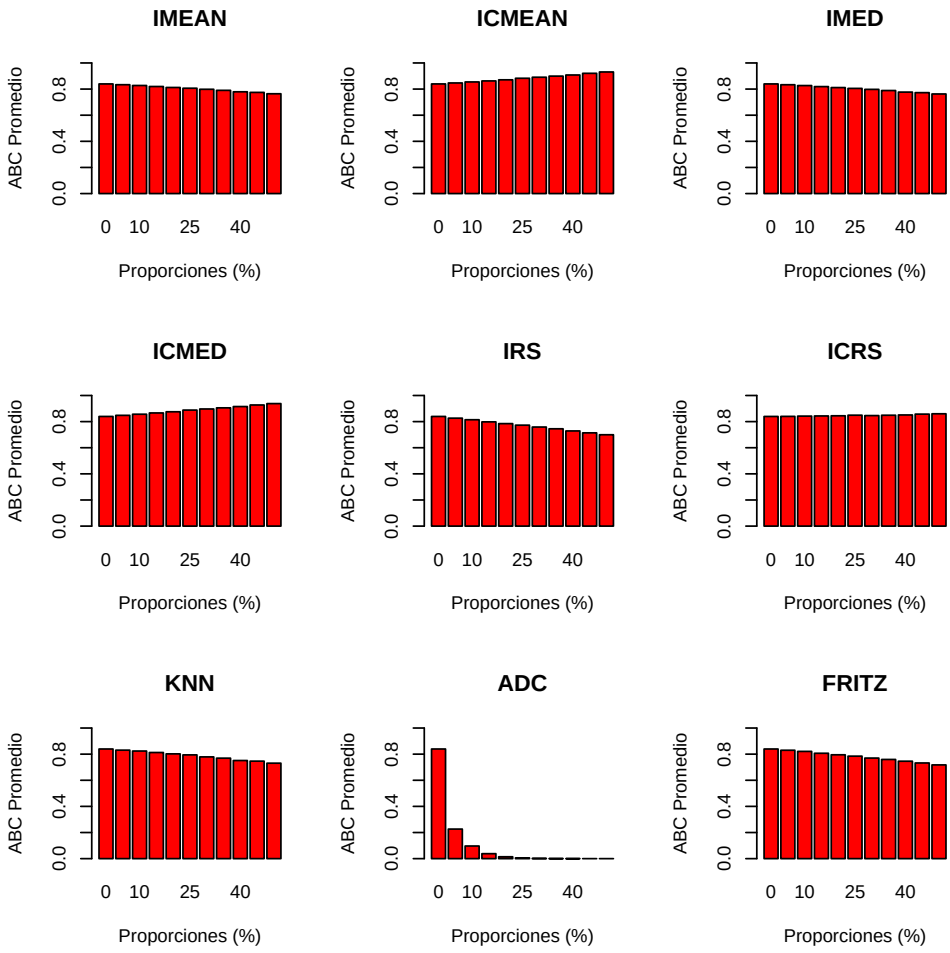


FIGURA 6.6: **Áreas Promedio vs. Proporción de datos faltantes para los datos *diabetes***

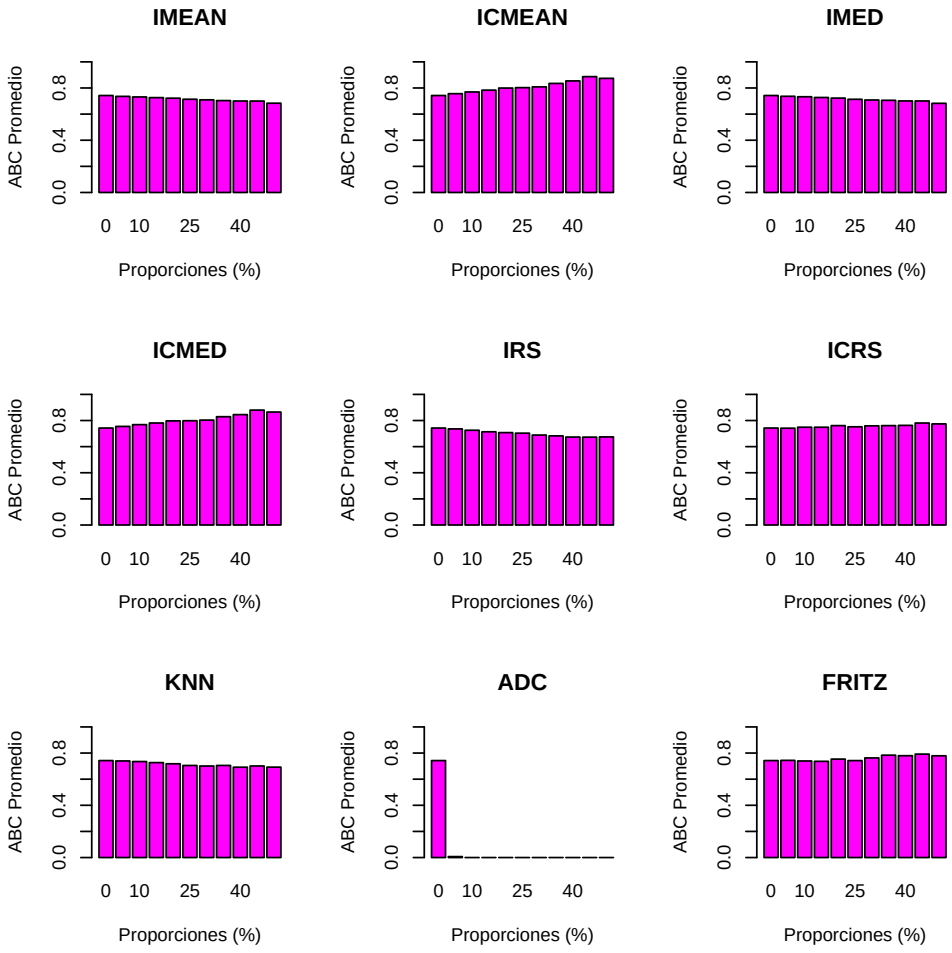


FIGURA 6.7: **Áreas Promedio vs. Proporción de datos faltantes para los datos *bajopeso***

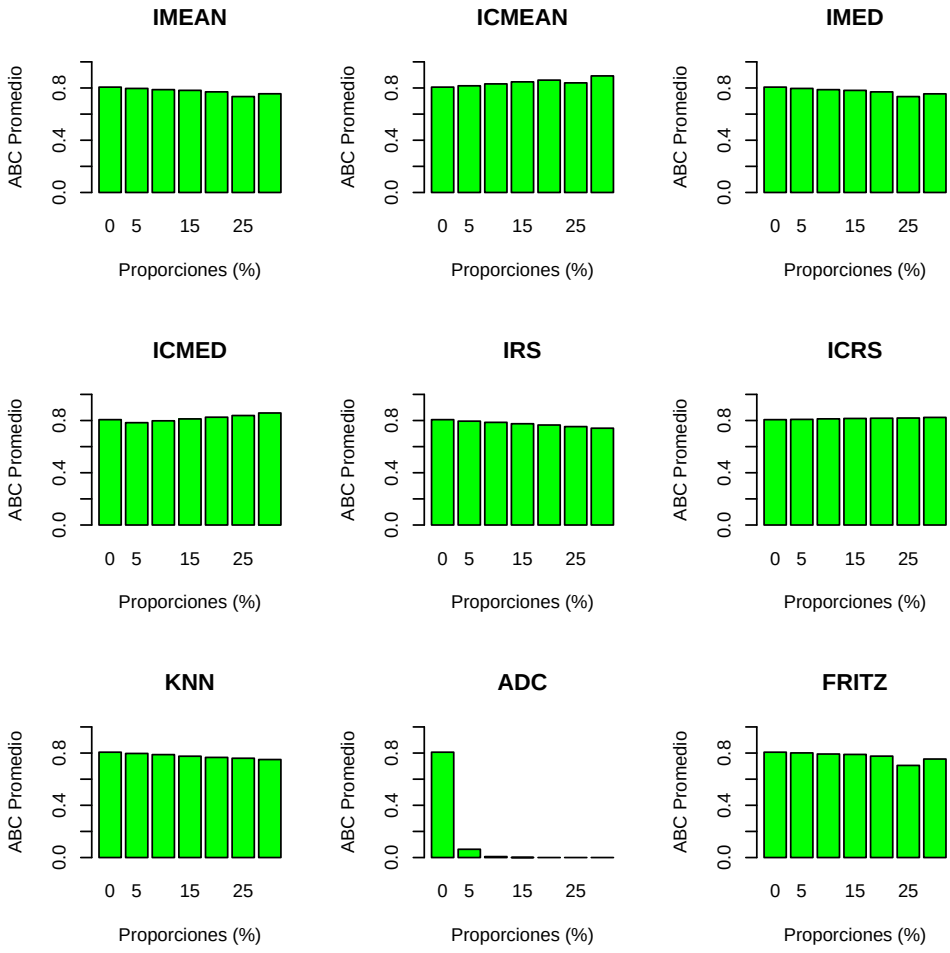


FIGURA 6.8: **Áreas Promedio vs. Proporción de datos faltantes para los datos *german***

ICRS, IMED, IMEAN, KNN Y FRITZ. Así sucesivamente la tercera y la cuarta fila y las otras filas de los demás conjuntos de datos. Los diagramas de comparación por porporción en la sección 6.5 se interpretan de la misma forma.

A continuación se muestran las sumas de rangos de las pruebas de Friedman globales y sus respectivos diagramas de las pruebas de significancia por conjunto de datos:

- **Resultados para los datos *Bupa***

  Suma de rangos

```
icmed icmean icrs imed imean knn fritz irs adc
   11     19   34   38    48  60    71  79  90
```

  Diagrama

```
       icmed icmean icrs imed imean knn fritz irs adc
        ----------------------------
                    -------------------------
                              -------------------
                                    -----------------
```

- **Resultados para los datos *Diabetes***

  Suma de Rangos

```
icrs imean imed icmean icmed knn fritz irs adc
  10    20   30     40    53  57    70  80  90
```

  Diagrama

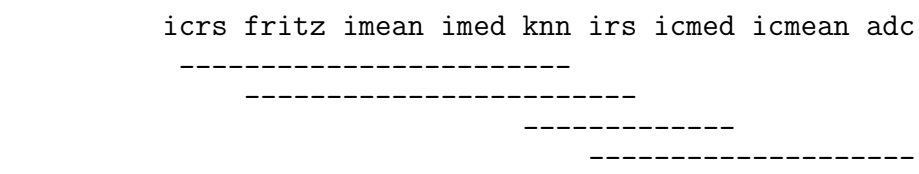
```
       icrs imean imed icmean icmed knn fritz irs adc
        --------------------
            --------------------------
                        ----------------------
                               ----------------------
```

- **Resultados para los datos *Bajopeso***

  Suma de rangos

```
icrs fritz imean imed knn irs icmed icmean adc
  19    30    31   31  39  60    70     80  90
```

Diagrama

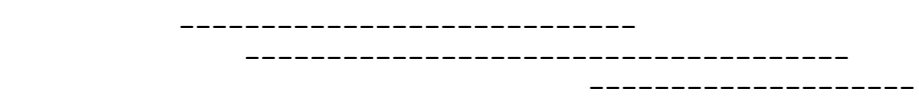


- **Resultados para los datos *German***

Suma de rangos

icrs	fritz	imean	knn	imed	irs	icmean	icmed	adc
6	20	21	23	25	34	41	46	54

Diagrama



6.5. Comparación de los métodos por proporción de datos faltantes

A continuación de muestra la información de las pruebas de Friedman para todas las bases de datos a través de todas las proporciones de datos faltantes. Nótese que los p-valores son menores de 0.05 en todos los casos, por lo que las pruebas indican que existe al menos un método que difiere significativamente de los demás con respecto a las sumas de rangos en todas las bases de datos y a través de las proporciones de datos faltantes. Luego se presentan los resultados de las pruebas de comparación múltiple medantes diagramas de rayas comoen la sección anterior (6.4). En el caso de los datos *bupa* y *diabetes*, dos métodos cuya diferencia en suma de rangos sea mayor o igual a 84.95 son significativamente diferentes. Para los datos *bajopeso* y *german* si las diferencias son mayores o iguales a 60.07 estas son significativamente diferentes.

Se muestra a continuación la información sobre las pruebas de Friedman por base de datos para cada proporción de datos faltanes. Se muestran el estdistíco de Chi Cuadrado y el P valor de las pruebas.

- **Resultados para los datos *Bupa***

  Información de la Prueba de Friedman por proporción de datos faltantes

	Chi cuadrado	P valor
5%	165.0022	1.436640e-31
10%	206.4480	2.794210e-40
15%	239.2972	3.189888e-47
20%	281.5147	3.518984e-56
25%	293.1200	1.198438e-58
30%	294.5464	5.958720e-59
35%	323.8933	3.353442e-65
40%	302.7253	1.082854e-60
45%	281.0133	4.497538e-56
50%	278.1973	1.784161e-55

- **Resultados para los datos *Diabetes***

  Información de la Prueba de Friedman por proporción de datos faltantes

	Chi cuadrado	P valor
5%	241.5307	1.073471e-47
10%	246.7627	8.363090e-49
15%	294.1387	7.276187e-59
20%	314.1067	4.082405e-63
25%	319.8560	2.431903e-64
30%	335.2320	1.281998e-67

35%	338.4160	2.683695e-68
40%	332.7627	4.310439e-67
45%	345.7813	7.198770e-70
50%	340.3040	1.061609e-68

- **Resultados para los datos *Bajopeso***

Información de la Prueba de Friedman por proporción de datos faltantes

	Chi cuadrado	P valor
5%	74.95565	5.034562e-13
10%	80.37333	4.111979e-14
15%	91.78667	2.016328e-16
20%	92.62863	1.359456e-16
25%	92.58667	1.386435e-16
30%	89.58933	5.634388e-16
35%	91.10340	2.775955e-16
40%	105.84533	2.713902e-19
45%	121.10933	1.956237e-22
50%	105.21641	3.652154e-19

- **Resultados para los datos *German***

Información de la Prueba de Friedman por proporción de datos faltantes

	Chi cuadrado	P valor
5%	99.2320	6.127342e-18
10%	125.1733	2.826366e-23
15%	131.5307	1.362309e-24
20%	153.3547	3.912070e-29

```
25%     138.5418 4.769476e-26
30%     152.2880 6.532162e-29
```

A continuación se presentan las sumas de rangos de Friedman y los diagramas de la comparaciones múltiple por base de datos y por proporción de datos faltantes.

- **Resultados para los datos *Bupa***

Suma de rangos del 5 %

```
     icmed icmean imean imed knn icrs fritz irs adc
     151.5  163.5   219  220 234  248   266 298 450
```

Diagrama del 5 %

```
     icmed icmean imean imed knn icrs fritz irs adc
     ---------------------------
           -------------------------
                 -----------------------------
                                          ---
```

Suma de rangos del 10 %

```
     icmed icmean imed imean icrs knn fritz irs adc
       130    143  199   207  251 255   292 323 450
```

Diagrama del 10 %

```
     icmed icmean imed imean icrs knn fritz irs adc
     ----------------------
                  -------------------
                            ------------------
                                          ---
```

Suma de rangos del 15 %

```
     icmed icmean icrs imed imean   knn fritz   irs adc
       121    139  192  203   217 268.5   307 352.5 450
```

Diagrama del 15 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     ----------------------
           ----------------------
                  -------------------
                                 -------------
                                              ---
```

Suma de rangos del 20 %

```
     icmed icmean icrs imed imean knn fritz irs adc
        88    115  202  203   234 286   325 347 450
```

Diagrama del 20 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     ------------
                  -------------------
                                 -------------
                                              ---
```

Suma de rangos del 25 %

```
     icmed icmean icrs imed imean knn fritz irs adc
        94    102  185  208   245 289   323 354 450
```

Diagrama del 25 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     ------------
           ------------
                  ---------------
                            ---------
                                 -------------
                                              ---
```

Suma de rangos del 30 %

```
     icmed icmean  icrs imed imean   knn fritz irs adc
       100    120 152.5  219   229 285.5   342 352 450
```

Diagrama del 30 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     -----------------
                  ---------------
                       --------------
                                 -------------
                                              ---
```

Suma de rangos del 35 %

```
     icmed icmean icrs imed imean knn fritz irs adc
        85    102  144  225   253 310   325 356 450
```

Diagrama del 35 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     -----------------
                 ---------
                      ----------
                           ----------------
                                 -------------
                                              ---
```

Suma de rangos del 40 %

```
     icmed icmean icrs imed imean knn fritz irs adc
       103    117  130  226   240 304   336 344 450
```

Diagrama del 40 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     -----------------
                      --------------
                                 -------------
                                              ---
```

Suma de rangos del 45 %

```
     icmed icmean icrs imed imean knn fritz irs adc
      116    122  127  230   251 285   333 336 450
```

Diagrama del 45 %

```
     icmed icmean icrs imed imean knn fritz irs adc
     -----------------
                      --------------
                           ----------------
                                 -------------
                                              ---
```

Suma de rangos del 50 %

```
     icrs icmed icmean imed imean knn fritz irs adc
      111   132    134  231   234 279   331 348 450
```

Diagrama del 50 %

```
icrs icmed icmean imed imean knn fritz irs adc
-----------------
                  --------------
                                -------------
                                              ---
```

- **Resultados para los datos *Diabetes***

Suma de rangos del 5 %

```
icrs imean imed icmean icmed knn fritz irs adc
 100   166  178    198   251 263   286 358 450
```

Diagrama del 5 %

```
icrs imean imed icmean icmed knn fritz irs adc
---------------
     -----------------
          -----------------
               ----------------
                     ---------------
                               ---------
                                         ---
```

Suma de rangos del 10 %

```
icrs imean imed icmean knn icmed fritz irs adc
  89   167  189    208 229   276   282 360 450
```

