Universidad de Valladolid

Grado en Estadística

TRABAJO FIN DE GRADO

ANÁLISIS DE DATOS PROVENIENTES DE PACIENTES INTERVENIDOS DE TAVI

Autor: Gonzalo Cruz González

Tutores:

Agustín Mayo Íscar

Itzíar Gómez Salvador

Agradecimientos

A todos los profesores del departamento de Estadística e Investigación Operativa de la Universidad de Valladolid, por transmitirme todo el conocimiento del cual puedo presumir con verdadero orgullo. A todos mis compañeros de carrera, haciendo una merecida mención especial a mis compañeras y amigas Paula y Celia, que sin ellas este camino habría sido muy diferente.

Agradecer de forma especial a mis tutores, Agustín e Itzíar, la dedicación, esfuerzo y ayuda que me han proporcionado a lo largo de este trabajo.

A todos ellos: Gracias

Índice general

Resumen: A partir de los pacientes intervenidos de TAVI (*Transcatether Aortic Valve Implantation*) en un hospital terciario del Sistema Nacional de Salud (SNS), se han clasificado utilizando técnicas estadísticas propias del análisis multivariante y modelado mediante la regresión logística múltiple. Con estas técnicas se ha conseguido elaborar un modelo predictivo para determinar la mortalidad en el seguimiento del paciente dependiendo de una serie de patologías y dar una aplicación clínica al modelo. Por otro lado, se han detectado los pacientes que son atípicos multivariantes mediante la técnica MCD y elaborado un perfil para poder comparar ese grupo de *outliers* con la muestra general. Todos los análisis han sido realizados con los paquetes estadísticos SAS® y R.

Palabras Clave: TAVI, Regresión Logística, Atípicos, MCD, SAS®.

Abstract: Beginnig with a data base of patients who were intervened of TAVI in a hospital of the National Health Insurance System, they have been clasified using statistical techniques such as multivariate analysis and a multiple logistic regresion model. With these techniques, a predictive model to determine whether a patient lives or dies depending on a range of pathologies has been developed, and also, to give a clinical application. On the other hand, the patients have been detected as multivariate outliers using the MCD algorithm and have given different profiles in order to compare the outliers group and the general sample. All statistical analysis have been perfomed with SAS® and R.

Key Words: TAVI, Logistic Regresion, Outliers, MCD, SAS®.

Capítulo 1

Introducción

> "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."
>
> *H.G. Wells*

La estenosis aórtica es una patología que afecta a la válvula aórtica reduciendo el orificio efectivo por el que pasa la sangre hasta la arteria aorta. Esta enfermedad se debe al depósito de calcio en los velos de la válvula.

Hasta el año 2002 el único tratamiento para esta enfermedad era la cirugía; sin embargo, a partir de este año aparecen las TAVIs, (*Transcatether Aortic Valve Implantation* de sus siglas en inglés), haciendo que la patología pudiese resolverse de una forma mínimamente invasiva.

Los modelos más utilizados en la actualidad son los desarrollados por los laboratorios Medtronic y Edwards Lifesciences, que crearon los CoreValve y SAPIEN respectivamente; los cuales, han ido perfeccionando para solventar algunos de los problemas inherentes al implante, como la regurgitación aórtica.

En el presente Trabajo Fin de Grado se analizaron un total de 132 pacientes intervenidos en un hospital terciario del Sistema Nacional de Salud (SNS). En ellos se han recogido variables demográficas, clínicas, ecocardiográficas, diagnósticas del procedimiento y del seguimiento antes y después del implante, y se intentará buscar respuestas a los siguientes problemas:

1. El primer problema a analizar será la mortalidad en el seguimiento de los pacientes. Tras el implante de TAVI, es necesario saber cuáles son los predictores independien-

tes de este evento. De esta forma se podrán conocer las patologías que más influyen en los pacientes y cuál es su efecto sobre la mortalidad en el seguimiento.

En el *Capítulo 2* se realizará un análisis de la mortalidad en el seguimiento de los pacientes utilizando técnicas propias del Análisis de Datos Categóricos y de los Modelos Lineales Generalizados. Para este capítulo se obtendrán diferentes variables asociadas con la variable mortalidad a largo plazo con las que se ajustará un modelo de regresión de tipo logístico binomial. El objetivo de este modelo será poder determinar en pacientes futuros cuál es la probabilidad de morir una vez se le ha hecho el implante de la válvula.

2. Otro de los problemas a analizar son las complicaciones vasculares una vez se ha realizado el implante, y la relación que ello tiene con la arteriopatía periférica[1].

   En el *Capítulo 3* se analizará si se puede predecir la complicación vascular a través de los pacientes con arteriopatía periférica. Además, se realizará un estudio de dichos pacientes utilizando técnicas propias del Análisis Multivariante, como las distancias de Mahalanobis y el Estimador del Determinante de Mínima Covarianza (MCD).

   Como parte final a este estudio, se realizarán diferentes simulaciones de datos para comprobar si los resultados obtenidos son fruto del azar, o si por el contrario, son característicos de la población.

Finalmente, en el *Capítulo 4*, se detallarán los resultados y conclusiones extraídos de los diferentes análisis.

[1]La arteriopatía periférica es una patología en la que las arterias se estrechan y endurecen debido a la acumulación de grasa en las paredes arteriales. Este endurecimiento produce una falta de irrigación sanguínea en piernas y pies provocando lesiones en tejidos y nervios.

Capítulo 2

Análisis Predictivo de la Mortalidad en el Seguimiento

> "All models are wrong, but some models are useful."
>
> *George E. P. Box (1979)*

En este capítulo se creará una regla para intentar predecir la mortalidad a largo plazo del paciente. Este análisis resulta de gran interés puesto que permite estimar después de la intervención la probabilidad de que un paciente viva o no.

2.1. Metodología

Para realizar este estudio se buscaron aquellas variables que podían tener una relación con la mortalidad en el seguimiento. Para ello se utilizaron para las variables categóricas los tests Ji-Cuadrado y de Fisher; éste último solo fue utilizado cuando la frecuencia esperada en alguna de las celdas era menor de 5. Para las variables continuas se utilizaron los test t de Student y el test de Wilcoxon, equivalente no paramétrico cuando la variable no era normal.

2.1.1. Regresión Logística

La regresión logística binaria es uno de los modelos más importantes a la hora de analizar datos con respuesta categórica. En el ambiente clínico desde el se está enfocando

este TFG, existen multitud de ejemplos en los que se aplican las técnicas de la regresión logística.

La regresión logística se define como un modelo que proviene de una transformación logit del parámetro binomial, en el que los parámetros entran de forma lineal. El logit lineal tiene la fórmula 2.1,

$$logit[\pi(x)] = \log \frac{\pi(x)}{1-\pi(x)} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \tag{2.1}$$

quedando el modelo en función de los predictores lineales. Si se deshace el logit de esta ecuación se obtiene el cálculo de las probabilidades:

$$\pi(x) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}. \tag{2.2}$$

De esta forma, el parámetro β_i se refiere al efecto de x_i en el logaritmo de la odds cuando $Y = 1$ controlando los otros x_j.

Otro de los aspectos que ha de ser explicado en un modelo de regresión logística son las Odds Ratio (OR) cuya interpretación es:

$$OR = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{e^{\alpha+\beta_1}}{e^{\alpha+\beta_0}} = \frac{e^{\beta_1}}{e^{\beta_0}} \tag{2.3}$$

De esta forma se puede considerar la odds ratio como el incremento de pasar de la categoría de referencia a otras categorías. En caso de que el cociente dé como resultado 1, se supondrá que no hay una diferencia significativa entre los dos efectos.

Inferencia en la regresión logística: Al utilizar la máxima verosimilitud para estimar los parámetros del modelo, podemos recurrir a los resultados de Wald (1943) suponiendo que los estimadores de los parámetros del modelo siguen una distribución asintótica normal.

Una de las hipótesis que primero han de ser contrastadas es $H_0 : \beta = 0$ para lo cual se utiliza la logverosimilitud en $\hat{\beta}$ con el estadístico test $z = \hat{\beta}/SE$ que tiene una distribución N(0,1), o utilizando su cuadrado z^2 que sigue una χ^2_1.

2.1.2. Métodos automáticos de Selección de Variables

Para realizar cualquier tipo de regresión no se puede intentar incluir todas las variables que se desee, puesto que en el ajuste de cualquier modelo se sigue el **principio de**

parsimonia, es decir, intentar explicar la mayor cantidad de información posible con un número mínimo de regresores o variables. Para seguir el principio de parsimonia se pueden utilizar métodos automáticos de selección de variables

Dentro de la estadística hay varios métodos automáticos para determinar qué variables deben pertenecer al modelo: backward, fordward, stepwise... En este estudio se utilizó de partida el algoritmo stepwise, que contiene los siguientes pasos:

Algoritmo Stepwise: Se parte del modelo nulo y se incluyen variables secuencialmente según unos criterios. En el paso 1 se incluye la variable i, la cual tiene el p-valor asociado a la estimación de su parámetro más pequeño, en el paso 2 se incluye la variable j, y se evalúa si la variable i tiene sentido seguir permaneciendo dentro de la regresión. En este algoritmo se determinan un p-valor máximo de entrada y un p-valor mínimo de permanencia; y éste parará cuando no existan variables que cumplan los requisitos.

2.1.3. Curva ROC

La curva ROC (*Receiver Operating Characteristic*, de sus siglas en inglés), es un gráfico que representa la capacidad predictiva del modelo. Está dibujado en el intervalo [0,1] donde el 0 representa el diagnóstico negativo y el 1 un diagnóstico positivo. Si se introduce un punto de corte en cualquier parte del intervalo, se obtiene un punto de la curva ROC. Esta curva tiene representada la sensibilidad en el eje vertical y 1-especificidad en el eje horizontal.

La sensibilidad y la especificidad, y consecuentemente el resultado obtenido, varían según el punto de corte. Cada punto de la curva corresponde a un par específico de sensibilidad y especificidad y la curva completa proporciona una visión general de la actuación del test. Para comparar las curvas ROC, cuanto más se acerque a la esquina superior izquierda mejor será; teniendo el peor de los casos en la línea diagonal que va del (0,0) al (1,1).

Finalmente, las curvas ROC podrían ser utilizadas para comparar sus resultados, si varios tests fuesen realizados sobre la misma muestra.

2.1.4. Bondad de Ajuste

En la práctica, no podremos asegurar al 100 % que el modelo esté bien ajustado, por lo que se necesitan de tests y gráficos para valorar este aspecto

Test de Hosmer-Lemeshow

Los estadísticos Hosmer y Lemeshow (1980) propusieron una forma de analizar la falta de ajuste en modelos de regresión logística binomial con variables categóricas, en las que la tabla de contingencia de todas ellas daba como resultado un contingente en la celda muy pequeño. La construcción de este estadístico parte de dividir la muestra en 10 partes (los deciles de una distribución) y comparar lo observado frente a lo predicho por el modelo. El valor predicho para cada corte (o decil) es la suma de las probabilidades estimadas para el resultado de todas las observaciones del grupo. Con esta base, propusieron un estadístico de Pearson que comparase lo observado frente a lo predicho para cada corte, quedando la siguiente fórmula:

$$HL = \sum_{i=1}^{g} \frac{(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij})^2}{(\sum_j \hat{\pi}_{ij})[1 - (\sum_j \hat{\pi}_{ij})/n_i]}. \tag{2.4}$$

Siendo y_{ij} el resultado binario para la observación j del grupo i de la partición $i = 1, \ldots, g$,$j = 1, \ldots, n_i$ y siendo $\hat{\pi}_{ij}$ la correspondiente probabilidad ajustada en el modelo para los datos sin agrupar. Agresti (2002) advierte que cuando muchas observaciones tienen la misma probabilidad estimada se produce cierta arbitrariedad a la hora de formar los grupos y diferentes softwares pueden dar resultados distintos para un mismo conjunto de datos.

Este estadístico no tiene una distribución en el límite chi-cuadrado porque las observaciones en un grupo no tienen un resultado idéntico, puesto que no comparten una probabilidad de éxito idéntica. Sin embargo, Hosmer y Lemeshow (1980) detectaron que cuando el número de patrones distintos en los valores de las variables era igual al tamaño de la muestra, la distribución bajo la hipótesis nula seguía una χ^2_{g-2}.

Aún así este estadístico, al igual que otros estadísticos para detectar la bondad de ajuste, no tiene suficiente potencia como para detectar diferentes tipos de falta de ajuste; puesto que sólo indica si existe o no.

AIC: Criterio de Información de Akaike

El Criterio de Información de Akaike es una medida que valora la bondad del ajuste y penaliza la inclusión de muchas variables en el modelo. El AIC sigue la fórmula: $AIC = 2k - 2\ln(L)$, siendo k el número de variables del modelo y L el máximo de la Logverosimilitud.

Esta medida sirve para comparar diferentes modelos, siendo el mejor aquel que tenga

un AIC menor. Sin embargo, el AIC no realiza un contraste de hipótesis, de forma que si un modelo no se ha ajustado correctamente a los datos, el AIC no va a dar evidencias de ello.

Análisis de los residuos

Los residuos más comunes a analizar en un modelo de regresión son: residuos de Pearson, residuos *deviance* y puntos con efecto palanca o *leverage*.

1. **Residuos de Pearson:** Los residuos de Pearson están diseñados para evaluar la existencia de posibles valores atípicos dentro de la muestra. Generalmente se utiliza el intervalo [-2,2] si se requiere ser estricto, o [-2.5,2.5] si se puede ser más laxo. En el caso de que un valor salga de este intervalo puede ser considerado como valor atípico.

   El cálculo de estos residuos siguen la siguiente fórmula:

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)]}} \tag{2.5}$$

2. **Residuos *deviance*:** Una alternativa a los residuos de Pearson son los residuos *deviance* cuyo objetivo es evaluar la falta de ajuste. Este tipo de residuos tienen la siguiente fórmula:

$$DR = \sqrt{d_i} \times sign(y_i - n_i\hat{\pi}_i) \tag{2.6}$$

   donde

$$d_i = 2\left(y_i \log\frac{y_i}{n_i\hat{\pi}_i} + (n_i - y_i)\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)$$

3. ***Leverage*:** También denominados puntos con efecto palanca. Cuando una observación tiene un *leverage* alto está provocando que el hiperplano de la regresión pase por ella o lo más próximo posible, alterando todas las estimaciones y, consecuentemente, la calidad del modelo. El cálculo de los puntos *leverage* viene a raíz de la matriz **H** o, comúmente conocida como **matriz hat**.

2.2. Descripción de las Variables Analizadas

Para cada paciente al que se le implantó una TAVI se han recogido variables demográficas, clínicas, ecocardiográficas, diagnósticas, del procedimiento y del seguimiento. Se

considerará la mortalidad en el seguimiento como la variable respuesta; siendo éste último de mediana 600 días (RI=(113.5, 1052)).

En el apéndice A, se encuentra una relación de las variables analizadas, con un resumen de sus propiedades: N, Media, SD, mínimo y máximo.

Las variables que obtuvieron asociación con la mortalidad en el seguimiento y que podrían ser susceptibles de ser incluidas en el modelo, están incluidas en la tabla 2.1.

Tabla 2.1: Tabla variables asociación con *Death*

Variable	N (A priori)	P-valor Asociación
IRC [1]	132	0.01
EPOC [2]	132	0.046
Fragilidad	132	0.01
STS[3] Score	132	0.011
Tabaquismo	131	0.029
IC[4] Previa	132	0.03
ICP [5] Previo	132	0.011
IAM [6] Previo	132	0.03
Clopidogrel	130	0.02
Insuf. Aórtica post [7]	130	0.04
Embolización de la prótesis	128	0.0028
IRA [8]	128	0.02
FADN [9]	123	0.01
Ritmo Cardiaco Inhosp[10]	124	0.03
Insuf. Aórtica Inhosp	118	0.02

2.3. Ajuste del modelo de Regresión Logística

2.3.1. Análisis del Modelo Preliminar

Atendiendo al principio de parsimonia, explicado en la *Sección 2.1.3*, y al disponer únicamente de 42 eventos en el conjunto de 132 observaciones, se podrán incluir 4 variables

[1]IRC: Insuficiencia Renal Crónica
[2]EPOC: Enfermedad Pulmonar Obstructiva Crónica
[3]STS: Society of Thoracic Surgeons
[4]IC: Insuficiencia Cardiaca
[5]ICP: Intervención Coronaria Percutánea
[6]IAM: Infarto Agudo de Miocardio
[7]Esta variable debido a sus características se utilizó categorizada con 4 niveles: 0, 1, 2 o 3
[8]IRA: Insuficiencia Renal Aguda
[9]FADN: Fibrilación Auricular de Novo
[10]Inhosp: Ingreso Hospitalario

como máximo, sin interacciones.

La ausencia de las interacciones en este modelo se debe a dos razones: a) a la regla empírica 1-10. Esta regla establece que se debe incluir una variable por cada 10 casos ocurridos en la variable respuesta, en nuestro modelo, que se produzca el fallecimiento del paciente. Y b) la complejidad a la hora de explicar el efecto que produce la interacción sobre la mortalidad en el seguimiento.

Para la regresión logística se hizo uso del SAS/STAT® PROC LOGISTIC, con el que se puede hacer la selección de variables, el ajuste del modelo y posterior obtención de plots y residuos de una forma sencilla.

En la tabla 2.2 se encuentran las variables candidatas a formar el modelo obtenidas mediante el algoritmo Stepwise.

Tabla 2.2: Resumen Stepwise

Parámetro	Clase Referencia	DF	Estimación	Error Estándar	Wald Chi-Square	Pr >ChiSq
Intercept		1	-0.0914	0.2813	0.1057	0.7451
IRC	0	1	0.5654	0.2537	4.9669	0.0258
EPOC	0	1	0.7394	0.2433	9.2348	0.0024
Fragilidad	0	1	0.5847	0.2172	7.2454	0.0071
PrevPCI	0	1	0.6032	0.2344	6.6251	0.0101

De las 4 variables candidatas, se ajustaron los modelos de 3 y 4 variables para comprobar si pasando de 3 variables a 4 la calidad del modelo mejoraba significativamente. En la tabla 2.3 se encuentran los diferentes aspectos medidos y se aprecia cómo el porcentaje de pacientes bien clasificados se reduce al aumentar la inclusión de variables.

Eligiendo un modelo en el que se diese un equilibrio entre el porcentaje de pacientes bien clasificados, un p-valor alto en el test de Hosmer-Lemeshow y un AIC bajo, las variables que se incluyeron en el modelo fueron: IRC, EPOC y Fragilidad.

Tabla 2.3: Comparación diferentes modelos

Modelo	Variables Incluidas	% Clasificados[11]	p-valor Hosmer-Lemeshow	Coeficientes Significativos	AIC[12]
1	IRC EPOC Fragilidad	75.76 %	0.7853	Sí	153.555
2	IRC EPOC PrevPCI	70.45 %	0.6187	Sí	153.897
3	EPOC Fragilidad PrevPCI	70.45 %	0.2934	Sí	151.597
4	IRC Fragilidad PrevPCI	72.73 %	0.8308	Sí	156.653
5	IRC EPOC Fragilidad PrevPCI	71.97 %	0.2695	Sí	148.468

2.3.2. Análisis del modelo ajustado

La tabla que resume el ajuste indicó que la regresión logística sí tenía sentido pues los p-valores en cualquiera de los tres estadísticos eran menores a 0.05.

Tabla 2.4: Test Global Hipótesis Nula $\beta = 0$

Test	Chi-Square	DF	Pr>ChiSq
Likelihood Ratio	19.5750	3	0.0002
Score	18.9564	3	0.0003
Wald	16.2155	3	0.0010

La tabla 2.5 de las estimaciones indicó que todos los parámetros eran significativos a nivel 0.05, a excepción del intercept.

Tabla 2.5: Estimaciones por máxima verosimilitud del modelo

Parameter	Clase Referencia	DF	Estimate	Standard Error	Wald Chi-Square	Pr >ChiSq
Intercept		1	0.1491	0.2605	0.3277	0.5670
IRC	0	1	0.6304	0.2472	6.5009	0.0108
EPOC	0	1	0.5864	0.2244	6.8258	0.0090
Fragilidad	0	1	0.5995	0.2112	8.0554	0.0045

Tabla 2.6: OR e Intervalos de Confianza de Wald

Efecto	Estimación Puntual	95 % Wald Intervalo de Confianza	
IRC 0 vs 1	3.528	1.339	9.229
EPOC 0 vs 1	3.231	1.340	7.788
Fragilidad 0 vs 1	3.317	1.449	7.590

En los intervalos de Wald (tabla 2.6), se confirma lo visto en la tabla 2.5: los intervalos no contienen al 1, por lo que el efecto de las diferentes variables es significativo. Respecto a la interpretación que merecen las odds ratio, se puede afirmar que la presencia de cada uno de los factores multiplica casi por tres el riesgo (en el sentido de las odds ratio) de mortalidad en el seguimiento.

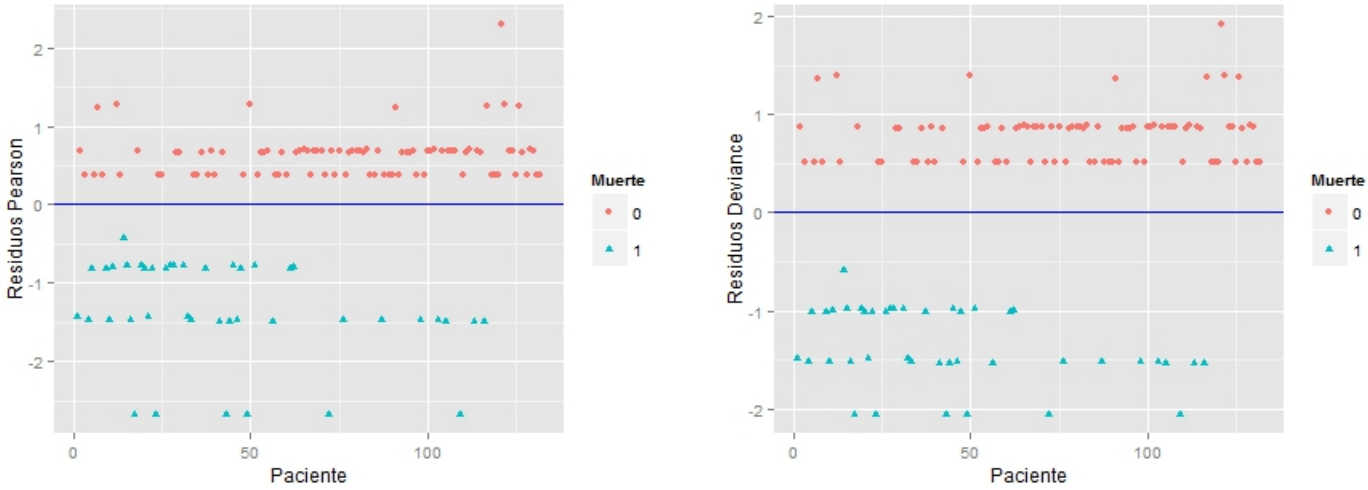
[11]Este porcentaje de clasificados se obtuvo para el punto (0.5,0.5) en la sensibilidad y especificidad
[12]Criterio de Información de Akaike

Plots de Diagnóstico

Como parte final del análisis del modelo, fueron analizados los plots de diagnóstico de la regresión para detectar la falta de ajuste.

Figura 2.1: Gráficos de Diagnóstico (I)

(a) Residuos Pearson (b) Residuos *deviance*

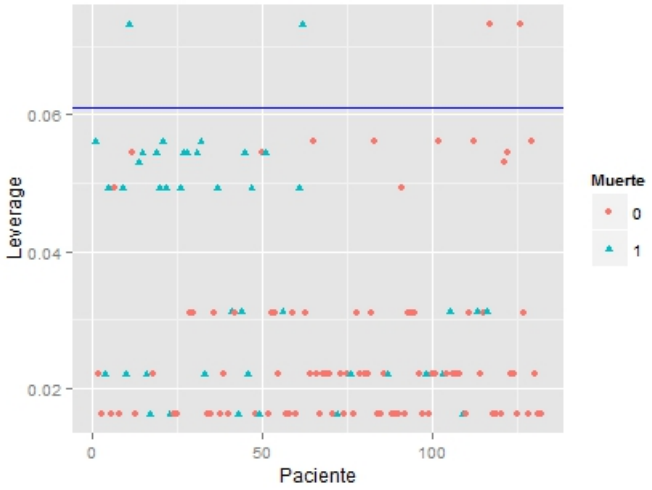


En la figura 2.1, en el primer gráfico, los Residuos de Pearson indican que no hay observaciones claramente atípicas. En los residuos *deviance*, situados en el gráfico de la derecha, tampoco se aprecian patrones de comportamiento, lo que indicó que no había falta de ajuste.