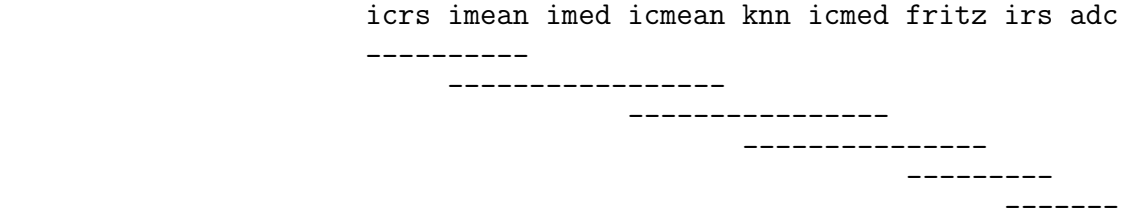
Diagrama del 10 %

```
icrs imean imed icmean knn icmed fritz irs adc
----------
     ---------------------
               ----------------------
                           ---------------
                                         ---
```

Suma de rangos del 15 %

```
icrs imean imed icmean knn icmed fritz irs adc
  76   146  169    196 258   268   313 374 450
```

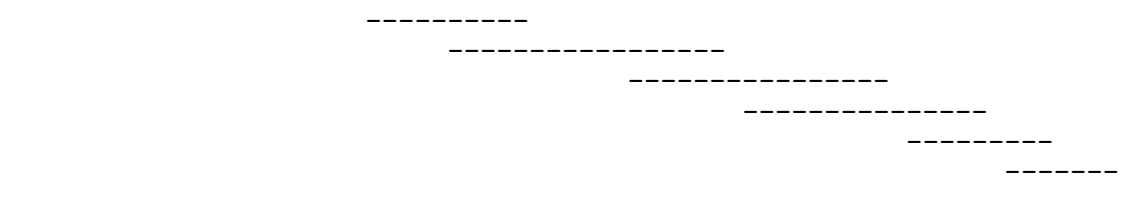
Diagrama del 15 %



Suma de rangos del 20 %

icrs	imean	imed	icmean	knn	icmed	fritz	irs	adc
63	146	171	193	262	263	321	381	450

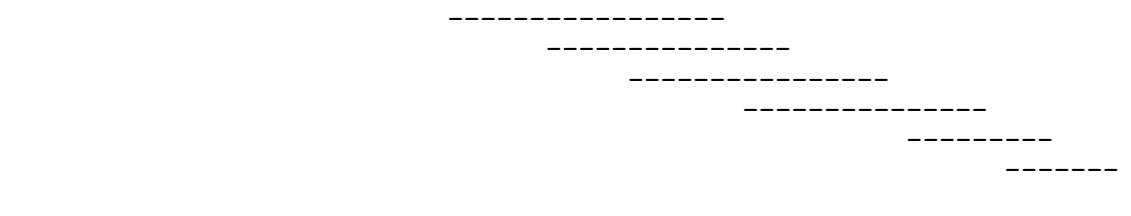
Diagrama del 20 %



Suma de rangos del 25 %

icrs	imean	imed	icmean	knn	icmed	fritz	irs	adc
63	130	173	210	244	290	304	386	450

Diagrama del 25 %



Suma de rangos del 30 %

icrs	imean	imed	icmean	knn	icmed	fritz	irs	adc
55	137	152	205	272	273	330	376	450

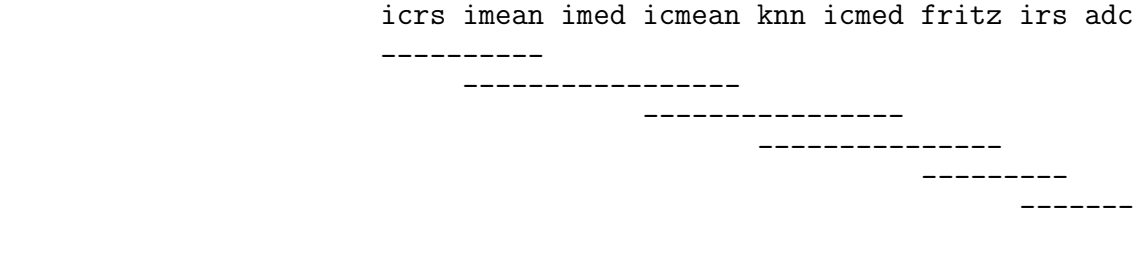
Diagrama del 30 %

```
     icrs imean imed icmean knn icmed fritz irs adc
     ----------
          -----------------
                     ----------------
                            ---------------
                                       ---------
                                            -------
```

Suma de rangos del 35 %

```
     icrs imean imed icmean knn icmed fritz irs adc
       52   136  156    202 270   277   331 376 450
```

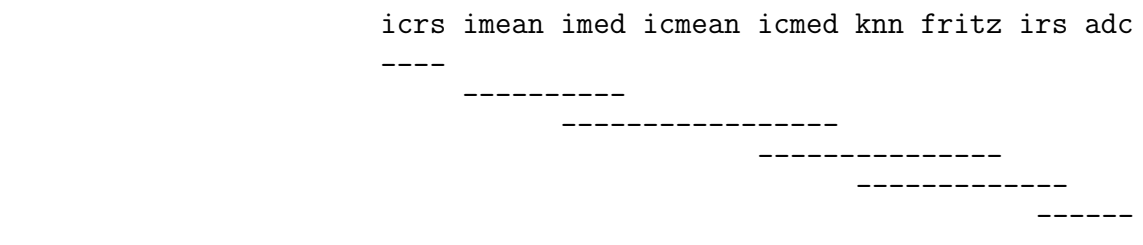
Diagrama del 35 %



Suma de rangos del 40 %

```
     icrs imean imed icmean icmed knn fritz irs adc
       52   140  177    182   255 294   322 378 450
```

Diagrama del 40 %



Suma de rangos del 45 %

```
     icrs imean imed icmean icmed knn fritz irs adc
       50   133  166    200   260 271   331 389 450
```

Diagrama del 45 %

```
icrs imean imed icmean icmed knn fritz irs adc
----------
     -----------------
                -----------------
                       ---------------
                                ---------
                                      -------
```

Suma de rangos del 50 %

```
icrs imean imed icmean icmed knn fritz irs adc
  54   136  159    199   258 280   330 384 450
```

Diagrama del 50 %

```
  icrs imean imed icmean icmed knn fritz irs adc
  ----------
       -----------------
                  -----------------
                         ---------------
                                  ---------
                                        -------
```

- **Resultados para los datos *Bajopeso***

Suma de rangos del 5 %

```
fritz icrs  imed knn irs imean icmed icmean adc
   85   96 103.5 105 112 117.5   136    145 225
```

Diagrama del 5 %

```
fritz icrs imed knn irs imean icmed icmean adc
----------------------------------------
                                       ---
```

Suma de rangos del 10 %

```
fritz imed icrs knn imean irs icmed icmean adc
   89   91  100 101   103 124   141    151 225
```

Diagrama del 10 %

```
fritz imed icrs knn imean irs icmed icmean adc
---------------------------------
      -----------------------------------
                                       ---
```

Suma de rangos del 15 %

```
     fritz icrs imed imean knn icmed irs icmean adc
        71   81  103   105 112   130 141    157 225
```

Diagrama del 15 %

```
     fritz icrs imed imean knn icmed irs icmean adc
     ------------------------------
           ----------------------------
                ------------------------------
                                          ---
```

Suma de rangos del 20 %

```
     imean imed icrs fritz knn irs icmed icmean adc
      85.5 86.5 87.5  95.5 109 128 147.5  160.5 225
```

Diagrama del 20 %

```
     imean imed icrs fritz knn irs icmed icmean adc
     ----------------------------
                ------------------------
                           --------------------
                                          ---
```

Suma de rangos del 25 %

```
     icrs fritz imean imed knn irs icmed icmean adc
       62    93    97  103 126 130   132    157 225
```

Diagrama del 25 %

```
     icrs fritz imean imed knn irs icmed icmean adc
     ---------------------
          ----------------------------
                ------------------------------
                                          ---
```

Suma de rangos del 30 %

```
     icrs imean imed knn fritz irs icmed icmean adc
       64    92  102 106   114 132   137    153 225
```

Diagrama del 30 %

```
     icrs imean imed knn fritz irs icmed icmean adc
     ------------------------
          ----------------------------
               ------------------------------
                                           ---
```

Suma de rangos del 35 %

```
     icrs imed  knn imean fritz irs icmed icmean adc
       74 92.5 93.5   100   104 140   142    154 225
```

Diagrama del 35 %

```
     icrs imed knn imean fritz irs icmed icmean adc
     ------------------------
          ----------------------------
                   ----------------------------
                                           ---
```

Suma de rangos del 40 %

```
     icrs imed imean fritz knn irs icmed icmean adc
       67   88    90    92 110 135   150    168 225
```

Diagrama del 40 %

```
     icrs imed imean fritz knn irs icmed icmean adc
     ------------------------
          ------------------------
               ------------------------
                          --------------------
                                           ---
```

Suma de rangos del 45 %

```
     knn imed icrs imean fritz irs icmed icmean adc
      81   82   85    87    96 120   167    182 225
```

Diagrama del 45 %

```
     knn imed icrs imean fritz irs icmed icmean adc
     ----------------------------
                               ---------
                                   ----------------
```

Suma de rangos del 50 %

```
     icrs fritz knn irs imean  imed icmed icmean adc
       79    84  89 104 105.5 107.5   155    176 225
```

Diagrama del 50 %

```
     icrs fritz knn irs imean imed icmed icmean adc
     ----------------------------
                   --------------------
                                  ------------
                                        ----------
```

- **Resultados para los datos *German***

Suma de rangos del 5 %

```
     icrs fritz knn imed imean icmed icmean irs adc
       51    72 122  123   126   126    137 144 224
```

Diagrama del 5 %

```
     icrs fritz knn imed imean icmed icmean irs adc
     ----------
          -------------------------
                ------------------------------
                                           ---
```

Suma de rangos del 10 %

```
     icrs fritz knn imean imed irs icmed icmean adc
       37    69 103   116  127 135   151    163 224
```

Diagrama del 10 %

```
     icrs fritz knn imean imed irs icmed icmean adc
      ----------
          --------------------
                ------------------------------
                                           ---
```

Suma de rangos del 15 %

```
     icrs fritz imean imed irs knn icmean icmed adc
       46    56    97  108 136 137    160   161 224
```

Diagrama del 15 %

```
     icrs fritz imean imed irs knn icmean icmed adc
     ----------------
          ----------------
               ------------------
                    -------------------------
                                           ---
```

Suma de rangos del 20 %

```
     icrs fritz imean imed knn irs icmed icmean adc
       28    68    97  101 128 130   174    175 224
```

Diagrama del 20 %

```
     icrs fritz imean imed knn irs icmed icmean adc
     ----------
          --------------------
               ------------------
                         --------------------
                                  ----------------
```

Suma de rangos del 25 %

```
     icrs imean imed fritz knn irs icmed icmean adc
       33  89.5 89.5  97.5 108 141   171  176.5 219
```

Diagrama del 25 %

```
     icrs imean imed fritz knn irs icmed icmean adc
     ---------------
          ------------------------
                              ----------------
                                  ----------------
```

Suma de rangos del 30 %

```
     icrs imean imed fritz knn irs icmed icmean adc
       30    85   89    93 107 147   175    175 224
```

Diagrama del 30 %

```
     icrs imean imed fritz knn irs icmed icmean adc
     ---------------
          --------------------
               ------------------
                              ----------------
                                  ----------------
```

6.6. Correlación entre ABC y Sesgo

En esta sección se presentan resultados adicionales con respecto al área bajo la curva **ROC** y el correspondiente sesgo absoluto. Se calcularon las correlaciones de *Spearman* entre el Sesgo y el ABC para observar alguna tendencia. Las correlaciones se calcularonó en cada base de datos y en términos globales. Se presentan a continuación las correlaciones de *Spearman* entre el Sesgo y el ABC para las cuatro bases de datos analizadas. Se construyeron dos vectores: uno de ABC's y otro de sesgos absolutos que constan de todos los estimados de todas las simulaciones a través de todos los métodos de imputación y a través de todas las proporciones de datos faltantes. En los datos *bupa* se llevaron a cabo 50 procesos recurrentes para cada uno de los 9 métodos de imputación y para cada una de las 10 proporciones de datos faltantes. Por lo que el vector de áreas ABC's y el vector de sesgos absolutos tienen longitudes de 4,500 elementos cada uno. En el caso del conjunto de datos *diabetes* es similar con ambos vectores de 4,500 elementos de longitud. Los datos *bajopeso* tienen ambos 2,250 elementos en cada vector de sesgos y de áreas, pues se llevaron a cabo 25 procesos por método y proporción en cada una de ellas. En los datos *german* los vectores tienen 1,350 elementos cada uno pues para esta base de datos se calcularon 25 procesos recurrentes para los 9 métodos, pero sólo hasta el 30 % de datos faltantes. La TABLA 6.3 muestra las correlaciones de *Spearman* entre el Sesgo y el ABC en cada base de datos y global y en la FIGURA 6.9 se presentan las gráficas de dispersión de Sesgo vs. ABC para cada conjunto de datos. La TABLA 6.4 muestra las correlaciones de *Spearman* entre el Sesgo y el ABC en cada base de datos por proporción de datos faltantes.

TABLA 6.3: **Correlación de *Spearman* entre el Sesgo y el ABC para cada conjunto de datos**

Datos	Correlación Área-Sesgo
Bupa	-0.9021752
Diabetes	-0.6069487
Bajopeso	-0.2511903
German	-0.4563674
Todos	-0.5294698

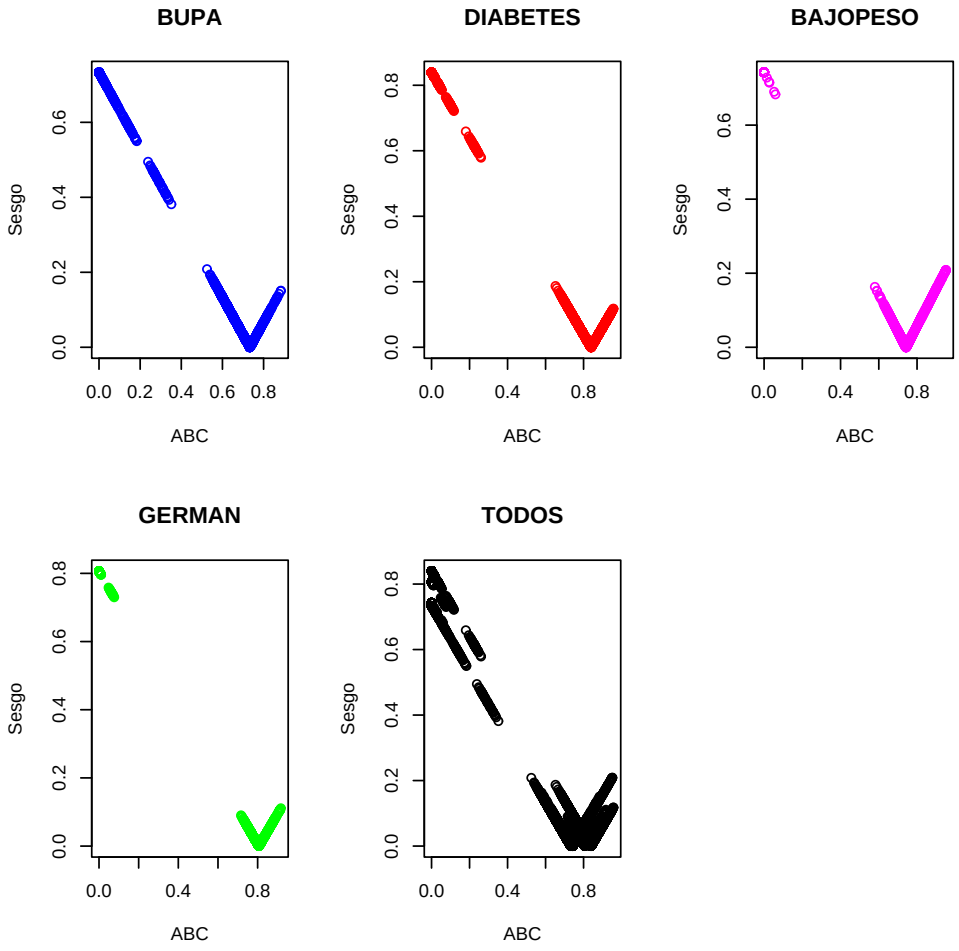


FIGURA 6.9: **Gráficos de dispersión de Sesgo vs. ABC por conjunto de datos y de todos en general**

TABLA 6.4: **Correlación de *Spearman* entre el Sesgo y el Área por proporción de datos faltantes en cada base de datos**

Proporción	Bupa	Diabetes	Bajopeso	German	Todos
5 %	-0.8451006	-0.633858	-0.254352	-0.57339	-0.5994328
10 %	-0.9133325	-0.5358602	-0.2494812	-0.3956912	-0.5713784
15 %	-0.9405083	-0.5376531	-0.3209705	-0.3375587	-0.5228137
20 %	-0.9769444	-0.5511793	-0.1750532	-0.2828601	-0.5204015
25 %	-0.9711033	-0.4728495	-0.35609	-0.3321372	-0.4851731
30 %	-0.951864	-0.5044845	-0.24715	-0.2623974	-0.4754941
35 %	-0.9720665	-0.5210082	-0.1741781		-0.5416833
40 %	-0.9420658	-0.5738538	-0.1747663		-0.5261037
45 %	-0.9101997	-0.5176986	-0.04328626		-0.4941487
50 %	-0.9021998	-0.5422803	-0.1288560		-0.4902549

6.7. Intervalos de confianza del poder de separación de los modelos para cada conjunto de datos

En esta sección se muestra los intervalos de confianza para los estimados del ABC. Éstos se llevaron a cabo directamente de los resultados de las simulaciones en cada base de datos y para cada método de imputación y proporción de datos faltantes. La idea es observar en cada base de datos, para cada método de imputación y proporción de datos faltantes cuáles de esos intervalos contienen al parámetro que en este caso es ABC de calculada de los datos originales sin haberles borrado información. Para esto, se llevó a cabo la prueba de normalidad de Shapiro a los estimados por conjunto de datos, método de imputación y proporción de datos faltantes. Luego se calculó el promedio y su error estándar para calcular las cotas de los intervalos de confianza. A continuación se muestran los P-valores de las pruebas de normalidad para cada conjunto de datos por método y proporción. Un P-valor mayor de 0.05 indica que los datos se distribuyen normalmente.

- **P valores de la prueba de normalidad de *Shapiro* en los datos *Bupa*.**

```
> spv_bupa
```

```
        imean     icmean       imed      icmed       irs       icrs       knn
5%  0.27881907 0.55710065 0.78618487 0.97899598 0.3991596 0.67036373 0.6198741
10% 0.51738223 0.40685535 0.30930299 0.19646970 0.6682233 0.04952454 0.8181375
15% 0.02700093 0.06667974 0.03207228 0.03259573 0.1044389 0.12189771 0.3149569
20% 0.20787740 0.09793709 0.19700867 0.23646005 0.2922742 0.44206931 0.5181043
25% 0.46921632 0.63517069 0.52917764 0.67968045 0.2436069 0.32119593 0.7047258
30% 0.86607382 0.98978864 0.83502229 0.84861906 0.8481687 0.59779660 0.9003088
35% 0.12429954 0.27194579 0.25926694 0.61846975 0.6246821 0.24482265 0.3340234
40% 0.28494311 0.65519028 0.01937662 0.55398909 0.4274258 0.29631810 0.8469630
45% 0.16141920 0.06680213 0.12042280 0.26939294 0.5463246 0.15070529 0.1709092
50% 0.83590218 0.44593012 0.84454182 0.10554331 0.5010270 0.83151291 0.1243628
            adc      fritz
5%  0.2272860655 0.41977948
10% 0.5157436379 0.71170888
15% 0.5431567164 0.03044966
20% 0.0588164560 0.01546865
25% 0.0525524018 0.37474394
30% 0.0205963095 0.75968635
35% 0.3412081707 0.35592884
40% 0.9623989129 0.74353340
45% 0.0076366501 0.02871399
50% 0.0000473583 0.90526212
```

- **P valores de la prueba de normalidad de *Shapiro* en los datos *Diabetes*.**

```
> spv_diabetes
```

```
         imean     icmean       imed       icmed       irs       icrs
5%  0.987005537 0.17837419 0.91737072 0.362816358 0.8078659 0.94510799
10% 0.442742153 0.11812405 0.31884553 0.603582581 0.5328138 0.47800057
15% 0.782399670 0.62898205 0.48507525 0.535445157 0.7634910 0.55160725
20% 0.337237991 0.18914554 0.55977351 0.503726019 0.9279757 0.80247774
25% 0.177013467 0.21967352 0.23581145 0.915655133 0.6280286 0.08181998
30% 0.489572685 0.54548188 0.50018459 0.350933699 0.5052716 0.61763637
35% 0.004346095 0.09235869 0.02262813 0.008880144 0.1114792 0.31094353
40% 0.395863714 0.11901045 0.69930666 0.698925707 0.9991941 0.05527022
45% 0.958678162 0.96528603 0.67624462 0.431575486 0.8656350 0.64376569
50% 0.489456494 0.11023887 0.51373846 0.068274623 0.2018001 0.13921052
          knn          adc     fritz
5%  0.08640372 5.526170e-01 0.8632872
10% 0.40649790 4.947158e-01 0.8361182
15% 0.49096014 5.858958e-03 0.4679413
20% 0.38034164 1.443277e-01 0.5598056
25% 0.27899315 6.230073e-01 0.3430755
30% 0.03008872 6.636612e-02 0.4088553
35% 0.16253060 3.117668e-01 0.1112705
40% 0.61769611 7.638619e-03 0.5256293
```

```
45% 0.84153705 1.071603e-05 0.1719305
50% 0.61943971 2.834205e-06 0.9660420
```

- **P valores de la prueba de normalidad de *Shapiro* en los datos *Bajopeso*.**

```
> spv_bajopeso
```

```
           imean     icmean        imed       icmed          irs       icrs
5%  0.4927692224 0.38793574 0.504153718 0.486722098 0.1133999675 0.36056868
10% 0.4157037374 0.15062439 0.569419254 0.145841126 0.8732709872 0.14138707
15% 0.9982493897 0.17035287 0.877832964 0.075350634 0.1519720078 0.41561573
20% 0.8780292312 0.45563017 0.937832835 0.776847026 0.8774620276 0.63596772
25% 0.9844148115 0.43599873 0.884119475 0.528430240 0.9644591325 0.04951649
30% 0.0004713936 0.39268489 0.000940736 0.374390982 0.0006157071 0.07365888
35% 0.5095946225 0.03069774 0.500598871 0.021735506 0.5833771778 0.58711184
40% 0.3686344316 0.06361536 0.627734510 0.031541480 0.2655057110 0.56673014
45% 0.2418836047 0.00467266 0.405188256 0.002755914 0.5328177817 0.12237292
50% 0.5814087648 0.14289443 0.456256110 0.018118777 0.1321070861 0.75020920
         knn      fritz
5%  0.6471567 0.01991147
10% 0.3490445 0.90353097
15% 0.4254001 0.38114042
20% 0.9612251 0.06477493
25% 0.1967671 0.52980921
30% 0.6533539 0.17757182
35% 0.3601628 0.45557877
40% 0.8989844 0.99365822
45% 0.3787787 0.31442747
50% 0.4988042 0.54980294
```

- **P valores de la prueba de normalidad de *Shapiro* en los datos *German*.**

```
> spv_german
```

```
           imean       icmean         imed        icmed        irs      icrs
5%  2.644417e-02 2.479381e-01 2.997527e-02 1.787529e-10 0.48860356 0.2734505
10% 7.678664e-01 4.948061e-01 8.059659e-01 1.996401e-10 0.03394938 0.4722143
15% 6.932965e-02 2.479580e-03 4.460605e-02 2.609236e-10 0.42114893 0.1599989
20% 2.307348e-01 6.407368e-01 3.334375e-01 2.337736e-10 0.05692817 0.2356521
25% 2.607253e-10 2.851234e-10 2.584362e-10 3.094729e-10 0.30925930 0.9420693
30% 7.818072e-01 2.726592e-02 9.167678e-01 3.747006e-10 0.85532079 0.8222676
           knn          adc        fritz
5%  0.54641714 9.265689e-01 7.800298e-01
10% 0.25966292 3.311748e-01 3.963794e-01
15% 0.88306390 1.123547e-03 9.397769e-02
20% 0.78159931 5.832289e-04 6.517112e-01
25% 0.05359598 3.480742e-04 1.609719e-09
30% 0.46318064 2.827071e-05 6.228435e-01
```

Ahora se muestran los intervalos de confianza por método de imputación y por proporción de datos faltantes en cada conjnto de datos. Además se muestra el valor del ABC para cada conjunto original de datos. Cabe señalar que para los datos Bupa y Diabetes se uilizó elestadístico $\boldsymbol{z_\alpha}$ para los intervalos de confianza con $\boldsymbol{\alpha = 0.05}$. Esto pues los estimados de distribuyen normalmente y la cantidad de estimados en cada método y proporción de datos faltantes es 50. Para los datos Bajopeso y Diabetes los estimados se distribuyen normalmente pero la cantidad de ellos es de 25 por lo que se utilizó el estdístico de $\boldsymbol{t}$ en los intervalos de confianza.

- **Intervalos de confianza para los datos Bupa**

```
> B1_bupa

     imean(L)  imean(U) icmean(L) icmean(U)   imed(L)   imed(U)
5%  0.7170317 0.7240448 0.7299147 0.7370818 0.7190471 0.7256467
10% 0.7049942 0.7140576 0.7297447 0.7395512 0.7068818 0.7160727
15% 0.6903948 0.7018831 0.7265495 0.7401837 0.6935274 0.7053126
20% 0.6739648 0.6862186 0.7239850 0.7376523 0.6780906 0.6894156
25% 0.6686649 0.6798241 0.7315640 0.7445077 0.6729709 0.6845381
30% 0.6605089 0.6735442 0.7382603 0.7557956 0.6623665 0.6754873
35% 0.6446368 0.6571267 0.7359671 0.7532909 0.6477419 0.6604029
40% 0.6392747 0.6518294 0.7505201 0.7698344 0.6413947 0.6539784
45% 0.6256543 0.6425960 0.7544393 0.7796269 0.6283128 0.6456079
50% 0.6227381 0.6354709 0.7715105 0.7944867 0.6260369 0.6386906

> B2_bupa

     icmed(L)  icmed(U)    irs(L)    irs(U)   icrs(L)   icrs(U)
5%  0.7300492 0.7366039 0.7059904 0.7153296 0.7174733 0.7267791
10% 0.7283481 0.7377512 0.6859204 0.6983451 0.7008226 0.7120291
15% 0.7254063 0.7379385 0.6661461 0.6781332 0.6940773 0.7069530
20% 0.7222669 0.7334980 0.6478431 0.6622569 0.6794845 0.6935458
25% 0.7306149 0.7431134 0.6415600 0.6552876 0.6803775 0.6931059
30% 0.7360530 0.7515525 0.6239243 0.6404184 0.6777631 0.6964818
35% 0.7380587 0.7514172 0.6126328 0.6242361 0.6716898 0.6876274
40% 0.7500865 0.7655439 0.6101100 0.6258438 0.6737520 0.6901204
45% 0.7576811 0.7788879 0.5953686 0.6102335 0.6671875 0.6893683
50% 0.7746687 0.7956424 0.5924035 0.6076503 0.6651949 0.6881161

> B3_bupa

       knn(L)    knn(U)       adc(L)       adc(U)  fritz(L)  fritz(U)
5%  0.7135081 0.7210850 2.831226e-01 0.2982091373 0.7117801 0.7183303
```

```
10% 0.6989849 0.7099869 1.376921e-01 0.1483472535 0.6915198 0.7015906
15% 0.6831521 0.6937707 6.772288e-02 0.0753598816 0.6725589 0.6846128
20% 0.6625112 0.6764357 3.022068e-02 0.0349731157 0.6562677 0.6697351
25% 0.6559050 0.6679874 1.443103e-02 0.0170882797 0.6474166 0.6587834
30% 0.6435408 0.6573296 4.730648e-03 0.0064052140 0.6316191 0.6418119
35% 0.6257078 0.6410274 1.902706e-03 0.0023552248 0.6183418 0.6340871
40% 0.6199964 0.6329491 8.101293e-04 0.0009981466 0.6104006 0.6242070
45% 0.6148167 0.6315040 3.149252e-04 0.0004567989 0.5978851 0.6130667
50% 0.6097371 0.6286450 9.901145e-05 0.0001616782 0.5940114 0.6075382
```

> A_bupa

```
[1] 0.7332241
```

- **Intervalos de confianza para los datos Diabetes**

> B1_diabetes

```
     imean(L)  imean(U) icmean(L) icmean(U)   imed(L)   imed(U)
5%  0.8324698 0.8343671 0.8458657 0.8479054 0.8323193 0.8342419
10% 0.8259032 0.8288730 0.8533505 0.8558209 0.8252534 0.8284325
15% 0.8177698 0.8214299 0.8606402 0.8643161 0.8171105 0.8207903
20% 0.8097525 0.8146844 0.8685514 0.8731923 0.8087967 0.8140217
25% 0.8040290 0.8091504 0.8805508 0.8848007 0.8024513 0.8076837
30% 0.7959855 0.8009142 0.8882181 0.8930369 0.7949742 0.8002455
35% 0.7876676 0.7928234 0.8966611 0.9015822 0.7863775 0.7914966
40% 0.7752610 0.7829829 0.9050214 0.9107928 0.7729187 0.7808379
45% 0.7714626 0.7775392 0.9179075 0.9233894 0.7693643 0.7759256
50% 0.7598399 0.7683863 0.9278170 0.9340733 0.7582669 0.7666733
```

> B2_diabetes

```
     icmed(L)  icmed(U)    irs(L)    irs(U)   icrs(L)   icrs(U)
5%  0.8467931 0.8488540 0.8250660 0.8276477 0.8388111 0.8410128
10% 0.8553175 0.8578840 0.8120472 0.8165842 0.8402124 0.8440212
15% 0.8643813 0.8681321 0.7951627 0.8007941 0.8407635 0.8458147
20% 0.8730267 0.8777078 0.7815035 0.7875610 0.8418895 0.8470402
25% 0.8860239 0.8904257 0.7698975 0.7755942 0.8461974 0.8520018
30% 0.8945222 0.8989773 0.7548851 0.7619364 0.8429921 0.8490991
35% 0.9028121 0.9080821 0.7408134 0.7493423 0.8457553 0.8518972
40% 0.9124548 0.9181447 0.7238541 0.7332810 0.8463721 0.8554872
45% 0.9244201 0.9302860 0.7094135 0.7183492 0.8535268 0.8610060
50% 0.9347591 0.9411141 0.6945775 0.7040401 0.8548834 0.8649794
```

> B3_diabetes

```
       knn(L)    knn(U)       adc(L)       adc(U)  fritz(L)  fritz(U)
5%  0.8293569 0.8318252 2.220275e-01 2.304949e-01 0.8285713 0.8310000
10% 0.8218421 0.8258132 9.391282e-02 9.941046e-02 0.8189480 0.8231195
```

```
15% 0.8100943 0.8148900 3.635146e-02 3.979048e-02 0.8043452 0.8090312
20% 0.7993864 0.8048953 1.375164e-02 1.504298e-02 0.7921646 0.7980032
25% 0.7914384 0.7972768 4.996002e-03 5.649520e-03 0.7810919 0.7882475
30% 0.7748808 0.7828672 1.752389e-03 2.150298e-03 0.7666113 0.7732619
35% 0.7656431 0.7727184 4.775219e-04 6.039706e-04 0.7556620 0.7628366
40% 0.7456885 0.7567554 1.355788e-04 1.926301e-04 0.7414810 0.7501939
45% 0.7415608 0.7513269 4.121035e-05 6.819264e-05 0.7283412 0.7365310
50% 0.7245660 0.7369958 9.259447e-06 1.611369e-05 0.7123234 0.7230389
```

> A_diabetes

```
[1] 0.839362
```

- **Intervalos de confianza para los datos Bajopeso**

> B1_bajopeso

```
     imean(L)  imean(U) icmean(L) icmean(U)   imed(L)   imed(U)
5%  0.7307758 0.7404993 0.7496576 0.7628118 0.7314624 0.7410591
10% 0.7239842 0.7389154 0.7584245 0.7807776 0.7254955 0.7402098
15% 0.7183577 0.7337048 0.7700943 0.7966175 0.7193878 0.7350372
20% 0.7159073 0.7278371 0.7849068 0.8141336 0.7158466 0.7292173
25% 0.7055008 0.7220716 0.7827985 0.8225627 0.7050965 0.7215111
30% 0.6978282 0.7207950 0.7865708 0.8305817 0.6968164 0.7193192
35% 0.6902584 0.7165108 0.8037879 0.8660218 0.6936158 0.7176593
40% 0.6880687 0.7126093 0.8249413 0.8832516 0.6891294 0.7124404
45% 0.6859324 0.7139320 0.8636768 0.9098721 0.6865432 0.7146458
50% 0.6734382 0.6934823 0.8522086 0.8952334 0.6723913 0.6927613
```

> B2_bajopeso

```
     icmed(L)  icmed(U)    irs(L)    irs(U)   icrs(L)   icrs(U)
5%  0.7480543 0.7615050 0.7301293 0.7412136 0.7365868 0.7457235
10% 0.7573608 0.7792754 0.7171505 0.7336396 0.7396260 0.7578134
15% 0.7678973 0.7941940 0.7071982 0.7200769 0.7410905 0.7559160
20% 0.7819650 0.8118707 0.6973033 0.7173669 0.7527153 0.7690161
25% 0.7783233 0.8184329 0.6932560 0.7130126 0.7424096 0.7613505
30% 0.7803190 0.8260695 0.6790760 0.6985459 0.7463330 0.7720034
35% 0.7966476 0.8616731 0.6665678 0.6983344 0.7424819 0.7792496
40% 0.8134730 0.8768373 0.6609914 0.6858352 0.7499372 0.7754031
45% 0.8548489 0.9046452 0.6573059 0.6882325 0.7640723 0.7975809
50% 0.8397987 0.8898075 0.6624657 0.6860923 0.7575134 0.7902988
```

> B3_bajopeso

```
       knn(L)    knn(U)       adc(L)     adc(U)  fritz(L)  fritz(U)
5%  0.7344528 0.7447128 0.0003249944 0.01405310 0.7389548 0.7487897
10% 0.7278970 0.7420430 0.0000000000 0.00000000 0.7332296 0.7456023
15% 0.7186041 0.7351012 0.0000000000 0.00000000 0.7284292 0.7446973
```

```
20% 0.7066223 0.7291899 0.0000000000 0.00000000 0.7405065 0.7667634
25% 0.6942635 0.7154731 0.0000000000 0.00000000 0.7270713 0.7572677
30% 0.6878255 0.7138642 0.0000000000 0.00000000 0.7416052 0.7837142
35% 0.6922445 0.7179771 0.0000000000 0.00000000 0.7658552 0.8015346
40% 0.6754700 0.7078833 0.0000000000 0.00000000 0.7579351 0.8014130
45% 0.6813096 0.7208078 0.0000000000 0.00000000 0.7677691 0.8159806
50% 0.6763025 0.7079009 0.0000000000 0.00000000 0.7586411 0.7991842
```

> A_bajopeso

```
[1] 0.742438
```

- **Intervalos de confianza para los datos German**

> B1_german

```
     imean(L)  imean(U) icmean(L) icmean(U)   imed(L)   imed(U)
5%  0.7948112 0.7982585 0.8145560 0.8185000 0.7947735 0.7983004
10% 0.7845042 0.7899187 0.8287943 0.8336349 0.7842513 0.7897925
15% 0.7777934 0.7858022 0.8426148 0.8510928 0.7774649 0.7856616
20% 0.7669728 0.7736868 0.8567998 0.8632430 0.7666746 0.7734770
25% 0.6710236 0.7975084 0.7667990 0.9113955 0.6708475 0.7972984
30% 0.7507840 0.7606181 0.8865099 0.8983360 0.7502444 0.7600308
```

> B2_german

```
     icmed(L)  icmed(U)    irs(L)    irs(U)   icrs(L)   icrs(U)
5%  0.7160374 0.8508446 0.7930334 0.7969292 0.8064898 0.8103106
10% 0.7290507 0.8663469 0.7837928 0.7878215 0.8106362 0.8147492
15% 0.7423016 0.8822493 0.7718031 0.7788312 0.8113451 0.8201962
20% 0.7542748 0.8963948 0.7620472 0.7690680 0.8146647 0.8202696
25% 0.7659435 0.9104394 0.7485812 0.7578904 0.8144311 0.8238422
30% 0.7835742 0.9315663 0.7353993 0.7468737 0.8188781 0.8284191
```

> B3_german

```
       knn(L)    knn(U)       adc(L)       adc(U)  fritz(L)  fritz(U)
5%  0.7946287 0.7990871 6.033233e-02 6.633167e-02 0.7992431 0.8029637
10% 0.7855728 0.7903078 6.451040e-03 7.899627e-03 0.7901969 0.7949861
15% 0.7715363 0.7800953 6.613823e-04 8.637606e-04 0.7859215 0.7930084
20% 0.7624358 0.7707715 3.928072e-05 8.243356e-05 0.7722672 0.7803383
25% 0.7552282 0.7652670 4.807824e-06 1.004932e-05 0.6174711 0.7928624
30% 0.7437283 0.7567587 2.009679e-06 3.133178e-06 0.7482488 0.7602796
```

> A_german

```
[1] 0.8064619
```

Se puede notar que para algunos métodos de imputación en algunas bases de datos los parámetros caen dentro de los intervalos de confianza correspondientes a proporciones de datos faltantes mayores del 15 %. Las FIGURAS 6.10, 6.11 y 6.12 presentan los casos en cada base de datos donde esto ocurre. En las gráficas las proporciones están representadas por índices; 5 % por 1, 10 % por 2 y así sucesivamente hasta 50 % por 10.