Finalmente para el *leverage* se utilizó un valor de corte para determinar si una observación hace de efecto palanca o no, siguiendo la siguiente fórmula: $h_{ii} > \frac{2(p+1)}{n} = \frac{2*(3+1)}{132} =$ 0,07. Se detectaron 4 observaciones que tenían un *leverage* de 0.08, superior al punto de corte, que podían afectar a la calidad del ajuste. Estas observaciones se tuvieron en cuenta hasta ver el test de Hosmer-Lemeshow y el p-valor de dicho test.

Figura 2.2: Gráfico de Diagnóstico (II)

(a) *leverage*



2.3.3. Validación del Modelo

Tal y como se mencionó en la teoría, *Sección 2.1.4*, para validar el modelo se hizo uso del Test de Hosmer-Lemeshow.

Para este test, generalmente se utilizan 10 cortes de la muestra; sin embargo, para no tener un contingente demasiado pequeño en cada corte, se decidió utilizar 6 en lugar de 10, quedando la tabla 2.7.

Tabla 2.7: Partición para el Test de Hosmer-Lemeshow

Grupo	**Total**	**Death = 0**		**Death = 1**	
		Observado	**Esperado**	**Observado**	**Esperado**
1	12	4	4.11	8	7.89
2	14	4	5.54	10	8.46
3	8	5	5.35	3	2.65
4	35	26	23.90	9	11.10
5	22	16	15.14	6	6.86
6	41	35	35.96	6	5.04

Se puede comprobar que no existen grandes diferencias entre lo observado y lo predicho, por lo que el ajuste a priori puede ser correcto. Aplicando la fórmula 2.4 se obtiene un valor del estadístico de 1.7297, que bajo la hipótesis nula sigue una χ^2_4, lo que da un p-valor de 0.7853. Este p-valor indicó que no había evidencias para rechazar la hipótesis

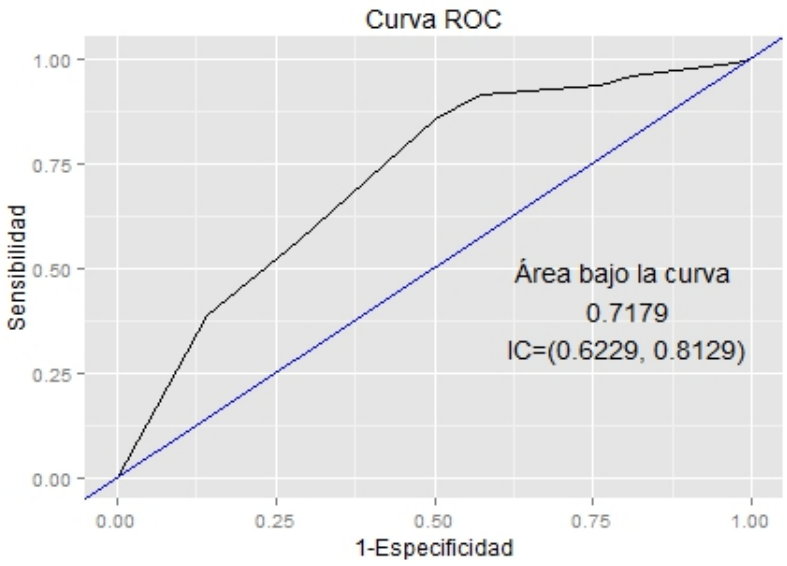
nula, por lo que se pudo afirmar que el modelo se ajustaba bien a los datos.

Este p-valor fue lo suficiente revelador como para no eliminar las 4 observaciones con *leverage* alto.

2.3.4. Predicciones del Modelo

El área bajo la curva ROC, que sirve como indicador de la capacidad predictiva del modelo fue del 71.79 %, lo que indicó que era un modelo de calidad media.

Figura 2.3: Curva ROC del modelo



Por otro lado, se obtuvo mediante SAS la predicción del evento de los diferentes pacientes, lo que proporcionó la tabla 2.8 de clasificación. Esta tabla se construyó con el punto de la curva ROC donde más se equilibraban la sensibilidad y la especificidad; dicho punto fue el (0.57,0.71).

Tabla 2.8: Tabla Clasificación

		Predicho		
		Vive	Muere	TOTAL
Obs	Vive	51	39	90
	Muere	12	30	42
	TOTAL	116	26	132

Para este punto concreto, el modelo es capaz de clasificar correctamente al 61 % de los

pacientes.

Para finalizar, se realizó el análisis con la predicción de las probabilidades de que un paciente viviese o falleciese. Sustituyendo en la ecuación 2.7 del modelo por los valores de las variables predictoras se obteniene la probabilidad predicha.

$$\widehat{P}(Death = 0) = \frac{\exp(0{,}1491 + 0{,}6304 * IRC + 0{,}5864 * EPOC + 0{,}5995 * Fragilidad)}{1 + \exp(0{,}1491 + 0{,}6304 * IRC + 0{,}5864 * EPOC + 0{,}5995 * Fragilidad)} \tag{2.7}$$

Tabla 2.9: Tabla probabilidades según modelo

	Patologías			Probabilidades	
Casos	IRC	EPOC	FRAGILIDAD	P (Vivir)	P (Morir)
1	0	0	0	0.877110041	0.122889959
2	1	0	0	0.669199375	0.330800625
3	0	1	0	0.688382449	0.311617551
4	0	0	1	0.682741438	0.317258562
5	1	1	0	0.385040229	0.614959771
6	0	1	1	0.399781385	0.600218615
7	1	0	1	0.378862805	0.621137195
8	1	1	1	0.158804121	0.841195879

La tabla 2.9 reveló que para el caso de tener una patología tienen más probabilidades de vivir aquellos que tienen EPOC que aquellos que tienen insuficiencia renal crónica. Del mismo modo, en el caso de dos patologías, tienen más probabilidad de sobrevivir los pacientes que tengan EPOC y Fragilidad que los que tengan problemas renales y fragilidad. Además se comprueba, que si un paciente no tiene ningún problema pulmonar, ni insuficiencia renal, ni fragilidad, la probabilidad que tiene de vivir es superior al 87 %. En cambio, un paciente que tenga las tres patología, tiene una probabilidad de vivir de tan solo un 15.8 %.

Aplicación Clínica

Este modelo cuenta con una singularidad, y es que las estimaciones de las odds ratio para las tres variables son prácticamente iguales, todas en torno al 3, por lo que las tres estimaciones de los parámetros beta podían ser reemplazadas por un valor común como la media, y en lugar de sustituir en las variables se contó en cuántas de las patologías el

paciente había sido positivo: en ninguna, 1, 2 ó 3.

$$\widehat{\pi(x)} = \frac{\exp(0{,}0559 + 0{,}5876 * X)}{1 + \exp(0{,}0559 + 0{,}5876 * X)}$$

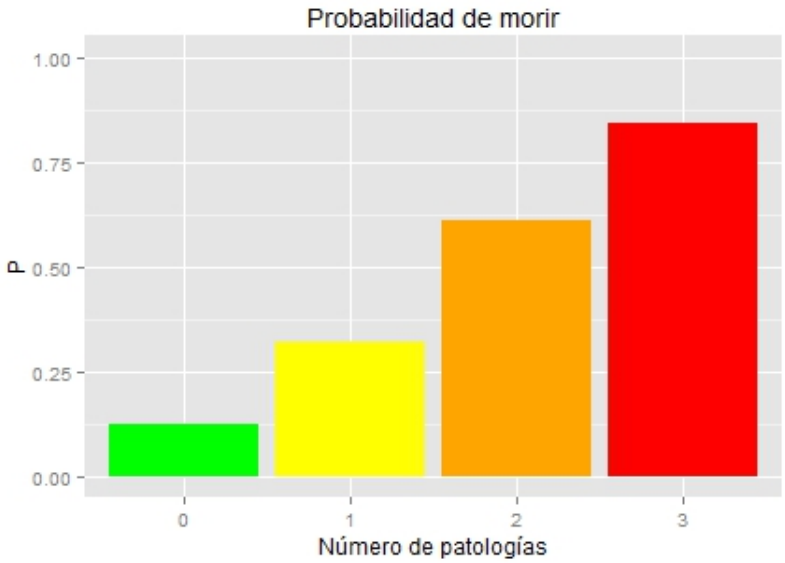
Sustituyendo por los diferentes valores de que puede tener un paciente, se obtiene la tabla de probabilidades 2.10.

Tabla 2.10: Tabla probabilidades según patologías

	P(Death=0)	P(Death=1)
Ninguna patología	0.8771054	0.1228946
1 patología	0.6801584	0.3198416
2 patologías	0.3878639	0.6121361
3 patologías	0.1588112	0.8411888

Finalmente, el gráfico de barras muestra las probabilidades de muerte de la tabla 2.10 dependiendo del número de patologías que tenga un paciente. Este gráfico sirve a los clínicos para tener en mente una idea de las probabilidades a priori.

Figura 2.4: Gráfico con las probabilidades de muerte



Capítulo 3

Análisis de las complicaciones vasculares

> "If you torture the data enough,
> nature will always confess."
>
> *Ronald Coase*

Uno de los problemas que conlleva el implante de una TAVI es el riesgo de complicación vascular, como un ictus. Esta complicación está asociada a la arteriopatía periférica, que es un estrechamiento y endurecimiento de las arterias del paciente, produciendo una falta de irrigación sanguínea en piernas y pies.

La arteriopatía periférica está relacionada con tamaños pequeños de las arterias femoral e iliaca. Estos individuos podrían identificarse utilizando criterios clínicos o criterios estadísticos relacionados con la atipicidad.

En esta parte del TFG se utilizaron los diferentes valores registrados en ocho variables: los diámetros máximos y mínimos de la femoral izquierda y derecha, y los diámetros máximos y mínimos de la ilíaca izquierda y derecha.

3.1. Metodología

Para identificar comportamientos atípicos y, en general, para obtener la posición relativa de cada individuo respecto a la muestra conjuntamente en una variables numéricas se obtuvieron las distancias de Mahalanobis clásicas y una versión robusta basada en el estimador del determinante de mínima covarianza (MCD) (Rousseeuw, 1986). Estas dos

técnicas permitieron identificar qué individuos estaban más alejados del centro de la nube de puntos, que en este análisis estaba contenida en $\mathbb{R}^8$.

3.1.1. Distancia de Mahalanobis

Las distancias de Mahalanobis de un punto al centro de los datos, fueron propuestas por Mahalanobis en el año 1936; y pueden ser entendidas como una generalización multivariante de la estandarización de los datos de datos univariante, dada por restar la media y dividir por la desviación típica muestrales. Estos valores estandarizados, al cuadrado, tendrían una distribución límite chi-cuadrado con un grado de libertad, cuando los datos provienen de un modelo normal. En el caso multivariante, la generalización viene dada por la fórmula 3.1 y el cuadrado de las distancias de Mahalanobis converge a una distribución χ^2_p.

$$\Delta^2 = (x_i - \bar{x})' \mathbf{S}^{-1} (x_i - \bar{x}). \tag{3.1}$$

La distancia de Mahalanobis está especialmente indicada en el análisis multivariante debido a que es equivariante frente a los cambios de escala y tiene en cuenta la estructura de correlación entre las variables. Al igual que sucedería para el caso univariante, se aplican reglas empíricas que clasifican como observaciones atípicas aquellas que disten de la media más que el percentil 97,5 de las distancias. En el caso de una normal univariante se tendrían que este 2.5 % de atípicos estarían repartidos en las dos colas de la distribución. En el caso multivariante estas desviaciones respecto de la media aparecen en todas las direcciones. Por todo ello, una definición de atípico vendría dada por: $DM(x) > \sqrt{\chi^2_{p,0,975}}$.

3.1.2. Estimador del Determinante de Mínima Covarianza (MCD)

Rousseeuw (2008) indica que debido a la existencia de contaminación en una muestra las estimaciones multivariantes pueden diferir sustancialmente de las estimaciones que se obtendría si no hubiera contaminación en la muestra y que el análisis de los residuales, incluso cuando estos son obtenidos mediante técnicas de *leave-one-out*, no funciona para descubrir la presencia de contaminación en la muestra. Ligados a la influencia de la contaminación aparecen dos fenómenos en la estimación como el *masking effect*, observaciones atípicas quedan enmascaradas y no son identificadas como tales y el *Swamping effect* correspondiente a que observaciones provenientes del modelo, a causa de la contaminación, quedan identificadas como atípicas.