6.8. Desviaciones estándar de las áreas bajo la curva ROC

En esta sección se presentan las desviaciones estándar de los estimados de las áreas bajo la curva **ROC** por bases de datos. En las FIGURAS 6.13 - 6.17 se comparan las desviaciones estándar de los estimados en cada método de imputación a través de todas las proporciones de datos. El orden de izquierda a derecha en que se presentan los métodos en la gráficas es: IMEAN, ICMEAN, IMED, IMED, IRS, ICRS, KNN, ADC y FRITZ.

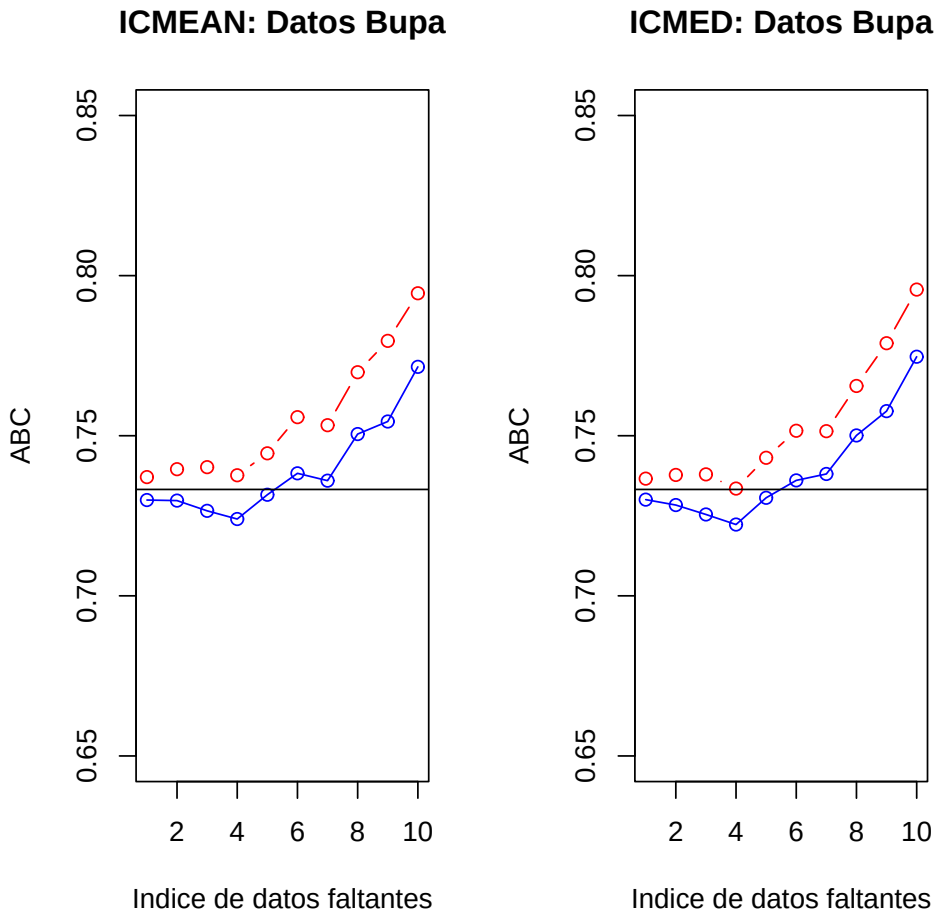


FIGURA 6.10: **La línea entrecortada con puntos representa las cotas superiores del intervalo y la otra línea entera con puntos sobrepuestos corresponde a las cotas inferiores. En ambos métodos de imputación ICMEAN e ICMED en los datos *Bupa*, el parámetro es representado por la línea recta y ésta pasa a través de los intervalos de confianza hasta el de 25 % de datos faltanes representado por el índice 5 en el eje horizontal.**

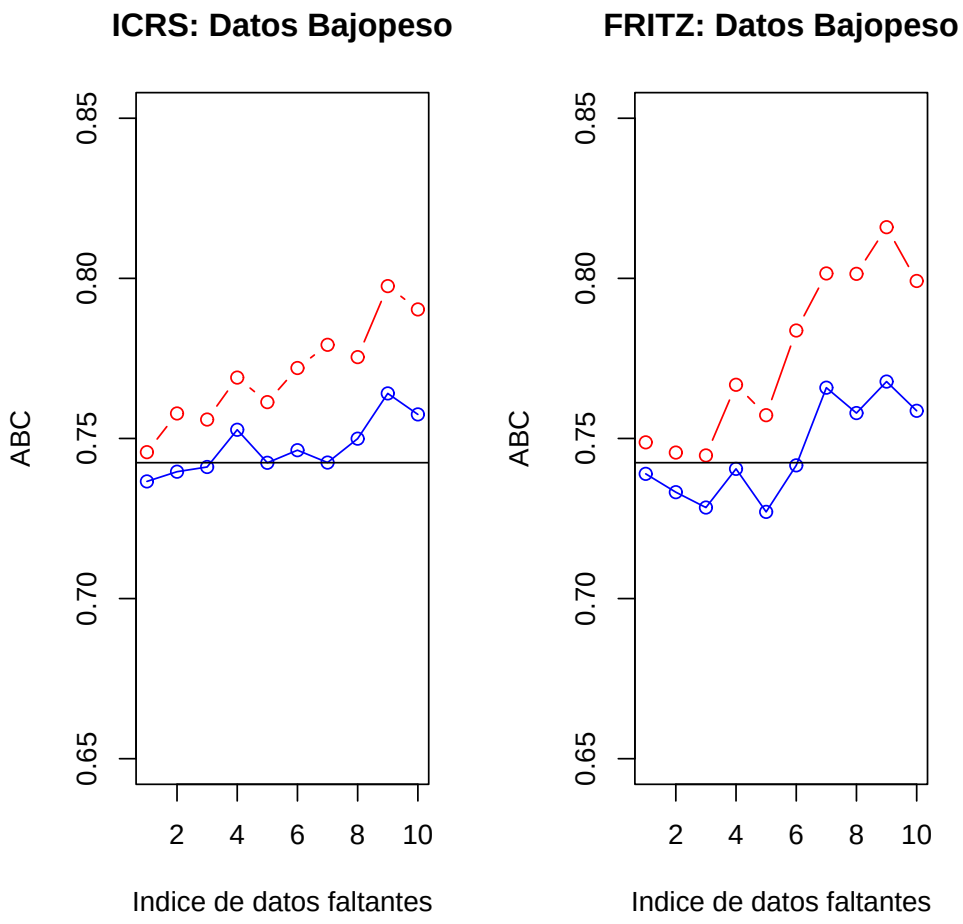


FIGURA 6.11: **La línea entrecortada con puntos representa las cotas superiores del intervalo y la otra línea entera con puntos sobrepuestos corresponde a las cotas inferiores. En ambos métodos de imputación ICRS y FRITZ en los datos *Bajopeso*, el parámetro es representado por la línea recta y ésta pasa a través de los intervalos de confianza hasta el 15 % de datos faltanes representado por el índice 5 en el eje horizontal para el método ICR y hasta el 30 % representado por el índice 6 para el método FRITZ.**

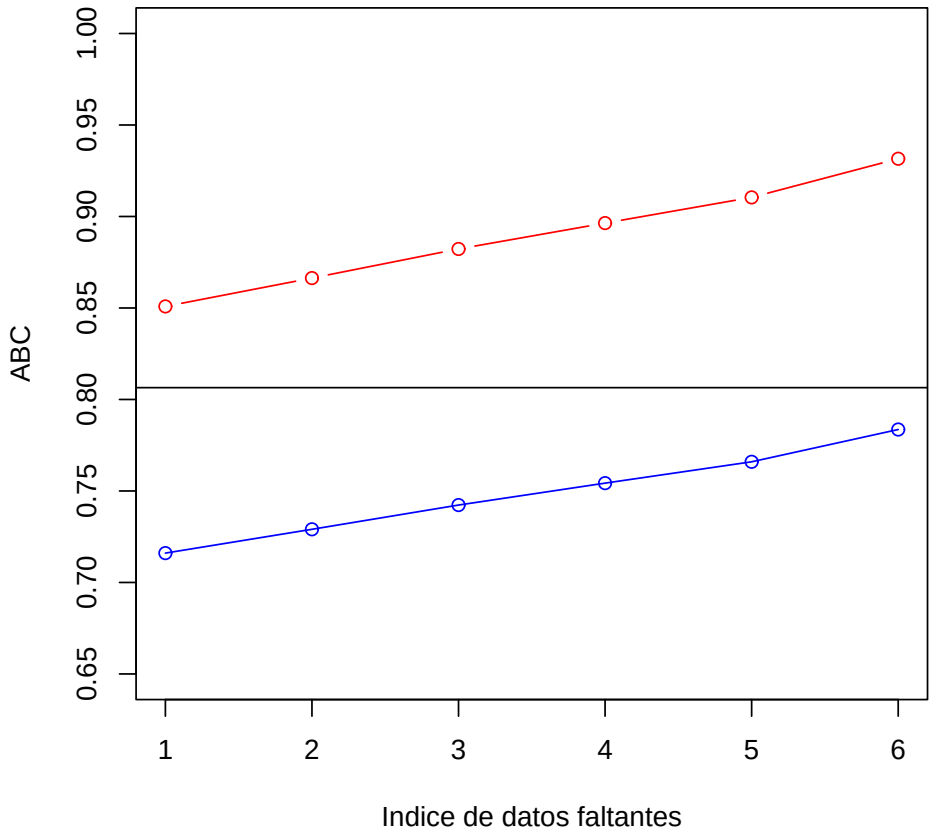


FIGURA 6.12: **La línea entrecortada con puntos representa las cotas superiores del intervalo y la otra línea entera con puntos sobrepuestos corresponde a las cotas inferiores. En el método de imputación ICMED en los datos *German*, el parámetro es representado por la línea recta y ésta pasa a través de los intervalos de confianza hasta el 30 % de datos faltanes representado por el índice 6 en el eje horizontal.**

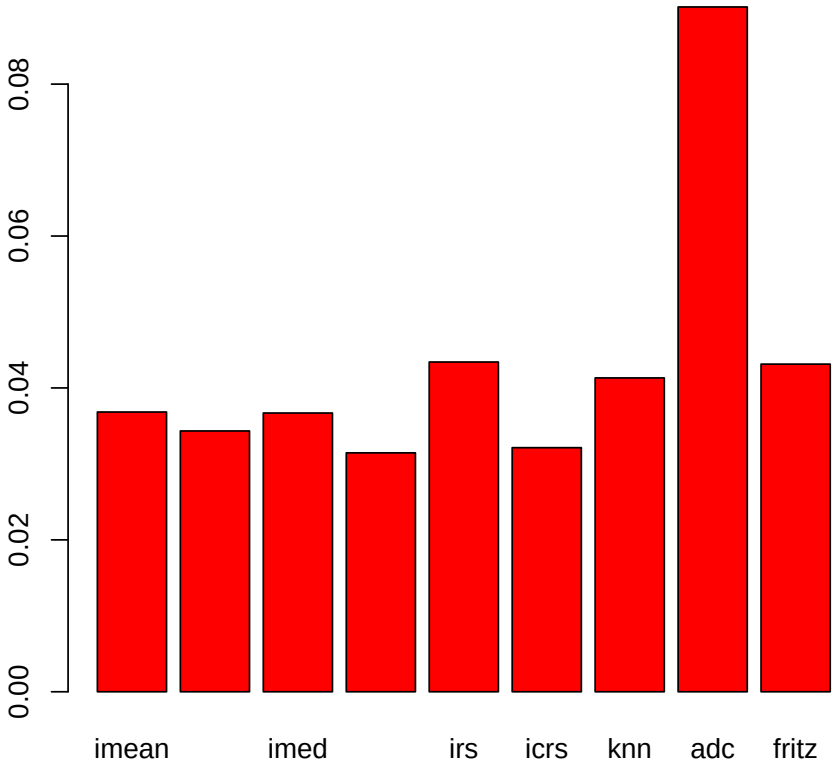


FIGURA 6.13: **Desviaciones estándar del las ABC's para los métodos de imputación en los datos *bupa* tomando en consideración todos los estimados de todas las proporciones de datos faltantes**

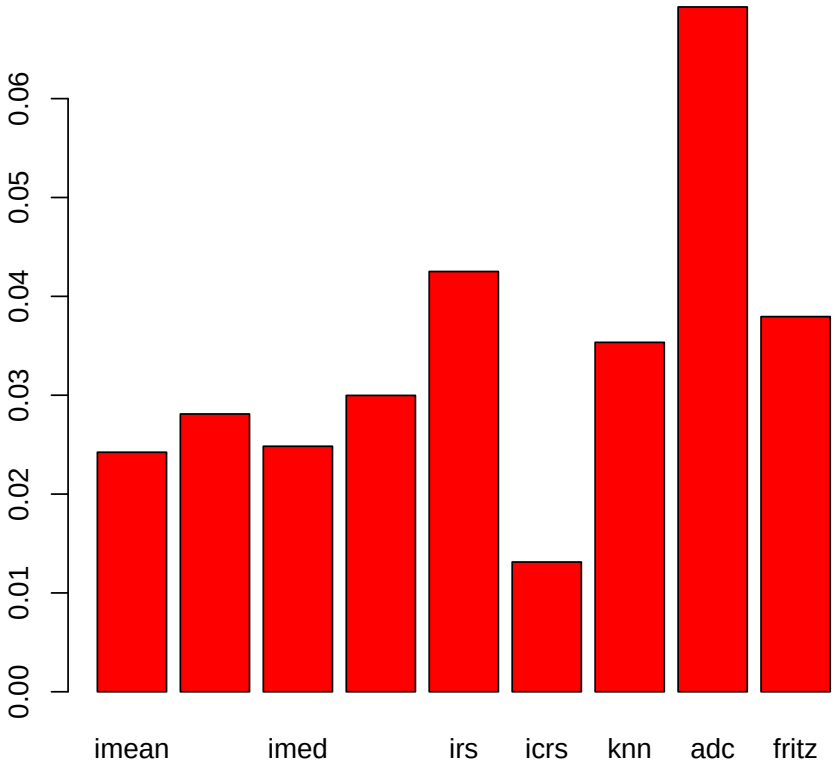


FIGURA 6.14: **Desviaciones estándar del las ABC's para los métodos de imputación en los datos *diabetes* tomando en consideración todos los estimados de todas las proporciones de datos faltantes**

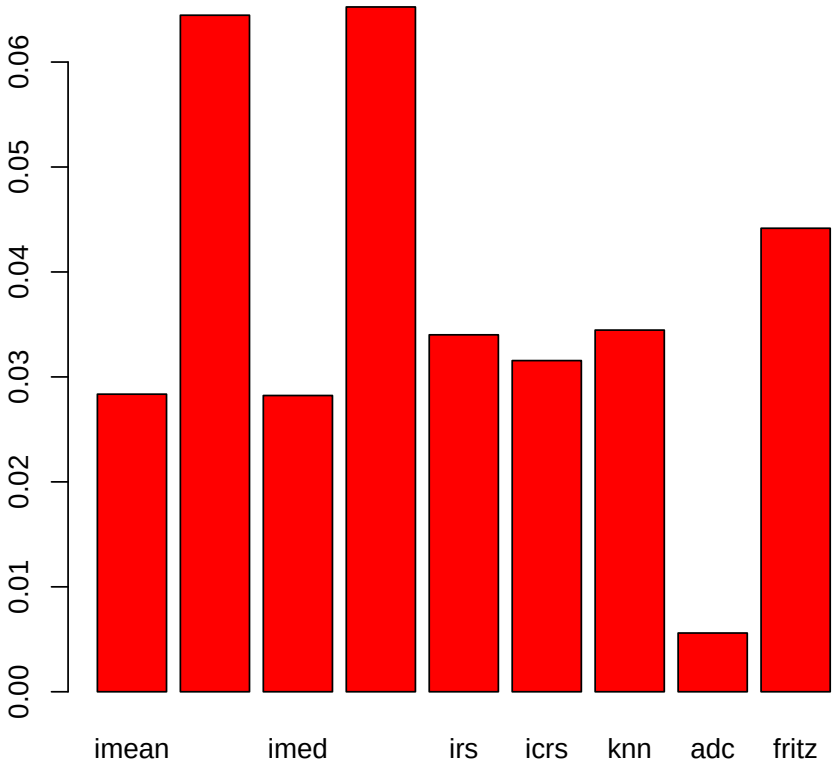


FIGURA 6.15: **Desviaciones estándar del las ABC's para los métodos de imputación en los datos *bajopeso* tomando en consideración todos los estimados de todas las proporciones de datos faltantes**

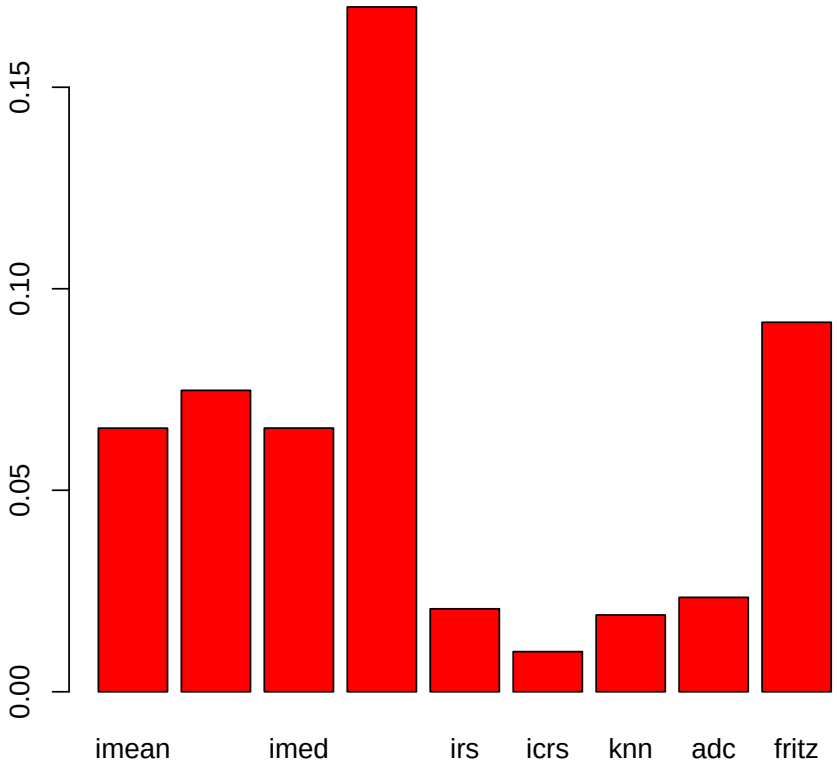


FIGURA 6.16: **Desviaciones estándar del las ABC's para los métodos de imputación en los datos *german* tomando en consideración todos los estimados de todas las proporciones de datos faltantes**

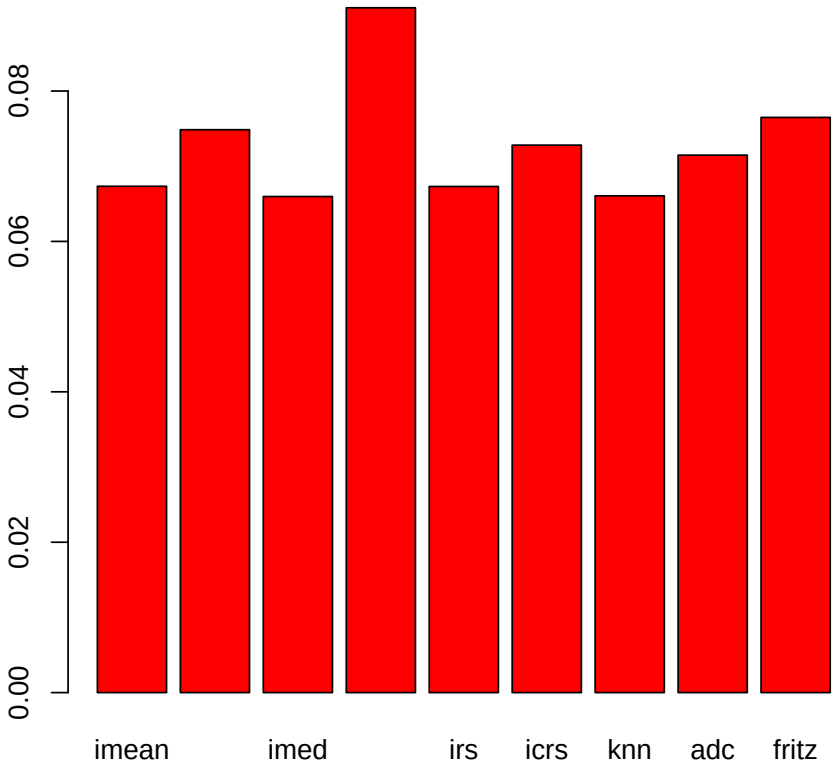


FIGURA 6.17: **Desviaciones estándar del las ABC's para los métodos de imputación de todos los conjuntos de datos tomando en consideración todos los estimados de todas las proporciones de datos faltantes**

Capítulo 7

Aplicación de los Métodos de Imputación

7.1. Introducción

En este capítulo se probaron los métodos de imputación en un conjunto de datos con valores faltantes originalmente. Se aplicaron las ideas de [Delong et al., 1998] para comparar los métodos de imputación mediante matrices de contrastes y mediante la matriz de covararianza calculada a base de los *estadísticos U*. El área bajo la curva **ROC** para estos modelos es calculado con el estimado de Mann-Whitney para el cual se programó la función AMW en el programado **R**. No obstante los intervalos de confianza se calcularon con el **ABC** de la Regla Trapezoidal. Las funciones que calculan las matrices de covarianzas y sus componentes fueron programadas en **R** y se muestran en el Apéndice.

7.2. Descripción del conjunto de datos

El conjunto de datos utilizado en esta parte de aplicación fue extraído del *Machine Learning Database Repository* de la Universidad de Califonia, Irvine. Se presentan las descripciones del conjunto de datos *hepatitis*. La variable de respuesta es binaria y no está incluída en el número de variables.

- Número de unidades: 155
- Número de variables explicaivas: 19
- Variables con datos faltantes (% datos faltantes en la variable): V4(0.65 %) V6(0.65 %) V7(0.65 %) V8(0.65 %) V9(6.5 %) V10(7.1 %) V11(3.23 %) V12(3.23 %) V13(3.23 %) V14(3.23 %) V15(3.87 %) V16(18.71 %) V17(2.58 %) V18(10.32 %) V19(43.23 %).
- % global de datos faltantes: 5.67 %
- % de unidades con datos faltantes: 48.38 %

Los datos *hepatitis* tienen dos tipos de variables que se dividen en 13 variables binarias de las cuales 11 contienen datos faltantes y 5 variables contínuas de las cuales 4 tienen datos faltantes. Los valores faltantes se muestran en la FIGURA 7.1 donde son representados por los espacios en blanco.

7.3. Estimados del área bajo la curva ROC

Los estimados del poder de separación del modelo de regresión logística con estos conjuntos de datos se calcularon con el estimado de Mann-Whitney discutido en el Capítulo 3 y con la Regla Trapezoidal. Las cotas para las áreas bajo la curva ROC se calcularon con las ideas expuestas en [Shapiro, 1998]. Esta información se muestra en la TABLA 7.1.

7.4. La matriz de varianza-covarianza para los estimados de AMW

En esta sección se mostrarán las matrices de varianza-covarianza de los estimados de las áreas bajo la curva ROC mediante Mann-Whitney para cada método de imputación. Nótese que el estimado de ADC no se tomará en consideración en los datos hepatitis debido a que no se pudo ajustar adecuadamente el modelo de regresión logística con los datos disponibles. No

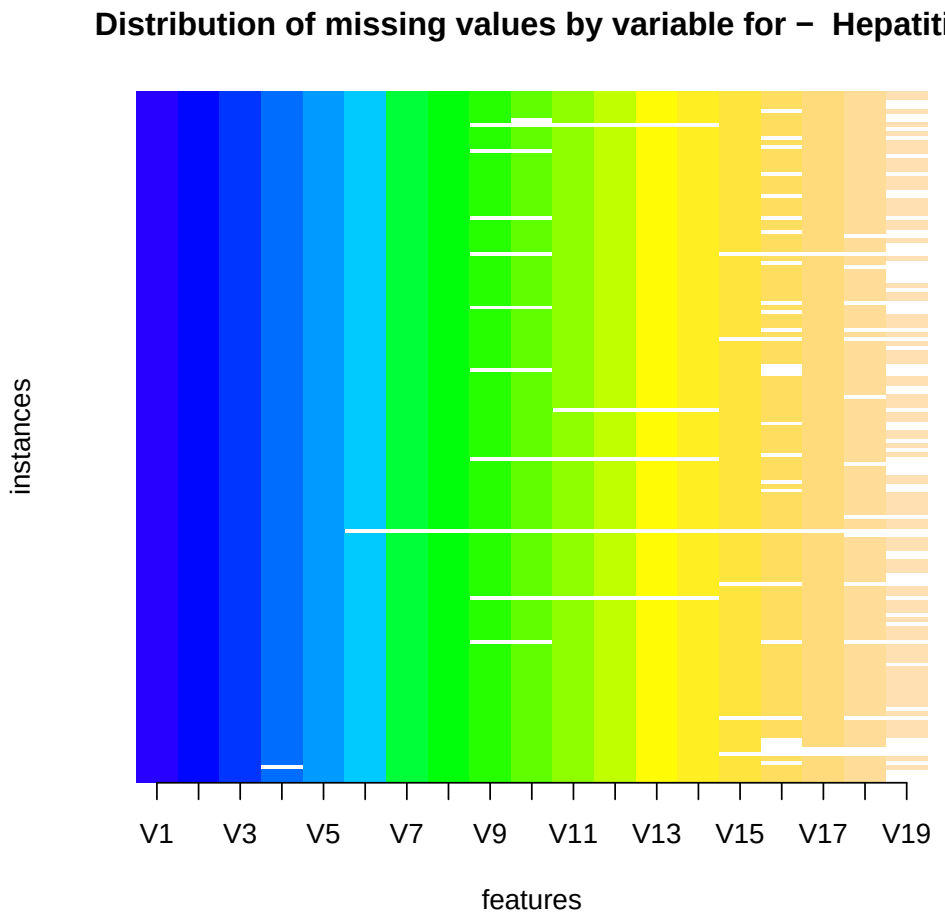


FIGURA 7.1: **Datos faltantes para los datos hepatitis**

TABLA 7.1: **Estimados de áreas bajo la curva ROC en los datos *hepatitis***

Método de Imputación	AMW	Regla Trapezoidal	Cotas de Shapiro
IMEAN	0.8221849	0.8217647	(0.7747899, 0.8888655)
ICMEAN	0.8635294	0.8621849	(0.8105042, 0.9293697)
IMED	0.8210084	0.8215966	(0.7747899, 0.901174)
ICMED	0.8705882	0.8693277	(0.8189076, 0.9341176)
IRS	0.8215126	0.8205882	(0.7571429, 0.8520658)
ICRS	0.8280672	0.827563	(0.7794118, 0.8712452)
KNN	0.8067227	0.807647	(0.7445378, 0.8591713)
ADC	NA	0.01378151	(0.5, NA)
FRITZ	0.7788235	0.7790756	(0.7420168, 0.8668423)

obstante, el estimado según la regla trapezoidal sí se pudo calcular y lo observamos en la TABLA 7.1, el mismo es demasiado pequeño. La gráfica de la FIGURA 7.2 muestra las desviaciones estándar de cada método de imputación calculadas directamente de la matriz de varianza-covarianza.
Matriz de varianza-covarianza de los métodos en los datos *hepatitis*.

```
> S_hepatitis
```

```
              IMEAN       ICMEAN         IMED        ICMED          IRS
IMEAN  0.0012173665 0.0009820831 0.0012172484 0.0009294229 0.0011046477
ICMEAN 0.0009820831 0.0009520845 0.0009769716 0.0009282815 0.0009366042
IMED   0.0012172484 0.0009769716 0.0012255106 0.0009255632 0.0011157604
ICMED  0.0009294229 0.0009282815 0.0009255632 0.0009154155 0.0008908591
IRS    0.0011046477 0.0009366042 0.0011157604 0.0008908591 0.0011692029
ICRS   0.0011222067 0.0009609853 0.0011240825 0.0009209694 0.0010841748
KNN    0.0011576976 0.0009219552 0.0011604813 0.0008752110 0.0010918223
FRITZ  0.0009843931 0.0007934540 0.0009884344 0.0007374422 0.0010037222
              ICRS          KNN        FRITZ
IMEAN  0.0011222067 0.0011576976 0.0009843931
ICMEAN 0.0009609853 0.0009219552 0.0007934540
IMED   0.0011240825 0.0011604813 0.0009884344
ICMED  0.0009209694 0.0008752110 0.0007374422
IRS    0.0010841748 0.0010918223 0.0010037222
ICRS   0.0011779079 0.0010645172 0.0010170110
KNN    0.0010645172 0.0012613172 0.0009616913
FRITZ  0.0010170110 0.0009616913 0.0014799392
```

Si calculamos el error estándar utilizando el estimado de Wilcoxon para las AMW's tenemos el gráfico de la FIGURA 7.3.

Como se puede observar en la TABLA 7.1 las áreas bajo la curva mayores corresponden a los métodos ICMEAN e ICMED y ambos métodos muestran tener la menor desviación estándar en los estimados de tales áreas. En resumen los métodos de imputación ICMEAN e ICMED son los que generan mayor área bajo la curva **ROC** y además poseen los errores estándar FIGURA (7.3) y varianzas de menor tamaño.

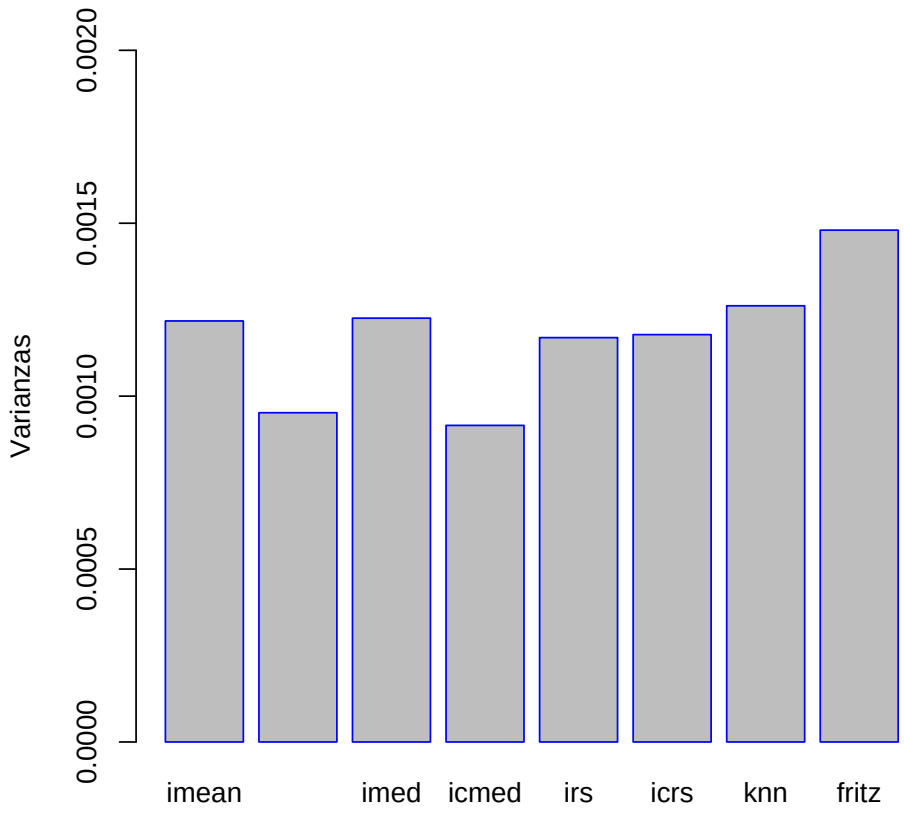


FIGURA 7.2: **Gráfica de barras que ilustra las desviaciones de los AMW's en cada método**

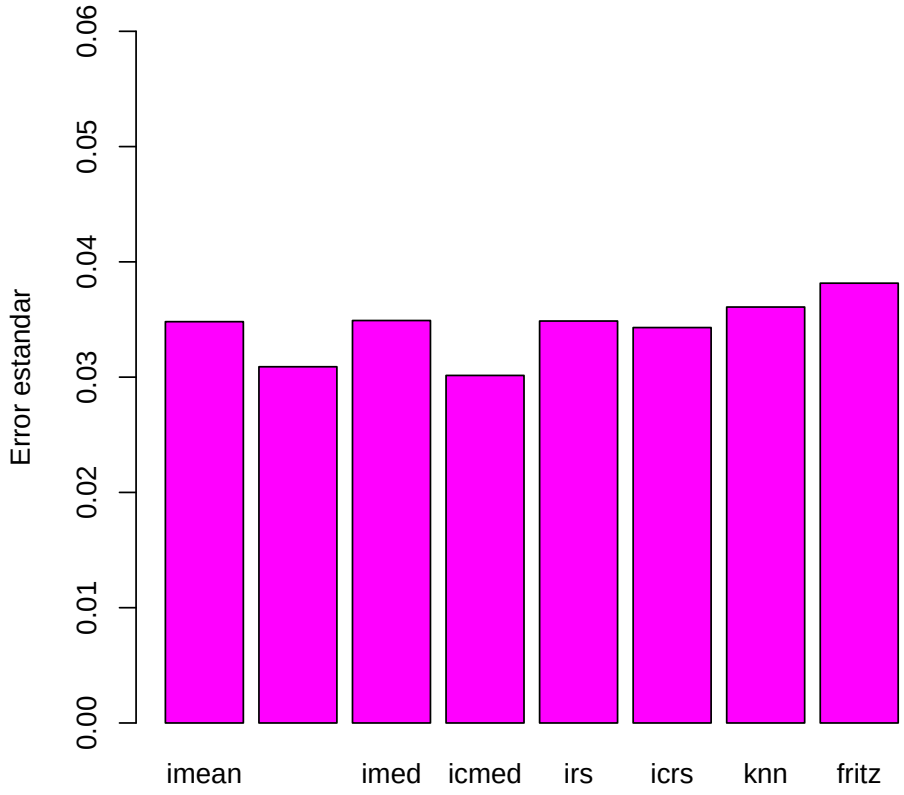


FIGURA 7.3: **Gráfica de barras que ilustra los errores estándar de los AMW's en cada método**

Capítulo 8

Análisis de los resultados, conclusiones y recomendaciones

En este capítulo se culmina con el análisis de los resultados de las simulaciones y de la parte de aplicación. Además se describen las conclusiones y las recomendaciones para investigaciones futuras.