Su propuesta para la identificación de atípicos corresponde al uso de distancias de Mahalanobis basadas en estimaciones robustas de los parámetros de localización y escala, para que la definición de atípico no esté condicionada por posibles observaciones contaminantes en la muestra. En el caso de las distancias de Mahalanobis, la necesidad de la robustez es clara, éstas están basadas en estimaciones de los parámetros de localización y escala, por lo que es importante que no estén afectadas por la contaminación. Esta robustez se consigue eliminando observaciones de la muestra a la hora de obtener estos resúmenes de los datos. El algoritmo que se utiliza para la localización y escala es el FAST-MCD propuesto por Rousseeuw y Van Driessen (1999).

Descripción del estimador MCD

Dado un porcentaje, α, superior al porcentaje de contaminación esperado en la muestra, el estimador MCD está basado en el porcentaje $1-\alpha$ de observaciones de la muestra más concentradas, en el sentido de que su matriz de covarianzas tenga el menor determinante posible. El estimador de la localización del MCD es la media y el estimador de escala la matriz de covarianzas de dichas observaciones más concentradas en el sentido anteriormente descrito. El estimador MCD es equivariante frente a cualquier transformación lineal de los datos. Esta propiedad produce que la identificación de los atípicos sea invariante frente a cambios de escala, traslaciones y/o rotaciones de los datos. Una forma de evidenciar la robustez de un estimador frente a observaciones contaminantes incluidas en la muestra es el punto de ruptura (Donoho and Huber, 1983). Dicho punto, se puede definir como el menor porcentaje de observaciones que al ser modificadas de manera caprichosa pueden producir que el estimador funcione de forma arbitraria. En el caso de la localización esto corresponde a que pudiéramos llevarnos su estimación a infinito y en el caso de la escala a que pudiéramos acercar tanto como quisiéramos los autovalores de la matriz de escala estimada a 0 o a infinito. Para los estimadores máximo verosímiles de la localización y la escala, la media y la matriz de covarianzas el punto de ruptura es próximo a 0 cuando el tamaño de la muestra es grande. Para el MCD, el punto de ruptura se sitúa alrededor de $\alpha\,\%$ tanto para el estimador de la localización como para el estimador de escala, y podría alcanzar valores cercanos al $50\,\%$ si escogiéramos $\alpha\,\%$ próximo a este valor. Esto significa, que incluso con un $50\,\%$ de observaciones aberrantes en la muestra las estimaciones MCD resistirían.

Algoritmo FAST-MCD

La estimación MCD no se puede obtener de forma explícita a partir de una muestra. Rousseeuw y Van Driessen (1999) desarrollaron el algoritmo FAST-MCD para calcular de forma eficiente el estimador MCD, basándose en un paso denominado de concentración. El algoritmo parte de un subconjunto inicial de observaciones aleatorio, del que se obtiene la media y la matriz de varianzas covarianzas y se toma como solución inicial. Las iteraciones vienen dadas por dos pasos:

1. Se calculan las distancias de Mahalanobis, basadas en la estimación actual, de los individuos de la muestra a la media estimada, se ordenan dichas distancias, quedándonos con el $1-\alpha\,\%$ de individuos más próximos a la media (paso de concentración).

2. Se toma como nueva estimación la media y la matriz de covarianzas de este subconjunto $1-\alpha\,\%$ de individuos.

Estos 2 pasos se podrían iterar hasta que el determinante de la matriz de covarianzas sea cero, o hasta que su valor no cambie de una iteración a la otra. La secuencia de determinantes, que es no creciente, debe converger en un número finito de pasos (puesto que sólo existe un conjunto finito de subconjuntos del tamaño elegido) a un mínimo local.

Una estrategia muy utilizada para aplicar el algoritmo corresponde a extraer muchas muestras de tamaño p+1, siendo p el número de variables, por m.a.s. para comenzar el algoritmo con sus resúmenes de localización y escala. Para cada una de ellas se realizan iteraciones de los 2 pasos anteriores hasta la convergencia. Se toma como estimación la solución que tenga un determinante menor (Hawkings y Olive (2002)). El hecho de empezar por conjuntos del mínimo tamaño posible p+1 para realizar estimaciones en un espacio de dimensión p se debe a Rousseeuw, y se basa en que la probabilidad de tener un subconjunto libre de *outliers* será mayor cuanto más pequeño sea este. Para muestras "grandes", el algoritmo utiliza submuestras del conjunto de datos, evitando así utilizar todas las observaciones en la obtención de la localización y la escala en cada paso.

Si se denota como $\hat{\boldsymbol{\mu}}_{opt}$ y $\hat{\boldsymbol{\Sigma}}_{opt}$ los estimadores óptimos, es decir, la media y la matriz de covarianzas de menor determinante, la aplicación habitual del algoritmo devuelve:

$$\hat{\boldsymbol{\mu}}_{MCD} = \hat{\boldsymbol{\mu}}_{opt} \text{ y } \hat{\boldsymbol{\Sigma}}_{MCD} = c_{\alpha,n}\hat{\boldsymbol{\Sigma}}_{opt}$$

donde $c_{\alpha,n}$ es un factor de consistencia y corrección para la matriz de varianzas-covarianzas (Pison et al. (2002)). Este factor de consistencia intenta corregir el sesgo bajo el modelo Normal de no contaminación que aparece en la estimación de la matriz de varianzas covarianzas, por haber eliminado las observaciones más alejadas.

Adicionalmente, la aplicación del algoritmo FAST-MCD, incluyen una estimación basada en un paso de repesado de las observaciones realizado tras la estimación. Este se realiza seleccionando las observaciones que están a distancia de Mahalanobis inferior al percentil 97.5 de la distribución chi-cuadrado cuando utilizamos la estimación que incluye el factor de consistencia anteriormente mencionado. De esta forma incluimos en la estimación de la media y la matriz de covarianzas más observaciones y mejoramos la eficiencia de la estimación. Hay que señalar que tanto el factor de consistencia como esta versión *reweighted* (repesado) funcionarán mejor cuando no haya contaminación en la muestra. El punto de ruptura del estimador *reweighted*, sigue siendo cercano a α, mientras que cuando no hay contaminación, la eficiencia del estimador *reweighted* es muy superior a la del MCD sin este paso adicional.

El algoritmo completo para la estimación del MCD se encuentra en la librería *robustbase* de R.

De ahora en adelante, se considerará como **DMC** a las distancias de Mahalanobis clásicas calculadas a partir de la media y la matriz de covarianzas; y como **DM-MCD** a las basadas en el MCD.

3.2. Selección y Descripción de las Variables

Tal y como se indicó en la introducción de este capítulo, las variables que fueron utilizadas para caracterizar a los pacientes que podían ser considerados atípicos potenciales son las que recogen los diámetros máximo y mínimo de las arterias femorales izquierda y derecha y la arteria iliaca izquierda y derecha.

Tabla 3.1: Descripción Variables

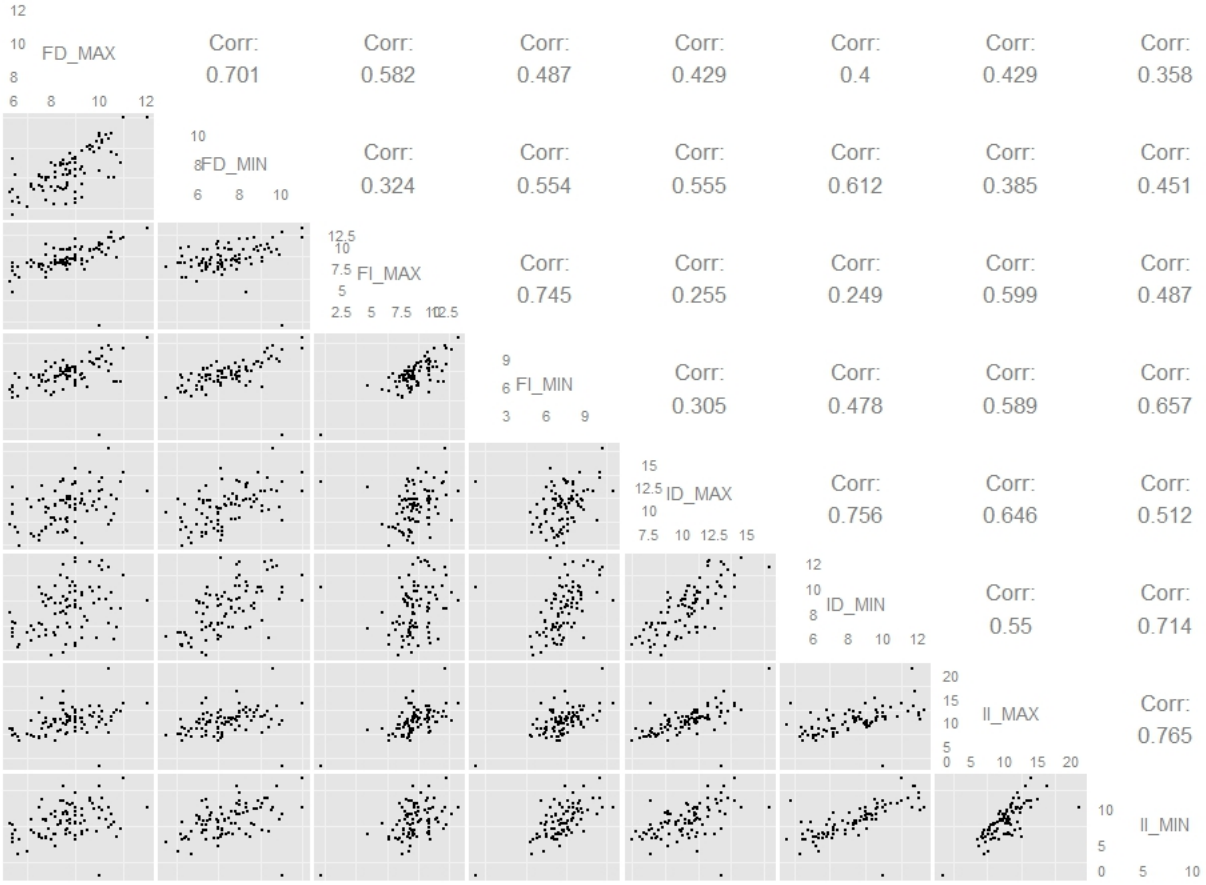
Variable	n	Media	Desv. Típica	Mediana	Mín	Máx	Skewness	Kurtosis
Femoral derecha MAX	88	8.69	1.21	8.70	6.3	12.0	0.03	-0.29
Femoral derecha MIN	88	7.44	1.40	7.25	4.5	11.0	0.33	-0.35
Femoral izquierda MAX	88	8.53	1.51	8.40	0.9	12.0	-1.29	6.20
Femoral izquierda MIN	88	7.24	1.41	7.20	0.8	11.0	-0.70	3.94
Iliaca derecha MAX	88	10.39	2.04	10.55	6.2	16.6	0.16	-0.19
Iliaca derecha MIN	88	8.33	1.90	8.40	4.6	12.3	0.17	-0.79
Iliaca izquierda MAX	88	10.18	2.62	10.30	0.9	21.1	0.43	3.41
Iliaca izquierda MIN	88	8.22	2.19	8.25	0.7	14.0	-0.12	0.75

En la tabla 3.1, se recogen las características de las 8 variables de estudio en los 88 pacientes que tenían completos los valores en cada una de ellas.

Con este conjunto de variables lo primero que se realizó fue una matriz de diagramas

de dispersión para intentar detectar en dos dimensiones posibles valores alejados de la nube de puntos. En la figura 3.1 se puede apreciar cómo algún valor sí que se aleja de la nube, pero únicamente lo hace en una de las variables.

Figura 3.1: Matriz Plots de Dispersión



A la vista del gráfico, no se podía determinar con precisión cuántos pacientes podían ser considerados atípicos, pues en dos dimensiones se pierde toda la percepción de ser atípico en ocho.