8.1. Análisis de las simulaciones

En esta sección se presenta el análisis de los hallazgos de las simulaciones y de todas las pruebas presentadas en el Capítulo 6 con el fin de cumplir con los objetivos establecidos para esta tesis. Los análisis se dividen en subsecciones, siguiendo el mismo orden en que aparecen los resultados en el Capítulo 6.

8.1.1. Análisis de las pruebas de Friedman

- En todos los conjunto de datos, existe por lo menos un método de imputación que difieren significativamente de los demás a través de todas las proporciones de datos faltantes. Refiérase a la TABLA 6.2.

- En los resultados de todos los conjuntos de datos, el **ADC** genera los estimados con el *mayor sesgo promedio* y el *menor poder de separación promedio*. Refiérase a las FIGURAS 6.1 - 6.4 con respecto a los sesgos y refiérase a las FIGURAS 6.5 - 6.8 con respecto al poder de separación.

- En el conjunto de datos *bupa* los métodos que menor sesgo produjeron fueron: ICMED, ICMEAN, ICRS, IMED e IMEAN. Entre éstos no hubo diferencia significativa.

- En el conjunto de datos *diabetes* los métodos que menor sesgo produjeron fueron: ICRS, IMEAN, IMED e ICMEAN. Entre éstos no existe diferencia significativa.

- En los datos *bajopeso* los que menor sesgo produjeron sin ser diferentes signifiativamente fueron: ICRS, FRITZ, IMEAN, IMED y KNN.

- Para los datos *german* los métodos que menor sesgo produjeron y que no mostraron diferencia significativa fueron: ICRS, FRITZ, IMEAN, KNN, IMED e IRS.

8.1.2. Análisis de las pruebas de Friedman por proporción de datos faltantes

A continuación se presentan comentarios acerca de las pruebas de Friedman por porporción de datos faltantes para todos los conjuntos de datos.

- Para todos los conjuntos de datos en todas las proporciones existen diferencias significativas con respecto al sesgo generado por los métodos de imputación.

- En todos los conjuntos de datos a través de todas las proporciones de datos faltantes, el ADC es diferente significativamente de todos los métodos de imputación. Además el ADC es el que mayor sesgo produce por lo que es el peor método sin importar las bases de datos y sin importar la proporción de datos faltantes.

- En los datos *bupa* los métodos ICMED e ICMEAN muestran ser los mejores a través de todas las proporciones de datos faltantes.

- En los datos *diabetes* el método ICRS es el mejor a través de todas las proporciones de datos faltantes.

- En datos *bajopeso* y en los datos *german* el método que menor sesgo genera a través de la mayoría de las proporciones de datos faltantes es el ICRS.

8.1.3. Análisis de las relaciones Sesgo - ABC

- En la TABLA 6.3 se muestra la correlación de Spearman para cada base de datos y para todos los datos en general y se encontró que estas correlaciones son negativas entre los rangos generados por el área bajo

la curva ROC y el sesgo, lo cual indica que a medida que aumenta el sesgo disminuye el área bajo la curva ROC (el poder de separación del modelo de regresión logística). La FIGURA 6.9 muestra que la relación entre el sesgo y el ABC no es lineal aunque se puede observar un conglomerado donde los métodos que generan mayor sesgo son los que tienen un ABC bajo y los métodos que generan menor sesgo son los que tienen un ABC mayor.

- En la TABLA 6.4 muestra que además de que por proporción de datos faltantes la relación entre sesgo y área es similar a la relación global.

- En los datos *Bupa* se obtuvieron las correlaciones de *Spearman* mayores. Las correlaciones más bajas se obtuvieron para los datos *Bajopeso*.

8.1.4. Los intervalos de confianza del área bajo la curva ROC

- En casi todos los conjuntos de datos, los estimados del **ABC** por porporción de datos falatantes y por método de imputación se distribuyen normalmente.

- En los datos *Bupa*, el parámetro del poder de separación cae en los intevalos de confianza generados con los estimados del **ABC** de los métodos de imputación **ICMEAN** e **ICMED** para las proporciones de datos faltantes desde el 5 % hasta el 25 %.

- En los datos *Diabetes*, el parámetro del poder de separación no caes dentro de ningún intervalo de confianza de los estimados.

- En los datos *Bajopeso* el parámetro cae dentro de los intervalos de confianza desde el 5 % hasta el 15 % para el método **ICRS**. Para el método **FRITZ** el parámetro cae dentro de los intervalos de confianza desde el 5 % hasta el 30 %.

- Para los datos *German*, en el método **ICMED** el parámetro cae en los intervalos de confianza desde el 5 % hasta el 30 %.

8.1.5. Comentarios acerca de las desviaciones estándar de las ABC's

Se presentan los comentarios de las desviaciones de los estimados de las ABC's de los métodos de imputación en cada conjunto de datos.

- Para los datos *bupa*, los métodos que menor desviación presentan a través de todas las porporciones de datos faltantes con respecto a las ABC's son ICMED, ICRS e IMEAN. Los que mayor desviación presentan son el ADC, FRITZ e IRS (Ver FIGURA 6.13).

- En los datos *diabetes*, el métodos ICRS presenta menor desviación estándar en los estimados. Le siguen IMEAN e IMED. Los que mayor desviación tienen son ADC, IRS, FRITZ y KNN (Ver FIGURA 6.14).

- En los datos bajopeso, el ADC muestra la menor desviación estándar. De los métodos de imputación, IMEAN e IMED muestran la menor desviación estándar. Las mayores desviaciones estándar corresponden a los métodos ICMEAN e ICMED (Ver FIGURA 6.15).

- Para los datos *german* los métodos ICRS, KNN e IRS corresponden a los métodos con menor desviación estándar en las ABC's. El ADC tambien muestra una desviación menor con respecto a la mayoría de los métodos de imputación. Los métodos de mayor desviación corresponden a ICMED y FRITZ (Ver FIGURA 6.16).

8.2. Conclusiones

Los resultados del experimento llevado a cabo en esta tesis induce a la conclusión con tres resultados generales finales de vital importancia:

1. El efecto de los métodos de imputación en el poder de separación del modelo de regresión logística varía dependiendo del conjunto de datos que se esté trabajando cuando los datos faltantes provienen de un mecanismo MCAR. Es decir que el poder de separación depende de las distribuciones de las variables que se pretenden imputar bajo un mecanismo **MCAR**. Además, la efectividad de los métodos es consistente en un conjunto de datos sin importar la cantidad de datos faltantes que éste posea.

2. La correlación entre el área bajo la curva o poder de separación del modelo de regresión logística y el sesgo del mismo es fuertemente negativa. Es decir que a mayor sesgo menor poder de separación. Por lo tanto al imputar un conjunto de datos y ajustar un modelo de regresión logística, el método que genere el mayor poder de separación es el que menor sesgo tendrá.

3. Bajo el supuesto de MCAR, existen varios modelos de imputación que generan estimados aceptables en el poder de separación para proporciones de datos faltantes mayores del 15 %. Esto contradice lo planteado por [Pyle, 1999] con respecto a que los resultados provenientes de conjuntos de datos con más del 15 % de datos faltantes son detrimentales.

4. Los resultados obtenidos en este experimento evidencian que el método de imputación por muestreo aleatorio de las observaciones por clase (**ICRS**) produce menos sesgo en la estimación del poder de separación del modelo de regresión logística. Una explicación para esto es que la información que se utiliza para imputar bajo este método en una clase particular proviene de esa misma clase, por lo que el nivel de separación permanece casi intacto. Además de que el método de seleccionar aleatoriamente las imputaciones no afecta grandemente en la distribución de los individuos en la clase que se imputó [Little and Rubin, 2002].

5. De los resultados obtenido no se pudo encontrar alguna relación clara entre los sesgos y la variabildad de los estimados del poder de separación.

8.3. Recomendaciones y proyecciones futuras

Aunque se pusieron a prueba los métodos de imputación de más popularidad, éstos no son los únicos que existen. Cada vez más en la investigación estadística se siguen creando nuevos métodos que contribuyen a un estimado deseado dependiendo el estadístico que se esté trabajando. En esta investigación se proveyó una guía del uso de la curva **ROC** para que en un futuro se puedan poner a prueba nuevos métodos destinados a manejar el problema de datos faltantes cuando se pretende ajustar un modelo de regresión logística con ua variable de respuesta binaria. En esta tesis se utilizó como métrica de comparación global el sesgo promedio de los estimados. Tal vez se puede utilizar otra métrica más robusta como por ejemplo el sesgo mediano y así eliminar el efecto de valores oestimados influenciales para el **ABC**.

Los métodos de imputación se pueden poner a prueba en otros métodos de clasificación con la ayuda de la curva ROC, como por ejemplo el análisis discriminante lineal, métodos de densidad por kernel, y otros métodos donde se utilicen conjuntos con datos faltantes. La metodología de esta investigación servirá de guía para este cometido.

En cuanto a los mecanismos de datos faltantes, se podría investigar el efecto de otros tipos, tales como **MAR** y **NMAR** los cuales son más frecuentes en la realidad. No obstante, como ya habíamos mencionado, simular estos métodos conlleva escoger dentro deun marco más amplio de formas para eliminar datos. Por lo tanto habría que estudiar con más detalle si un método que resulta apropiado dentro de un mecanismo **MCAR** pueda ser también apropiado dentro de mecanismos menos restrictivos como lo son **MAR** y **NMAR**.

Apéndice A

Códigos de programas de funciones generales diseñados en R

A.1. Función que genera datos faltantes con una proporción dada

```
> GenPat

function(X,prob)
{
  n=dim(X)[1]
  p=dim(X)[2]
  gmv=matrix(0,n,p)

  for(j in 1:p)
     {
      gmv[,j]=rbinom((n),1,prob)
      w=which(gmv[,j]==1)

      X[w,j]=NA
     }
  X
}
```

A.2. Función para calcular la curva ROC

```
> ROC

function(Y,modelo)
{
  phat<-as.vector(fitted.values(modelo))
  nobs<-length(Y)
  p<-seq(0,1,by=.01)
  sensit<-rep(0,length(p))
  umespecif<-rep(0,length(p))
  paso=length(which(Y==1))
  nopaso=length(which(Y==0))

  for(j in 1:length(p))
     {
       clases<-rep(0,nobs)
       c1<-which(phat>=p[j])

       clases[c1]=1

       bcc1=length(which(Y==1 & clases==1))
       bcc2=length(which(Y==0 & clases==0))

       sensit[j]=(bcc1/paso)
       umespecif[j]=(1-(bcc2/nopaso))
     }
  tabla<-as.data.frame(cbind(p,sensit,umespecif))
}
```

A.3. Función para calcular el área bajo la curva ROC utilizando la Regla Trapezoidal

```
> Area

function(roc)
{
  Y<-roc$sensit
  X<-roc$umespecif

  n=length(Y)
```

```
  S=rep(0,n-1)

  for(i in 1:n-1)
     {
      S[i]=(X[i+1]-X[i])*(Y[i]+Y[i+1])/2
     }

  area=abs(sum(S[1:length(S)]))
  area
}
```

A.4. Función para calcular el área bajo la curva ROC utilizando el estimado de Mann-Whitney o Wilcoxon

```
> AMW

function(R,F) # R=real response, F=fitted response
{
  w1=which(R==1)
  w0=which(R==0)

  m=length(w1)
  n=length(w0)

  X=F[w1]
  Y=F[w0]

  K=rep(0,(m*n))
  count=0

  for(i in 1:m)
     {
       for(j in 1:n)
          {
            count=count+1

            K[count]=kernel(X[i],Y[j])
          }
     }
```

```
   A=(sum(K))/(m*n)

   return(A)
}
```

A.5. Función para el kernel utilizada en AMW

```
> kernel

function(X,Y)
{
  if(Y<X){k=1}
  if(Y==X){k=0.5}
  if(Y>X){k=0}

  return(k)
}
```

A.6. Función para calcular el punto de corte óptimo utilizando la curva ROC

```
> Poptimo

function(roc)
{
  y<-roc$sensit
  x<-roc$umespecif
  p<-roc$p

  dist=(((y-1)^2)+((x)^2))^(1/2)
  distop<-min(dist)
  w<-which(dist==distop)
  Po<-p[w]
  Po=median(Po)
  Po
}
```

A.7. Función para calcular el intervalo de confianza del área bajo la curva ROC según Shapiro

```
> Bounds

function(roc)
{
  y<-roc$sensit
  x<-roc$umespecif
  p<-roc$p

  dist=(((y-1)^2)+((x)^2))^(1/2)
  distop<-min(dist)
  w<-which(dist==distop)

  Po<-median(p[w])
  xo<-median(x[w])
  yo<-median(y[w])

  x1<-(x[min(w)-4])
  y1<-(y[min(w)-4])

  x2<-(x[max(w)+4])
  y2<-(y[max(w)+4])

  m1<-(y1-yo)/(x1-xo)
  m2<-(y2-yo)/(x2-xo)

  m<-(m1+m2)/2

  yp<-yo-m*xo
  xq<-((1-yo)/m)+xo

  At1<-((1-yp)*xq)/2
  At2<-((1-xo)*(1-yo))/2
  At3<-(yo*xo)/2
  AR<-(1-xo)*yo

  Ub<-1-At1
  Lb<-At2+At3+AR
```

```
  Bounds<-cbind(Lb,Ub)
  Bounds
}
```

A.8. Función para calcular el error estándar de área bajo la curva ROC

```
> StdError

function(Y,A)
{
  Q1 = A/(2-A)
  Q2 = (2*A*A)/(1+A)

  n1 = length(which(Y==1))

  n0 = length(which(Y==0))

  SE = sqrt((A*(1-A)+(n1-1)*(Q1-A*A)+(n0-1)*(Q2-A*A))/(n1*n0))

  SE
}
```

A.9. Función para calcular el promedio por columna de una matriz

```
> PromCols

function(data)
{
  c=dim(data)[2]

  proms<-rep(0,c)

  for(j in 1:c)
     {
       proms[j]=mean(data[,j])
      }
```

```
  proms
}
```

Apéndice B

Códigos de programas de funciones de imputación diseñados en R

B.1. Función para imputación por la media muestral (IMEAN)

```
> IMEAN

function(datmiss)
{
  impute(datmiss,what="mean")
}
```

B.2. Función para imputación por la media muestral condicionada a las clases de la variable de respuesta (ICMEAN)

```
> ICMEAN

function(Y,datmiss)
{
 c1=which(Y==1)
 c0=which(Y==0)

 dmc1<-datmiss[c1,]
```

```
 dmc0<-datmiss[c0,]

 datimpc1<-impute(dmc1,what="mean")
 datimpc0<-impute(dmc0,what="mean")

 datmiss[c1,]=datimpc1
 datmiss[c0,]=datimpc0

 datmiss
}
```

B.3. Función para imputación por la mediana (IMED)

```
> IMED

function(datmiss)
{
  impute(datmiss,what="median")
}
```

B.4. Función para imputación por la mediana condicionada a las clases de la variable de respuesta(ICMED)

```
> ICMED

function(Y,datmiss)
{
 c1=which(Y==1)
 c0=which(Y==0)

 dmc1<-datmiss[c1,]
 dmc0<-datmiss[c0,]

 datimpc1<-impute(dmc1,what="median")
 datimpc0<-impute(dmc0,what="median")

 datmiss[c1,]=datimpc1
```

```
 datmiss[c0,]=datimpc0

 datmiss
}
```

B.5. Función para imputación por muestreo aleatorio de los valores observados (IRS)

```
> IRS

function(datmiss)
{
  n=dim(datmiss)[1]
  p=dim(datmiss)[2]
  for(i in 1:p)
     {
      count=which(is.na(datmiss[,i]))
      Xobs=na.omit(datmiss[,i])
      S=sample(Xobs,size=length(count),replace=T)

      datmiss[count,i]=S
     }
  datmiss
}
```

B.6. Función para imputación por muestreo aleatorio de los valores observados condicionados a las clases de la variable de respuesta (ICRS)

```
> ICRS

function(Y,datmiss)
{
 c1=which(Y==1)
 c0=which(Y==0)

 dmc1<-datmiss[c1,]
```

```
 dmc0<-datmiss[c0,]

 datimpc1<-IRS(dmc1)
 datimpc0<-IRS(dmc0)

 datmiss[c1,]=datimpc1
 datmiss[c0,]=datimpc0

 datimp=as.data.frame(datmiss)
}
```

B.7. Función para imputación por la moda (IMOD)

> IMOD

```
function(datmiss)
{
  n=dim(datmiss)[1]
  m=dim(datmiss)[2]

  for(j in 1:m)
    {
     mod=moda(datmiss[,j],na.rm=TRUE)
     wna<-which(is.na(datmiss[,j]))

     if(length(mod)>1)
      {
        for(i in 1:length(wna))
          {
            datmiss[wna[i],j]=sample(mod,1,replace=TRUE)
          }
      }

     else
      {
        datmiss[wna,j]=mod
      }
     }
```

```
  datmiss
}
```

B.8. Función para imputación por la moda condicionada a las clases de la variable de respuesta(ICMOD)

```
> ICMOD

function(Y,datmiss)
{
 c1=which(Y==1)
 c0=which(Y==0)

 datc1<-datmiss[c1,]
 datc0<-datmiss[c0,]

 datimpc1<-IMOD(datc1)
 datimpc0<-IMOD(datc0)

 datmiss[c1,]=datimpc1
 datmiss[c0,]=datimpc0

 datmiss
}
```

B.9. Función para imputación por los k^{th} vecinos más cercanos (KNN)

La función principal **ec.knnimp** de esta función proviene de la librería **dprep** [Rodríguez, 2004].

```
> KNN

function(Y,datmiss)
{
  Ync=rep(0,length(Y))
  w1=which(Y==0)
  Ync[w1]=1
```

```
  w2=which(Y==1)
  Ync[w2]=2
  datimp=ec.knnimp(as.matrix(cbind(datmiss,Ync)),k=1)

  datimpx=datimp[,1:dim(datimp)[2]-1]
}
```

B.10. Función para imputación múltiple, el algoritmo FRITZ para variables contínuas en la primera iteración