3.3. Detección Multivariante

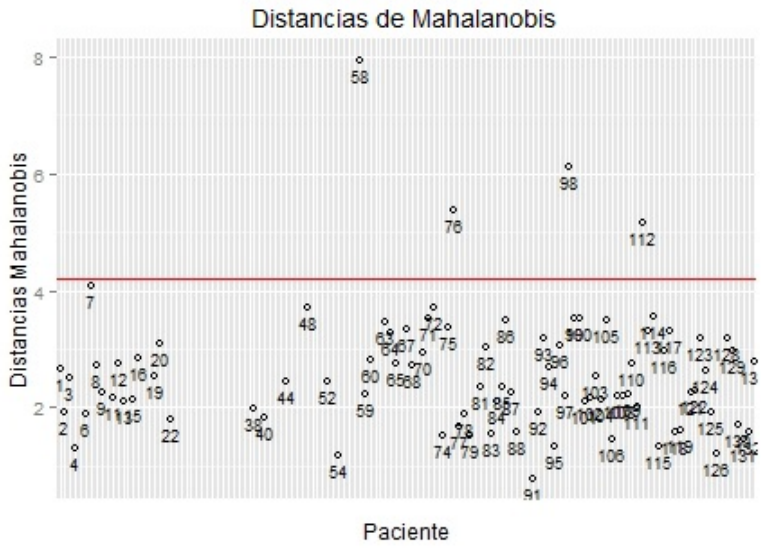
El plot de las distancias de Mahalanobis clásicas, reveló que había cuatro pacientes que destacaban en su atipicidad; puesto que si se hace uso del resultado visto en la *Sección 3.1.2*, hay que considerar como *outlier* toda aquella observación que tenga una distancia de Mahalanobis superior a $\sqrt{\chi^2_{8,0,975}} = 4{,}19$. Dichos pacientes se encuentran en la tabla 3.2.

Tabla 3.2: Pacientes atípicos con Mahalanobis

Paciente	Distancia M	FD MAX	FD MIN	FI MAX	FI MIN	ID MAX	ID MIN	II MAX	II MIN
58	7.932	10.0	10.0	0.9	0.8	13.0	10.0	0.9	0.7
76	5.357	8.5	6.2	8.2	5.6	7.6	4.8	13.8	10.9
98	6.116	10.4	9.6	10.9	10.2	16.6	11.6	21.1	10.1
112	5.144	8.3	6.4	6.0	4.7	12.5	6.7	8.3	6.5

En primer lugar el paciente 58, se caracteriza por tener valores extremos tanto en los valores máximos como en los mínimos; lo que indica que tiene unas arterias muy grandes en el lado derecho y muy pequeñas en el izquierdo. En el caso del paciente 98, tiene valores altos en todos los diámetros. Sin embargo, el paciente 112, tiene valores muy bajos, lo que indica que el tamaño de sus arterias es pequeño, y puede conllevar a tener problemas futuros, lo que puede considerarse factor pronóstico de problemas futuros.

Figura 3.2: Plot con las Distancias de Mahalanobis Clásicas



Haciendo uso de la librería *robustbase* de R se realizó la estimación robusta. En la figura 3.3 está el gráfico con la distancia robusta calculada de cada observación. Se puede apreciar de forma llamativa como ha aumentado el número de *outliers*, pasando de cuatro pacientes en el método anterior a tener un grupo de 22 pacientes, lo que supone haber aumentado del 4.54 % al 25 %.

Tal y como se notaba en el plot de las distancias de Mahalanobis, los pacientes 7, 48 y 75 estaban cerca del umbral; y ahora con el estimador MCD lo superan.

Figura 3.3: Plot las distancias Mahalanobis - MCD

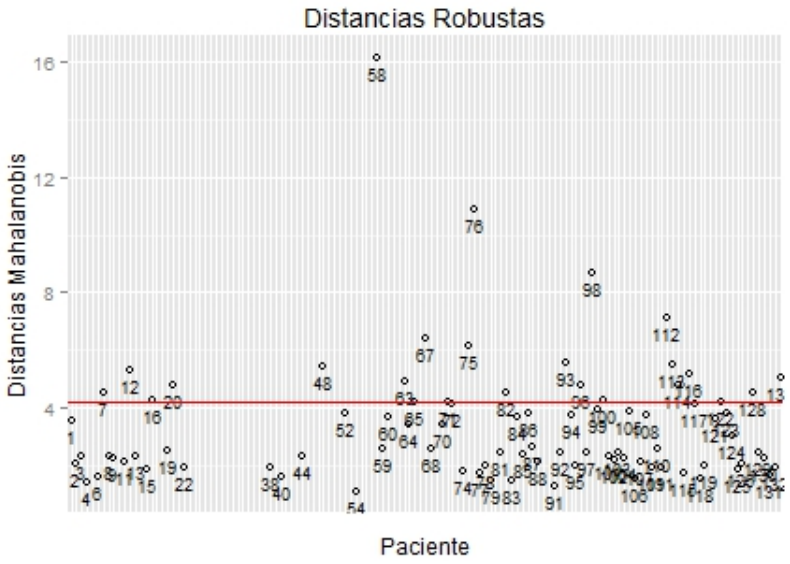
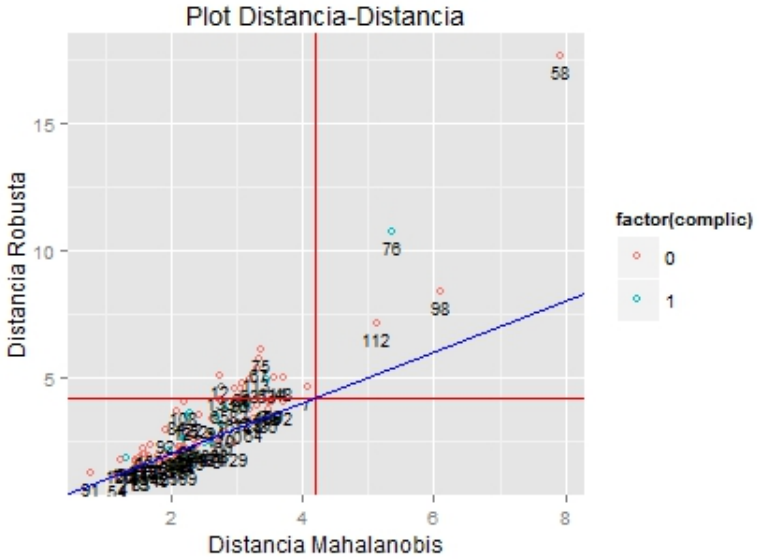


Figura 3.4: Plot Comparativo Distancias de Mahalanobis y Robustas

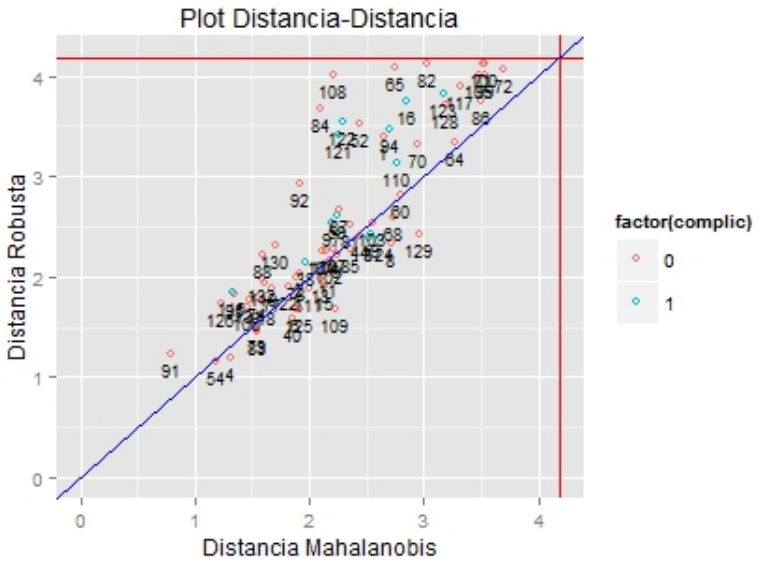


En la figura 3.4 están representadas las distancias de Mahalanobis clásicas y las distancias robustas basadas en el MCD, separadas en dos grupos: Rojo si el paciente tuvo complicaciones vasculares y azul en caso contrario. En caso de estar en el cuadrante inferior izquierda, las observaciones no son consideradas como atípicas. Por el contrario, las observaciones que se sitúan en el cuadrante superior derecha son las observaciones que con los dos procedimientos son consideradas como atípicas. Finalmente, en el cuadrante

superior izquierda están situadas las observaciones que son consideradas *outliers* por el estimador MCD.

En este plot ya se pudo observar, que salvo dos individuos, los pacientes determinados como atípicos por los dos métodos, pertenecían al grupo que no había tenido problemas vasculares.

Figura 3.5: Plot Distancia Distancia Ampliado



Por otro lado se apreció como en el cuadrante inferior izquierda, se encontraban la mayoría de los pacientes con problemas vasculares.

3.4. Comparación de Perfiles

En la *Sección 3.1* se describieron los pacientes que habían sido detectados como atípicos tanto en la distancia de Mahalanobis como en el estimador MCD. En la tabla 3.3 están los nuevos pacientes atípicos detectados únicamente con el estimador robusto.

Si se compara la tabla 3.3 con la tabla 3.1 se comprueba como en los nuevos *outliers* detectados con el MCD hay pacientes con todas sus medidas muy superiores a la media general como el paciente 7 o el 48, o el 114 que todos los valores máximos les tienen claramente superiores a la media. Por otro lado hay pacientes como el 128 que es totalmente opuesto a los anteriores: valores muy bajos en todas las variables medidas. El resto de pacientes destacan en alguna de ellas, lo que indica que se alejan de la nube en tan solo alguna de las 8 marginales.

Tabla 3.3: Pacientes atípicos con MCD

Paciente	Distancia	FD MAX	FD MIN	FI MAX	FI MIN	ID MAX	ID MIN	II MAX	II MIN
7	4.074	11.0	11.0	11.0	10.0	14.0	11.0	14.0	14.0
12	2.756	10.1	9.5	9.3	9.1	12.1	12.1	10.5	9.7
20	3.082	7.6	6.4	7.8	6.6	12.6	9.5	14.9	11.2
48	3.711	12.0	11.0	12.0	11.0	12.0	9.0	14.0	10.0
63	3.454	9.3	5.7	8.3	7.3	6.5	6.0	8.7	6.7
65	2.746	9.0	5.6	9.8	6.6	7.0	5.6	8.1	5.6
67	3.336	8.1	5.2	9.7	5.3	9.1	6.4	10.5	6.4
75	3.362	10.6	7.0	10.4	6.5	9.6	5.9	12.0	8.0
82	3.025	9.5	9.1	8.2	5.9	9.9	7.4	7.9	7.0
96	3.063	9.0	7.1	8.4	7.5	10.1	8.6	11.1	6.2
93	3.199	6.4	5.4	8.1	5.3	10.6	6.4	8.5	5.8
100	3.527	10.9	8.0	10.8	6.4	13.0	8.3	12.3	7.2
113	3.309	9.0	5.3	8.8	6.0	8.7	5.4	9.0	8.0
114	3.562	10.7	8.5	8.7	6.5	11.1	6.8	12.5	6.1
116	2.971	6.4	4.5	7.5	4.8	9.7	6.1	7.2	5.4
117	3.317	9.0	8.4	10.7	7.5	12.9	10.1	11.2	9.5
128	3.188	6.6	5.7	7.5	4.9	7.6	5.3	7.0	3.7
133	2.780	8.0	6.4	7.2	5.0	8.4	4.6	9.3	5.6

Para comparar los perfiles de los pacientes detectados como *outliers* frente a los pacientes normales, se separaron los atípicos en dos grupos: los atípicos que tenían valores en los diámetros muy altos y los que tenían diámetros muy bajos; siendo de especial interés comparar aquellos que tenían unas arterias más pequeñas frente a los pacientes no considerados *outliers*. Esta separación responde a que un paciente con diámetros grandes en las arterias va a tener menos riesgo de complicación vascular que un paciente con un diámetro pequeño.

Esta comparación se hizo mediante dos criterios: 1) criterio clínico: se cogieron aquellos pacientes *outliers* cuyos diámetros mínimos estaban por debajo de 6.5 en 3 o 4 de las 4 medidas mínimas; y 2) criterio estadístico: se han considerado *outliers* por abajo aquellos que eran menores a la media menos dos veces la desviación típica en cada una de las 8 variables. Estos pacientes son los resaltados en azul en las tablas 3.2 y 3.3 para el primer criterio, rojo para el segundo criterio y magenta cuando coincidieron con los dos criterios. Como variables a comparar se utilizaron: Edad, Altura, Peso, Complicación Vascular, Tabaquismo, Fallo Cardíaco Previo, IRC, EPOC, Fragilidad y Muerte.

En ninguna de las dos tablas, 3.4 y 3.5, se encontraron evidencias de que hubiese diferencias significativas entre los atípicos con menor tamaño de arteria y los pacientes que conforman el núcleo de la nube. Esto se debe a que la muestra de atípicos no es lo suficientemente potente como para detectar diferencias significativas.

Tabla 3.4: Comparación de pacientes Criterio 1

Variable	Atípicos N=8	Normales N=65	P-valor
Tabaquismo	44.44 %	20.59 %	0.24
Fallo Previo	66.67 %	52.94 %	0.67
IRC	22.22 %	14.71 %	0.92
EPOC	44.44 %	17.65 %	0.50
Fragilidad	11.11 %	42.65 %	0.14
Complicación	22.22 %	5.88 %	0.29
Death	11.11 %	22.06 %	0.75
Edad	78.33 ± 5.68	80.46 ± 6.93	0.71
Peso	70.44 ± 10.24	69.10 ± 13.73	0.81
Altura	1.62 ± 0.09	1.58 ± 0.08	0.28

Tabla 3.5: Tabla comparación de pacientes criterio 2

Variable	Atípicos N=5	Normales N=65	P-valor
Tabaquismo	40.00 %	20.59 %	0.65
Fallo Previo	40.00 %	52.94 %	0.92
IRC	0.00 %	14.71 %	0.8
EPOC	60.00 %	17.65 %	0.09
Fragilidad	0.00 %	42.65 %	0.16
Complicación	0.00 %	5.88 %	1
Death	0.00 %	22.06 %	0.54
Edad	81.2 ± 4.32	80.46 ± 6.93	0.82
Peso	73.8 ± 12.07	69.10 ± 13.73	0.55
Altura	1.634 ± 0.08	1.58 ± 0.08	0.41

3.5. Nuevos pacientes en la muestra

Posteriormente a la detección de los *Outliers* se añadieron a la base de datos 10 nuevos registros. Con ellos se calcularon las distancias de Mahalanobis utilizando como vector de medias y matriz de covarianzas los obtenidos en la *Sección 3.3*. El resultado fue que de los 10 pacientes intervenidos de TAVI, 3 de ellos han resultado ser atípicos basados en el estimador del MCD anterior.

Figura 3.6: Plot con las Distancias de Mahalanobis. Nuevas observaciones.

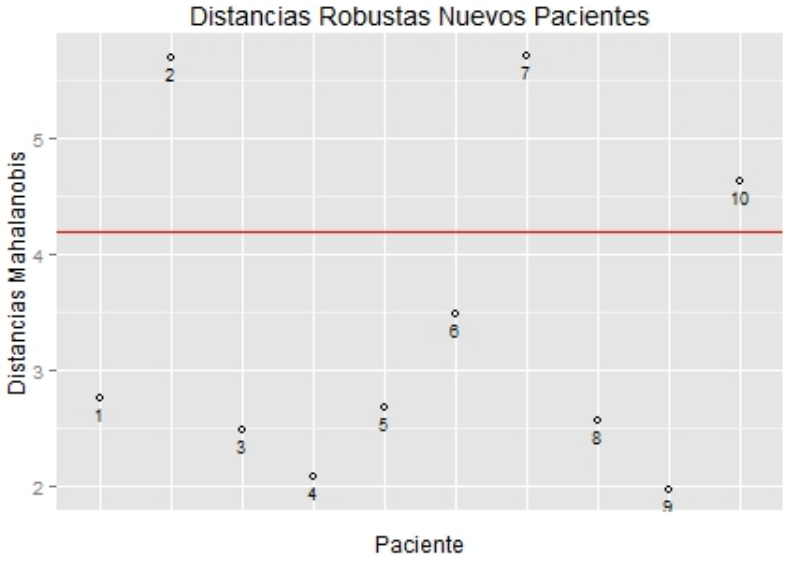


Tabla 3.6: Tabla *Outliers* nuevos pacientes

Paciente	Distancia	FD MAX	FD MIN	FI MAX	FI MIN	ID MAX	ID MIN	II MAX	II MIN
2	5,68768	7	5	8	4.3	7	4.5	6.5	5.1
7	5,70388	9.5	5.9	9.2	7.5	8.6	5.5	10	3.7
10	4,62234	7.7	6.6	8.8	6.8	7.9	6.6	11.3	6.6

En este caso se apreció que los 3 pacientes declarados cómo *Outliers* destacaban por tener unos valores diametrales muy bajos. Ademas, esta tabla sirvió como instrumento para probar si de cualquier conjunto de datos se obtenía un porcentaje similar de atípicos, dando como resultado un 30 %, muy en sintonía al 25 % encontrado en la *Sección 3.2*.

Este hecho fue un motivo para incluir estos pacientes en la muestra e intentar mejorar la estimación del MCD.

3.5.1. Reestimación

Como sucede en todos los modelos, cuando se incluye un nuevo volumen de datos es necesaria la reestimación de todos los parámetros calculados. Para realizar este cálculo se hará uso de la opción de repesado del algoritmo FAST-MCD; si los *outliers* que previamente habían sido detectados ahora no se encuentran será indicativo de que en realidad son contaminación y se están incluyendo en el núcleo de la nube.

Con los 88 pacientes del conjunto original, más los 10 del segundo conjunto, se obtuvo una muestra de tamaño 98 para volver a hacer todos los cálculos.

Tras realizar el algoritmo MCD de nuevo, esta vez con la opción de repesado, se obtuvieron 14 atípicos (14 % del total), de los cuales 12 habían sido detectados en las secciones anteriores.

Tabla 3.7: Tabla Atípicos Reestimación

Atípicos DM	Atípicos DM-MCD	Atípicos Sin Repesado	Atípicos con Repesado
58	7	12	71
76	20	63	105
98	48	65	
112	58	67	
	75	82	
	96	93	
	113	100	
	114	116	
		117	
		128	
		133	

En la tabla 3.7 se aprecia como los pacientes de la primera columna son los que aparecen en todas las detecciones, lo que indica que son pacientes muy alejados de la nube en cualquier dirección y que son detectados por cualquier método. La segunda columna de la tabla muestra los pacientes que son detectados únicamente de forma robusta, eso sí, en los dos procedimientos.

La tercera columna cobra especial relevancia, porque son los pacientes que suponen una contaminación en la nube de puntos. Estos pacientes son atípicos que, con la reestimación no se detectan como tal, cuando en secciones anteriores sí habían sido considerados como atípicos.

Finalmente, la última columna son dos pacientes que en procesos anteriores no habían sido detectados como *outliers* y en este sí.

3.6. Simulación de Datos

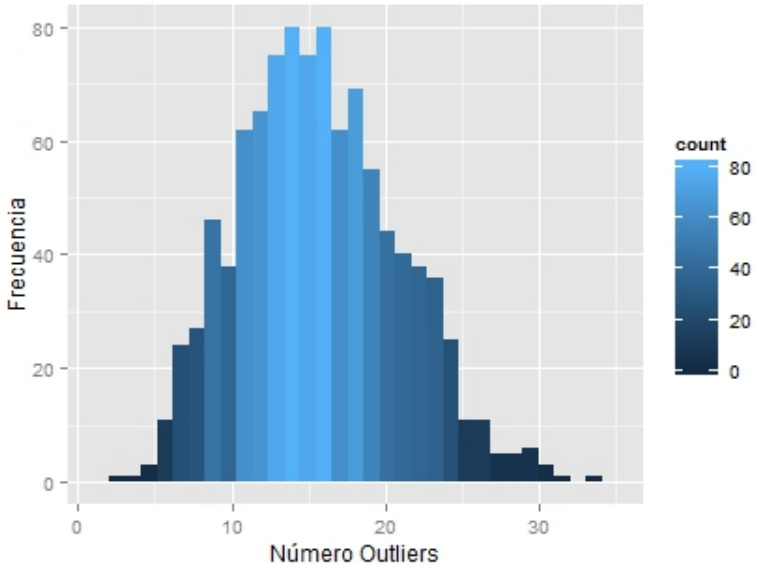
En la aplicación del método robusto, para la identificación de atípicos, en la muestra aparecían hasta un 25 % de atípicos. Como el tamaño muestral no parecía demasiado grande para el número de dimensiones, 8, con el que se estaba trabajando, cabe preguntarse si podía aparecer el mismo porcentaje de atípicos en datos provenientes de una normal, porque quizá el azar pudiera explicar el inusual porcentaje obtenido en nuestra muestra. Para comprobarlo y para intentar entender la dependencia del porcentaje de individuos identificado como atípicos por el MCD del tamaño muestral, el número de dimensiones y el porcentaje de recorte, bajo el modelo normal, se simularon cinco escenarios distintos: (1) una muestra de $n = 88$ datos de una normal de ocho dimensiones con vector de medias $\vec{0}$ y matriz de varianzas covarianzas $\mathbf{I}$ con un $\alpha = 0{,}5$ en el algoritmo MCD; (2) una muestra de las mismas características pero con un $\alpha = 0{,}25$; (3) con un parámetro $\alpha = 0{,}5$ y aumentando la muestra hasta $n = 300$; (4) una muestra de 200 observaciones de una normal identidad de 4 dimensiones y un $\alpha = 0{,}5$; y (5) en el que tomarán diferentes dimensiones, diferentes tamaños muestrales y diferentes alfas.

Se aplicó el MCD a 1000 conjuntos de datos correspondientes a cada uno de los escenarios. El escenario (1) y (2) permite estimar como es de probable obtener un 25 % de atípicos cuando se aplicaba el MCD 50 % a una matriz de datos del tamaño de la nuestra si proviniese de una distribución normal. En definitiva, el plan consiste en la obtención del número de *outliers* identificados por las distancias de Mahalanobis clásicas y por las correspondientes al MCD, para 1000 conjuntos de datos obtenidos por el método de Monte Carlo de cada uno de los diferentes escenarios.

3.6.1. Escenario 1: Datos de una $N_8(\vec{0}, \mathbf{I})$ y $\alpha = 0{,}5$

En este primer escenario se procedió a simular una muestra de $n = 88$ de una normal identidad de ocho dimensiones; con ella se calculó el estimador MCD por defecto, que realiza 500 subgrupos y un $\alpha = 0{,}5$. El alfa determina el porcentaje de observaciones que se eliminarán de la estimación. Los resultados fueron los siguientes: de las 1000 muestras aleatorias se obtuvo que en 105 de ellas el número de atípicos era igual o superior a 22, de lo que se deduce una probabilidad estimada de obtener un porcentaje de atípicos, al aplicar el MCD, como el obtenido (25 %) o superior del 0.105. Con este resultado, al nivel de significación habitual del 5 %, no se podría rechazar, en el sentido del porcentaje de atipicidad, que nuestros datos provinieran de un modelo normal sin contaminación. Por otro lado el histograma reveló que la media y la mediana se encontraban situadas en los 15 (17 %) atípicos; esto es, en un 50 % de las muestras aleatorias se obtuvieron 15 atípicos

Figura 3.7: Histograma *Outliers* Escenario 1

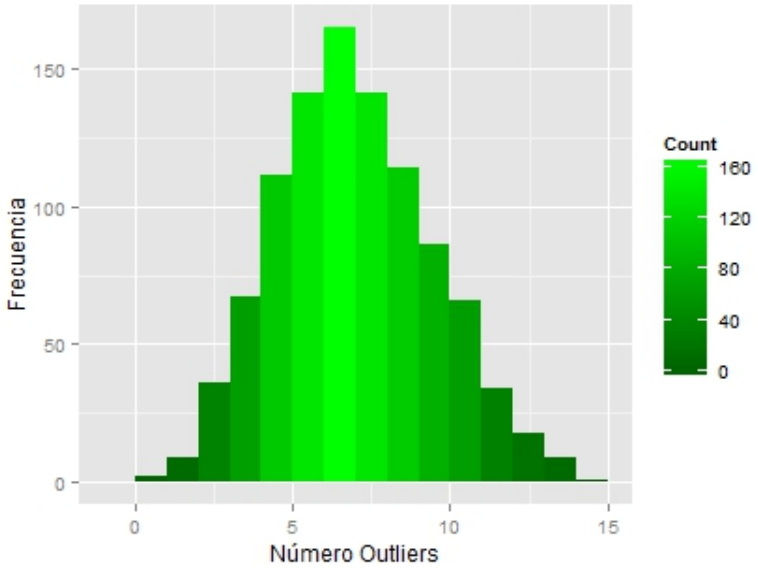


o más.