> FC_1

```
function(Xmiss)
{
   Xmc=Xmiss
   c=dim(Xmc)[2]

   for(j in 1:c)
      {
        wmiss=which(is.na(Xmc[,j]))

        if(length(wmiss)!=0)
           {
            for(i in 1:length(wmiss))
           {
                 wavai=which(!is.na(Xmiss[wmiss[i],]))

                 if(length(wavai)!=0)
                   {
                    Xpred=Xmiss[,wavai]
                Xr=Xmc[,j]
                    data=as.data.frame(cbind(Xr,Xpred))

                datavai=as.data.frame(na.exclude(cbind(Xr,Xpred)))

                    model<-glm(Xr~.,data=datavai,family=gaussian)

                    Xmc[wmiss[i],j]=predict(model,newdata=data[wmiss[i],],
                    type="response")+sample(as.vector(model$residuals),1,
```

```
                    replace=TRUE)
                   }

             else
                   {
                    Xmc[wmiss[i],j] = mean(Xmc[,j],na.rm=TRUE)
                   }
               }
             }
         }
   Xmc
}
```

B.11. Función para imputación múltiple, el algoritmo FRITZ para variables binarias en la primera iteración

> FD_1

```
function(Xmiss,wbin)
{
   Xmb=Xmiss[,wbin]

   c=dim(Xmb)[2]

   for(j in 1:c)
      {
        wmiss=which(is.na(Xmb[,j]))

        if(length(wmiss)!=0)
           {
            for(i in 1:length(wmiss))
           {
                 wavai=which(!is.na(Xmiss[wmiss[i],]))

                 if(length(wavai)!=0)
                   {
                    Xpred=Xmiss[,wavai]
                Xr=Xmb[,j]
                data=as.data.frame(cbind(Xr,Xpred))
```

```
                datavai=as.data.frame(na.exclude(cbind(Xr,Xpred)))

                    model<-glm(Xr~.,data=datavai,family=binomial)
                    fit<-predict(model,newdata=data[wmiss[i],],
                    type="response")

                    roc<-ROC(Xr,model)
                    O<-Poptimo(roc)

                       if(fit < O){Xmb[wmiss[i],j]=0}
                       else{Xmb[wmiss[i],j]=1}

                   }

             else
                   {
                    Xmb[wmiss[i],j] = moda(Xmb[,j],na.rm=TRUE)
                   }
               }
             }
        }
    Xmiss[,wbin]=Xmb
}
```

B.12. Función para imputación múltiple, el algoritmo FRITZ para variables contínuas en la iteración t

```
> FC_T

function(Xmiss,Ximptm1)
{
   Xmc=Xmiss

   c=dim(Xmc)[2]

   for(j in 1:c)
      {
        wmiss=which(is.na(Xmc[,j]))
```

```
        if(length(wmiss)!=0)
           {
            for(i in 1:length(wmiss))
           {
                     Xpred=Ximptm1[,-j]
                Xr=Xmc[,j]
                     data=as.data.frame(Xpred)

                datavai=as.data.frame(na.exclude(cbind(Xr,Xpred)))

                     model<-glm(Xr~.,data=datavai,family=gaussian)

                     Xmc[wmiss[i],j]=predict(model,newdata=data[wmiss[i],],
                     type="response")+sample(as.vector(model$residuals),
                     1,replace=TRUE)

                }
             }
        }
    Xmc
}
```

B.13. Función para imputación múltiple, el algoritmo FRITZ para variables binarias en la iteración t

> FD_T

```
function(Xmiss,Ximptm1,wbin)
{
   Xmb=Xmiss[,wbin]

   c=dim(Xmb)[2]

   for(j in 1:c)
      {
        wmiss=which(is.na(Xmb[,j]))

        if(length(wmiss)!=0)
```

```
            {
             for(i in 1:length(wmiss))
            {
                   Xpred=Ximptm1[,-wbin[j]]
               Xr=Xmb[,j]
               data=as.data.frame(Xpred)

               datavai=as.data.frame(na.exclude(cbind(Xr,Xpred)))

                   model<-glm(Xr~.,data=datavai,family=binomial)
                   fit<-predict(model,newdata=data[wmiss[i],],
                   type="response")

                   roc<-ROC(Xr,model)
                   O<-Poptimo(roc)

                   if(fit < O){Xmb[wmiss[i],j]=0}
                   else{Xmb[wmiss[i],j]=1}

                 }
              }
         }
    Xmiss[,wbin]=Xmb
}
```

B.14. Función para imputación múltiple, el algoritmo FRITZ para conjuntos mixtos en la primera iteración

```
> FRITZ_ONE

function(Y,Xmiss,wcont,wbin,word)
{
  f1c=FC_1(Xmiss,wcont)
  f1b=FD_1(Xmiss,wbin)
  f1o=KNN(Y,Xmiss[,word])

  Ximp1=Xmiss
  Ximp1[,wcont]=f1c
  Ximp1[,wbin]=f1b
```

```
  Ximp1[,word]=f1o

  Ximp1
}
```

B.15. Función para imputación múltiple, el algoritmo FRITZ para conjuntos mixtos en la iteración t

```
> FRITZ_T

function(Y,Xmiss,Ximptm1,wcont,wbin,word)
{
  ftc=FC_T(Xmiss,Ximptm1,wcont)
  ftb=FD_T(Xmiss,Ximptm1,wbin)
  fto=KNN(Y,Xmiss[,word])

  Ximpt=Xmiss
  Ximpt[,wcont]=ftc
  Ximpt[,wbin]=ftb
  Ximpt[,word]=fto

  Ximpt
}
```

Apéndice C

Códigos de programas de funciones relacionadas a las pruebas no paramétricas

C.1. Función para calcular los rangos de la prueba de Friedman

```
> Score

function(data)
{
  t=dim(data)[2]
  n=dim(data)[1]

  Scores=matrix(0,n,t)

  for(j in 1:n)
     {
       Scores[j,]=rank(data[j,],ties.method="average")
     }

  Scores
}
```

C.2. Función para calcular los rangos de la prueba de Kruskall-Wallis

```
> ScoreAll

function(vector,numtrat,dimtrat) #Entrada:vector de observaciones,
                                 #numero de tratamientos,
                                 #dimension del tratamiento
 {

   Scores<-matrix(0,dimtrat,numtrat)
   R=rank(vector,ties.method="average")

   for(j in 1:numtrat)
       {
        Scores[,j]=R[((j*dimtrat)-(dimtrat-1)):(j*dimtrat)]
       }
   Scores
}
```

C.3. Función para calcular las sumas de rangos de la prueba de Friedman

```
> Rsums

function(score)
{
   R=rep(0,9)

   for(i in 1:length(R))
      {
        R[i]=sum(score[,i])
      }
R
}
```

C.4. Función para calcular los rangos promedio de la prueba de Kruskall-Wallis

```
> Rmean
```

```
function(score)
{
   R=rep(0,9)

   for(i in 1:length(R))
      {
        R[i]=mean(score[,i])
      }
R
}
```

C.5. Función para calcular las diferencias en los rangos

> Differences

```
function(R)
{
   D=matrix(0,length(R),length(R))
   cont<-c(1,2,3,4,5,6,7,8,9)

   for(i in 1:length(R))
      {
        for(j in cont[i]:length(R))
           {
             D[i,j]=abs(R[i]-R[j])
           }
      }

   n=dim(D)[1]
   m=dim(D)[2]

   for(i in 1:n)
      {
        for(j in 1:m)
           {
             if(i==j){D[i,j]=NA}
             if(i>j){D[i,j]=NA}
           }
       }
```

```
   D
}
```

C.6. Función para identificar cuáles diferencias de rangos son significativas

```
> Significant

function(D,q)
{
   n=dim(D)[1]
   m=dim(D)[2]

   for(j in 1:m)
      {
        for(i in 1:n)
           {
             if(i<j)
             {
               if(D[i,j]>=q){D[i,j]=1}
               else {D[i,j]=0}
             }
           }
      }
   D
}
```

Apéndice D

Matrices de Significancia

Son aquellas matrices triangulares superiores donde los 0's indican que dos pares de métodos no son diferentes ignificativamente con respecto a las sumas de rangos de *Friedman* y los 1's indican donde si existe diferencia significativa con respecto a la suma de rangos de *Friedman*.

D.1. Matrices de significancia para las pruebas globales de Friedman en los conjuntos de datos

> Sig_bupa

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    0    0     0   1     1   1   1
icmean    NA     NA    0    0     0   1     1   1   1
icrs      NA     NA   NA    0     0   0     0   1   1
imed      NA     NA   NA   NA     0   0     0   1   1
imean     NA     NA   NA   NA    NA   0     0   0   1
knn       NA     NA   NA   NA    NA  NA     0   0   0
fritz     NA     NA   NA   NA    NA  NA    NA   0   0
irs       NA     NA   NA   NA    NA  NA    NA  NA   0
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> Sig_diabetes

```
       icrs imean imed icmean icmed knn fritz irs adc
icrs     NA     0    0      0     1   1     1   1   1
imean    NA    NA    0      0     0   0     1   1   1
imed     NA    NA   NA      0     0   0     1   1   1
icmean   NA    NA   NA     NA     0   0     0   1   1
icmed    NA    NA   NA     NA    NA   0     0   0   0
knn      NA    NA   NA     NA    NA  NA     0   0   0
```

```
fritz    NA    NA   NA     NA    NA  NA    NA   0   0
irs      NA    NA   NA     NA    NA  NA    NA  NA   0
adc      NA    NA   NA     NA    NA  NA    NA  NA  NA
```

> Sig_bajopeso

```
       icrs fritz imean imed knn irs icmed icmean adc
icrs     NA     0     0    0   0   1     1       1   1
fritz    NA    NA     0    0   0   0     1       1   1
imean    NA    NA    NA    0   0   0     1       1   1
imed     NA    NA    NA   NA   0   0     1       1   1
knn      NA    NA    NA   NA  NA   0     0       1   1
irs      NA    NA    NA   NA  NA  NA     0       0   0
icmed    NA    NA    NA   NA  NA  NA    NA       0   0
icmean   NA    NA    NA   NA  NA  NA    NA      NA   0
adc      NA    NA    NA   NA  NA  NA    NA      NA  NA
```

> Sig_german

```
       icrs fritz imean knn imed irs icmean icmed adc
icrs     NA     0     0   0    0   0      1     1   1
fritz    NA    NA     0   0    0   0      0     0   1
imean    NA    NA    NA   0    0   0      0     0   1
knn      NA    NA    NA  NA    0   0      0     0   1
imed     NA    NA    NA  NA   NA   0      0     0   1
irs      NA    NA    NA  NA   NA  NA      0     0   0
icmean   NA    NA    NA  NA   NA  NA     NA     0   0
icmed    NA    NA    NA  NA   NA  NA     NA    NA   0
adc      NA    NA    NA  NA   NA  NA     NA    NA  NA
```

D.2. Matrices de significancia para las pruebas de Friedman por proporción en los datos *Bupa*

> Sig05_bupa

```
       icmed icmean imean imed knn icrs fritz irs adc
icmed     NA      0     0    0   0    1     1   1   1
icmean    NA     NA     0    0   0    0     1   1   1
imean     NA     NA    NA    0   0    0     0   0   1
imed      NA     NA    NA   NA   0    0     0   0   1
knn       NA     NA    NA   NA  NA    0     0   0   1
icrs      NA     NA    NA   NA  NA   NA     0   0   1
fritz     NA     NA    NA   NA  NA   NA    NA   0   1
irs       NA     NA    NA   NA  NA   NA    NA  NA   1
adc       NA     NA    NA   NA  NA   NA    NA  NA  NA
```

> Sig10_bupa

```
       icmed icmean imed imean icrs knn fritz irs adc
icmed     NA      0    0     0    1   1     1   1   1
icmean    NA     NA    0     0    1   1     1   1   1
imed      NA     NA   NA     0    0   0     1   1   1
imean     NA     NA   NA    NA    0   0     1   1   1
icrs      NA     NA   NA    NA   NA   0     0   0   1
knn       NA     NA   NA    NA   NA  NA     0   0   1
fritz     NA     NA   NA    NA   NA  NA    NA   0   1
irs       NA     NA   NA    NA   NA  NA    NA  NA   1
adc       NA     NA   NA    NA   NA  NA    NA  NA  NA
```

> *Sig15_bupa*

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    0    0     1   1     1   1   1
icmean    NA     NA    0    0     0   1     1   1   1
icrs      NA     NA   NA    0     0   0     1   1   1
imed      NA     NA   NA   NA     0   0     1   1   1
imean     NA     NA   NA   NA    NA   0     1   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> *Sig20_bupa*

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    1    1     1   1     1   1   1
icmean    NA     NA    1    1     1   1     1   1   1
icrs      NA     NA   NA    0     0   0     1   1   1
imed      NA     NA   NA   NA     0   0     1   1   1
imean     NA     NA   NA   NA    NA   0     1   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> *Sig25_bupa*

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    1    1     1   1     1   1   1
icmean    NA     NA    0    1     1   1     1   1   1
icrs      NA     NA   NA    0     0   1     1   1   1
imed      NA     NA   NA   NA     0   0     1   1   1
imean     NA     NA   NA   NA    NA   0     0   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> *Sig30_bupa*

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    0    1     1   1     1   1   1
icmean    NA     NA    0    1     1   1     1   1   1
icrs      NA     NA   NA    0     0   1     1   1   1
imed      NA     NA   NA   NA     0   0     1   1   1
imean     NA     NA   NA   NA    NA   0     1   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> Sig35_bupa

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    0    1     1   1     1   1   1
icmean    NA     NA    0    1     1   1     1   1   1
icrs      NA     NA   NA    0     1   1     1   1   1
imed      NA     NA   NA   NA     0   1     1   1   1
imean     NA     NA   NA   NA    NA   0     0   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> Sig40_bupa

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    0    1     1   1     1   1   1
icmean    NA     NA    0    1     1   1     1   1   1
icrs      NA     NA   NA    1     1   1     1   1   1
imed      NA     NA   NA   NA     0   0     1   1   1
imean     NA     NA   NA   NA    NA   0     1   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> Sig45_bupa

```
       icmed icmean icrs imed imean knn fritz irs adc
icmed     NA      0    0    1     1   1     1   1   1
icmean    NA     NA    0    1     1   1     1   1   1
icrs      NA     NA   NA    1     1   1     1   1   1
imed      NA     NA   NA   NA     0   0     1   1   1
imean     NA     NA   NA   NA    NA   0     0   1   1
knn       NA     NA   NA   NA    NA  NA     0   0   1
fritz     NA     NA   NA   NA    NA  NA    NA   0   1
irs       NA     NA   NA   NA    NA  NA    NA  NA   1
adc       NA     NA   NA   NA    NA  NA    NA  NA  NA
```

> Sig50_bupa

```
       icrs icmed icmean imed imean knn fritz irs adc
icrs     NA     0      0    1     1   1     1   1   1
icmed    NA    NA      0    1     1   1     1   1   1
icmean   NA    NA     NA    1     1   1     1   1   1
imed     NA    NA     NA   NA     0   0     1   1   1
imean    NA    NA     NA   NA    NA   0     1   1   1
knn      NA    NA     NA   NA    NA  NA     0   0   1
fritz    NA    NA     NA   NA    NA  NA    NA   0   1
irs      NA    NA     NA   NA    NA  NA    NA  NA   1
adc      NA    NA     NA   NA    NA  NA    NA  NA  NA
```

D.3. Matrices de significancia para las pruebas de Friedman por proporción en los datos *diabetes*

> Sig05_diabetes

```
       icrs imean imed icmean icmed knn fritz irs adc
icrs     NA     0    0      1     1   1     1   1   1
imean    NA    NA    0      0     1   1     1   1   1
imed     NA    NA   NA      0     0   1     1   1   1
icmean   NA    NA   NA     NA     0   0     1   1   1
icmed    NA    NA   NA     NA    NA   0     0   1   1
knn      NA    NA   NA     NA    NA  NA     0   1   1
fritz    NA    NA   NA     NA    NA  NA    NA   0   1
irs      NA    NA   NA     NA    NA  NA    NA  NA   1
adc      NA    NA   NA     NA    NA  NA    NA  NA  NA
```

> Sig10_diabetes

```
       icrs imean imed icmean knn icmed fritz irs adc
icrs     NA     0    1      1   1     1     1   1   1
imean    NA    NA    0      0   0     1     1   1   1
imed     NA    NA   NA      0   0     1     1   1   1
icmean   NA    NA   NA     NA   0     0     0   1   1
knn      NA    NA   NA     NA  NA     0     0   1   1
icmed    NA    NA   NA     NA  NA    NA     0   0   1
fritz    NA    NA   NA     NA  NA    NA    NA   0   1
irs      NA    NA   NA     NA  NA    NA    NA  NA   1
adc      NA    NA   NA     NA  NA    NA    NA  NA  NA
```

> Sig15_diabetes

```
       icrs imean imed icmean knn icmed fritz irs adc
icrs     NA     0    1      1   1     1     1   1   1
imean    NA    NA    0      0   1     1     1   1   1
imed     NA    NA   NA      0   1     1     1   1   1
icmean   NA    NA   NA     NA   0     0     1   1   1
```

```
knn      NA    NA   NA     NA  NA     0     0   1   1
icmed    NA    NA   NA     NA  NA    NA     0   1   1
fritz    NA    NA   NA     NA  NA    NA    NA   0   1
irs      NA    NA   NA     NA  NA    NA    NA  NA   0
adc      NA    NA   NA     NA  NA    NA    NA  NA  NA
```

> Sig20_diabetes

```
       icrs imean imed icmean icmed knn fritz irs adc
icrs     NA     0    1      1     1   1     1   1   1
imean    NA    NA    0      0     1   1     1   1   1
imed     NA    NA   NA      0     1   1     1   1   1
icmean   NA    NA   NA     NA     0   0     1   1   1
icmed    NA    NA   NA     NA    NA   0     0   1   1
knn      NA    NA   NA     NA    NA  NA     0   1   1
fritz    NA    NA   NA     NA    NA  NA    NA   0   1
irs      NA    NA   NA     NA    NA  NA    NA  NA   0
adc      NA    NA   NA     NA    NA  NA    NA  NA  NA
```

> Sig25_diabetes

```
       icrs imean imed icmean knn icmed fritz irs adc
icrs     NA     0    1      1   1     1     1   1   1
imean    NA    NA    0      0   1     1     1   1   1
imed     NA    NA   NA      0   0     1     1   1   1
icmean   NA    NA   NA     NA   0     0     1   1   1
knn      NA    NA   NA     NA  NA     0     0   1   1
icmed    NA    NA   NA     NA  NA    NA     0   1   1
fritz    NA    NA   NA     NA  NA    NA    NA   0   1
irs      NA    NA   NA     NA  NA    NA    NA  NA   0
adc      NA    NA   NA     NA  NA    NA    NA  NA  NA
```