3.6.2. Escenario 2: Datos de una $N_8(\vec{0}, \mathbf{I})$ y $\alpha = 0{,}25$

En el segundo escenario se procedió a simular una muestra del mismo tamaño que la anterior, de una normal, utilizando como $\alpha = 0{,}25$. En el conjunto inicial, si disminuía el α hasta el 0,25 también se obtenían 22 *outliers*, por lo que hubo interés de nuevo en conocer como de probable es esta cantidad de atípicos al aplicar el MCD con un 25 % de observaciones eliminadas de la estimación. . La solución obtenida difiere a la correspondiente al escenario (1): en ninguna de los conjuntos de datos se obtuvieron 22 *outliers* o más, de hecho, el máximo de *outliers* encontrado fue de 16. La media se situó en 7 *outliers* y la mediana también. A la vista del histograma se puede deducir que la obtención de 10 *outliers* o más es muy poco frecuente. En este caso, la estimación a partir de las 1000 réplicas de la probabilidad de obtener un numero de atípicos tan al to como el observado o superior es 0, lo que nos llevaría a rechazar que nuestros datos provengan de una distribución normal sin contaminación. Por otra parte, al comparar los resultados obtenidos al aplicar el MCD con dos niveles de recorte distinto a datos normales, observamos que el porcentaje de atípicos identificado disminuye lo que podría deberse a que la mayor eficiencia correspondiente al MCD con menor recorte concentra más la distribución del porcentaje de atípicos en torno al 2,5 % nominal.

Figura 3.8: Histograma *Outliers* Escenario 2



3.6.3. Escenario 3: Datos de una $N_8(\vec{0}, \mathbf{I})$ n=300 y $\alpha = 0{,}5$

En este tercer escenario se cambiaron aún mas las condiciones de generación de los datos, determinando tanto el tamaño muestral como el α. Los resultados fueron que la media se situó en torno a los 17 atípicos (5.6 %), al igual que la mediana. El histograma reveló que un porcentaje muy bajo de las 1000 muestras independientes se obtienen 30 o más atípicos.

Al comparar los resultados con los obtenidos del primer escenario observamos que al aumentar la muestra, el porcentaje de atípicos se concentra más en torno al 2.5 %; lo que, también, puede ser explicado por la mayor eficiencia que proporciona el aumento del tamaño muestral.

3.6.4. Escenario 4: Datos de una $N_4(\vec{0}, \mathbf{I})$ n=200 y $\alpha = 0{,}5$

En el escenario 4 se redujo la dimensión de los conjuntos de datos a la mitad. Los resultados obtenidos evidencian que al reducir la dimensión, en una muestra de 200 datos y un alfa de 0.5, se obtuvo un valor medio de 12 atípicos (5.5 %) y una mediana de 11, para un valor de referencia situado en el 2.5 %. En este caso también es esperable una mayor eficiencia al estimar la localización y la escala con datos en un menor número de dimensiones, lo que puede explicar la mayor proximidad de los porcentajes de atípicos obtenidos al 2.5 %.

Figura 3.9: Histograma *Outliers* Escenario 3

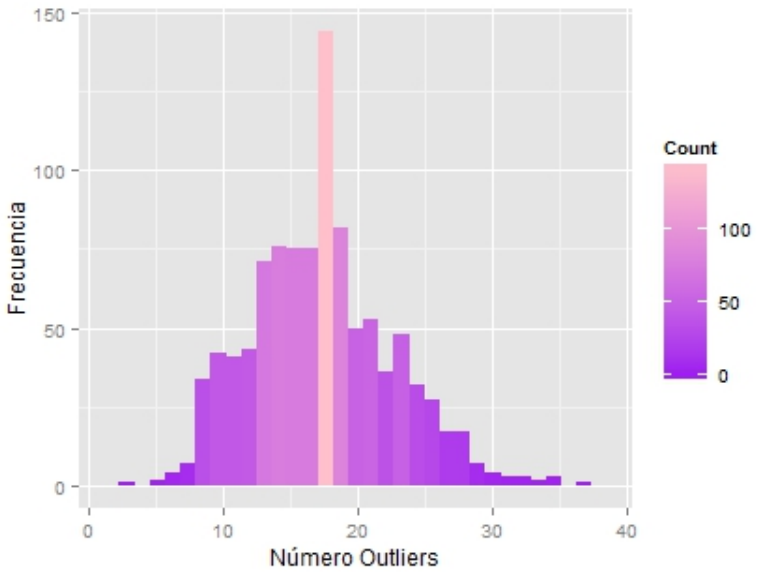
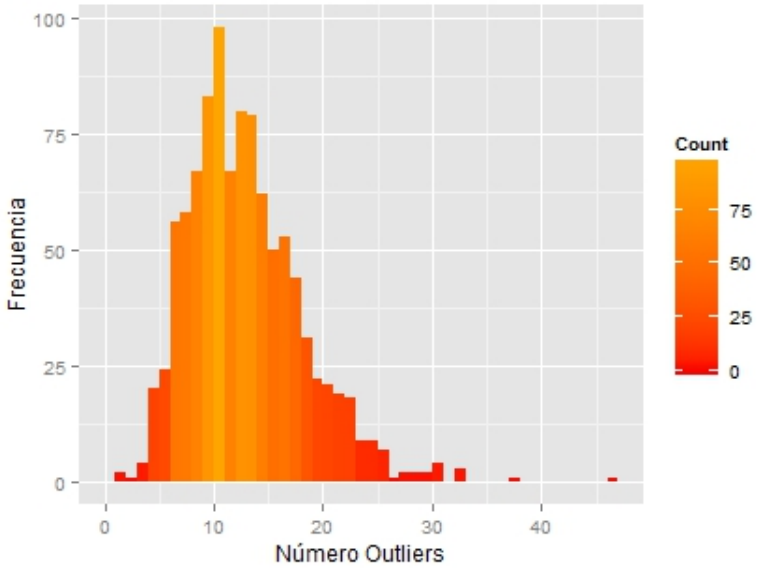


Figura 3.10: Histograma *Outliers* Escenario 4



3.6.5. Escenario 5: Simulación para varios tamaños, varias dimensiones y varios α

Para entender mejor la relación entre los factores anteriores y el porcentaje de atípicos generamos dos subescenarios: el primero (5.1) en el que se estudió la distribución de la cantidad de outliers al variar el número de dimensiones $d = (2, 3, 4, 6)$ y el tamaño muestral $n = (50, 100, 500, 1000)$ y al mantener el nivel alfa en 0.5; y el subescenario (5.2) variando los tamaños muestrales como antes, el nivel de recorte alfas en α =(0.5,0.6,0.75,0.99) para un numero de dimensiones fijado en 5. Para cada uno de ellos representamos una matriz de histogramas y una matriz con la mediana del número de outliers encontrados en cada caso.

Subescenario 5.1

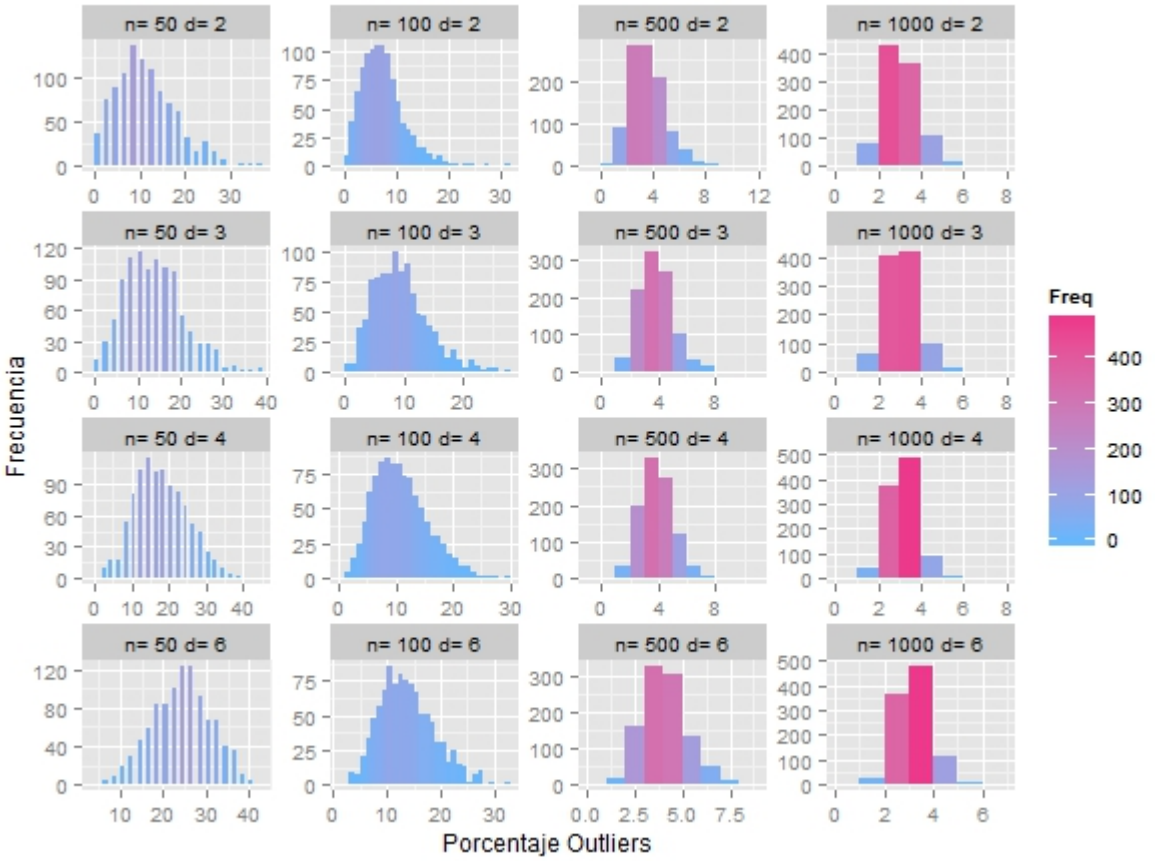
Al observar los resultados, se comprueba en la tabla 3.8 que en muestras pequeñas de muchas dimensiones se encontraron más outliers que en muestras grandes de muchas dimensiones, como era de esperar, como ya señalamos por el aumento del nivel de eficiencia. Al aumentar el tamaño muestral la mediana se situaba en valores en torno al 3-4 %, un porcentaje muy cercano al 2.5 % especificado.

Tabla 3.8: Tabla mediana *Outliers* Subescenario 5.1

		n			
		50	100	500	1000
Dimensión	2	5 10 %	6 6 %	17 3 %	29 3 %
	3	6 12 %	8 8 %	18 4 %	30 3 %
	4	9 18 %	10 10 %	18 4 %	31 3 %
	6	12 24 %	12,5 13 %	20 4 %	32 3 %

De este subescenario se extrae una conclusión importante: al utilizar pocos datos en muestras de varias dimensiones los porcentajes de atípicos pueden llegar a ser en mediana del 24 %, lo que indica que el estimador MCD bajo ciertas condiciones puede encontrar más atípicos de los que realmente existen.

Figura 3.11: Matriz histogramas Subescenario 5.1



Subescenario 5.2

El resultado obtenido indicó que para muestras pequeñas, un alfa de 0.5 hace que se detecten más *outliers* que con un alfa de 0.01, pasando de un 20 % a un 2 %.

Por otro lado, de la tabla 3.9 se extrajo que para muestras grandes el valor de alfa deja de ser determinante y se encuentran un porcentaje similar de valores atípicos.

Figura 3.12: Matriz histogramas Subescenario 5.2

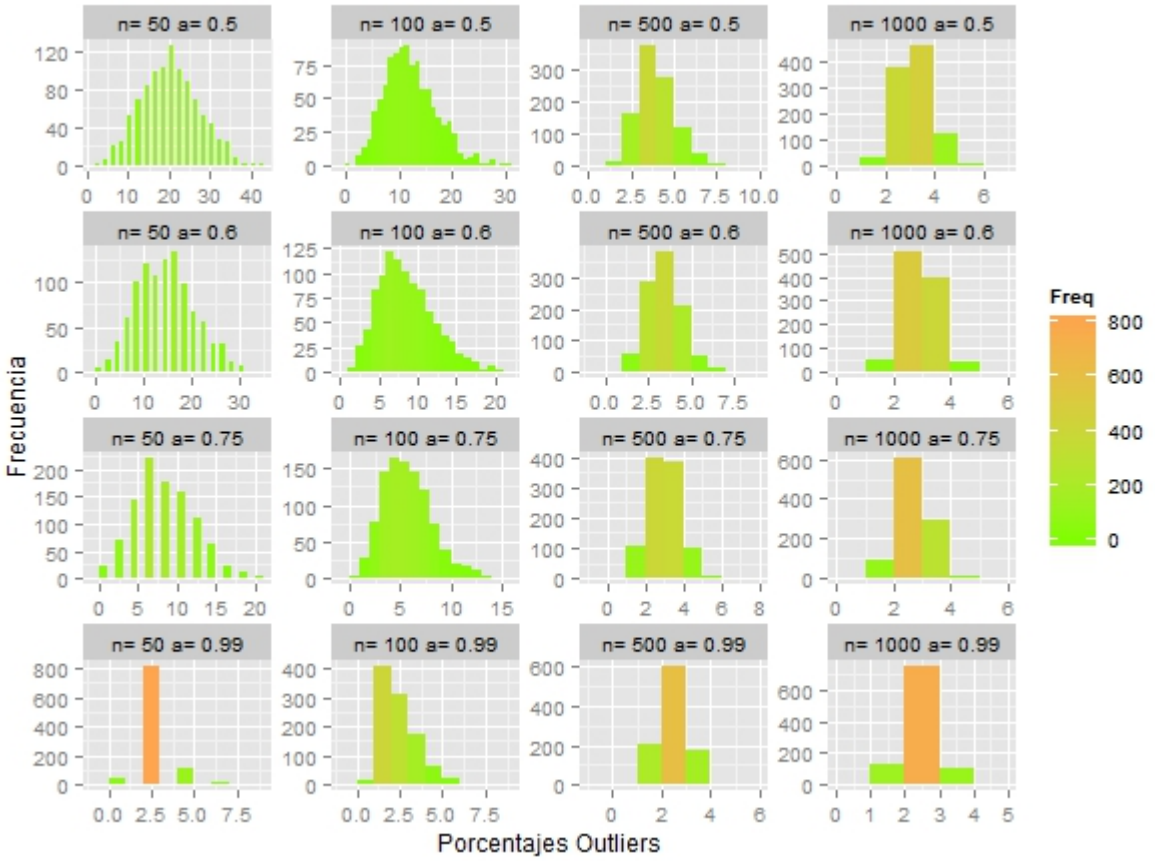


Tabla 3.9: Tabla mediana *Outliers* en Subescenario 5.2

		n			
		50	100	500	1000
α	0,5	10 20 %	11 11 %	19 4 %	31 3 %
	0,6	7 14 %	7 7 %	16 3 %	28 3 %
	0,75	4 8 %	5 5 %	15 3 %	27 3 %
	0,99	1 2 %	2 2 %	12 2 %	24 2 %

Capítulo 4

Resultados Generales

> "An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem."
>
> *John Tukey*

Al finalizar los diferentes análisis y técnicas de clasificación de los pacientes, se pueden establecer los siguientes resultados del bloque del análisis de la mortalidad a largo plazo:

1. En pacientes con Implante de TAVI, la IRC, EPOC y Fragilidad se han revelado como predictores independientes de la mortalidad en el seguimiento.

2. Los diferentes procedimientos diagnósticos indicaron que la calidad del modelo era buena, y que no existía falta de ajuste. De hecho, el p-valor en el test de Hosmer-Lemeshow era superior al 0.05; en concreto de 0.73.

3. Con la regresión logística se ha conseguido determinar cuatro grupos de pacientes, según el número de patologías. Estas probabilidades revelaron una escala de gravedad de las patologías dependiendo de la probabilidad de supervivencia. En el caso de tener una sola patología, los problemas renales eran los más graves, seguidos de la fragilidad y, en último lugar, los problemas pulmonares.

4. Dado que las OR del modelo eran muy similares entre sí, se pudo hacer una aplicación clínica sustituyendo las OR por su media y creando cuatro grupos de pacientes: con ninguna patología, una, dos o tres. Asociadas a estas probabilidades se creó un gráfico que puede dar una idea rápida de la gravedad del paciente según sea su número de patologías.

5. Las limitaciones que tiene este modelo son la poca cantidad de eventos positivos, la muestra reducida y consecuentemente, la simplicidad del modelo. En un futuro, la inclusión de nuevos pacientes a la base, permitirá aumentar la potencia en los contrastes para identificar más variables relacionadas con la variable respuesta que pudieran ser incluidas en el modelo.

En el segundo bloque, de detección de *outliers* multivariantes, se pueden establecer los siguientes resultados:

1. Para detectar los posibles pacientes que destacaban en la muestra se eligieron las distancias de Mahalanobis al centro de la nube y el estimador robusto MCD. Con las distancias de Mahalanobis clásicas se obtuvieron 4 pacientes que estaban muy alejados del resto, y por el segundo 18 atípicos más, elevando el número hasta 22 (25 %). Los primeros 4 pacientes eran claramente clasificables como aquellos que tienen unos valores muy superiores a la media en la mayoría de las variables medidas; y por tanto tenían menos riesgo de sufrir una complicación vascular tras el implante de TAVI. En los otros 18 pacientes no se apreció ningún patrón sobre el tamaño de las arterias: había pacientes con diámetros grandes, pequeños o mixtos.

2. Con todos los atípicos identificados, se procedió a comparar los perfiles entre los pacientes no atípicos y los atípicos con valores más bajos en los diámetros. En ninguno de los casos se obtuvo una diferencia significativa, lo que puede explicarse por la baja potencia de los contrastes causada por el bajo número de individuos detectados con pequeños valores en los diámetros.

3. En un momento posterior del estudio se contó con 10 pacientes nuevos intervenidos de TAVI y 3 resultaron ser atípicos. Este 30 % seguía en línea con los porcentajes de *outliers* presentados en pasos anteriores.

4. El análisis de los datos obtenidos al aplicar la identificación de atípicos a conjuntos de datos pequeños provenientes de distribuciones normales en altas dimensiones nos indica que debemos ser prudentes a la hora de considerar como provenientes de contaminación todos los atípicos que procedimientos robustos como el MCD encuentran, ya que podrían ser debidos a la baja eficiencia en las estimaciones.

Bibliografía

[1] Schaff, H. V. (2011) TAVI - At what price?. The New Englad journal of Medicine. Vol. 364, No. 23, 2256-2258.

[2] Smith, C. R. et al. (2011) Transcatheter versus Surgical Aortic-Valve Replacement in High-Risk Patients. The New Englad journal of Medicine. Vol. 364, No. 23, 2187-2198.

[3] Leon, M. B. et al. (2010) Transcatheter Aortic-Valve Implantation for Aortic Stenosis in Patients Who Cannot Undergo Surgery. The New Englad journal of Medicine. Vol. 363, No. 17, 1597-1607.

[4] Rodès-Cabau, J. (2012) Transcatheter aortic valve implantation: current and future approaches. Natural Reviews Cardiology. Vol. 9, 15-29.

[5] Agresti, A. (2002) Categorical Data Analysis. Second Edition. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN 0-471-36093-7.

[6] Kallin Westin, L. (2001) Reciever Operating characteristic (ROC) analysis. Department of Computer Science, Umea University, Sweden.

[7] Cuadras, C. M. (2012) Nuevos métodos de análisis multivariante. Sin publicar. Barcelona, España.

[8] Rencher, A. C. (2002) Methods of Multivariate Analysis. Second Edition. John Wiley & Sons, Inc., Hoboken, New Jersey. ISBN 0-471-41889-7.

[9] Hubert, M., Rousseeuw, P. J., Van Aelst, S. (2008) High-Breakdown Robust Multivariate Methods. Statistical Science. Vol. 23, No. 1, 92-119. Institute of Mathematical Statistics.

[10] MacQueen, J. B. (1967), Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.

[11] Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0.

[12] SAS 9.2 (SAS Institute, Cary NC).

[13] Juergen Gross and bug fixes by Uwe Ligges (2012). nortest: Tests for Normality. R package version 1.0-2. http://CRAN.R-project.org/package=nortest.

[14] Torsten Hothorn and Kurt Hornik (2015). exactRankTests: Exact Distributions for Rank and Permutation Tests. R package version 0.8-28. http://CRAN.R-project.org/package=exactRankTests.

[15] David Meyer, Achim Zeileis, and Kurt Hornik (2014). vcd: Visualizing Categorical Data. R package version 1.3-2.

[16] Gregory R. Warnes. Includes R source code and/or documentation contributed by Ben Bolker, Thomas Lumley, Randall C Johnson. Contributions from Randall C. Johnson are Copyright SAIC-Frederick, Inc. Funded by the Intramural Research Program, of the NIH, National Cancer Institute and Center for Cancer Research under NCI Contract NO1-CO-12400. (2013). gmodels: Various R programming tools for model fitting. R package version 2.15.4.1. http://CRAN.R-project.org/package=gmodels .

[17] Torsten Hothorn, Kurt Hornik, Mark A. van de Wiel and Achim Zeileis (2006). A Lego System for Conditional Inference. The American Statistician 60(3), 257-263.

[18] Adrian A. Dragulescu (2014). xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.5.7. http://CRAN.R-project.org/package=xlsx .

[19] R Core Team (2014). foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ... R package version 0.8-61. http://CRAN.R-project.org/package=foreign .

[20] Revelle, W. (2015) psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, http://CRAN.R-project.org/package=psych Version = 1.5.1.

[21] Peter Rousseeuw, Christophe Croux, Valentin Todorov, Andreas Ruckstuhl, Matias Salibian-Barrera, Tobias Verbeke, Manuel Koller, Martin Maechler (2015). robustbase: Basic Robust Statistics. R package version 0.92-4. URL http://CRAN.R-project.org/package=robustbase.

Apéndice A

Tablas resumen con las variables utilizadas

Tabla A.1: Muestra Variables Continuas

Nombre	Descripción	N	Media	Sd	Min	Max
Edad		132	80,56	6,54	52	91
Peso		132	68,97	12,50	40	100
Altura		132	1,60	0,09	1,35	1,75
BMI	Body Mass Index	131	26,91	4,60	17,63	42,17
ASC..m2.	Area superficie corporal	130	1,75	0,18	1,29	2,12
STS Score	Society of Thoracic Surgeons Score	132	6,77	5,32	1,2	27,2
LogEuroSCORE	European System for Cardiac Operative Risck Evaluation	132	14,65	8,79	2,3	46,4
Femoral.derecha_MAX		97	8,73	1,21	6,3	12
Femoral.derecha_MIN		95	7,42	1,42	4,5	11
Femoral.izquierda_MAX		93	8,56	1,49	0,9	12
Femoral.izquierda_MIN		91	7,26	1,41	0,8	11
Iliaca.derecha_MAX		99	10,35	2,08	4	16,6
Iliaca.derecha_MIN		97	8,34	1,98	4	12,9
Iliaca.izquierda_MAX		94	10,09	2,58	0,9	21,1
Iliaca.izquierda_MIN		92	8,15	2,19	0,7	14

Tabla A.2: Muestra Variables Categóricas

Nombre	Descripción	N	Valores que toma	n
Gender	Sexo	132	1 (Hombre) 2 (Mujer)	79 53
Smoking	Historial Tabaquismo	131	0 (Nunca) 1(Fum) 2(Ex)	96 12 23
PrevHeartFailure	Insuficiencia Cardiaca Previa	132	0 (no) 1 (si)	45 87
PrevPCI	Previous percutaneous coronary intervention	132	0 (no) 1 (si)	95 37
PrevMI	Previous myocardial infarction	132	0 (no) 1 (si)	103 29
IRC	Problemas Renales Crónicos	132	0 (no) 1 (si)	108 24
EPOC	Obstrucción pulmonar crónica	132	0 (no) 1 (si)	94 38
Fragilidad	Fragilidad	105	0 (no) 1 (si)	63 42
Ao.Regurg Post	Grado de Regurgitación Aórtica	129	0 1 2 3 4 No realizado	46 41 26 9 4 6

Tabla A.3: Muestra variables categóricas (Cont.)

Nombre	Descripción	N	Valores que toma	n
Calcio_ETT	Calcio en ecocardiograma transtorácico	122	0 1 2 3	3 3 42 74
Cardiac.rithm	Ritmo Cardiaco	130	1 2 3	97 26 7
ValveEmbolization	Embolia	128	0 (no) 1 (si)	123 5
NOAF	New onset atrial fibrillation	123	0 (no) 1 (si)	109 14
IRA	Insuficiencia Renal Aguda	128	0 (no) 1 (si)	114 14
Clopidogrel	Fármaco Copidogrel	130	0 (no) 1 (si)	102 28
Death	Muerte del paciente	131	0 1	89 42

Índice de figuras

Índice de tablas