> Sig30_diabetes

```
       icrs imean imed icmean knn icmed fritz irs adc
icrs     NA     0    1      1   1     1     1   1   1
imean    NA    NA    0      0   1     1     1   1   1
imed     NA    NA   NA      0   1     1     1   1   1
icmean   NA    NA   NA     NA   0     0     1   1   1
knn      NA    NA   NA     NA  NA     0     0   1   1
icmed    NA    NA   NA     NA  NA    NA     0   1   1
fritz    NA    NA   NA     NA  NA    NA    NA   0   1
irs      NA    NA   NA     NA  NA    NA    NA  NA   0
adc      NA    NA   NA     NA  NA    NA    NA  NA  NA
```

> Sig35_diabetes

```
       icrs imean imed icmean knn icmed fritz irs adc
icrs     NA     0    1      1   1     1     1   1   1
imean    NA    NA    0      0   1     1     1   1   1
imed     NA    NA   NA      0   1     1     1   1   1
icmean   NA    NA   NA     NA   0     0     1   1   1
```

```
knn      NA    NA   NA     NA  NA     0     0   1   1
icmed    NA    NA   NA     NA  NA    NA     0   1   1
fritz    NA    NA   NA     NA  NA    NA    NA   0   1
irs      NA    NA   NA     NA  NA    NA    NA  NA   0
adc      NA    NA   NA     NA  NA    NA    NA  NA  NA
```

> Sig40_diabetes

```
       icrs imean imed icmean icmed knn fritz irs adc
icrs     NA     1    1      1     1   1     1   1   1
imean    NA    NA    0      0     1   1     1   1   1
imed     NA    NA   NA      0     0   1     1   1   1
icmean   NA    NA   NA     NA     0   1     1   1   1
icmed    NA    NA   NA     NA    NA   0     0   1   1
knn      NA    NA   NA     NA    NA  NA     0   0   1
fritz    NA    NA   NA     NA    NA  NA    NA   0   1
irs      NA    NA   NA     NA    NA  NA    NA  NA   0
adc      NA    NA   NA     NA    NA  NA    NA  NA  NA
```

> Sig45_diabetes

```
       icrs imean imed icmean icmed knn fritz irs adc
icrs     NA     0    1      1     1   1     1   1   1
imean    NA    NA    0      0     1   1     1   1   1
imed     NA    NA   NA      0     1   1     1   1   1
icmean   NA    NA   NA     NA     0   0     1   1   1
icmed    NA    NA   NA     NA    NA   0     0   1   1
knn      NA    NA   NA     NA    NA  NA     0   1   1
fritz    NA    NA   NA     NA    NA  NA    NA   0   1
irs      NA    NA   NA     NA    NA  NA    NA  NA   0
adc      NA    NA   NA     NA    NA  NA    NA  NA  NA
```

> Sig50_diabetes

```
       icrs imean imed icmean icmed knn fritz irs adc
icrs     NA     0    1      1     1   1     1   1   1
imean    NA    NA    0      0     1   1     1   1   1
imed     NA    NA   NA      0     1   1     1   1   1
icmean   NA    NA   NA     NA     0   0     1   1   1
icmed    NA    NA   NA     NA    NA   0     0   1   1
knn      NA    NA   NA     NA    NA  NA     0   1   1
fritz    NA    NA   NA     NA    NA  NA    NA   0   1
irs      NA    NA   NA     NA    NA  NA    NA  NA   0
adc      NA    NA   NA     NA    NA  NA    NA  NA  NA
```

D.4. Matrices de significancia para las pruebas de Friedman por proporción en los datos *Bajopeso*

> Sig05_bajopeso

```
       fritz icrs imed knn irs imean icmed icmean adc
fritz     NA    0    0   0   0     0     0      0   1
icrs      NA   NA    0   0   0     0     0      0   1
imed      NA   NA   NA   0   0     0     0      0   1
knn       NA   NA   NA  NA   0     0     0      0   1
irs       NA   NA   NA  NA  NA     0     0      0   1
imean     NA   NA   NA  NA  NA    NA     0      0   1
icmed     NA   NA   NA  NA  NA    NA    NA      0   1
icmean    NA   NA   NA  NA  NA    NA    NA     NA   1
adc       NA   NA   NA  NA  NA    NA    NA     NA  NA
```

> Sig10_bajopeso

```
       fritz imed icrs knn imean irs icmed icmean adc
fritz     NA    0    0   0     0   0     0      1   1
imed      NA   NA    0   0     0   0     0      0   1
icrs      NA   NA   NA   0     0   0     0      0   1
knn       NA   NA   NA  NA     0   0     0      0   1
imean     NA   NA   NA  NA    NA   0     0      0   1
irs       NA   NA   NA  NA    NA  NA     0      0   1
icmed     NA   NA   NA  NA    NA  NA    NA      0   1
icmean    NA   NA   NA  NA    NA  NA    NA     NA   1
adc       NA   NA   NA  NA    NA  NA    NA     NA  NA
```

> Sig15_bajopeso

```
       fritz icrs imed imean knn icmed irs icmean adc
fritz     NA    0    0     0   0     0   1      1   1
icrs      NA   NA    0     0   0     0   0      1   1
imed      NA   NA   NA     0   0     0   0      0   1
imean     NA   NA   NA    NA   0     0   0      0   1
knn       NA   NA   NA    NA  NA     0   0      0   1
icmed     NA   NA   NA    NA  NA    NA   0      0   1
irs       NA   NA   NA    NA  NA    NA  NA      0   1
icmean    NA   NA   NA    NA  NA    NA  NA     NA   1
adc       NA   NA   NA    NA  NA    NA  NA     NA  NA
```

> Sig20_bajopeso

```
       imean imed icrs fritz knn irs icmed icmean adc
imean     NA    0    0     0   0   0     1      1   1
imed      NA   NA    0     0   0   0     1      1   1
icrs      NA   NA   NA     0   0   0     0      1   1
fritz     NA   NA   NA    NA   0   0     0      1   1
knn       NA   NA   NA    NA  NA   0     0      0   1
irs       NA   NA   NA    NA  NA  NA     0      0   1
icmed     NA   NA   NA    NA  NA  NA    NA      0   1
icmean    NA   NA   NA    NA  NA  NA    NA     NA   1
adc       NA   NA   NA    NA  NA  NA    NA     NA  NA
```

> Sig25_bajopeso

```
       icrs fritz imean imed knn irs icmed icmean adc
icrs     NA     0     0    0   1   1     1       1   1
fritz    NA    NA     0    0   0   0     0       1   1
imean    NA    NA    NA    0   0   0     0       0   1
imed     NA    NA    NA   NA   0   0     0       0   1
knn      NA    NA    NA   NA  NA   0     0       0   1
irs      NA    NA    NA   NA  NA  NA     0       0   1
icmed    NA    NA    NA   NA  NA  NA    NA       0   1
icmean   NA    NA    NA   NA  NA  NA    NA      NA   1
adc      NA    NA    NA   NA  NA  NA    NA      NA  NA
```

> *Sig30_bajopeso*

```
       icrs imean imed knn fritz irs icmed icmean adc
icrs     NA     0    0   0     0   1     1       1   1
imean    NA    NA    0   0     0   0     0       1   1
imed     NA    NA   NA   0     0   0     0       0   1
knn      NA    NA   NA  NA     0   0     0       0   1
fritz    NA    NA   NA  NA    NA   0     0       0   1
irs      NA    NA   NA  NA    NA  NA     0       0   1
icmed    NA    NA   NA  NA    NA  NA    NA       0   1
icmean   NA    NA   NA  NA    NA  NA    NA      NA   1
adc      NA    NA   NA  NA    NA  NA    NA      NA  NA
```

> *Sig35_bajopeso*

```
       icrs imed knn imean fritz irs icmed icmean adc
icrs     NA    0   0     0     0   1     1       1   1
imed     NA   NA   0     0     0   0     0       1   1
knn      NA   NA  NA     0     0   0     0       1   1
imean    NA   NA  NA    NA     0   0     0       0   1
fritz    NA   NA  NA    NA    NA   0     0       0   1
irs      NA   NA  NA    NA    NA  NA     0       0   1
icmed    NA   NA  NA    NA    NA  NA    NA       0   1
icmean   NA   NA  NA    NA    NA  NA    NA      NA   1
adc      NA   NA  NA    NA    NA  NA    NA      NA  NA
```

> *Sig40_bajopeso*

```
       icrs imed imean fritz knn irs icmed icmean adc
icrs     NA    0     0     0   0   1     1       1   1
imed     NA   NA     0     0   0   0     1       1   1
imean    NA   NA    NA     0   0   0     0       1   1
fritz    NA   NA    NA    NA   0   0     0       1   1
knn      NA   NA    NA    NA  NA   0     0       0   1
irs      NA   NA    NA    NA  NA  NA     0       0   1
icmed    NA   NA    NA    NA  NA  NA    NA       0   1
icmean   NA   NA    NA    NA  NA  NA    NA      NA   0
adc      NA   NA    NA    NA  NA  NA    NA      NA  NA
```

> *Sig45_bajopeso*

```
       knn imed icrs imean fritz irs icmed icmean adc
knn     NA    0    0     0     0   0     1       1   1
imed    NA   NA    0     0     0   0     1       1   1
icrs    NA   NA   NA     0     0   0     1       1   1
imean   NA   NA   NA    NA     0   0     1       1   1
fritz   NA   NA   NA    NA    NA   0     1       1   1
irs     NA   NA   NA    NA    NA  NA     0       1   1
icmed   NA   NA   NA    NA    NA  NA    NA       0   0
icmean  NA   NA   NA    NA    NA  NA    NA      NA   0
adc     NA   NA   NA    NA    NA  NA    NA      NA  NA
```

> Sig50_bajopeso

```
       icrs fritz knn irs imean imed icmed icmean adc
icrs     NA     0   0   0     0    0     1       1   1
fritz    NA    NA   0   0     0    0     1       1   1
knn      NA    NA  NA   0     0    0     1       1   1
irs      NA    NA  NA  NA     0    0     0       1   1
imean    NA    NA  NA  NA    NA    0     0       1   1
imed     NA    NA  NA  NA    NA   NA     0       1   1
icmed    NA    NA  NA  NA    NA   NA    NA       0   1
icmean   NA    NA  NA  NA    NA   NA    NA      NA   0
adc      NA    NA  NA  NA    NA   NA    NA      NA  NA
```

D.5. Matrices de significancia para las pruebas de Friedman por proporción en los datos *German*

> Sig05_german

```
       icrs fritz knn imed imean icmed icmean irs adc
icrs     NA     0   1    1     1     1       1   1   1
fritz    NA    NA   0    0     0     0       1   1   1
knn      NA    NA  NA    0     0     0       0   0   1
imed     NA    NA  NA   NA     0     0       0   0   1
imean    NA    NA  NA   NA    NA     0       0   0   1
icmed    NA    NA  NA   NA    NA    NA       0   0   1
icmean   NA    NA  NA   NA    NA    NA      NA   0   1
irs      NA    NA  NA   NA    NA    NA      NA  NA   1
adc      NA    NA  NA   NA    NA    NA      NA  NA  NA
```

> Sig10_german

```
       icrs fritz knn imean imed irs icmed icmean adc
icrs     NA     0   1     1    1   1     1       1   1
fritz    NA    NA   0     0    0   1     1       1   1
knn      NA    NA  NA     0    0   0     0       0   1
imean    NA    NA  NA    NA    0   0     0       0   1
```

```
imed     NA    NA  NA    NA   NA   0     0      0   1
irs      NA    NA  NA    NA   NA  NA     0      0   1
icmed    NA    NA  NA    NA   NA  NA    NA      0   1
icmean   NA    NA  NA    NA   NA  NA    NA     NA   1
adc      NA    NA  NA    NA   NA  NA    NA     NA  NA
```

> Sig15_german

```
       icrs fritz imean imed irs knn icmean icmed adc
icrs     NA     0     0    1   1   1      1     1   1
fritz    NA    NA     0    0   1   1      1     1   1
imean    NA    NA    NA    0   0   0      1     1   1
imed     NA    NA    NA   NA   0   0      0     0   1
irs      NA    NA    NA   NA  NA   0      0     0   1
knn      NA    NA    NA   NA  NA  NA      0     0   1
icmean   NA    NA    NA   NA  NA  NA     NA     0   1
icmed    NA    NA    NA   NA  NA  NA     NA    NA   1
adc      NA    NA    NA   NA  NA  NA     NA    NA  NA
```

> Sig20_german

```
       icrs fritz imean imed knn irs icmed icmean adc
icrs     NA     0     1    1   1   1     1      1   1
fritz    NA    NA     0    0   0   1     1      1   1
imean    NA    NA    NA    0   0   0     1      1   1
imed     NA    NA    NA   NA   0   0     1      1   1
knn      NA    NA    NA   NA  NA   0     0      0   1
irs      NA    NA    NA   NA  NA  NA     0      0   1
icmed    NA    NA    NA   NA  NA  NA    NA      0   0
icmean   NA    NA    NA   NA  NA  NA    NA     NA   0
adc      NA    NA    NA   NA  NA  NA    NA     NA  NA
```

> Sig25_german

```
       icrs imean imed fritz knn irs icmed icmean adc
icrs     NA     0    0     1   1   1     1      1   1
imean    NA    NA    0     0   0   0     1      1   1
imed     NA    NA   NA     0   0   0     1      1   1
fritz    NA    NA   NA    NA   0   0     1      1   1
knn      NA    NA   NA    NA  NA   0     1      1   1
irs      NA    NA   NA    NA  NA  NA     0      0   1
icmed    NA    NA   NA    NA  NA  NA    NA      0   0
icmean   NA    NA   NA    NA  NA  NA    NA     NA   0
adc      NA    NA   NA    NA  NA  NA    NA     NA  NA
```

> Sig30_german

```
       icrs imean imed fritz knn irs icmed icmean adc
icrs     NA     0    0     1   1   1     1      1   1
imean    NA    NA    0     0   0   1     1      1   1
imed     NA    NA   NA     0   0   0     1      1   1
fritz    NA    NA   NA    NA   0   0     1      1   1
```

```
knn      NA    NA   NA    NA  NA   0     1      1   1
irs      NA    NA   NA    NA  NA  NA     0      0   1
icmed    NA    NA   NA    NA  NA  NA    NA      0   0
icmean   NA    NA   NA    NA  NA  NA    NA     NA   0
adc      NA    NA   NA    NA  NA  NA    NA     NA  NA
```

Bibliografía

[Bradley, 1996] Bradley, A. P. (1996). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*.

[Delong et al., 1998] Delong, E. R., Delong, D. M., and Clarke-Pearson, D. L. (1998). Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics*, 44:837–845.

[Dowdy and Wearden, 1991] Dowdy, S. and Wearden, S. (1991). *Statistics for Research Segunda Edición*. John Wiley and Sons.

[Hanley and McNeil, 1982] Hanley, J. and McNeil, B. (1982). The meaning and use of the area under the receiver operating characteristic (roc) curve. *Radiology*, 143:29–36.

[Hanley and McNeil, 1983] Hanley, J. and McNeil, B. (1983). A method of comparing the areas under the receiver operating characteristic curves derived from the same cases. *Radiology*, 148:839–843.

[Herzog and Rubin, 1983] Herzog, T. and Rubin, D. (1983). Using multiple imputations to handle nonresponse in sample surveys. Chapter in Incomplete Data in Sample Surveys, Vol.2, W.G. Madow, I.Olkin, and D.B. Rubin.

[Hollander and Wolfe, 1973] Hollander, M. and Wolfe, D. A. (1973). *Nonparametric Statistical Methods*. John Wiley and Sons.

[Hosmer and Lemeshow, 1989] Hosmer, D. W. and Lemeshow, S. (1989). *Applied Logistic Regression*. John Wiley and Sons.

[Iannacchione, 1999] Iannacchione, V. G. (1999). Location and response propensity modeling for the 1995 national survey of family growth. *Research Triangle Institute*.

[Jinn, 2000] Jinn, J.-H. (2000). The effect of different imputation methods on analytical statistics of simple linear regression. *Unknow*.

[Kalton and Kasprzyk, 1982] Kalton, G. and Kasprzyk, D. (1982). Imputing for missing survey responses. *American Statistical Association*.

[Kennickell, 1991] Kennickell, A. B. (1991). Imputation of the 1989 survey of consumer finances: Stochastic relaxation and multiple imputation. *Redactado para el Annual Meetings of the American Statistical Association, Atlanta Georgia*.

[Kennickell, 1998] Kennickell, A. B. (1998). Multiple imputation in the survey of consumer finances. Redactado para el *Joint Statistical Meetings*, Dallas, Texas.

[Le, 2003] Le, C. T. (2003). *Introductory Biostatistics*. John Wiley and Sons.

[Little and Raghunathan, 2004] Little, R. J. and Raghunathan, T. (2004). *Statistical Analysis with Missing Data*. JPSM.

[Little and Rubin, 2002] Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data, Segunda Edición*. John Wiley and Sons.

[Perez et al., 2002] Perez, A., Dennis, R. J., Gil, J. F., Rondón, M. A., and López, A. (2002). Use of the mean hot deck and multiple imputation techniques to predict outcome in intensive care unit patients in colombia. *Statistics in Medicine*, 21:3885–3896.

[Pyle, 1999] Pyle, D. (1999). *Data Preparation for Data Mining*. Morgan Kaufmann, San Francisco.

[Rodríguez, 2004] Rodríguez, C. (2004). A computational enviroment for data preprocessing in supervised classification.

[Rubin, 1987] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. John Wiley and Sons.

[Santos, 1981] Santos, R. (1981). Effects of imputation on regression coefficients. *American Statistical Association*.

[Shaefer, 1997] Shaefer, J. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall.

[Shapiro, 1998] Shapiro, J. H. (1998). Bounds on the area under the roc curve. *J. Opt. Soc.*, 15.

[Vach and Blettner, 1999] Vach, W. and Blettner, M. (1999). Logistic regression with incompletely observed categorical covariates ... *Statistics in Medicine*.